Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127192	14	01010011001111100100	AGCGCGTACA	10	232
C1127192	14	01011110100110011000	AGTACTGTGT	10	233
C1127192	14	01001110101001100110	ATACTCACAC	10	234
C1127192	14	00100110010111011100	CACAGTAGTA	10	235
C1127192	14	00110101001001110011	CGAGCACGCG	10	236
C1127192	14	00111001010101010110	CGTGAGAGAC	10	237
C1127192	14	00111001101010101100	CGTGTCTCTA	10	238
C1127192	14	00101010010010101111	CTCATCTACG	10	239
C1127192	14	10110010110010010101	TCGCTATGAG	10	240
C1127192	14	10101100111001001100	TCTATACATA	10	241
C1127192	14	10101011001100111000	TCTCGCGCGT	10	242
C1127192	14	10010111011011001000	TGACGACTAT	10	243
C1127192	14	10010100110110110010	TGATAGTCGC	10	244
C1127193	14	01101001100101001011	ACTGTGATCG	10	245
C1127193	14	01010011001111100100	AGCGCGTACA	10	246
C1127193	14	01011110100110011000	AGTACTGTGT	10	247
C1127193	14	00100110010111011100	CACAGTAGTA	10	248
C1127193	14	00110101001001110011	CGAGCACGCG	10	249
C1127193	14	00111001010101010110	CGTGAGAGAC	10	250
C1127193	14	00111001101010101100	CGTGTCTCTA	10	251
C1127193	14	00101010010010101111	CTCATCTACG	10	252
C1127193	14	11010010101001101010	TAGCTCACTC	10	253
C1127193	14	10110010110010010101	TCGCTATGAG	10	254
C1127193	14	10101100111001001100	TCTATACATA	10	255
C1127193	14	10101011001100111000	TCTCGCGCGT	10	256
C1127193	14	10010111011011001000	TGACGACTAT	10	257
C1127193	14	10010100110110110010	TGATAGTCGC	10	258
C1127194	14	01101001100101001011	ACTGTGATCG	10	259
C1127194	14	01010011001111100100	AGCGCGTACA	10	260
C1127194	14	01011100111001011000	AGTATACAGT	10	261
C1127194	14	00100100111110011001	CATACGTGTG	10	262
C1127194	14	00110101001001110011	CGAGCACGCG	10	263
C1127194	14	00111001010101010110	CGTGAGAGAC	10	264
C1127194	14	00111001101010101100	CGTGTCTCTA	10	265
C1127194	14	00101010010010101111	CTCATCTACG	10	266
C1127194	14	11100100110101001100	TACATAGATA	10	267
C1127194	14	11010010101001101010	TAGCTCACTC	10	268
C1127194	14	10110010110010010101	TCGCTATGAG	10	269
C1127194	14	10101011001100111000	TCTCGCGCGT	10	270
C1127194	14	10010111011011001000	TGACGACTAT	10	271
C1127194	14	10010100110110110010	TGATAGTCGC	10	272
C1127195	14	01101110101001010100	ACTACTCAGA	10	273
C1127195	14	01101001100101001011	ACTGTGATCG	10	274
C1127195	14	01010011001111100100	AGCGCGTACA	10	275
C1127195	14	00100100111110011001	CATACGTGTG	10	276

IDT Exhibit 1020.5027

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127195	14	00110101001001110011	CGAGCACGCG	10	277
C1127195	14	00111001010101010110	CGTGAGAGAC	10	278
C1127195	14	00111001101010101100	CGTGTCTCTA	10	279
C1127195	14	00101010010010101111	CTCATCTACG	10	280
C1127195	14	11100100110101001100	TACATAGATA	10	281
C1127195	14	11010010101001101010	TAGCTCACTC	10	282
C1127195	14	10110010110010010101	TCGCTATGAG	10	283
C1127195	14	10101011001100111000	TCTCGCGCGT	10	284
C1127195	14	10010111011011001000	TGACGACTAT	10	285
C1127195	14	10010100110110110010	TGATAGTCGC	10	286
C1127196	14	01100101011010011010	ACAGACTGTC	10	287
C1127196	14	01101011001100101001	ACTCGCGCTG	10	288
C1127196	14	01010011001111100100	AGCGCGTACA	10	289
C1127196	14	01011100111001001001	AGTATACATG	10	290
C1127196	14	01001110010110110100	ATACAGTCGA	10	291
C1127196	14	00100111110111001000	CACGTAGTAT	10	292
C1127196	14	00110101001001110011	CGAGCACGCG	10	293
C1127196	14	00111001010101010110	CGTGAGAGAC	10	294
C1127196	14	00111001101010101100	CGTGTCTCTA	10	295
C1127196	14	00101010010010101111	CTCATCTACG	10	296
C1127196	14	11010010101001101010	TAGCTCACTC	10	297
C1127196	14	11001001100110010011	TATGTGTGCG	10	298
C1127196	14	10110100101110010100	TCGATCGTGA	10	299
C1127196	14	10010110100101001101	TGACTGATAG	10	300
C1127197	14	01100101011001001101	ACAGACATAG	10	301
C1127197	14	01101011001100101001	ACTCGCGCTG	10	302
C1127197	14	01010011001111100100	AGCGCGTACA	10	303
C1127197	14	01011100111001011000	AGTATACAGT	10	304
C1127197	14	01001110010110110100	ATACAGTCGA	10	305
C1127197	14	00100111110111001000	CACGTAGTAT	10	306
C1127197	14	00110101001001110011	CGAGCACGCG	10	307
C1127197	14	00111001010101010110	CGTGAGAGAC	10	308
C1127197	14	00111001101010101100	CGTGTCTCTA	10	309
C1127197	14	00101010010010101111	CTCATCTACG	10	310
C1127197	14	11010100110010101010	TAGATATCTC	10	311
C1127197	14	11001001100110010011	TATGTGTGCG	10	312
C1127197	14	10110100101110010100	TCGATCGTGA	10	313
C1127197	14	10011110100101001001	TGTACTGATG	10	314
C1127198	14	01100101011001001101	ACAGACATAG	10	315
C1127198	14	01101011001100101001	ACTCGCGCTG	10	316
C1127198	14	01010011001111100100	AGCGCGTACA	10	317
C1127198	14	01011100111001011000	AGTATACAGT	10	318
C1127198	14	01001110010110110100	ATACAGTCGA	10	319
C1127198	14	00100111110111001000	CACGTAGTAT	10	320
C1127198	14	00110101001001110011	CGAGCACGCG	10	321

IDT Exhibit 1020.5028

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127198	14	00111001010101010110	CGTGAGAGAC	10	322
C1127198	14	00111001101010101100	CGTGTCTCTA	10	323
C1127198	14	00101010010010101111	CTCATCTACG	10	324
C1127198	14	11010010110100101010	TAGCTAGCTC	10	325
C1127198	14	11001001100110010011	TATGTGTGCG	10	326
C1127198	14	10110100101110010100	TCGATCGTGA	10	327
C1127198	14	10011110100101001001	TGTACTGATG	10	328
C1127199	14	01100101011001001101	ACAGACATAG	10	329
C1127199	14	01101011001100101001	ACTCGCGCTG	10	330
C1127199	14	01010011001111100100	AGCGCGTACA	10	331
C1127199	14	01011100111001011000	AGTATACAGT	10	332
C1127199	14	01001110010110110100	ATACAGTCGA	10	333
C1127199	14	00100111110111001000	CACGTAGTAT	10	334
C1127199	14	00110101001001110011	CGAGCACGCG	10	335
C1127199	14	00111001010101010110	CGTGAGAGAC	10	336
C1127199	14	00111001101010101100	CGTGTCTCTA	10	337
C1127199	14	00101010010010101111	CTCATCTACG	10	338
C1127199	14	11010010101001101010	TAGCTCACTC	10	339
C1127199	14	11001001100110010011	TATGTGTGCG	10	340
C1127199	14	10110100101110010100	TCGATCGTGA	10	341
C1127199	14	10011110100101001001	TGTACTGATG	10	342
C1127200	14	01100101011001001101	ACAGACATAG	10	343
C1127200	14	01101011001100101001	ACTCGCGCTG	10	344
C1127200	14	01010011001111100100	AGCGCGTACA	10	345
C1127200	14	01001110010110110100	ATACAGTCGA	10	346
C1127200	14	00100111110111001000	CACGTAGTAT	10	347
C1127200	14	00110101001001110011	CGAGCACGCG	10	348
C1127200	14	00111001010101010110	CGTGAGAGAC	10	349
C1127200	14	00111001101010101100	CGTGTCTCTA	10	350
C1127200	14	00101010010010101111	CTCATCTACG	10	351
C1127200	14	11010100110010101010	TAGATATCTC	10	352
C1127200	14	11010011011001011000	TAGCGACAGT	10	353
C1127200	14	11001001100110010011	TATGTGTGCG	10	354
C1127200	14	10110100101110010100	TCGATCGTGA	10	355
C1127200	14	10011110100101001001	TGTACTGATG	10	356
C1127201	14	01101011001100101001	ACTCGCGCTG	10	357
C1127201	14	01010011001111100100	AGCGCGTACA	10	358
C1127201	14	01001110010110110100	ATACAGTCGA	10	359
C1127201	14	01001010111001100110	ATCTACACAC	10	360
C1127201	14	00100111110111001000	CACGTAGTAT	10	361
C1127201	14	00110101001001110011	CGAGCACGCG	10	362
C1127201	14	00111001010101010110	CGTGAGAGAC	10	363
C1127201	14	00111001101010101100	CGTGTCTCTA	10	364
C1127201	14	00101010010010101111	CTCATCTACG	10	365
C1127201	14	11010100110010101010	TAGATATCTC	10	366

IDT Exhibit 1020.5029

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127201	14	11010011011001011000	TAGCGACAGT	10	367
C1127201	14	11001001100110010011	TATGTGTGCG	10	368
C1127201	14	10110100101110010100	TCGATCGTGA	10	369
C1127201	14	10011110100101001001	TGTACTGATG	10	370
C1127202	14	01100101011001001101	ACAGACATAG	10	371
C1127202	14	01101011001100101001	ACTCGCGCTG	10	372
C1127202	14	01101010110010010110	ACTCTATGAC	10	373
C1127202	14	01010011001111100100	AGCGCGTACA	10	374
C1127202	14	01011100111001011000	AGTATACAGT	10	375
C1127202	14	01001110010110110100	ATACAGTCGA	10	376
C1127202	14	00100111110111001000	CACGTAGTAT	10	377
C1127202	14	00110101001001110011	CGAGCACGCG	10	378
C1127202	14	00111001010101010110	CGTGAGAGAC	10	379
C1127202	14	00111001101010101100	CGTGTCTCTA	10	380
C1127202	14	11010100110010101010	TAGATATCTC	10	381
C1127202	14	11001001100110010011	TATGTGTGCG	10	382
C1127202	14	10110100101110010100	TCGATCGTGA	10	383
C1127202	14	10011110100101001001	TGTACTGATG	10	384
C1127203	14	01100101011001001101	ACAGACATAG	10	385
C1127203	14	01101011001100101001	ACTCGCGCTG	10	386
C1127203	14	01101010110010010110	ACTCTATGAC	10	387
C1127203	14	01010011001111100100	AGCGCGTACA	10	388
C1127203	14	01011100111001011000	AGTATACAGT	10	389
C1127203	14	01001110010110110100	ATACAGTCGA	10	390
C1127203	14	00100111110111001000	CACGTAGTAT	10	391
C1127203	14	00110101001001110011	CGAGCACGCG	10	392
C1127203	14	00111001010101010110	CGTGAGAGAC	10	393
C1127203	14	00111001101010101100	CGTGTCTCTA	10	394
C1127203	14	11010010110100101010	TAGCTAGCTC	10	395
C1127203	14	11001001100110010011	TATGTGTGCG	10	396
C1127203	14	10110100101110010100	TCGATCGTGA	10	397
C1127203	14	10011110100101001001	TGTACTGATG	10	398
C1127204	14	01100101011001001101	ACAGACATAG	10	399
C1127204	14	01101011001100101001	ACTCGCGCTG	10	400
C1127204	14	01101010110010010110	ACTCTATGAC	10	401
C1127204	14	01010011001111100100	AGCGCGTACA	10	402
C1127204	14	01011100111001011000	AGTATACAGT	10	403
C1127204	14	01001110010110110100	ATACAGTCGA	10	404
C1127204	14	00100111110111001000	CACGTAGTAT	10	405
C1127204	14	00110101001001110011	CGAGCACGCG	10	406
C1127204	14	00111001010101010110	CGTGAGAGAC	10	407
C1127204	14	00111001101010101100	CGTGTCTCTA	10	408
C1127204	14	11010010101001101010	TAGCTCACTC	10	409
C1127204	14	11001001100110010011	TATGTGTGCG	10	410
C1127204	14	10110100101110010100	TCGATCGTGA	10	411

IDT Exhibit 1020.5030

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127204	14	10011110100101001001	TGTACTGATG	10	412
C1127205	14	01100101011001001101	ACAGACATAG	10	413
C1127205	14	01101011001100101001	ACTCGCGCTG	10	414
C1127205	14	01101010110010010110	ACTCTATGAC	10	415
C1127205	14	01010011001111100100	AGCGCGTACA	10	416
C1127205	14	01001110010110110100	ATACAGTCGA	10	417
C1127205	14	00100111110111001000	CACGTAGTAT	10	418
C1127205	14	00110101001001110011	CGAGCACGCG	10	419
C1127205	14	00111001010101010110	CGTGAGAGAC	10	420
C1127205	14	00111001101010101100	CGTGTCTCTA	10	421
C1127205	14	11010100110010101010	TAGATATCTC	10	422
C1127205	14	11001011011001011000	TATCGACAGT	10	423
C1127205	14	11001001100110010011	TATGTGTGCG	10	424
C1127205	14	10110100101110010100	TCGATCGTGA	10	425
C1127205	14	10011110100101001001	TGTACTGATG	10	426
C1127206	14	01100101011001001101	ACAGACATAG	10	427
C1127206	14	01101011001100101001	ACTCGCGCTG	10	428
C1127206	14	01101010110010010110	ACTCTATGAC	10	429
C1127206	14	01010011001111100100	AGCGCGTACA	10	430
C1127206	14	01001110010110110100	ATACAGTCGA	10	431
C1127206	14	00100111110111001000	CACGTAGTAT	10	432
C1127206	14	00110101001001110011	CGAGCACGCG	10	433
C1127206	14	00111001010101010110	CGTGAGAGAC	10	434
C1127206	14	00111001101010101100	CGTGTCTCTA	10	435
C1127206	14	11010010110100101010	TAGCTAGCTC	10	436
C1127206	14	11001011011001011000	TATCGACAGT	10	437
C1127206	14	11001001100110010011	TATGTGTGCG	10	438
C1127206	14	10110100101110010100	TCGATCGTGA	10	439
C1127206	14	10011110100101001001	TGTACTGATG	10	440
C1127207	14	01100101011001001101	ACAGACATAG	10	441
C1127207	14	01101011001100101001	ACTCGCGCTG	10	442
C1127207	14	01101010110010010110	ACTCTATGAC	10	443
C1127207	14	01010011001111100100	AGCGCGTACA	10	444
C1127207	14	01001110010110110100	ATACAGTCGA	10	445
C1127207	14	00100111110111001000	CACGTAGTAT	10	446
C1127207	14	00110101001001110011	CGAGCACGCG	10	447
C1127207	14	00111001010101010110	CGTGAGAGAC	10	448
C1127207	14	00111001101010101100	CGTGTCTCTA	10	449
C1127207	14	11010100110010101010	TAGATATCTC	10	450
C1127207	14	11010011011001011000	TAGCGACAGT	10	451
C1127207	14	11001001100110010011	TATGTGTGCG	10	452
C1127207	14	10110100101110010100	TCGATCGTGA	10	453
C1127207	14	10011110100101001001	TGTACTGATG	10	454
C1127208	14	01100100110011110010	ACATATACGC	10	455
C1127208	14	01101011001100101001	ACTCGCGCTG	10	456

IDT Exhibit 1020.5031

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127208	14	01010011001111100100	AGCGCGTACA	10	457
C1127208	14	01011110100110011000	AGTACTGTGT	10	458
C1127208	14	00100110010111011100	CACAGTAGTA	10	459
C1127208	14	00110101001001110011	CGAGCACGCG	10	460
C1127208	14	00111001010101010110	CGTGAGAGAC	10	461
C1127208	14	00111001101010101100	CGTGTCTCTA	10	462
C1127208	14	00101010111001001011	CTCTACATCG	10	463
C1127208	14	11010010110100101010	TAGCTAGCTC	10	464
C1127208	14	11001011011001011000	TATCGACAGT	10	465
C1127208	14	11001001100110010011	TATGTGTGCG	10	466
C1127208	14	10110100101110010100	TCGATCGTGA	10	467
C1127208	14	10010010011010110110	TGCACTCGAC	10	468
C1127209	14	01101011001100101001	ACTCGCGCTG	10	469
C1127209	14	01010011001111100100	AGCGCGTACA	10	470
C1127209	14	01001101101011011000	ATAGTCTAGT	10	471
C1127209	14	01001010010011100111	ATCATACACG	10	472
C1127209	14	00100111110111001000	CACGTAGTAT	10	473
C1127209	14	00110101001001110011	CGAGCACGCG	10	474
C1127209	14	00111001010101010110	CGTGAGAGAC	10	475
C1127209	14	00111001101010101100	CGTGTCTCTA	10	476
C1127209	14	00101110100101110100	CTACTGACGA	10	477
C1127209	14	11010100110010101010	TAGATATCTC	10	478
C1127209	14	11001111001001001100	TATACGCATA	10	479
C1127209	14	11001001100110010011	TATGTGTGCG	10	480
C1127209	14	10110100101110010100	TCGATCGTGA	10	481
C1127209	14	10010010011010110110	TGCACTCGAC	10	482
C1127210	14	01100101011001001101	ACAGACATAG	10	483
C1127210	14	01101011001100101001	ACTCGCGCTG	10	484
C1127210	14	01010011001111100100	AGCGCGTACA	10	485
C1127210	14	01011100111001011000	AGTATACAGT	10	486
C1127210	14	01001110010110110100	ATACAGTCGA	10	487
C1127210	14	00100111110111001000	CACGTAGTAT	10	488
C1127210	14	00110101001001110011	CGAGCACGCG	10	489
C1127210	14	00111001010101010110	CGTGAGAGAC	10	490
C1127210	14	00111001101010101100	CGTGTCTCTA	10	491
C1127210	14	00101010010010101111	CTCATCTACG	10	492
C1127210	14	11010100110010101010	TAGATATCTC	10	493
C1127210	14	11001001100110010011	TATGTGTGCG	10	494
C1127210	14	10110010101101011000	TCGCTCGAGT	10	495
C1127210	14	10101110100100100110	TCTACTGCAC	10	496
C1127211	14	01100101011001001101	ACAGACATAG	10	497
C1127211	14	01101011001100101001	ACTCGCGCTG	10	498
C1127211	14	01010011001111100100	AGCGCGTACA	10	499
C1127211	14	01011100111001011000	AGTATACAGT	10	500
C1127211	14	01001110010110110100	ATACAGTCGA	10	501

IDT Exhibit 1020.5032

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127211	14	00100111110111001000	CACGTAGTAT	10	502
C1127211	14	00110101001001110011	CGAGCACGCG	10	503
C1127211	14	00111001010101010110	CGTGAGAGAC	10	504
C1127211	14	00111001101010101100	CGTGTCTCTA	10	505
C1127211	14	00101010010010101111	CTCATCTACG	10	506
C1127211	14	11010100110010101010	TAGATATCTC	10	507
C1127211	14	11001001100110010011	TATGTGTGCG	10	508
C1127211	14	10110010101101011000	TCGCTCGAGT	10	509
C1127211	14	10011110100101001001	TGTACTGATG	10	510
C1127212	14	01100101011001001101	ACAGACATAG	10	511
C1127212	14	01101011001100101001	ACTCGCGCTG	10	512
C1127212	14	01101010110010010110	ACTCTATGAC	10	513
C1127212	14	01010011001111100100	AGCGCGTACA	10	514
C1127212	14	01011100111001011000	AGTATACAGT	10	515
C1127212	14	01001110010110110100	ATACAGTCGA	10	516
C1127212	14	00100111110111001000	CACGTAGTAT	10	517
C1127212	14	00110101001001110011	CGAGCACGCG	10	518
C1127212	14	00111001010101010110	CGTGAGAGAC	10	519
C1127212	14	00111001101010101100	CGTGTCTCTA	10	520
C1127212	14	11010100110010101010	TAGATATCTC	10	521
C1127212	14	11001001100110010011	TATGTGTGCG	10	522
C1127212	14	10110010101101011000	TCGCTCGAGT	10	523
C1127212	14	10011110100101001001	TGTACTGATG	10	524
C1127213	14	01100111011010010010	ACACGACTGC	10	525
C1127213	14	01011100111001011000	AGTATACAGT	10	526
C1127213	14	01001110101001001011	ATACTCATCG	10	527
C1127213	14	01001011010011101100	ATCGATACTA	10	528
C1127213	14	00110101001001110011	CGAGCACGCG	10	529
C1127213	14	00111001010101010110	CGTGAGAGAC	10	530
C1127213	14	00111001101010101100	CGTGTCTCTA	10	531
C1127213	14	00101110010110011100	CTACAGTGTA	10	532
C1127213	14	00101010111100100110	CTCTACGCAC	10	533
C1127213	14	11010010101010010110	TAGCTCTGAC	10	534
C1127213	14	10100110010011010101	TCACATAGAG	10	535
C1127213	14	10101001100110010011	TCTGTGTGCG	10	536
C1127213	14	10010100110110110010	TGATAGTCGC	10	537
C1127213	14	10011011001101001001	TGTCGCGATG	10	538
C1127214	14	01100110101010011010	ACACTCTGTC	10	539
C1127214	14	01011100111001011000	AGTATACAGT	10	540
C1127214	14	01001101001110100101	ATAGCGTCAG	10	541
C1127214	14	01001011010011101100	ATCGATACTA	10	542
C1127214	14	00110101001001110011	CGAGCACGCG	10	543
C1127214	14	00111001010101010110	CGTGAGAGAC	10	544
C1127214	14	00111001101010101100	CGTGTCTCTA	10	545
C1127214	14	00101111010110011000	CTACGAGTGT	10	546

IDT Exhibit 1020.5033

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127214	14	00101010111100100110	CTCTACGCAC	10	547
C1127214	14	11101001011001001010	TACTGACATC	10	548
C1127214	14	10100110010011010101	TCACATAGAG	10	549
C1127214	14	10101001100110010011	TCTGTGTGCG	10	550
C1127214	14	10010100110110110010	TGATAGTCGC	10	551
C1127214	14	10011011001101001001	TGTCGCGATG	10	552
C1127215	14	01100100111010111000	ACATACTCGT	10	553
C1127215	14	01010010101111010100	AGCTCGTAGA	10	554
C1127215	14	01011100100110010011	AGTATGTGCG	10	555
C1127215	14	01001010100101111010	ATCTGACGTC	10	556
C1127215	14	00110101001001110011	CGAGCACGCG	10	557
C1127215	14	00111001010101010110	CGTGAGAGAC	10	558
C1127215	14	00111001101010101100	CGTGTCTCTA	10	559
C1127215	14	00101111010110011000	CTACGAGTGT	10	560
C1127215	14	00101010111001100101	CTCTACACAG	10	561
C1127215	14	11100100110101001100	TACATAGATA	10	562
C1127215	14	11010011100100100110	TAGCGTGCAC	10	563
C1127215	14	10100110100110101001	TCACTGTCTG	10	564
C1127215	14	10111010010011101000	TCGTCATACT	10	565
C1127215	14	10011011001101001001	TGTCGCGATG	10	566
C1127216	14	01100100111010110010	ACATACTCGC	10	567
C1127216	14	01010010101111010100	AGCTCGTAGA	10	568
C1127216	14	01011100100110010011	AGTATGTGCG	10	569
C1127216	14	01001010100101111010	ATCTGACGTC	10	570
C1127216	14	00110101001001110011	CGAGCACGCG	10	571
C1127216	14	00111001010101010110	CGTGAGAGAC	10	572
C1127216	14	00111001101010101100	CGTGTCTCTA	10	573
C1127216	14	00101111010110011000	CTACGAGTGT	10	574
C1127216	14	00101010111001100101	CTCTACACAG	10	575
C1127216	14	11100100110101001100	TACATAGATA	10	576
C1127216	14	11010011100100100110	TAGCGTGCAC	10	577
C1127216	14	10100110100110101001	TCACTGTCTG	10	578
C1127216	14	10111010010011101000	TCGTCATACT	10	579
C1127216	14	10011011001101001001	TGTCGCGATG	10	580
C1127217	14	01100110011001101010	ACACACACTC	10	581
C1127217	14	01100100101010011101	ACATCTGTAG	10	582
C1127217	14	01110011001110011000	ACGCGCGTGT	10	583
C1127217	14	01010011010011001110	AGCGATATAC	10	584
C1127217	14	01011101100101010100	AGTAGTGAGA	10	585
C1127217	14	00110101010100110011	CGAGAGCGCG	10	586
C1127217	14	00111010010011111000	CGTCATACGT	10	587
C1127217	14	00111001101010101100	CGTGTCTCTA	10	588
C1127217	14	00101110110110101000	CTACTAGTCT	10	589
C1127217	14	00101011001101100101	CTCGCGACAG	10	590
C1127217	14	11010110101101001000	TAGACTCGAT	10	591

IDT Exhibit 1020.5034

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127217	14	11001001010101011010	TATGAGAGTC	10	592
C1127217	14	10010010100111010011	TGCTGTAGCG	10	593
C1127217	14	10011111001010010010	TGTACGCTGC	10	594
C1127218	14	01100100101010011101	ACATCTGTAG	10	595
C1127218	14	01110011001110011000	ACGCGCGTGT	10	596
C1127218	14	01010011010011001110	AGCGATATAC	10	597
C1127218	14	01011101100101010100	AGTAGTGAGA	10	598
C1127218	14	00110101010100110011	CGAGAGCGCG	10	599
C1127218	14	00111010010011111000	CGTCATACGT	10	600
C1127218	14	00111001101010101100	CGTGTCTCTA	10	601
C1127218	14	00101110110110101000	CTACTAGTCT	10	602
C1127218	14	00101011001101100101	CTCGCGACAG	10	603
C1127218	14	11010110101101001000	TAGACTCGAT	10	604
C1127218	14	11001100111001100100	TATATACACA	10	605
C1127218	14	10100110010101011100	TCACAGAGTA	10	606
C1127218	14	10010010100111010011	TGCTGTAGCG	10	607
C1127218	14	10011111001010010010	TGTACGCTGC	10	608
C1127219	14	01110011001110011000	ACGCGCGTGT	10	609
C1127219	14	01010101011001001110	AGAGACATAC	10	610
C1127219	14	01001101110110101000	ATAGTAGTCT	10	611
C1127219	14	01001010100101110101	ATCTGACGAG	10	612
C1127219	14	00100111001011001011	CACGCTATCG	10	613
C1127219	14	00110101010100110011	CGAGAGCGCG	10	614
C1127219	14	00111010010011111000	CGTCATACGT	10	615
C1127219	14	00111001101010101100	CGTGTCTCTA	10	616
C1127219	14	00101010011110100110	CTCACGTCAC	10	617
C1127219	14	11010010101010010110	TAGCTCTGAC	10	618
C1127219	14	11001110010010011001	TATACATGTG	10	619
C1127219	14	10100110100101011100	TCACTGAGTA	10	620
C1127219	14	10101100111001100100	TCTATACACA	10	621
C1127219	14	10010011110101001001	TGCGTAGATG	10	622
C1127220	14	01100110010011001101	ACACATATAG	10	623
C1127220	14	01110011001110011000	ACGCGCGTGT	10	624
C1127220	14	01010111001001101010	AGACGCACTC	10	625
C1127220	14	01001100111110100100	ATATACGTCA	10	626
C1127220	14	01001001101011010011	ATGTCTAGCG	10	627
C1127220	14	00110101010100110011	CGAGAGCGCG	10	628
C1127220	14	00110010101101100101	CGCTCGACAG	10	629
C1127220	14	00111001101010101100	CGTGTCTCTA	10	630
C1127220	14	00101011110101001010	CTCGTAGATC	10	631
C1127220	14	11101001010100101100	TACTGAGCTA	10	632
C1127220	14	11010011010010010110	TAGCGATGAC	10	633
C1127220	14	10101100100110011001	TCTATGTGTG	10	634
C1127220	14	10101010011010110010	TCTCACTCGC	10	635
C1127220	14	10010101111001011000	TGAGTACAGT	10	636

IDT Exhibit 1020.5035

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127221	14	01100110010011001101	ACACATATAG	10	637
C1127221	14	01110011001110011000	ACGCGCGTGT	10	638
C1127221	14	01010111001001101010	AGACGCACTC	10	639
C1127221	14	01001100111110100100	ATATACGTCA	10	640
C1127221	14	01001001101011010011	ATGTCTAGCG	10	641
C1127221	14	00110101010100110011	CGAGAGCGCG	10	642
C1127221	14	00110010101101100101	CGCTCGACAG	10	643
C1127221	14	00111001101010101100	CGTGTCTCTA	10	644
C1127221	14	00101011110101001010	CTCGTAGATC	10	645
C1127221	14	11101001010100101100	TACTGAGCTA	10	646
C1127221	14	11010011010010010110	TAGCGATGAC	10	647
C1127221	14	10101100100110011001	TCTATGTGTG	10	648
C1127221	14	10101010011010110010	TCTCACTCGC	10	649
C1127221	14	10010101111001010100	TGAGTACAGA	10	650
C1127222	14	01100100110011100101	ACATATACAG	10	651
C1127222	14	01110011001110011000	ACGCGCGTGT	10	652
C1127222	14	01010111001001101010	AGACGCACTC	10	653
C1127222	14	01001001011011010011	ATGACTAGCG	10	654
C1127222	14	00100110010111011100	CACAGTAGTA	10	655
C1127222	14	00110101010100110011	CGAGAGCGCG	10	656
C1127222	14	00110010111010010110	CGCTACTGAC	10	657
C1127222	14	00111001101010101100	CGTGTCTCTA	10	658
C1127222	14	00101011001101100101	CTCGCGACAG	10	659
C1127222	14	11100100101101010010	TACATCGAGC	10	660
C1127222	14	11011110101110000000	TAGTACTCGT	10	661
C1127222	14	10101110011010100100	TCTACACTCA	10	662
C1127222	14	10101100100110011001	TCTATGTGTG	10	663
C1127222	14	10011010010101001110	TGTCAGATAC	10	664
C1127223	14	01100110011001101010	ACACACACTC	10	665
C1127223	14	01100100101010011101	ACATCTGTAG	10	666
C1127223	14	01110011001110011000	ACGCGCGTGT	10	667
C1127223	14	01010011010011001110	AGCGATATAC	10	668
C1127223	14	01011101100101010100	AGTAGTGAGA	10	669
C1127223	14	00110101010100110011	CGAGAGCGCG	10	670
C1127223	14	00111001101010101100	CGTGTCTCTA	10	671
C1127223	14	00101110110110101000	CTACTAGTCT	10	672
C1127223	14	00101011001101100101	CTCGCGACAG	10	673
C1127223	14	11010110101101001000	TAGACTCGAT	10	674
C1127223	14	11001010010010100111	TATCATCACG	10	675
C1127223	14	11001001010101011010	TATGAGAGTC	10	676
C1127223	14	10010010100111010011	TGCTGTAGCG	10	677
C1127223	14	10011111001010010010	TGTACGCTGC	10	678
C1127224	14	01100110101010100101	ACACTCTCAG	10	679
C1127224	14	01110011001110011000	ACGCGCGTGT	10	680
C1127224	14	01111101011001001000	ACGTAGACAT	10	681

IDT Exhibit 1020.5036

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127224	14	01010110010111001100	AGACAGTATA	10	682
C1127224	14	01011100100110010011	AGTATGTGCG	10	683
C1127224	14	00110101001100100111	CGAGCGCACG	10	684
C1127224	14	00111001101010101100	CGTGTCTCTA	10	685
C1127224	14	00101111010010101010	CTACGATCTC	10	686
C1127224	14	00101010111101010100	CTCTACGAGA	10	687
C1127224	14	11011010011100100100	TAGTCACGCA	10	688
C1127224	14	11001101001001011001	TATAGCAGTG	10	689
C1127224	14	10111010100101001010	TCGTCTGATC	10	690
C1127224	14	10101001010011001101	TCTGATATAG	10	691
C1127224	14	10010010011011111000	TGCACTACGT	10	692
C1127225	14	01100101101100100110	ACAGTCGCAC	10	693
C1127225	14	01110011001110011000	ACGCGCGTGT	10	694
C1127225	14	01010110100110010101	AGACTGTGAG	10	695
C1127225	14	01011100111001001001	AGTATACATG	10	696
C1127225	14	00100110111011010010	CACTACTAGC	10	697
C1127225	14	00110101001001110011	CGAGCACGCG	10	698
C1127225	14	00111001101010101100	CGTGTCTCTA	10	699
C1127225	14	00101111001101001100	CTACGCGATA	10	700
C1127225	14	00101010010110011011	CTCAGTGTCG	10	701
C1127225	14	11001011001010100101	TATCGCTCAG	10	702
C1127225	14	11001001110110010100	TATGTAGTGA	10	703
C1127225	14	10110010010011110100	TCGCATACGA	10	704
C1127225	14	10010101010011001110	TGAGATATAC	10	705
C1127225	14	10011011010101101000	TGTCGAGACT	10	706
C1127226	14	01100100111011001001	ACATACTATG	10	707
C1127226	14	01110011001110011000	ACGCGCGTGT	10	708
C1127226	14	01011101010011011000	AGTAGATAGT	10	709
C1127226	14	01001001100101110110	ATGTGACGAC	10	710
C1127226	14	00100101010010111110	CAGATCGTAC	10	711
C1127226	14	00110101001001110011	CGAGCACGCG	10	712
C1127226	14	00111001101010101100	CGTGTCTCTA	10	713
C1127226	14	00101111001101001100	CTACGCGATA	10	714
C1127226	14	00101010010110011011	CTCAGTGTCG	10	715
C1127226	14	11010110011010010100	TAGACACTGA	10	716
C1127226	14	11001010110100111000	TATCTAGCGT	10	717
C1127226	14	10111100100101010100	TCGTATGAGA	10	718
C1127226	14	10101011001010100110	TCTCGCTCAC	10	719
C1127226	14	10010010110011101010	TGCTATACTC	10	720
C1127227	14	01100101101100100110	ACAGTCGCAC	10	721
C1127227	14	01110011001110011000	ACGCGCGTGT	10	722
C1127227	14	01011100110101100100	AGTATAGACA	10	723
C1127227	14	01001001011011100101	ATGACTACAG	10	724
C1127227	14	00100111110101001001	CACGTAGATG	10	725
C1127227	14	00110101010100110101	CGAGAGCGAG	10	726

IDT Exhibit 1020.5037

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127227	14	00111010010011111000	CGTCATACGT	10	727
C1127227	14	00111001101010101100	CGTGTCTCTA	10	728
C1127227	14	00101010011100101011	CTCACGCTCG	10	729
C1127227	14	11110100101001101000	TACGATCACT	10	730
C1127227	14	11010010010010011101	TAGCATGTAG	10	731
C1127227	14	11001110010100110010	TATACAGCGC	10	732
C1127227	14	10110101010011001010	TCGAGATATC	10	733
C1127227	14	10101001100110010011	TCTGTGTGCG	10	734
C1127228	14	01100101101100100110	ACAGTCGCAC	10	735
C1127228	14	01110011001110011000	ACGCGCGTGT	10	736
C1127228	14	01010110100101011010	AGACTGAGTC	10	737
C1127228	14	01001010111001010101	ATCTACAGAG	10	738
C1127228	14	00100111110101001001	CACGTAGATG	10	739
C1127228	14	00110101010100110101	CGAGAGCGAG	10	740
C1127228	14	00111010010011111000	CGTCATACGT	10	741
C1127228	14	00111001101010101100	CGTGTCTCTA	10	742
C1127228	14	00101010011100101011	CTCACGCTCG	10	743
C1127228	14	11110100101001101000	TACGATCACT	10	744
C1127228	14	11010010010010011101	TAGCATGTAG	10	745
C1127228	14	11001111001001001100	TATACGCATA	10	746
C1127228	14	10110101010011001010	TCGAGATATC	10	747
C1127228	14	10101001100110010011	TCTGTGTGCG	10	748
C1127229	14	01100110110010101001	ACACTATCTG	10	749
C1127229	14	01110011001110011000	ACGCGCGTGT	10	750
C1127229	14	01010011010011001110	AGCGATATAC	10	751
C1127229	14	01011010101101110000	AGTCTCGACG	10	752
C1127229	14	01001110010101011010	ATACAGAGTC	10	753
C1127229	14	00100111100101110010	CACGTGACGC	10	754
C1127229	14	00110101010100110101	CGAGAGCGAG	10	755
C1127229	14	00111001101010101100	CGTGTCTCTA	10	756
C1127229	14	00101010110111010100	CTCTAGTAGA	10	757
C1127229	14	11011010010111001000	TAGTCAGTAT	10	758
C1127229	14	11001100111001100100	TATATACACA	10	759
C1127229	14	10101011001010100110	TCTCGCTCAC	10	760
C1127229	14	10101001100110010011	TCTGTGTGCG	10	761
C1127229	14	10010101001011111000	TGAGCTACGT	10	762
C1127230	14	01110011001110010100	ACGCGCGTGA	10	763
C1127230	14	01101001110101010010	ACTGTAGAGC	10	764
C1127230	14	01010101011001001110	AGAGACATAC	10	765
C1127230	14	01001111100100101010	ATACGTGCTC	10	766
C1127230	14	00100110111001101001	CACTACACTG	10	767
C1127230	14	00110101010100110011	CGAGAGCGCG	10	768
C1127230	14	00111010010011111000	CGTCATACGT	10	769
C1127230	14	00111001101010101100	CGTGTCTCTA	10	770
C1127230	14	00101010010110011101	CTCAGTGTAG	10	771

IDT Exhibit 1020.5038

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127230	14	11010100110110011000	TAGATAGTGT	10	772
C1127230	14	11001011011100100100	TATCGACGCA	10	773
C1127230	14	10101111010011001000	TCTACGATAT	10	774
C1127230	14	10101100101010010011	TCTATCTGCG	10	775
C1127230	14	10010010011010110110	TGCACTCGAC	10	776
C1127231	14	01110011001110010100	ACGCGCGTGA	10	777
C1127231	14	01010111001001101010	AGACGCACTC	10	778
C1127231	14	01011101010111001000	AGTAGAGTAT	10	779
C1127231	14	01001011100101100101	ATCGTGACAG	10	780
C1127231	14	01001001011011010011	ATGACTAGCG	10	781
C1127231	14	00100110011010011101	CACACTGTAG	10	782
C1127231	14	00110101010100110011	CGAGAGCGCG	10	783
C1127231	14	00111010010011111000	CGTCATACGT	10	784
C1127231	14	00111001101010101100	CGTGTCTCTA	10	785
C1127231	14	11011010011100100100	TAGTCACGCA	10	786
C1127231	14	11001010110101001010	TATCTAGATC	10	787
C1127231	14	10100111010011100100	TCACGATACA	10	788
C1127231	14	10101100100110011001	TCTATGTGTG	10	789
C1127231	14	10010010010110110110	TGCAGTCGAC	10	790
C1127232	14	01110011001110010100	ACGCGCGTGA	10	791
C1127232	14	01101001010111101000	ACTGAGTACT	10	792
C1127232	14	01010100101111001010	AGATCGTATC	10	793
C1127232	14	01001110101010100101	ATACTCTCAG	10	794
C1127232	14	01001001111001010011	ATGTACAGCG	10	795
C1127232	14	00110101010100110011	CGAGAGCGCG	10	796
C1127232	14	00110010010111001101	CGCAGTATAG	10	797
C1127232	14	00111001101010101100	CGTGTCTCTA	10	798
C1127232	14	00101111011001001100	CTACGACATA	10	799
C1127232	14	11010101011010011000	TAGAGACTGT	10	800
C1127232	14	11010011100100100110	TAGCGTGCAC	10	801
C1127232	14	10111010110101001000	TCGTCTAGAT	10	802
C1127232	14	10101100100110011001	TCTATGTGTG	10	803
C1127232	14	10101010011010110010	TCTCACTCGC	10	804
C1127233	14	01100101101100100110	ACAGTCGCAC	10	805
C1127233	14	01101001011001011001	ACTGACAGTG	10	806
C1127233	14	01010100110101101100	AGATAGACTA	10	807
C1127233	14	01010011001011100101	AGCGCTACAG	10	808
C1127233	14	01001010110010101011	ATCTATCTCG	10	809
C1127233	14	00110101010100110011	CGAGAGCGCG	10	810
C1127233	14	00111010010011111000	CGTCATACGT	10	811
C1127233	14	00111001101010101100	CGTGTCTCTA	10	812
C1127233	14	00101111100101100100	CTACGTGACA	10	813
C1127233	14	11011110100100101000	TAGTACTGCT	10	814
C1127233	14	11001010101001010110	TATCTCAGAC	10	815
C1127233	14	10100101010111010100	TCAGAGTAGA	10	816

IDT Exhibit 1020.5039

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127233	14	10110010100110011001	TCGCTGTGTG	10	817
C1127233	14	10010011011101001010	TGCGACGATC	10	818
C1127234	14	01100110011010011001	ACACACTGTG	10	819
C1127234	14	01110010110101001100	ACGCTAGATA	10	820
C1127234	14	01101001100100111010	ACTGTGCGTC	10	821
C1127234	14	01001111001001110100	ATACGCACGA	10	822
C1127234	14	00110101010100110011	CGAGAGCGCG	10	823
C1127234	14	00110010011110100110	CGCACGTCAC	10	824
C1127234	14	00111010010011111000	CGTCATACGT	10	825
C1127234	14	00111001101010101100	CGTGTCTCTA	10	826
C1127234	14	00101110110110101000	CTACTAGTCT	10	827
C1127234	14	11100100101101010010	TACATCGAGC	10	828
C1127234	14	11011001110011100000	TAGTGTATAC	10	829
C1127234	14	10101100100101100101	TCTATGACAG	10	830
C1127234	14	10010100111001101010	TGATACACTC	10	831
C1127234	14	10011011010110010100	TGTCGAGTGA	10	832
C1127235	14	01110100101010010101	ACGATCTGAG	10	833
C1127235	14	01101011001100101001	ACTCGCGCTG	10	834
C1127235	14	01011100110101100100	AGTATAGACA	10	835
C1127235	14	01001110101001001011	ATACTCATCG	10	836
C1127235	14	01001001100101111010	ATGTGACGTC	10	837
C1127235	14	00100111110111001000	CACGTAGTAT	10	838
C1127235	14	00110101010100110011	CGAGAGCGCG	10	839
C1127235	14	00110010011110100110	CGCACGTCAC	10	840
C1127235	14	00111010010011111000	CGTCATACGT	10	841
C1127235	14	00111001101010101100	CGTGTCTCTA	10	842
C1127235	14	11010100110010101010	TAGATATCTC	10	843
C1127235	14	11010010011101011000	TAGCACGAGT	10	844
C1127235	14	11001101100110010100	TATAGTGTGA	10	845
C1127235	14	10101010100101001110	TCTCTGATAC	10	846
C1127236	14	01100100101010111001	ACATCTCGTG	10	847
C1127236	14	01110010100101100101	ACGCTGACAG	10	848
C1127236	14	01010111001001101010	AGACGCACTC	10	849
C1127236	14	01011110110010011000	AGTACTATGT	10	850
C1127236	14	01001011010101011100	ATCGAGAGTA	10	851
C1127236	14	00100101001110011110	CAGCGTGTAC	10	852
C1127236	14	00110101010100110011	CGAGAGCGCG	10	853
C1127236	14	00111001101010101100	CGTGTCTCTA	10	854
C1127236	14	00101010011001110110	CTCACACGAC	10	855
C1127236	14	11100101010011010010	TACAGATAGC	10	856
C1127236	14	11011010011100100100	TAGTCACGCA	10	857
C1127236	14	10101110100100101010	TCTACTGCTC	10	858
C1127236	14	10010011001110111000	TGCGCGTCGT	10	859
C1127236	14	10011100101001010101	TGTATCAGAG	10	860
C1127237	14	01101010110010100110	ACTCTATCAC	10	861

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127237	14	01101001010011011001	ACTGATAGTG	10	862
C1127237	14	01010111001001101010	AGACGCACTC	10	863
C1127237	14	01011101100101010100	AGTAGTGAGA	10	864
C1127237	14	00100101001110011110	CAGCGTGTAC	10	865
C1127237	14	00110101010100110011	CGAGAGCGCG	10	866
C1127237	14	00111001101010101100	CGTGTCTCTA	10	867
C1127237	14	00101010101101010101	CTCTCGAGAG	10	868
C1127237	14	11100101011101100000	TACAGACGAC	10	869
C1127237	14	11011101010010101000	TAGTAGATCT	10	870
C1127237	14	11001010011010010011	TATCACTGCG	10	871
C1127237	14	10101110011001001100	TCTACACATA	10	872
C1127237	14	10010100100111001101	TGATGTATAG	10	873
C1127237	14	10010011001110111000	TGCGCGTCGT	10	874
C1127238	14	01101010110010100110	ACTCTATCAC	10	875
C1127238	14	01101001010011011001	ACTGATAGTG	10	876
C1127238	14	01010111001001101010	AGACGCACTC	10	877
C1127238	14	01011101100101010100	AGTAGTGAGA	10	878
C1127238	14	00100101001110011110	CAGCGTGTAC	10	879
C1127238	14	00110101010100110011	CGAGAGCGCG	10	880
C1127238	14	00111001101010101100	CGTGTCTCTA	10	881
C1127238	14	00101010101101010101	CTCTCGAGAG	10	882
C1127238	14	11100110100100101001	TACACTGCTG	10	883
C1127238	14	11100101011101100000	TACAGACGAC	10	884
C1127238	14	11011101010010101000	TAGTAGATCT	10	885
C1127238	14	10101110011001001100	TCTACACATA	10	886
C1127238	14	10010100100111001101	TGATGTATAG	10	887
C1127238	14	10010011001110111000	TGCGCGTCGT	10	888
C1127239	14	01110010010010101110	ACGCATCTAC	10	889
C1127239	14	01101011001100101001	ACTCGCGCTG	10	890
C1127239	14	01010110100110011010	AGACTGTGTC	10	891
C1127239	14	01001100111100100110	ATATACGCAC	10	892
C1127239	14	00100111110110010100	CACGTAGTGA	10	893
C1127239	14	00110101010100110011	CGAGAGCGCG	10	894
C1127239	14	00111011001001010110	CGTCGCAGAC	10	895
C1127239	14	00111001101010101100	CGTGTCTCTA	10	896
C1127239	14	00101010111001001011	CTCTACATCG	10	897
C1127239	14	11100101010011010010	TACAGATAGC	10	898
C1127239	14	11011010100101100100	TAGTCTGACA	10	899
C1127239	14	10101100101101011000	TCTATCGAGT	10	900
C1127239	14	10010100110011101001	TGATATACTG	10	901
C1127239	14	10010011001110111000	TGCGCGTCGT	10	902
C1127240	14	01101011100110010100	ACTCGTGTGA	10	903
C1127240	14	01101001011001111000	ACTGACACGT	10	904
C1127240	14	01010010011011101010	AGCACTACTC	10	905
C1127240	14	01001110010101001110	ATACAGATAC	10	906

IDT Exhibit 1020.5041

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127240	14	01001100100110111010	ATATGTCGTC	10	907
C1127240	14	00110101010100110011	CGAGAGCGCG	10	908
C1127240	14	00111011001001010110	CGTCGCAGAC	10	909
C1127240	14	00111001101010101100	CGTGTCTCTA	10	910
C1127240	14	00101010111111001000	CTCTACGTAT	10	911
C1127240	14	11011010011100100100	TAGTCACGCA	10	912
C1127240	14	10100110011011010100	TCACACTAGA	10	913
C1127240	14	10101100101010010011	TCTATCTGCG	10	914
C1127240	14	10010100100111001101	TGATGTATAG	10	915
C1127240	14	10010011001110111000	TGCGCGTCGT	10	916
C1127241	14	01101011100110010100	ACTCGTGTGA	10	917
C1127241	14	01101001001010100111	ACTGCTCACG	10	918
C1127241	14	01011100111001011000	AGTATACAGT	10	919
C1127241	14	01001110010111100100	ATACAGTACA	10	920
C1127241	14	01001101101101001010	ATAGTCGATC	10	921
C1127241	14	00100111010011101010	CACGATACTC	10	922
C1127241	14	00110101010100110011	CGAGAGCGCG	10	923
C1127241	14	00111011001001010110	CGTCGCAGAC	10	924
C1127241	14	00111001101010101100	CGTGTCTCTA	10	925
C1127241	14	00101010111111001000	CTCTACGTAT	10	926
C1127241	14	11001010110100110010	TATCTAGCGC	10	927
C1127241	14	11001001010101001101	TATGAGATAG	10	928
C1127241	14	10100110011010010101	TCACACTGAG	10	929
C1127241	14	10010011001110111000	TGCGCGTCGT	10	930
C1127242	14	01101011100110010100	ACTCGTGTGA	10	931
C1127242	14	01101001001010100111	ACTGCTCACG	10	932
C1127242	14	01011100111001011000	AGTATACAGT	10	933
C1127242	14	01001101101101001010	ATAGTCGATC	10	934
C1127242	14	00100111010011101010	CACGATACTC	10	935
C1127242	14	00110101010100110011	CGAGAGCGCG	10	936
C1127242	14	00111011001001010110	CGTCGCAGAC	10	937
C1127242	14	00111001101010101100	CGTGTCTCTA	10	938
C1127242	14	00101010111111001000	CTCTACGTAT	10	939
C1127242	14	11001010110100110010	TATCTAGCGC	10	940
C1127242	14	11001001010101001101	TATGAGATAG	10	941
C1127242	14	10100110011011010100	TCACACTAGA	10	942
C1127242	14	10100101110010011001	TCAGTATGTG	10	943
C1127242	14	10010011001110111000	TGCGCGTCGT	10	944
C1127243	14	01100100111010100110	ACATACTCAC	10	945
C1127243	14	01110010011101001100	ACGCACGATA	10	946
C1127243	14	01101001100110010101	ACTGTGTGAG	10	947
C1127243	14	01011100111001011000	AGTATACAGT	10	948
C1127243	14	01001111110010010010	ATACGTATGC	10	949
C1127243	14	00110101010100110011	CGAGAGCGCG	10	950
C1127243	14	00111011001001010110	CGTCGCAGAC	10	951

IDT Exhibit 1020.5042

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127243	14	00111001101010101100	CGTGTCTCTA	10	952
C1127243	14	00101010111001001011	CTCTACATCG	10	953
C1127243	14	11100110100100101001	TACACTGCTG	10	954
C1127243	14	11001001010101101010	TATGAGACTC	10	955
C1127243	14	10100101001111100100	TCAGCGTACA	10	956
C1127243	14	10101011001100111000	TCTCGCGCGT	10	957
C1127243	14	10010110110111001000	TGACTAGTAT	10	958
C1127244	14	01110010110101101000	ACGCTAGACT	10	959
C1127244	14	01011011100101010100	AGTCGTGAGA	10	960
C1127244	14	01001100101001001111	ATATCATACG	10	961
C1127244	14	00100100110111011001	CATAGTAGTG	10	962
C1127244	14	00110101010101001110	CGAGAGATAC	10	963
C1127244	14	00111001101010101100	CGTGTCTCTA	10	964
C1127244	14	00101011001110011010	CTCGCGTGTC	10	965
C1127244	14	00101010111100100101	CTCTACGCAG	10	966
C1127244	14	11100110011001001100	TACACACATA	10	967
C1127244	14	11101001100111001000	TACTGTGTAT	10	968
C1127244	14	11010010101010110010	TAGCTCTCGC	10	969
C1127244	14	10101101010010110100	TCTAGATCGA	10	970
C1127244	14	10010110010100110011	TGACAGCGCG	10	971
C1127244	14	10010101111001011000	TGAGTACAGT	10	972
C1127245	14	01110101100100111000	ACGAGTGCGT	10	973
C1127245	14	01110010101101100100	ACGCTCGACA	10	974
C1127245	14	01010111001001101010	AGACGCACTC	10	975
C1127245	14	01010010011010011101	AGCACTGTAG	10	976
C1127245	14	01001100101100110101	ATATCGCGAG	10	977
C1127245	14	01001010010101100111	ATCAGACACG	10	978
C1127245	14	00100100110111011001	CATAGTAGTG	10	979
C1127245	14	00110101010101001110	CGAGAGATAC	10	980
C1127245	14	00111001101010101100	CGTGTCTCTA	10	981
C1127245	14	00101011001110011010	CTCGCGTGTC	10	982
C1127245	14	11101001010100101100	TACTGAGCTA	10	983
C1127245	14	11011100101001010010	TAGTATCAGC	10	984
C1127245	14	10101010110010010011	TCTCTATGCG	10	985
C1127245	14	10010100111110101000	TGATACGTCT	10	986
C1127246	14	01100110100111001100	ACACTGTATA	10	987
C1127246	14	01100101001101111000	ACAGCGACGT	10	988
C1127246	14	01100100111010100110	ACATACTCAC	10	989
C1127246	14	01001011001100101101	ATCGCGCTAG	10	990
C1127246	14	01001001110011011010	ATGTATAGTC	10	991
C1127246	14	00110110010010111001	CGACATCGTG	10	992
C1127246	14	00110101010101001110	CGAGAGATAC	10	993
C1127246	14	00111001101010101100	CGTGTCTCTA	10	994
C1127246	14	00101010111101010100	CTCTACGAGA	10	995
C1127246	14	11101101010010101000	TACTAGATCT	10	996

IDT Exhibit 1020.5043

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127246	14	11010100100110010101	TAGATGTGAG	10	997
C1127246	14	10101001101001010011	TCTGTCAGCG	10	998
C1127246	14	10010011001110011010	TGCGCGTGTC	10	999
C1127246	14	10011011111001001000	TGTCGTACAT	10	1000
C1127247	14	01100110100111001100	ACACTGTATA	10	1001
C1127247	14	01010110011010111000	AGACACTCGT	10	1002
C1127247	14	01001011001100101101	ATCGCGCTAG	10	1003
C1127247	14	01001001110011011010	ATGTATAGTC	10	1004
C1127247	14	00110101010101001110	CGAGAGATAC	10	1005
C1127247	14	00111010100101010011	CGTCTGAGCG	10	1006
C1127247	14	00111001101010101100	CGTGTCTCTA	10	1007
C1127247	14	00101111101101100000	CTACGTCGAC	10	1008
C1127247	14	11101101001010100100	TACTAGCTCA	10	1009
C1127247	14	11010100100110010101	TAGATGTGAG	10	1010
C1127247	14	10100101101001110010	TCAGTCACGC	10	1011
C1127247	14	10101110010101011000	TCTACAGAGT	10	1012
C1127247	14	10010011001110011010	TGCGCGTGTC	10	1013
C1127247	14	10011011111001001000	TGTCGTACAT	10	1014
C1127248	14	01101010101101100100	ACTCTCGACA	10	1015
C1127248	14	01010100111001101001	AGATACACTG	10	1016
C1127248	14	01001111110010101000	ATACGTATCT	10	1017
C1127248	14	01001001100101111010	ATGTGACGTC	10	1018
C1127248	14	00100111011010010101	CACGACTGAG	10	1019
C1127248	14	00110101010101001110	CGAGAGATAC	10	1020
C1127248	14	00110010011110111000	CGCACGTCGT	10	1021
C1127248	14	00111001101010101100	CGTGTCTCTA	10	1022
C1127248	14	00101001110011010011	CTGTATAGCG	10	1023
C1127248	14	11010100100110011100	TAGATGTGTA	10	1024
C1127248	14	11001010010110100110	TATCAGTCAC	10	1025
C1127248	14	10110110111010100000	TCGACTACTC	10	1026
C1127248	14	10111101001001010100	TCGTAGCAGA	10	1027
C1127248	14	10010101001100110011	TGAGCGCGCG	10	1028
C1127249	14	01100101001011100101	ACAGCTACAG	10	1029
C1127249	14	01010010101010011011	AGCTCTGTCG	10	1030
C1127249	14	01001111100100101010	ATACGTGCTC	10	1031
C1127249	14	01001001110011111000	ATGTATACGT	10	1032
C1127249	14	00100111010110011001	CACGAGTGTG	10	1033
C1127249	14	00110101010101001110	CGAGAGATAC	10	1034
C1127249	14	00111001101010101100	CGTGTCTCTA	10	1035
C1127249	14	00101010011111010100	CTCACGTAGA	10	1036
C1127249	14	11010110011010010100	TAGACACTGA	10	1037
C1127249	14	11001010110101001100	TATCTAGATA	10	1038
C1127249	14	10100100110010110011	TCATATCGCG	10	1039
C1127249	14	10101001100110010110	TCTGTGTGAC	10	1040
C1127249	14	10010011101101011000	TGCGTCGAGT	10	1041

IDT Exhibit 1020.5044

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127249	14	10011011001001100110	TGTCGCACAC	10	1042
C1127250	14	01100110011001111000	ACACACACGT	10	1043
C1127250	14	01101010101101001010	ACTCTCGATC	10	1044
C1127250	14	01010011001110011001	AGCGCGTGTG	10	1045
C1127250	14	01001001011011010011	ATGACTAGCG	10	1046
C1127250	14	00110101010101001110	CGAGAGATAC	10	1047
C1127250	14	00111010010010010111	CGTCATGACG	10	1048
C1127250	14	00111001101010101100	CGTGTCTCTA	10	1049
C1127250	14	00101111101010010010	CTACGTCTGC	10	1050
C1127250	14	11100101001100101100	TACAGCGCTA	10	1051
C1127250	14	11011010010111001000	TAGTCAGTAT	10	1052
C1127250	14	10100110100111010100	TCACTGTAGA	10	1053
C1127250	14	10101001100100110011	TCTGTGCGCG	10	1054
C1127250	14	10010100111010011010	TGATACTGTC	10	1055
C1127250	14	10011011001001100110	TGTCGCACAC	10	1056
C1127251	14	01100110011001111000	ACACACACGT	10	1057
C1127251	14	01101010101101001010	ACTCTCGATC	10	1058
C1127251	14	01010011001110011001	AGCGCGTGTG	10	1059
C1127251	14	01001001011011010011	ATGACTAGCG	10	1060
C1127251	14	00110101010101001110	CGAGAGATAC	10	1061
C1127251	14	00111010010010010111	CGTCATGACG	10	1062
C1127251	14	00111001101010101100	CGTGTCTCTA	10	1063
C1127251	14	00101111101010010010	CTACGTCTGC	10	1064
C1127251	14	11100101001100101100	TACAGCGCTA	10	1065
C1127251	14	11011010010111001000	TAGTCAGTAT	10	1066
C1127251	14	10110100100111010100	TCGATGTAGA	10	1067
C1127251	14	10101001100100110011	TCTGTGCGCG	10	1068
C1127251	14	10010100111010011010	TGATACTGTC	10	1069
C1127251	14	10011011001001100110	TGTCGCACAC	10	1070
C1127252	14	01100110011001111000	ACACACACGT	10	1071
C1127252	14	01101010101101001010	ACTCTCGATC	10	1072
C1127252	14	01010011001110011001	AGCGCGTGTG	10	1073
C1127252	14	01001001011011010011	ATGACTAGCG	10	1074
C1127252	14	00110101010101001110	CGAGAGATAC	10	1075
C1127252	14	00111010010010010111	CGTCATGACG	10	1076
C1127252	14	00111001101010101100	CGTGTCTCTA	10	1077
C1127252	14	00101111101010010010	CTACGTCTGC	10	1078
C1127252	14	11100101001100101100	TACAGCGCTA	10	1079
C1127252	14	11011010010111001000	TAGTCAGTAT	10	1080
C1127252	14	10101100110011010100	TCTATATAGA	10	1081
C1127252	14	10101001100100110011	TCTGTGCGCG	10	1082
C1127252	14	10010100111010011010	TGATACTGTC	10	1083
C1127252	14	10011011001001100110	TGTCGCACAC	10	1084
C1127253	14	01100110011001111000	ACACACACGT	10	1085
C1127253	14	01101010101101001010	ACTCTCGATC	10	1086

IDT Exhibit 1020.5045

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127253	14	01010011001110011001	AGCGCGTGTG	10	1087
C1127253	14	01001001011011010011	ATGACTAGCG	10	1088
C1127253	14	00110101010101001110	CGAGAGATAC	10	1089
C1127253	14	00111010010010010111	CGTCATGACG	10	1090
C1127253	14	00111001101010101100	CGTGTCTCTA	10	1091
C1127253	14	00101111101010010010	CTACGTCTGC	10	1092
C1127253	14	11100101001100101100	TACAGCGCTA	10	1093
C1127253	14	11011010010111001000	TAGTCAGTAT	10	1094
C1127253	14	11001100110011010100	TATATATAGA	10	1095
C1127253	14	10101001100100110011	TCTGTGCGCG	10	1096
C1127253	14	10010100111010011010	TGATACTGTC	10	1097
C1127253	14	10011011001001100110	TGTCGCACAC	10	1098

Example 4- Exemplary computer code for representing and manipulating nucleotide sequences for UID identification

```
package com.fourfivefour.amplicons;

import java.util.HashSet;
import java.util.Set;

/**
 * Code to implement common operations on Nucleotide Sequences
 *
 *
 *
 */
public class Sequence implements Comparable<Sequence> {
        private String sequence;
        static final char possibleBases[] = { 'A', 'C', 'T', 'G' };

        public Sequence(String sequence) {
                this.sequence = sequence.toUpperCase();
        }

        public String getSequence() {
                return sequence;
        }

        public int hashCode() {
                return sequence.hashCode();
        }

        public boolean equals(Object obj) {
                return ((this == obj) ||
```

```
                ((obj instanceof Sequence) &&
                 sequence.equals(((Sequence) obj).sequence)));
        }

        public int compareTo(Sequence obj) {
                return sequence.compareTo(obj.sequence);
        }

        public String toString() {
                return sequence;
        }


        /**
         * Generate the set of all single base insertions for the
         * Sequence.
         *
         * @return          A set of Sequences representing all single base
         *          insertions of the Sequence.
         */
        public Set<Sequence> generateSingleInsertions() {
                Set<Sequence> insertions = new HashSet<Sequence>();

                int seqLen = sequence.length();
                for (int insertIdx = 0; insertIdx <= seqLen; insertIdx++) {
                        String prefixString = sequence.substring(0, insertIdx);
                        String suffixString = sequence.substring(insertIdx,seqLen);
                        for (char insertBase : possibleBases) {
                                insertions.add(new Sequence(prefixString + insertBase +
suffixString));
                        }
                }
                return insertions;
        }

        /**
         * Generate the set of all single base substitutions for the
         * Sequence.
         *
         * @return          A set of Sequences representing all single base
         *          substitutions of the Sequence.
         */
        public Set<Sequence> generateSingleSubstitutions() {
                Set<Sequence> substitutions = new HashSet<Sequence>();

                int seqLen = sequence.length();
                for (int substBaseIdx = 0; substBaseIdx < seqLen; substBaseIdx++) {
                        String prefixString =
                                sequence.substring(0, substBaseIdx);
                        String suffixString =
```

```
                        sequence.substring(substBaseIdx + 1, seqLen);
                    char originalBase =
                        sequence.charAt(substBaseIdx);
                    for (char substBase : possibleBases) {
                        if (substBase != originalBase) {
                            substitutions.add(
                             new Sequence(prefixString + substBase + suffixString)
                            );
                        }
                    }
                }
                return substitutions;
            }

            /**
             * Generate the set of all single base deletions for the
             * Sequence.
             *
             * @return              A set of sequences representing all single base
             *          deletions of the Sequence.
             */
            public Set<Sequence> generateSingleDeletions() {
                Set<Sequence> deletions = new HashSet<Sequence>();

                int seqLen = sequence.length();
                for (int deleteBaseIdx = 0; deleteBaseIdx < seqLen; deleteBaseIdx++) {
                    String prefixString =
                        sequence.substring(0, deleteBaseIdx);
                    String suffixString =
                        sequence.substring(deleteBaseIdx + 1, seqLen);
                    deletions.add(new Sequence(prefixString + suffixString));
                }
                return deletions;
            }

            /**
             * Generate all 1-base mutations starting from each of the sequences in
             * the input set of sequences.
             *
             * @param inputSeqs  The input set of sequences.
             * @return      A set of sequences that are exactly one mutation
             *          away from each of the sequences in the input set
             *          of sequences.
             */
            public static Set<Sequence> generateSingleMutations(Set<Sequence> inputSeqs) {
                Set<Sequence> mutatedSequences = new HashSet<Sequence>();
                for (Sequence inputSeq : inputSeqs) {
                    mutatedSequences.addAll(inputSeq.generateSingleDeletions());
                    mutatedSequences.addAll(inputSeq.generateSingleInsertions());
                    mutatedSequences.addAll(inputSeq.generateSingleSubstitutions());
```

```
            }
            return mutatedSequences;
        }
}
```

As stated previously, it will be appreciated that the foregoing computer code is provided for the purposes of example, and that numerous alternative methods and code structures may be employed. It will also be appreciated that the exemplary code provided herein is not intended to execute as a stand alone application or to run perfectly without additional computer code or modification.

Having described various embodiments and implementations, it should be apparent to those skilled in the relevant art that the foregoing is illustrative only and not limiting, having been presented by way of example only. Many other schemes for distributing functions among the various functional elements of the illustrated embodiment are possible. The functions of any element may be carried out in various ways in alternative embodiments.

IDT Exhibit 1020.5049

CLAIMS

What is claimed is:

1. An identifier element for identifying an origin of a template nucleic acid molecule, comprising:

a nucleic acid element comprising a sequence composition that enables detection of an introduced error in sequence data generated from the nucleic acid element and correction of the introduced error, wherein the nucleic acid element is constructed to couple with the end of a template nucleic acid molecule and identifies an origin of the template nucleic acid molecule.

2. The identifier element of claim 1, wherein:

the sequence composition enables detection of up to three of the introduced errors and correction for up to two of the introduced errors.

3. The identifier element of claim 1, wherein:

The sequence composition comprises 10 sequence positions.

4. The identifier element of claim 1, wherein:

the introduced error is selected from the group consisting of an insertion error, a deletion error, and a substitution error.

5. The identifier element of claim 1, wherein:

the sequence composition comprises a design based upon a set of parameters selected from the group consisting of minimum sequence length, minimum number of flow cycles, sequence distinctiveness, and monomer repeats.

6. The identifier element of claim 1, wherein:

the sequence composition comprises a design based upon a set of parameters selected from the group consisting of melting temperature, Gibbs free energy, hairpin formation, and dimer formation.

7. The identifier element of claim 1, wherein:

IDT Exhibit 1020.5050

the nucleic acid element is incorporated into an adaptor comprising a primer element, wherein the adaptor couples with the end of the template nucleic acid molecule.

8. The identifier element of claim 7, wherein:
the nucleic acid element is in a known position relative to the primer element.

9. The identifier element of claim 7, wherein:
the primer element is selected from the group consisting of an amplification primer, a sequencing primer, or a bipartite amplification – sequencing primer.

10. The identifier element of claim 7, wherein:
the adaptor comprises a quality control element.

11. The identifier element of claim 7, wherein:
the nucleic acid element is in a known position relative to the quality control element.

12. The identifier element of claim 1, wherein:
the origin of the template nucleic acid molecule comprises an experimental sample or diagnostic sample.

13. The identifier element of claim 1, wherein:
the nucleic acid element belongs to a set comprising a plurality of compatible nucleic acid elements each comprising a distinctive sequence composition, wherein the detection of the introduced error is relative to the sequence composition of the compatible nucleic acid elements of the set.

14. The identifier element of claim 13, wherein:
the set comprises 14 of the compatible nucleic acid elements.

15. A method for identifying an origin of a template nucleic acid molecule, comprising the steps of:
identifying a first identifier sequence from sequence data generated from a template nucleic acid molecule;

IDT Exhibit 1020.5051

detecting an introduced error in the first identifier sequence;

correcting the introduced error in the first identifier sequence;

associating the corrected first identifier sequence with a first identifier element coupled to the template molecule; and

identifying an origin of the template molecule using the association of the corrected first identifier sequence with the first identifier element.

16. The method of claim 15, further comprising:

sequencing a template nucleic acid molecule to generate the sequence data.

17. The method of claim 15, wherein:

the template nucleic acid molecule is included in a multiplex sample comprising a plurality of template molecules from a plurality of different origins.

18. The method of claim 15, further comprising:

detecting up to three of the introduced errors in the first identifier sequence; and

correcting up to two of the introduced errors in the first identifier sequence.

19. The method of claim 15, wherein:

the introduced error is selected from the group consisting of an insertion error, a deletion error, and a substitution error.

20. The method of claim 15, wherein the step of detecting comprises:

measuring one or more characteristics of sequence composition in one or more sequence regions that flank the identifier sequence; and

detecting the introduced error using one or more assumptions derived from the measured characteristics.

21. The method of claim 15, wherein:

the first identifier element is incorporated into an adaptor comprising a primer element, wherein the adaptor is coupled to the template nucleic acid molecule.

22. The method of claim 21, wherein:

IDT Exhibit 1020.5052

the first identifier element is in a known position relative to the primer element.

23. The method of claim 21, wherein:

the primer element is selected from the group consisting of an amplification primer, a sequencing primer, or a bipartite amplification – sequencing primer.

24. The method of claim 21, wherein:

the adaptor comprises a quality control element.

25. The method of claim 21, wherein:

the first identifier element is in a known position relative to the quality control element.

26. The method of claim 15, wherein:

the origin of the template nucleic acid molecule comprises an experimental sample or diagnostic sample.

27. The method of claim 15, further comprising the steps of:

identifying a second identifier sequence from the sequence data generated from the template nucleic acid molecule;

detecting an introduced error in the second identifier sequence;

correcting the introduced error in the second identifier sequence;

associating the corrected second identifier sequence with a second identifier element coupled with the template nucleic acid molecule; and

identifying an origin of the template nucleic acid molecule using the association of the corrected second identifier sequence with the second identifier element combinatorially with the association of the corrected first identifier sequence with the first identifier element.

28. The method of claim 27, further comprising:

detecting up to three of the introduced errors in the second identifier sequence; and

correcting up to two of the introduced errors in the second identifier sequence.

29. The method of claim 15, wherein:

IDT Exhibit 1020.5053

the introduced error is selected from the group consisting of an insertion error, a deletion error, and a substitution error.

30. The method of claim 15, wherein:

the first identifier belongs to at least one set of compatible identifiers of a plurality of sets of identifiers.

31. The method of claim 15, wherein:

the set of compatible identifiers comprise 14 identifiers that enable the detection and the correction of the introduced error.

32. A kit for identifying an origin of a template nucleic acid molecule comprising:

a set of nucleic acid elements each comprising a distinctive sequence composition that enables detection of an introduced error in sequence data generated from each nucleic acid element and correction of the introduced error, wherein each of the nucleic acid elements is constructed to couple with the end of a template nucleic acid molecule and identifies the origin of the template nucleic acid molecule.

34. The kit of claim 32, wherein:

the distinctive sequence composition enables detection of up to three of the introduced errors and correction for up to two of the introduced errors.

35. The kit of claim 32, wherein:

the introduced error is selected from the group consisting of an insertion error, a deletion error, and a substitution error.

36. The kit of claim 32, wherein:

each nucleic acid element is incorporated into an adaptor comprising a primer element, wherein the adaptor couples with the end of the template nucleic acid molecule.

37. The kit of claim 36, wherein:

the nucleic acid element is in a known position relative to the primer element.

IDT Exhibit 1020.5054

38. The kit of claim 36, wherein:

the primer element is selected from the group consisting of an amplification primer, a sequencing primer, or a bipartite amplification – sequencing primer.

39. The kit of claim 36, wherein:

the adaptor comprises a quality control element.

40. The kit of claim 36, wherein:

the nucleic acid element is in a known position relative to the quality control element.

41. The kit of claim 32, wherein:

the detection of the introduced error in each of the nucleic acid elements is relative to the distinctive sequence composition of the other nucleic acid elements of the set.

42. The kit of claim 41, wherein:

the set comprises 14 of the nucleic acid elements.

43. A computer, comprising executable code stored thereon, wherein the executable code performs a method for identifying an origin of a template nucleic acid molecule, comprising the steps of:

identifying an identifier sequence from sequence data generated from a template nucleic acid molecule;

detecting an introduced error in the identifier sequence;

correcting the introduced error in the identifier sequence;

associating the corrected identifier sequence with an identifier element coupled with the template molecule; and

identifying an origin of the template molecule using the association of the corrected identifier sequence with the identifier element.

44. The method of claim 43, wherein:

the template nucleic acid molecule is included in a multiplex sample comprising a plurality of template molecules from a plurality of different origins.

IDT Exhibit 1020.5055

45. The method of claim 43, further comprising:

detecting up to three of the introduced errors in the first identifier sequence; and

correcting up to two of the introduced errors in the first identifier sequence.

46. The method of claim 43, wherein:

the introduced error is selected from the group consisting of an insertion error, a deletion error, and a substitution error.

48. The method of claim 43, wherein the step of identifying further comprises:

determining a position for the identifier sequence using a known positional relationship of one or more elements in the sequence data.

49. The method of claim 48, wherein:

the one or more elements include a primer sequence.

50. The method of claim 43, wherein the step of detecting further comprises:

measuring one or more characteristics of sequence composition in one or more sequence regions that flank the identifier sequence; and

detecting the introduced error using one or more assumptions derived from the measured characteristics.

51. The method of claim 43, further comprising:

identifying a second identifier sequence from the sequence data generated from the template nucleic acid molecule;

detecting an introduced error in the second identifier sequence;

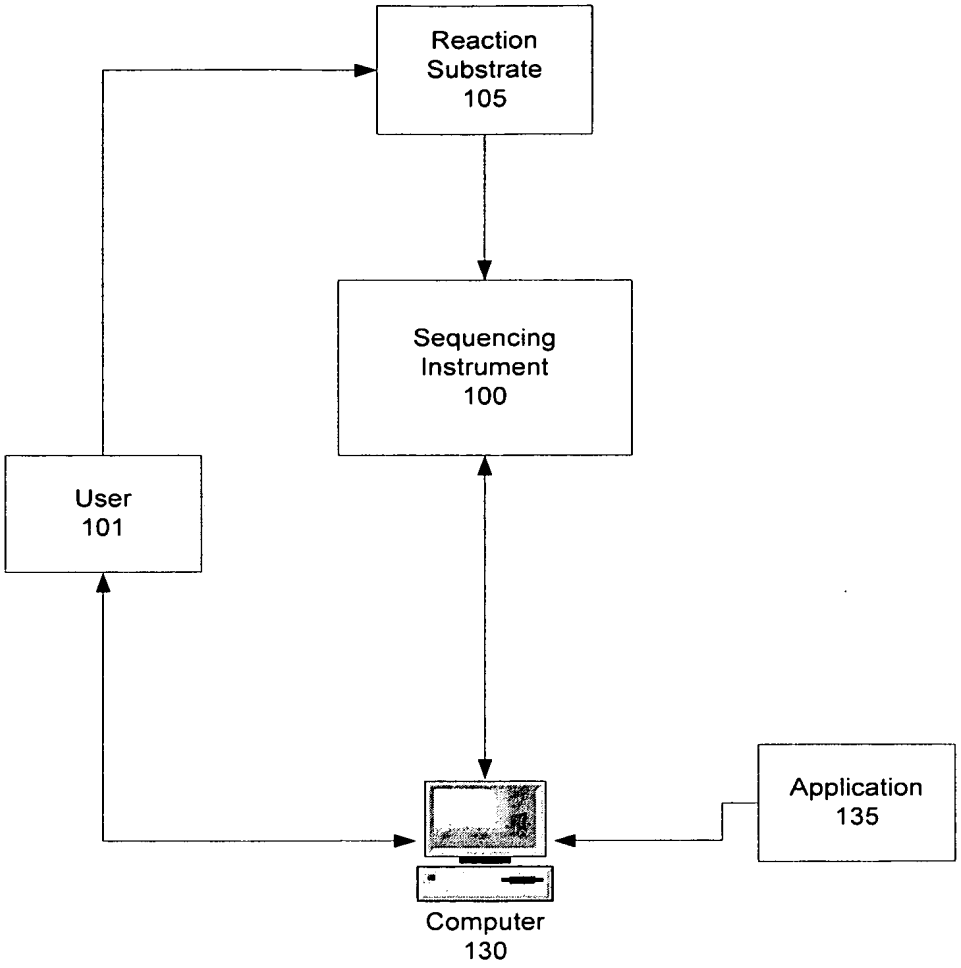
correcting the introduced error in the second identifier sequence;

associating the corrected second identifier sequence with a second identifier element coupled with the template molecule; and

identifying an origin of the template molecule using the association of the corrected second identifier sequence with the second identifier element combinatorially with the association of the corrected first identifier sequence with the first identifier element.

IDT Exhibit 1020.5056

FIGURE 1



IDT Exhibit 1020.5057

FIGURE 2A

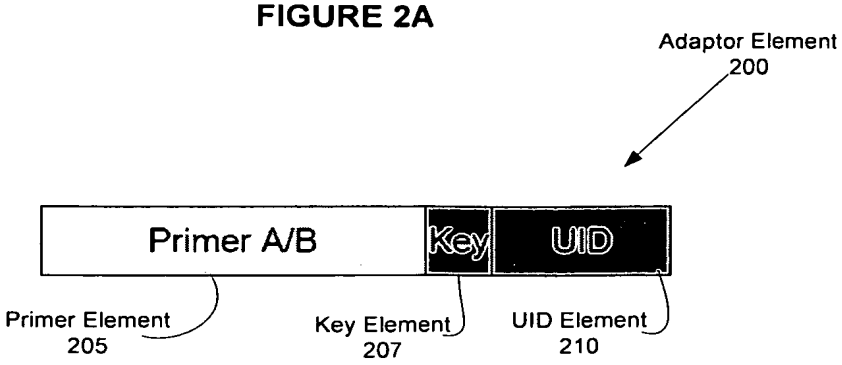
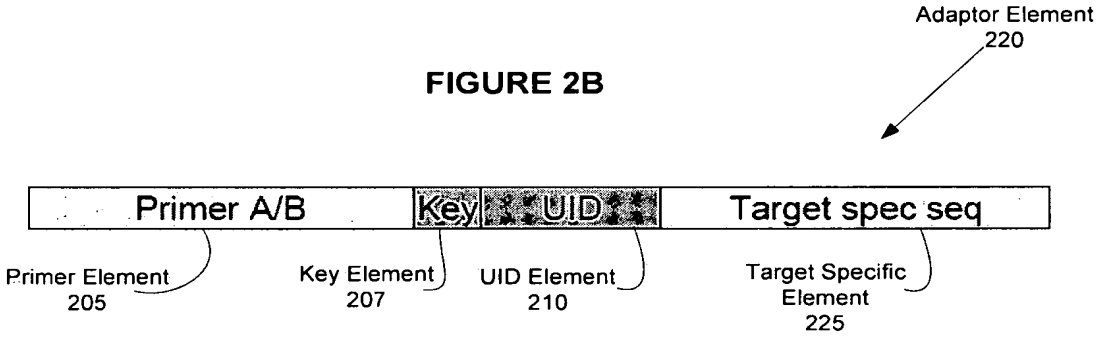
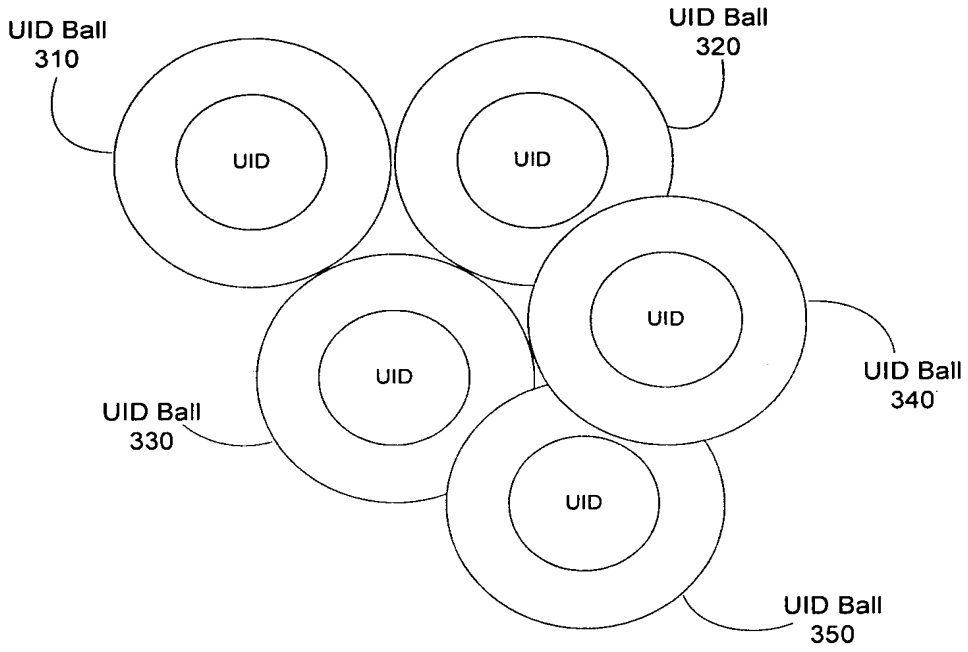


FIGURE 2B



IDT Exhibit 1020.5058

FIGURE 3



IDT Exhibit 1020.5059

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 08/06822

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68 (2008.04)
USPC - 435/6
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC - 435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC - 435/91.2; 536/23.1; 536/25.32; 536/25.4; 536/26.6

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
WEST(PGPB,USPT,USOC,EPAB,JPAB), Google Scholar: (sequencing adaptors origin identify error, "Patrick S. Schnable" origin, "Daniel A. Ashlock" algorithm, sequencing adaptors quality control element error, sequencing adaptors "quality control" element error, sequencing adaptors primer "quality control element" error)

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X --- Y	QIU et al. DNA Sequence-Based "Bar Codes" for Tracking the Origins of Expressed Sequence Tags from a Maize cDNA Library Constructed Using Multiple mRNA Sources. Plant Physiology, 2003, vol 133, pp 475-481; Abstract; pg 475, col 2 para 3; pg 476, col 1 para 1, ln 1-8, para 3; col 2 para 1, ln 1-2; pg 480, col 1, para 3, ln 5-11,14-17, col 2 para 2, ln 7-13, para 3, 6, ln 2-3, para 7, ln 4-5; pg 481, col 1, para 2, 4, 8, ln 1-6	1-8, 12-22, 26-31, 43-46, 48-51 --- 9-11, 23-25, 32, 34-42
Y	US 2007/00772208 A1 (DRMANAC) 29 Mar 2007 (29.03.2007); para [0015], [0053]	9-11, 23-25, 32, 34-42
Y	US 2005/0221341 A1 (SHIMKETS et al) 6 Oct 2005 (06.10.2005); para [0050], [0053]	10-11, 24-25, 39-40

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
01 September 2008 (01.09.2008)	08 SEP 2008
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (April 2007)

Espacenet

Bibliographic data: WO2007136717 (A1) — 2007-11-29

NUCLEIC ACID SEPARATION AND PURIFICATION METHOD BASED ON REVERSIBLE CHARGE INTERACTIONS

Inventor(s): KURN NURITH [US]; HEATH JOE DON [US] ± (KURN, NURITH, ; HEATH, JOE, DON)

Applicant(s): NUGEN TECHNOLOGIES INC [US]; KURN NURITH [US]; HEATH JOE DON [US] ± (NUGEN TECHNOLOGIES, INC, ; KURN, NURITH, ; HEATH, JOE, DON)

Classification: - international: ***C12N15/10; C12Q1/68***
- cooperative: **C12N15/1006; C12Q1/6806** more

Application number: WO2007US11856 20070516 Global Dossier

Priority number (s): US20060801088P 20060516

Also published as: CA2652052 (A1) EP2079838 (A1) GB2451993 (A) US2009275486 (A1)

Abstract of WO2007136717 (A1)

The invention provides a method for purifying nucleic acids using a polycationic reagent and an anionic substrate to form a complex with a nucleic acid to be purified. The complex may be separated from other components of a mixture and the nucleic acid eluted from the complex with a high ionic strength solution or an anionic reagent.

IDT Exhibit 1020.5061

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
29 November 2007 (29.11.2007)

PCT

(10) International Publication Number
WO 2007/136717 A1

(51) International Patent Classification:
C12N 15/10 (2006.01) *C12Q 1/68* (2006.01)

(21) International Application Number:
PCT/US2007/011856

(22) International Filing Date: 16 May 2007 (16.05.2007)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/801,088 16 May 2006 (16.05.2006) US

(71) Applicant *(for all designated States except US)*: **NUGEN TECHNOLOGIES, INC.** [US/US]; 821 Industrial Road, Unit A, San Carlos, CA 94070 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **KURN, Nurith** [IL/US]; 2876 Ramona Street, Palo Alto, CA 94306 (US). **HEATH, Joe, Don** [US/US]; 856 Grape Avenue, Sunnyvale, CA 94087 (US).

(74) Agent: **REAMEY, Robert, H.**; Wilson Sonsini Goodrich & Rosati, Patent and Innovation Strategies Group, 650 Page Mill Road, Palo Alto, CA 94304 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HU, IE, IS, IT, LT, LU, LV, MC, MT, NL, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report*
— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments*

For two-letter codes and other abbreviations, refer to the "Guidance Notes on Codes and Abbreviations" appearing at the beginning of each regular issue of the PCT Gazette.

WO 2007/136717 A1

(54) Title: NUCLEIC ACID SEPARATION AND PURIFICATION METHOD BASED ON REVERSIBLE CHARGE INTERACTIONS

(57) Abstract: The invention provides a method for purifying nucleic acids using a polycationic reagent and an anionic substrate to form a complex with a nucleic acid to be purified. The complex may be separated from other components of a mixture and the nucleic acid eluted from the complex with a high ionic strength solution or an anionic reagent.

IDT Exhibit 1020.5062

Nucleic Acid Separation and Purification Method Based on Reversible Charge Interactions

CROSS-REFERENCE TO RELATED APPLICATIONS

This application claims the benefit of U.S. Provisional Application No. 60/801,088, filed on May 16, 2006, the disclosure of which is incorporated by reference herein in its entirety.

FIELD OF THE INVENTION

The invention relates to nucleic acid purification, in particular involving a method that includes reversibly binding a nucleic acid to an anionic surface using a polycation to mediate binding between the nucleic acid and the anionic substrate.

BACKGROUND

Various methods for nucleic acid isolation and purification have developed in recent years. It is desirable to obtain nucleic acids that are substantially free of contaminants which could interfere with analysis or further processing. For example, contaminants include substances that interfere with hybridization or enzyme-catalyzed reactions, substances that degrade nucleic acids, and substances that interfere with detection of a nucleic acid of interest.

Early techniques for isolation of nucleic acids from a complex mixture, such as a cell extract or amplification reaction mixture, included multiple organic extraction and precipitation steps, using hazardous chemicals, such as chloroform and phenol. More recently, methods employing non-sequence specific binding of polynucleotides to solid surfaces have been developed. For example, matrices have been developed involving ion-exchange chromatography (*e.g.*, *J. of Chromatography* (1990) 508:61-73; *Nucleic Acids Research* (1993) 21(12):2913-2915; U.S. Patent Nos. 5,856,192, 5,660,984, and 4,699,717), reverse phase chromatography (*e.g.*, Hirbayashi et al. (1996) *J. Chromatography* (1996) 722:135-142; U.S. Patent No. 5,057,426), affinity chromatography (*e.g.*, U.S. Patent No. 5,712,383), or a combination thereof (*e.g.*, U.S. Patent No. 5,652,348; *J. Chromatography* (1983) 270:117-126).

IDT Exhibit 1020.5063

Solid phases developed for use in nucleic acid isolation include silica-based resins. Silica gel particles developed for HPLC are functionalized with anion-exchange moieties which can exchange with nucleic acids under certain salt and pH conditions (*e.g.*, U.S. Patent No. 4,699,717 and 5,057,426). Modern silica-based systems include controlled pore glass, filters embedded with silica particles, silica gel particles, resins containing silica in the form of diatomaceous earth, glass fibers, or mixtures thereof, and are configured to reversibly bind nucleic acids in the presence of chaotropic agents or a high ionic strength buffer. The nucleic acids remain bound to the solid phase during processing steps such as centrifugation or vacuum filtration, and are then eluted by exposing the solid phase to an elution solution such as a low ionic strength buffer.

An advantage of purification methods involving binding of nucleic acid to a solid surface is the ability to wash the bound material using solutions that retain the bound molecules on the solid surface while removing other non-related components, thus resulting in isolation and purification of the polynucleotides of interest from the sample solution. The use of solid surfaces for binding of polynucleotides of interest is desirable as these are easy to manipulate and are amenable for use in routine laboratory procedures, and do not involve the use of hazardous chemicals. Columns containing such solid surfaces are commercially available and are commonly configured as "spin columns" for use with a centrifugation step. Beads are also commercially available, including magnetic beads with a glass/silica based coating, as well as other configurations, including vacuum based devices.

Solid phases in the form of magnetically responsive particles have also been developed for nucleic acid isolation. Such particles bind directly or indirectly to nucleic acids. An example of a system which utilizes direct binding includes magnetically responsive porous glass beads (*e.g.*, U.S. Patent Nos. 4,233,169; 4,395,271; 4,297,337). However, nucleic acids bind very tightly to glass and may be difficult to remove once bound. Elution efficiency is higher from porous silica based magnetic particles designed to reversibly bind nucleic acids (*e.g.*, MagneSil™ particles from Promega or BioMAG™ particles from PerSeptive Biosystems). Magnetic systems utilizing indirect binding require magnetic particles, an intermediary, and a medium containing nucleic acid to be isolated. A disadvantage with currently available indirect binding magnetic systems is

IDT Exhibit 1020.5064

that different solution and/or temperature conditions may be required for intermediary/nucleic acid and intermediary/particle binding reactions, increasing risk of contamination of the isolated nucleic acid end product. An example of an indirect binding system is a streptavidin coated magnetically responsive microsphere with an oligo dT moiety covalently attached to the streptavidin. The streptavidin-oligo dT molecules act as intermediaries for hybridization to the poly A tail of mRNA (*e.g.*, ProActive™ from Bangs Laboratories (Carmel, IN) or PolyATract™ from Promega Corp. (Madison, WI)).

Most methods developed to date are optimized for purification of high molecular weight polynucleotide, and result in low recovery of low molecular weight polynucleotides. There is a growing need for the isolation and purification of low molecular weight polynucleotides, such small RNA, for example microRNA, fragmented DNA, fragmented and labeled target generated for hybridization to microarrays, and other sequence detection methods.

Methods for non-specific binding of nucleic acid to magnetic particles induced by precipitation using PEG (polyethylene glycol) and other polyalkylene glycols, have also been described (U.S. Patent No. http://patft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=5,898,071.PN.&OS=PN/5,898,071&RS=PN/5,898,071 - h0http://patft1.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=5,898,071.PN.&OS=PN/5,898,071&RS=PN/5,898,071 - h25,898,071; *Biotechniques* (2002) 32:1296-1302; Hawkins, et. al. (1995) *Nucleic Acids Res.* 23: 4742-4743), and are commercially available. Although effective and amenable for automation, these methods are not suitable for effective purification of small fragments of polynucleotides such as oligonucleotides less than 100 nucleotides in length.

The recent developments in very large scale gene expression analysis using massively parallel analytical tools such as microarray technologies for whole transcriptome or whole genome analysis require the employment of nucleic acid amplification and subsequent target preparation for further analysis. Targets suitable for

IDT Exhibit 1020.5065

microarray analysis of whole genomes or whole transcriptome often require fragmentation and labeling of amplification products. Methods for isothermal amplification of either DNA or RNA samples and subsequent fragmentation and labeling of amplification products have been described. Single primer isothermal amplification of DNA and RNA using chimeric DNA/RNA primers has been described (U.S. Patent Nos. 6,251,639, 6,692,918, and 6,946,251, and U.S. Patent Application No. 2005/0019793). These methods produce single stranded amplification products. Methods for fragmentation and labeling of amplification products which employ the use of non canonical nucleotides have also been described (U.S. Patent Application No. 2004/0005614). Similarly, methods for linear amplification of RNA which employ in-vitro transcription to generate RNA amplification products have also been described (U.S. Patent Nos. 6,686,156, 5,545,222, and 5,716,785). Purification of fragmented and labeled amplification products for subsequent analysis on microarrays requires the development of suitable methods which will be highly efficient, fast and amenable for automation. Magnetic particle based technologies are commonly used for automated separation of analytes and are commonly employed in automated immunoassays. The attachment of analyte binding entities to magnetic particles is known in the art and is commonly used in laboratory practice (*e.g.*, U.S. Patent No. 4,935,147). Methods for particle separation which enable the separation of non-magnetic particulate material, by complexation through interaction with polycations has been described (U.S. Patent No. 4,935,147). The complexation of negatively-charged non-magnetic particles and other particulate materials such as cells or liposomes, with negatively charged magnetic particles through the interaction with polycations, such as polyamines has been demonstrated. The charge based interactions used in these methods are reversible and are not related to the nature of the particulate material to be separated other than the charge distribution. The reversal of complexation can be induced in the presence of polyanions (such as citrate), by changes in the ionic strength of the solution in which the complexes are suspended or by cleavage of the polycation used for the complexation (U.S. Patent No. 5,405,743).

The major drawback of the various methods developed thus far is their inefficiency with respect to the purification and recovery of low molecular weight

IDT Exhibit 1020.5066

polynucleotides as compared with high molecular weight polynucleotides. Most methods developed thus far were optimized for the purification of nucleic acids from biological samples such as genomic DNA and RNA, or purification of amplification products, where amplification products, which are normally shorter than the genomic DNA, are purified and recovered while short primers and probes are preferentially not bound onto the separation matrix. However, there is a need for purification of short oligonucleotides from either biological samples or reaction mixtures, where the short polynucleotides of interest are in a size range up to about 200 nucleotides, using methods which provide high recovery rates, are easy to manipulate and especially suitable for high throughput application and automation. These needs are mostly driven by the fast developments in molecular biology in general and the rapid adoption of genomic technologies in basic research, drug development, and molecular diagnostics.

BRIEF SUMMARY OF THE INVENTION

The invention provides methods, compositions, and kits for nucleic acid purification.

In one aspect, the invention provides a method for nucleic acid purification, comprising contacting a sample or reaction mixture comprising nucleic acid with a polycationic reagent and an anionic substrate, wherein a nucleic acid-polycation-anionic substrate complex is formed. In one embodiment, the polycationic reagent is polybrene. In one embodiment, the anionic substrate is a carboxylated substrate, such as, for example, carboxylated polystyrene. In one embodiment, the anionic substrate is in the form of microparticles, such as, for example, carboxylated microparticles. In one embodiment, the anionic substrate is in the form of magnetically responsive microparticles, such as, for example, magnetically responsive carboxylated microparticles. The nucleic acid-polycation-anionic substrate complex may be separated from other components of the sample or reaction mixture. In one embodiment, a magnetically responsive anionic substrate is used and separation comprises application of a magnetic field. Nucleic acid may be eluted from the anionic substrate with a high ionic strength solution or an anionic reagent. In one embodiment, citrate is used for elution. In another embodiment, the nucleic acid is eluted in a high ionic strength solution which is

IDT Exhibit 1020.5067

suitable for hybridization of the eluted nucleic acid to probes which are immobilized on a solid surface such as a microarray.

In one embodiment, the method for nucleic acid purification comprises separating a nucleic acid-polycation, anionic substrate complex from a sample or reaction mixture, wherein the complex is produced by contacting a sample or reaction mixture comprising nucleic acid with a polycationic reagent and an anionic substrate.

In some embodiments, the nucleic acid to be purified is the product of an amplification reaction and/or a reaction in which a nucleic acid is synthesized from a template, such as, for example, PCR, primer extension, reverse transcription, DNA replication, strand displacement amplification (SDA), multiple displacement amplification (MDA), or template-independent synthesis, such as chemical synthesis or synthesis using a template independent nucleic acid polymerase. In one embodiment, the nucleic acid is the product of a linear isothermal amplification method comprising (a) hybridizing a single stranded DNA template comprising a target sequence with a composite primer, said composite primer comprising an RNA portion and a 3' DNA portion; (b) extending the composite primer with DNA polymerase; and (c) cleaving the RNA portion of the annealed composite primer with an enzyme that cleaves RNA from an RNA/DNA hybrid such that another composite primer hybridizes to the template and repeats primer extension by strand displacement, whereby multiple copies of the complementary sequence of the target sequence are produced. In one embodiment, the amplification reaction comprises a method comprising (a) extending a first primer hybridized to a target RNA with at least one enzyme comprising RNA-dependent DNA polymerase activity, wherein the first primer is a composite primer comprising an RNA portion and a 3' DNA portion, whereby a complex comprising a first primer extension product and the target RNA is produced; (b) cleaving RNA in the complex of step (a) with at least one enzyme that cleaves RNA from an RNA/DNA hybrid; (c) extending a second primer hybridized to the first primer extension product with at least one enzyme comprising DNA-dependent DNA polymerase activity and at least one enzyme comprising RNA-dependent DNA polymerase activity, whereby a second primer extension product is produced to form a complex of first and second primer extension products; (d) cleaving RNA from the first primer in the complex of first and second

IDT Exhibit 1020.5068

primer extension products with at least one enzyme that cleaves RNA from an RNA/DNA hybrid such that a composite amplification primer hybridizes to the second primer extension product, wherein the composite amplification primer comprises an RNA portion and a 3′ DNA portion; and (e) extending the composite amplification primer hybridized to the second primer extension product with at least one enzyme comprising DNA-dependent DNA polymerase activity; whereby said first primer extension product is displaced, RNA is cleaved from the composite amplification primer and another composite amplification primer hybridizes such that primer extension and strand displacement are repeated, and whereby multiple copies of a polynucleotide sequence complementary to the RNA sequence of interest are generated.

In some embodiments, the nucleic acid to be purified is a nucleic acid fragment, for example, a nucleic acid fragment of about 20 to about 100 nucleotides, about 50 to about 150, about 100 to about 200, about 150 to about 300, or about 250 to a about 500 nucleotides in length. In one embodiment, the fragment is about 10 to about 500 nucleotides in length. In another embodiment, the fragment is about 10 to about 200 nucleotides in length. In some embodiment, a population of nucleic acid fragments is purified, wherein the population comprises fragments of at least about 10 to any of about 20, 50, 100, 150, 200, 250, 300, 350, 400, 450, or 500 nucleotides in length. Often, a population of nucleic fragment of about 100 to about 200 nucleotides in length is purified. In one embodiment, the nucleic acid fragment is a labeled nucleic acid fragment. In one embodiment, the labeled nucleic acid fragment is prepared by a method comprising (a) synthesizing a polynucleotide from a polynucleotide template in the presence of a non-canonical nucleotide, whereby a polynucleotide comprising the non-canonical nucleotide is generated;(b) cleaving a base portion of the non-canonical nucleotide from the synthesized polynucleotide with an enzyme capable of cleaving the base portion of the non-canonical nucleotide, whereby an abasic site is generated; (c) cleaving a phosphodiester backbone of the polynucleotide comprising the abasic site at or near the abasic site; and (d) labeling the polynucleotide at the abasic site; whereby a labeled polynucleotide fragment is generated. In one embodiment, labeling the polynucleotide at the abasic site comprises labeling with terminal deoxynucleotidyl transferase. In one embodiment, labeling the polynucleotide at the abasic site comprises

IDT Exhibit 1020.5069

labeling with a label capable of reacting with an aldehyde residue at the abasic site. In one embodiment, trifluoracetic acid salt (ARP) is used for labeling. In one embodiment, the labeled nucleic acid fragment comprises a biotin label. Other methods for fragmentation and/or labeling of the fragmented nucleic acid are known in the art, including methods for labeling during fragmentation, and the resulting nucleic acid fragments may be purified using the methods described herein.

In one embodiment, the nucleic acid to be purified is a fragment of DNA prepared by digestion with an enzyme selected from the group consisting of DNase and a restriction endonuclease. In one embodiment, the nucleic acid to be purified is a fragment prepared by chemical cleavage. In one embodiment, the nucleic acid to be purified is a fragment of RNA prepared by heating the RNA at a temperature suitable to produce RNA fragments.

In another aspect, the invention provides compositions. In some embodiments, the invention provides a composition comprising a reaction mixture for nucleic acid purification according to a method as described herein. In one embodiment, the composition comprises: a polycationic reagent, for example, polybrene; an anionic substrate, for example, a carboxylated substrate, such as a magnetically responsive carboxylated substrate; and a nucleic acid of interest. In one embodiment, the composition comprises a nucleic acid-polycation-anionic substrate complex. In one embodiment, the composition comprises a nucleic acid-polycation-anionic substrate complex and a high ionic strength solution or anionic reagent for elution of the nucleic acid from the complex. In one embodiment, the composition comprises a nucleic acid purified by a method as described herein.

In one aspect, the invention provides kits for nucleic acid purification according to methods as described herein. In some embodiments, kits comprise a polycationic reagent and an anionic substrate, and/or a buffer suitable for formation of a nucleic acid-polycation-anionic substrate complex, and/or a solution suitable for elution of nucleic acid from a nucleic acid-polycation-anionic substrate complex, in packaging. Kits optionally further comprise instructions for use in a method as described herein.

IDT Exhibit 1020.5070

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1. Composition of reaction mixture for enabling binding of nucleic acid (DNA) molecules to negatively charged particles (magnetic particles) A: polycation (for example polybrene); B: Nucleic acid (any size including small fragments and particularly to the examples below, fragmented and labeled single stranded DNA amplification product; C: negatively charge particles (carboxylate magnetic beads).

Figure 2. Charge based binding of nucleic acid molecules to negatively charged particles through interaction with a polycation as binding mediator. The complexes of nucleic acid-polycation-particles can be isolated from the reaction mixture, for example, by magnetic separation (in embodiments in which magnetic particles are utilized), by centrifugation, or by filtration.

DETAILED DESCRIPTION OF THE INVENTION

The invention provides methods, compositions, and kits for isolation of nucleic acids from other components of a reaction mixture or sample. In some embodiments, the nucleic acids are oligonucleotide fragments of less than about 200 nucleotides, generally less than about 100 nucleotides in length. Methods of the invention involve formation of a nucleic acid-polycation-anionic solid substrate complex, and separation of the complex from other components in a reaction mixture or sample. In one embodiment, the polycation is polybrene. In one embodiment, the solid substrate is magnetically responsive, and separation includes magnetic separation. Nucleic acids may be bound to the polycation and/or in a nucleic acid-polycation-anionic solid substrate complex in a low ionic strength buffer, and eluted from the complex, after separation of the complex from other components of the reaction mixture or sample, with a high ionic strength solution, or a solution comprising a polyanion such as, for example, citrate. Anions in the elution solution interfere with the charge interaction between the polycation and the nucleic acid. Generally, the cationic charges of the polycation are neutralized by anionic charges in the elution solution. Nucleic acid isolation in accordance with the invention is primarily based on charge interaction between the nucleic acid and polycation, rather than size or sequence of the nucleic acid(s) to be isolated.

IDT Exhibit 1020.5071

Definitions

"Polynucleotide," or "nucleic acid," as used interchangeably herein, refer to polymers of nucleotides of any length, and include DNA, RNA, PNA, or modified forms thereof. The nucleotides can be deoxyribonucleotides, modified nucleotides or bases, and/or their analogs, or any substrate that can be incorporated into a polymer by a polymerase or synthetically. Nucleotides include canonical and non-canonical nucleotides and a polynucleotide can comprise canonical and non-canonical nucleotides. A polynucleotide may comprise modified (altered) nucleotides, such as, for example, modification to the nucleotide structure and or modification to the phosphodiester backbone. Modified nucleotides can be canonical nucleotide or non-canonical (cleavable) nucleotides. A polynucleotide may be further modified after polymerization, such as by conjugation with a labeling component. Other types of modifications include, for example, "caps", substitution of one or more of the naturally occurring nucleotides with an analog, internucleotide modifications such as, for example, those with uncharged linkages (e.g., methyl phosphonates, phosphotriesters, phosphoamidates, carbamates, etc.) and with charged linkages (e.g., phosphorothioates, phosphorodithioates, etc.), those containing pendant moieties, such as, for example, proteins (e.g., nucleases, toxins, antibodies, signal peptides, ply-L-lysine, etc.), those with intercalators (e.g., acridine, psoralen, etc.), those containing chelators (e.g., metals, radioactive metals, boron, oxidative metals, etc.), those containing alkylators, those with modified linkages (e.g., alpha anomeric nucleic acids, etc.), as well as unmodified forms of the polynucleotide(s). It is understood that internucleotide modifications may, e.g., alter the efficiency and/or kinetics of cleavage of the phosphodiester backbone. Further, any of the hydroxyl groups ordinarily present in the sugars may be replaced, for example, by phosphonate groups, phosphate groups, protected by standard protecting groups, or activated to prepare additional linkages to additional nucleotides. The 5' and 3' terminal OH can be phosphorylated or substituted with amines or organic capping groups moieties of from 1 to 20 carbon atoms. Other hydroxyls may also be derivatized to standard protecting groups. Polynucleotides can also contain analogous forms of ribose or deoxyribose sugars that are generally known in the art, including, for example, 2'--O-methyl-, 2'-O-

IDT Exhibit 1020.5072

allyl, 2'-fluoro- or 2'-azido-ribose, carboxylic sugar analogs, α-anomeric sugars, epimeric sugars such as arabinose, xyloses or lyxoses, pyranose sugars, furanose sugars, sedoheptuloses, acyclic analogs and abasic nucleoside analogs. One or more phosphodiester linkages may be replaced by alternative linking groups. These alternative linking groups include, but are not limited to, embodiments wherein phosphate is replaced by P(O)S("thioate"), P(S)S ("dithioate"), "(O)NR_2 ("amidate"), P(O)R, P(O)OR', CO or CH_2 ("formacetal"), in which each R or R' is independently H or substituted or unsubstituted alkyl (1-20 C) optionally containing an ether (-O-) linkage, aryl, alkenyl, cycloalkyl, cycloalkenyl or araldyl. Not all linkages in a polynucleotide need be identical. The preceding description applies to all polynucleotides referred to herein, including DNA.

"Oligonucleotide," as used herein, generally refers to short, generally single stranded, generally synthetic polynucleotides that are generally, but not necessarily, less than about 200 nucleotides in length. The terms "oligonucleotide" and "polynucleotide" and "nucleic acid" are not mutually exclusive. The description above for polynucleotides and nucleic acids is equally and fully applicable to oligonucleotides.

A "primer," as used herein, refers to a nucleotide sequence (a polynucleotide), generally with a free 3'-OH group, that hybridizes with a template sequence (such as a template RNA, or a primer extension product) and is capable of promoting polymerization of a polynucleotide complementary to the template. A "primer" can be, for example, an oligonucleotide. It can also be, for example, a sequence of the template (such as a primer extension product or a fragment of an RNA template created following RNase cleavage of a template RNA-DNA complex) that is hybridized to a sequence in the template itself (for example, as a hairpin loop), and that is capable of promoting nucleotide polymerization. Thus, a primer can be an exogenous (e.g., added) primer or an endogenous (e.g., template fragment) primer.

A "complex" is an assembly of components. A complex in accordance with the methods described herein may comprise a nucleic acid, a polycationic molecule, and a negatively-charged solid substrate (magnetically charged in methods in which magnetic separation of the complex is used), a nucleic acid and a polycationic molecule, or a

IDT Exhibit 1020.5073

polycationic molecule and a negatively-charged solid substrate (magnetically charged in methods in which separation of the complex is used).

A "reaction mixture" is an assemblage of components, which, under suitable conditions, react to form a complex (which may be an intermediate) and/or a product(s). A reaction mixture encompasses any type of suitable components and includes any sample for which separation of nucleic acid is desired, including biological samples.

A "fragment" of a polynucleotide or oligonucleotide is a contiguous sequence of 2 or more bases. In some embodiments, a fragment (also termed "region" or "portion") is any of about 20, about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 or more nucleotides in length. In some embodiments, the fragments can be at least about 20, about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 or more nucleotides in length. In other embodiments, the fragments can be less than about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 or more nucleotides in length. In some embodiment, these fragment lengths represent an average size in the population of fragments generated using the methods of the invention.

"A", "an" and "the", and the like, unless otherwise indicated include plural forms. "A" fragment means one or more fragments.

Conditions that "allow" an event to occur or conditions that are "suitable" for an event to occur, such as polynucleotide binding to a polycationic molecule, and the like, or "suitable" conditions, are conditions that do not prevent such events from occurring. Thus, these conditions permit, enhance, facilitate, and/or are conducive to the event. Such conditions, known in the art and described herein, depend upon, for example, the nature of the polynucleotide sequence, temperature, and buffer conditions. These conditions also depend on what event is desired, such as binding of a polynucleotide to a

IDT Exhibit 1020.5074

cationic molecule, formation of a polynucleotide, polycationic molecule, anionic solid substrate complex.

"Microarray" and "array," as used interchangeably herein, comprise a surface with an array, preferably ordered array, of putative binding (e.g., by hybridization) sites for a biochemical sample (target) which often has undetermined characteristics. In a preferred embodiment, a microarray refers to an assembly of distinct polynucleotide or oligonucleotide probes immobilized at defined positions on a substrate. Arrays are formed on substrates fabricated with materials such as paper, glass, plastic (e.g., polypropylene, nylon, polystyrene), polyacrylamide, nitrocellulose, silicon and other metals, optical fiber or any other suitable solid or semi-solid support, and configured in a planar (e.g., glass plates, silicon chips) or three-dimensional (e.g., pins, fibers, beads, particles, microtiter wells, capillaries) configuration. Probes forming the arrays may be attached to the substrate by any number of ways including (i) *in situ* synthesis (e.g., high-density oligonucleotide arrays) using photolithographic techniques (see, Fodor et al., *Science* (1991), 251:767-773; Pease et al., *Proc. Natl. Acad. Sci. U.S.A.* (1994), 91:5022-5026; Lockhart et al., *Nature Biotechnology* (1996), 14:1675; U.S. Pat. Nos. 5,578,832; 5,556,752; and 5,510,270); (ii) spotting/printing at medium to low-density (e.g., cDNA probes) on glass, nylon or nitrocellulose (Schena et al, *Science* (1995), 270:467-470, DeRisi et al, *Nature Genetics* (1996), 14:457-460; Shalon et al., *Genome Res.* (1996), 6:639-645; and Schena et al., *Proc. Natl. Acad. Sci. U.S.A.* (1995), 93:10539-11286); (iii) by masking (Maskos and Southern, *Nuc. Acids. Res.* (1992), 20:1679-1684) and (iv) by dot-blotting on a nylon or nitrocellulose hybridization membrane (see, e.g., Sambrook et al., Eds., 1989, Molecular Cloning: A Laboratory Manual, 2nd ed., Vol. 1-3, Cold Spring Harbor Laboratory (Cold Spring Harbor, N.Y.)). Probes may also be noncovalently immobilized on the substrate by hybridization to anchors, by means of magnetic beads, or in a fluid phase such as in microtiter wells or capillaries. The probe molecules are generally nucleic acids such as DNA, RNA, PNA, and cDNA but may also include proteins, polypeptides, oligosaccharides, cells, tissues and any permutations thereof which can specifically bind the target molecules.

IDT Exhibit 1020.5075

The term "3'" generally refers to a region or position in a polynucleotide or oligonucleotide 3' (downstream) from another region or position in the same polynucleotide or oligonucleotide.

The term "5'" generally refers to a region or position in a polynucleotide or oligonucleotide 5' (upstream) from another region or position in the same polynucleotide or oligonucleotide.

The term "3'-DNA portion," "3'-DNA region," "3'-RNA portion," and "3'-RNA region," refer to the portion or region of a polynucleotide or oligonucleotide located towards the 3' end of the polynucleotide or oligonucleotide, and may or may not include the 3' most nucleotide(s) or moieties attached to the 3' most nucleotide of the same polynucleotide or oligonucleotide. The 3' most nucleotide(s) can be preferably from about 1 to about 50, more preferably from about 10 to about 40, even more preferably from about 20 to about 30 nucleotides.

As used herein, "canonical" nucleotide means a nucleotide comprising one the four common nucleic acid bases adenine, cytosine, guanine and thymine that are commonly found in DNA. The term also encompasses the respective deoxyribonucleosides, deoxyribonucleotides or 2'-deoxyribonucleoside-5'-triphosphates that contain one of the four common nucleic acid bases adenine, cytosine, guanine and thymine (though as explained herein, the base can be a modified and/or altered base as discussed, for example, in the definition of polynucleotide). As used herein, the base portions of canonical nucleotides are generally not cleavable under the conditions used in the methods of the invention.

As used herein, "non-canonical nucleotide" (interchangeably called "non-canonical deoxyribonucleoside triphosphate") refers to a nucleotide comprising a base other than the four canonical bases. The term also encompasses the respective deoxyribonucleosides, deoxyribonucleotides or 2'-deoxyribonucleoside-5'-triphosphates that contain a base other than the four canonical bases. In the context of this invention, nucleotides containing uracil (such as dUTP), or the respective deoxyribonucleosides, deoxyribonucleotides or 2'-deoxyribonucleoside-5'-triphosphates, are a non-canonical nucleotides. As used herein, the base portions of non-canonical nucleotides are capable of being, generally, specifically or selectively cleaved (such that a nucleotide comprising

IDT Exhibit 1020.5076

an abasic site is created) under the reaction conditions used in the methods of the invention. Non-canonical nucleotides are generally also capable of being incorporated into a polynucleotide during synthesis of a polynucleotide (during e.g., primer extension and/or replication); capable of being generally, specifically or selectively cleaved by an agent that cleaves a base portion of a nucleotide, such that a polynucleotide comprising an abasic site is generated; comprise a suitable internucleotide connection (when incorporated into a polynucleotide) such that a phosphodiester backbone at an abasic site (i.e., the non-canonical nucleotide following cleavage of a base portion) is capable of being cleaved by an agent capable of such cleavage; capable of being labeled (following generation of an abasic site); and/or capable of immobilization to a surface (following generation of an abasic site), according to the methods described herein. It is understood that the non-canonical nucleotide may, but does not necessarily, require all of the features described above, depending on the particular method of the invention in which the non-canonical nucleotide is to be used. In some embodiments, non-canonical nucleotides are altered and/or modified nucleotides as described herein. Non-canonical nucleotide refers to a nucleotide that is incorporated into a polynucleotide as well as to a single nucleotide.

An "isolated" or "purified" nucleic acid is one that is substantially free of the materials with which it is associated in nature and/or substantially free of the materials with which it is associated in a reaction mixture or sample. "Purifying" a nucleic acid may include separating the nucleic acid from other components of a mixture and/or concentrating the nucleic acid. Generally, the nucleic acid is at least any of about 80, 85, 90, 95, 99, 99.5, or 99.9% pure.

Methods of the invention

Methods of the invention provide a fast and effective means for nucleic acid purification, which is based on reversible charge interactions and thus is effective for the isolation and purification of short polynucleotide molecules. Effective purification and isolation of short polynucleotides from solutions, either samples to be analyzed or from reaction mixtures, using currently available methods is ineffective and time consuming. In particular, the use of negatively charged magnetic particles in methods of the invention provides methods suitable for automation as required for large scale genomic analysis.

IDT Exhibit 1020.5077

Methods of the invention may be used to purify DNA and RNA polynucleotides of various lengths, including, but not limited to, small fragments which are generated by laboratory procedures such as amplification, fragmentation and labeling, restriction fragmentation, *in vitro* transcription, *in vitro* synthesis, and the like. The methods are particularly useful, but not limited to, the purification of fragmented and labeled targets generated by amplification of DNA or RNA and subsequent analysis, in particular analysis using microarrays. The exemplary use of the method of the invention for the purification steps during the generation of fragmented and biotin labeled targets for gene expression analysis from total RNA isolated from very small biological samples is described in the Examples below. The Ovation™ RNA amplification system is useful for isothermal linear amplification of all transcripts in minute biological samples. The amplification results in the generation of amplified single stranded cDNA which is subsequently fragmented and biotin labeled for gene expression analysis on microarrays, such as the GeneChip™ high density oligonucleotide arrays (Affymetrix). Insofar as labeling of the targets is carried out by a reactive biotin conjugate that is capable of reacting with the surface of the array, it is necessary to purify the short fragmented and labeled targets from the reaction mixture components. The methods of the invention are effective for purification of such fragments and are also advantageously easily adapted for automation, as required for high throughput systems.

Methods of the invention include binding of polynucleotides to a negatively charged solid substrate, such as negatively-charged particles, wherein binding is mediated through charge-based interaction with a polycationic reagent. Various negatively charged particles are known in the art. In one embodiment, beads comprising carboxyl groups are used, for example, carboxylate-derivatized acrylate, latex, or polyacrolein microparticles. (See, *e.g.*, U.S. Patent Nos. 4,678,814 and 4,935,147.) Beads comprising other negatively charged groups may be used. In some embodiments, negatively-charged magnetically responsive particles are used, which provides advantages for assay procedures due to the ease of automation of particle separation. Such particles are known in the art. (See, *e.g.*, U.S. Patent Nos. 6,872,578, 6,534,262, and 6,958,372.) Suitable paramagnetic microparticles for use in the instant invention can be obtained, for example,

IDT Exhibit 1020.5078

from Bangs Laboratories, Inc., Fishers, IN (*e.g.*, Estapor™ carboxylate-modified encapsulated magnetic microspheres) or from Cortex (MagaPhase™ product line).

In methods of the invention, a reaction mixture or sample that contains nucleic acid(s) of interest is contacted with a polycationic reagent a negatively-charged solid matrix to form a nucleic acid-polycation-anionic solid substrate complex. After formation of the complex, the complex is separated from other components of the sample or reaction mixture. Separation may be effected, for example, by washing the substrate to remove unbound materials, by centrifugation, or by filtration. After separation, nucleic acid may be dissociated from the complex by altering conditions to favor dissociation (*e.g.*, by altering pH, temperature, or ionic strength of the buffer) and/or by adding a releasing agent, wherein nucleic acid is released from the complex.

In some embodiments, nucleic acids isolated in accordance with the methods of the invention are fragments, optionally labeled fragments. Polynucleotide fragments may be prepared and/or labeled, for example, as described in U.S. Patent Application No. 2004/0005614. In some embodiments, polynucleotide fragments of any of about 20, about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 or more nucleotides in length are isolated. In some embodiments, the fragments can be at least about 20, about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 or more nucleotides in length. In other embodiments, the fragments can be less than about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 or more nucleotides in length. In some embodiments, the fragments are any of at least about 20, about 25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125, about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600 nucleotides in length, with an upper limit of any of about25, about 30 about 35 about 40, about 50, about 65, about 75, about 85, about 100, about 125,

IDT Exhibit 1020.5079

about 150, about 175, about 200, about 225, about 250, about 300, about 350, about 400, about 450, about 500, about 550, about 600, about 650 nucleotides in length. In some embodiments, these fragment lengths represent an average size in the population of fragments generated using a method for fragmenting nucleic acids.

A "polycationic reagent" or "polycation" as used herein refers to a compound, composition, or material, either inorganic or organic, naturally occurring or synthetic, having at least two positive charged groups. In some embodiments, the polycationic reagent contains at least 10 positively charged groups. In methods of the invention, a polycationic reagent serves as a bridge between a negatively-charged nucleic acid to be purified and a negatively-charged substrate, permitting the nucleic acid to bind indirectly to the substrate. Examples of polycationic reagents include, but are not limited to, polyalkylene amines, such as polyethyleneimine and polypropyleneimine and their lower alky ammonium salts such as Polybrene®, metal ions such as calcium and barium ion, aminodextrans, protamine, positively charged liposomes, and polylysine.

In one embodiment, a cleavable polycation is used. Cleavable polycations are described, for example, in U.S. Patent No. 5,405,743. The use of a cleavable polycation for mediation of binding of the polynucleotides to negatively charged particles according to the method of the invention permits the dissociation of the isolated and purified polynucleotide from the particles by means of cleaving the polycation and thus does not require the reversal of ionic interaction for elution of the bound polynucleotide.

The negatively-charged substrate to which the polycationic reagent binds may be in the form of particles or beads, either magnetic or non-magnetic. (See, *e.g.*, U.S. Patent No. 4,935,147.) The particles may be inherently negatively charged or may be treated chemically or physically to introduce a negative charge. Negatively-charged functional groups may be incorporated to render the substrate material anionic, such as carboxylate groups or other anionic groups. In one embodiment, carboxylated polystyrene particles are used.

In some embodiments, magnetic particles are used. The terms "magnetic particles" and "magnetically responsive particles" are used interchangeably herein, and refer to particles that are intrinsically magnetically responsive or have been rendered magnetically responsive, for example, by attachment to a magnetically responsive

IDT Exhibit 1020.5080

substance or by incorporation of such a substance into the particles. The magnetic particles can be paramagnetic, ferromagnetic, or superparamagnetic, usually paramagnetic and will have magnetic susceptibilities (X) of at least 5×10^{-5} emu/Oecm^3, usually at least 4×10^{-4} emu/Oecm^3. The diameter of the particles should be small, generally in the range from about 5 nm to 1 micron, preferably from about 10 to 250 nm, more preferably from about 20 to 100 nm, most preferably colloidal.

Exemplary of the magnetic component of particles that are intrinsically magnetically responsive are complex salts and oxides, borides, and sulfides of iron, cobalt, nickel and rare earth elements having high magnetic susceptibility, e.g. hematite, ferrite. The magnetic component of other such particles includes pure metals or alloys comprising one or more of these elements.

For the most part the magnetic particles will contain a core of the magnetic component with surface functional groups such as carboxylate groups. Alternatively, the magnetic component can be incorporated into a particle such as, for example, impregnating the substance in a polymeric matrix. However, this procedure frequently yields larger particles. For a more in-depth discussion of the preparation of magnetic particles by this method, see Whitesides, et al. (1983) *Trends in Biotechnology* 1(5):144-148 and references cited therein.

Magnetic particles of less than a hundred nanometers in diameter can be made by precipitating iron oxides in the presence or absence of a coating such as a polysaccharide or protein. Magnetic particles of a few microns diameter can also be made by a ball milling process and removing material which is not in the size range of interest. Typically, magnetic particles formed by this latter process are quite polydisperse, and not as generally useful. More useful monodisperse metal oxide suspensions can be prepared by careful control of pH, temperature and concentrations during the precipitation process. Coating the magnetic particles with macromolecules can increase their colloidal stability. This can be done by direct adsorption of high molecular weight polymers or by functionalizing the surface of the particle and then binding macromolecules to the functional groups. Emulsion polymerization and grafting techniques provide a means for coating magnetic particles with polymers.

In methods of the invention using magnetic particles, a reaction mixture or sample

IDT Exhibit 1020.5081

containing nucleic acid to be purified is contacted with a polycationic reagent and anionic magnetic particles to form a nucleic acid-polycation-anionic magnetic particle complex. The complexes are separated from other components of the reaction mixture or sample by application of a magnetic current and separation of the magnetic particles from unbound components.

After separation of a complex comprising nucleic acid, polycationic reagent, and anionic substrate, nucleic acid may be eluted from the complex. The elution of the purified polynucleotides from the complex, following binding to the negatively charged particles, separation and washing of the complex, is affected by the choice of ionic strength of the elution solution. High ionic strength solutions, or polyanions, such as citrate, or polymeric anions such as polyacrylate, are effective in dissociation of the complexes. Polyanions suitable for use in elution of nucleic acids from the complexes include, but are not limited to, heparin, dextran sulfate, negatively charged phospholipids vesicles, polycarboxylic acids, such as polyacrylate or polyglutamate. High ionic strength solutions which are suitable for subsequent analysis of the eluted nucleic acids by hybridization can also be used.

In some embodiments, a nucleic acid is eluted from the negatively complex with a releasing agent. As used herein, "releasing agent" refers to a compound, composition, or material, either naturally occurring or synthetic, organic or inorganic, capable of reversing the non-specific binding between negatively-charged molecules, *i.e.*, dissociating such molecules. The releasing agent acts upon the non-specific bond between the molecules. For example, the releasing agent can act to change the pH of the medium to one which is unfavorable or incompatible with the charge interactions between the molecules. The releasing agent can, therefore, be an acid such as a mineral acid or an organic acid or a base such as a mineral base or an organic base. Alternatively, the releasing agent can act to shield ionic interactions and thus can be a high ionic strength solution or a solution of a neutral polymer such as dextran. Alternatively, the releasing agent can have a charge which disrupts the non-specific binding between the oppositely-charged molecules. Exemplary of the latter are be polyelectrolyte salts such as citrate, polyacrylate, dextran sulfate, and the like, as well as other cations, such as, for example, Mg^{2+}.

IDT Exhibit 1020.5082

Compositions

The invention provides compositions comprising reaction mixtures, complexes, and/or end products of the nucleic acid purification procedures described herein.

In some embodiments, the invention provides a reaction mixture for nucleic acid purification, comprising a nucleic acid to be purified, a polycationic reagent, and a negatively charged substrate. In one embodiment, the polycationic reagent is polybrene. In one embodiment, the negatively charged substrate is a carboxylate derivatized substrate, such as a carboxylated polystyrene particle. In one embodiment, the negatively charged substrate is a magnetically responsive particle.

In one embodiment, the invention provides a nucleic acid-polycation-anionic substrate complex. In one embodiment, the invention provides a reaction mixture comprising a nucleic acid-polycation-anionic substrate complex and a high ionic strength solution or a releasing agent.

In one embodiment, the invention provides a nucleic acid purified by a method as described herein.

Kits

The invention provides kits for carrying out the methods of the invention. Kits comprise components for performing the methods for nucleic acid purification described herein in suitable packaging, optionally further comprising instructions for use in a method as described herein.

The kits of the invention comprise one or more containers comprising any combination of the components described herein, such as one or more polycationic reagents, one or more negatively charged substrates, and/or buffer suitable for binding nucleic acid and polycationic reagent to a negatively charged substrate, and/or elution buffer or releasing agent to elute the nucleic acid from a nucleic acid-polycation-anionic substrate complex.

One or more reagents in the kit can be provided as a dry powder, usually lyophilized, including excipients, which on dissolution will provide for a reagent solution having the appropriate concentrations for performing any of the methods described

IDT Exhibit 1020.5083

herein. Each component can be packaged in separate containers or some components can be combined in one container where cross-reactivity and shelf life permit.

The kits of the invention may optionally include a set of instructions, generally written instructions, although electronic storage media (*e.g.*, magnetic diskette or optical disk) containing instructions are also acceptable, relating to the use of components of the methods of the invention for the intended nucleic acid purification methods of the invention, and/or, as appropriate, for using the purified nucleic acid products for purposes such as, for example preparing a hybridization probe, expression profiling, preparing a microarray, or characterizing a nucleic acid. The instructions included with the kit generally include information regarding reagents (whether included or not in the kit) necessary for practicing the methods of the invention, instructions on how to use the kit, and/or appropriate reaction conditions.

The component(s) of the kit may be packaged in any convenient, appropriate packaging. The components may be packaged separately, or in one or multiple combinations.

The relative amounts of the various components in the kits can be varied widely to provide for concentrations of the reagents that substantially optimize the reactions that need to occur to practice the methods disclosed herein and/or to further optimize the sensitivity of any assay.

The following examples are intended to illustrate, but not limit, the invention.

IDT Exhibit 1020.5084

EXAMPLES

Example 1

Polybrene and AMPure™ Bead Purification of Fragmentation and Labeling (F&L) Product

Concept

Polybrene will bind negatively charges molecules (DNA and carboxylated beads) which can then be placed on magnetic stand and washed with H2O. Resuspension of beads in 0.2M citrate should remove the DNA from the beads. The following experiment was performed to test this concept.

Materials

Polybrene
Unpurified fragmented and biotin labeled cDNA
H2O
SSC (0.3M citrate)
AMPure™ beads (Agencourt)

1 Dilutions of polybrene in H20

	polybrene	H20
1/5	10	40
1/10	5	45
1/50	5 (1/5)	45
1/100	5 (1/10)	45
1/500	5 (1/50)	45
1/1000	5 (1/100)	45

2 Bead preparation: AMPure beads were washed (1 ml in 1.5 ml tube, spin 16K rpm 1 min, remove supernate and resuspend in 1 ml H2O.

IDT Exhibit 1020.5085

Set up binding solutions (add polybrene and EtOH to fragmented and biotin labeled cDNA targets before beads)

tube	F&L cDNA	beads	H2O	polybrene	polybrene dilution	100% EtOH
1	50	90 (washed)	5	0		
2	50	90 (washed)	0	5	undil	
3	50	90 (washed)	0	5	1/5	
4	50	90 (washed)	0	5	1/10	
5	50	90 (washed)	0	5	1/50	
6	50	90 (washed)	0	5	1/100	
7	50	90 (washed)	0	5	1/500	
8	50	90 (washed)	0	5	1/1000	
9	25	90	0			25
10	25	90	12.5			12.5
11	25	90	22.5			2.5
12	25	90	25			0

Incubate on desktop 5 minutes

Place tubes on magnetic rack, wait until solution clears

Remove clear solution and save for possible analysis

Wash beads on magnetic rack with 70% EtOH twice

Remove EtOH and air dry five minutes

IDT Exhibit 1020.5086

9 Resuspend samples 1-8 in 50 µl 20X SSC

10 Resuspend samples 9-12 in 50 µl H2O

11 Measure recovery by Nanodrop

Results

Sample ID	ng/µl	260/280	Yield (ug)	%recovery
1	0.07	-2.89	0.00	0.1%
2	72.93	1.9	3.65	68.8%
3	68.45	1.91	3.42	64.6%
4	48.75	1.92	2.44	46.0%
5	0.22	-1.69	0.01	0.2%
6	0.43	1.75	0.02	0.4%
7	0.82	2.76	0.04	0.8%
8	-0.07	0.35	0.00	-0.1%
9	16.89	2.21	1.69	31.9%
10	13.09	2.01	1.31	24.7%
11	13.08	2.08	1.31	24.7%
12	11.19	2.87	1.12	21.1%

Conclusions

- Polybrene significantly improved recovery of fragmentation and labeling cDNA using AMPure™ beads.

Example 2

The experiment described in Example 1 was repeated using the full concentration of polybrene (10 mg/ml in water); wash complex of particles and targets twice with 70% ethanol, and resuspend particles for target elution with 1X hybridization buffer, 2X hybridization buffer or 20X SSC. In addition, performance of the method was assessed using washed particles or unwashed particles (Agnecourt magnetic particles in binding buffer).

IDT Exhibit 1020.5087

	Beads	Elution	ng/µl	Yield (ug)	%recovery
2	washed	1X hybe	70.26	3.51	66.3%
6	unwashed	1X hybe	25.83	1.29	24.4%
3	washed	2X hybe	69.57	3.48	65.6%
7	unwashed	2X hybe	20.26	1.01	19.1%
4	washed	20X SSC	59.23	2.96	55.9%
8	unwashed	20X SSC	23.09	1.15	21.8%
	DyeEx1		97.74	4.40	83.0%
	DyeEx2		96.77	4.35	82.2%

to 21

Conclusions

- AMPure beads should be removed from binding buffer and resuspended in water.
- 1X hybridization buffer is sufficient to elute target

Comparison of Method of the Invention to Spin Column Purification Procedure for Binding to GeneChip™ Array

GeneChip™ arrays (U133A v2) data summary:

	Scale Factor	Background	%P
DyeEx purified target	0.991	Avg: 44.48	71.00%
Polybrene magnetic particles based target purification	0.798	Avg: 46.52	72.00%

Parameters that are important for performance with respect to gene expression analysis on an array using fragmented and biotin labeled targets include the effective

IDT Exhibit 1020.5088

recovery yield of the purified fragmented and labeled targets and the removal of the free reactive biotin conjugate. These performance parameters will affect GeneChip™ array results as reflected by background signals (background due to the reaction of the reactive biotin conjugate with the array surface) and the percent of transcripts detected by hybridization to the high density oligonucleotide array (GeneChip™ arrays, Affymetrix), as denoted by %P. The results of comparison of the performance of targets prepared by the Ovation™ system and purified using either size exclusion spin column, DyeEx, or the method of the invention, demonstrate equal of better performance of the target purified by the method of the invention. The method is easy to perform, highly reproducible and suitable for automation.

Example 3 - Purification of Fragmented and Biotin Labeled Targets Generated by RNA Amplification using the Ovation™ System with Polybrene and AMPure™ Magnetic Beads

Concept

Attempt to increase recovery of fragmented and labeled cDNA by increasing the amount of polybrene, increase beads, increase volume of washed beads thereby decreasing the ionic strength of the cDNA-polybrene mix. The following experiment was performed to test this concept.

Materials

Polybrene
(10ug/ml)
unpurified F&L
cDNA
H2O
1X hybe cocktail
AMPure beads

IDT Exhibit 1020.5089

1 Make 1X washed beads
spin down 0.5 mL beads at 16K for 1'
Remove sup
Resuspend in 0.5 ml H20

2 Make 2X washed beads
spin down 0.5 mL beads at 16K for 1'
Remove sup
Resuspend in 0.25 ml H20

3 Set up purifications

	Mix cDNA and polybrene by pipette
a	mixing
b	Add beads and mix by pipette mixing
c	Incubate at RT for 5 minutes
d	Place on mag stand
e	After 5-10 minutes or clear, remove sup
f	wash beads 200 µl 70% EtOH, sit on mag stand 2 minute, remove EtOH
g	repeat
h	remove all residual EtOH reasonably possible
i	add 50 µl 1X hybe cocktail. Resuspend. Sit 5 minutes RT.
j	Place on mag stand 5-10 minutes until clear
k	remove sup and measure cDNA on Nanodrop

well #	F&L cDNA µl	polybrene µl	1X beads µl	2X beads µl	ng/µl	260/280	Effective yield	% recovery
1	25	2.5	45		35.39	1.84	3.539	66.8%
2	25	2.5	75		41.35	1.84	4.135	78.0%
3	25	2.5		22.5	41.02	1.87	4.102	77.4%
4	25	2.5		45	43.32	1.84	4.332	81.7%

IDT Exhibit 1020.5090

5	25	5	45		38.73	1.89	3.873	73.1%
6	25	5	75		37.74	1.88	3.774	71.2%
7	25	5		22.5	37.52	1.91	3.752	70.8%
8	25	5		45	37.88	1.83	3.788	71.5%

Example 4 - Protocol for Polybrene-aided Bead Cleanup of Ovation™ Biotin cDNA and Fragmented and Labeled cDNA Target (Ovation™ RNA amplification and Biotin Labeling System product)

The protocol is designed for purification of amplified cDNA, prior to fragmentation and labeling, using AMPure™ magnetic particles, according to Agnecourt's protocol, elution of purified cDNA in fragmentation reaction buffer, performing fragmentation and labeling according to the Ovation™ protocol, in the presence of the magnetic beads, and purifying the fragmented and labeled target according to the method described herein, for further analysis of gene expression by hybridization to GeneChip™ high density arrays.

Materials

Ovation Biotin amplified cDNA (unfragmented and unpurified)
Polybrene (10 ug/µl) (cat# TR-1003-G, Specialty Media 1-888-209-8870)
Agencourt AMPure beads
H2O
Fresh 70% EtOH
SPRIPlate 96R magnet (Agencourt)
1X GeneChip Hybridization Buffer

Procedure

1. Swirl or vigorously shake AMPure bottle to mix beads thoroughly.
2. Add 144 µl of resuspended AMPure beads to each 80 µl SPIA reaction

IDT Exhibit 1020.5091

3. Mix the samples completely by pipetting up and down.
4. Let sit at RT for 5 minutes.
5. Place one of the pair of strip tubes or PCR plates on magnetic plate.
6. Wait 10 minutes or until solution is clear (beads will form a donut ring).
7. Carefully remove and discard clear supernatant without disturbing donut ring of beads.
8. Transfer bead-sample mix from second plate into correct wells/tubes of 1st plate.
9. Wait 10 minutes or until solution is clear (beads will form a donut ring).
10. Carefully remove and discard clear supernatant without disturbing donut ring of beads.
11. Leaving the tubes or plate on the SPRIPlate 96R, added 200 µl 70% EtOH.
12. Allow solution to clear before carefully removing and discarding EtOH without disturbing donut ring of beads.
13. Repeat steps 11-12.
14. Allow to air dry no more than 5 minutes.
15. Remove strip tubes or PCR plate from magnetic rack.
16. Add 35 µl 1X Fragmentation Master Mix 1 using components from Ovation Biotin kit (25 µl H2O, 5 µl F1, 5 µl F2) to each tube.
17. Pipette up and down to resuspend beads.
18. Close strip tube caps or seal PCR plate.
19. Incubate strip tubes or PCR plate in thermal cycler at 50 degrees for 30 minutes.
20. Remove from thermal cycler.
21. Add 5 µl F3 and 2.5 µl F4 to each well. Mix thoroughly by pipetting.
22. Reseal and incubate in thermal cycler at 50 degrees for 30 minutes.
23. Remove from thermal cycler.
24. Add 5 µl of 10mg/ml polybrene to each well and mix completely by pipetting.
25. Let sit at RT for 5 minutes.
26. Place strip tube or PCR plate on magnetic plate.
27. Wait 5 minutes or until solution is clear (beads will form a donut ring).
28. Carefully remove and discard clear supernatant without disturbing donut ring of beads.

IDT Exhibit 1020.5092

29. Add 200 µl 70% EtOH to each well. Wait 1 minute, leaving strip tube or PCR plate on magnetic plate.
30. Carefully remove EtOH without disturbing donut ring of beads.
31. Repeat steps 8-9.
32. If necessary, remove residual EtOH from bottom of tube with small multichannel pipette.
33. Let air dry for no more than 5 minutes.
34. Remove strip tubes or PCR plate from magnetic rack.
35. Add 100 µl 1X GeneChip Hybridization Buffer to each well.
36. Pipette mix or seal/cap and gently vortex to return beads to solution.
37. Allow to sit at RT for 5 minutes.
38. Place strip tubes or plate on magnetic rack.
39. Wait 5 minutes or until solution is clear (beads will form a donut ring).
40. Carefully transfer clear supernatant to a fresh container (tube or PCR plate) without disturbing donut ring of beads.
41. Measure concentration of fragmented cDNA in supernatant on Nanodrop using 1X GeneChip Hybridization Buffer as blank.

IDT Exhibit 1020.5093

All publications, patents, and patent applications cited herein are hereby incorporated by reference in their entireties for all purposes and to the same extent as if each individual publication, patent, or patent application were specifically and individually indicated to be so incorporated by reference.

Although the foregoing invention has been described in some detail by way of illustration and examples for purposes of clarity of understanding, it will be apparent to those skilled in the art that certain changes and modifications may be practiced without departing from the spirit and scope of the invention. Therefore, the description should not be construed as limiting the scope of the invention, which is delineated by the appended claims.

IDT Exhibit 1020.5094

CLAIMS

We claim:

1. A method for purifying a nucleic acid, comprising:

contacting a sample or reaction mixture comprising said nucleic acid with a polycationic reagent and an anionic substrate, wherein a nucleic acid-polycation-anionic substrate complex is formed;

separating said nucleic acid-polycation-anionic substrate complex from other components of said sample or reaction mixture; and

eluting said nucleic acid from said nucleic acid-polycation-anionic substrate complex.

2. A method according to claim 1, wherein said polycationic reagent is polybrene.

3. A method according to claim 1, wherein said anionic substrate is magnetically responsive, and said separating comprises application of a magnetic field.

4. A method according to claim 1, wherein said anionic substrate is a carboxylated substrate.

5. A method according to claim 4, wherein said carboxylated substrate is carboxylated polystyrene.

6. A method according to claim 1, wherein said nucleic acid is eluted from said anionic substrate with a high ionic strength solution.

7. A method according to claim 6, wherein said high ionic strength solution is suitable for hybridization of the eluted nucleic acid to probes.

IDT Exhibit 1020.5095

8. A method according to claim7, wherein said probes are immobilized on a substrate.

9. A method according to claim 8, wherein said probes are in the form of a microarray.

10. A method according to claim 1, wherein said nucleic acid is eluted from said anionic substrate with an anionic reagent.

11. A method according to claim 10, wherein said anionic reagent is citrate.

12. A method according to claim 1, wherein said nucleic acid is produced in an amplification reaction.

13. A method according to claim 1, wherein said nucleic acid is synthesized from a nucleic acid template comprising DNA or RNA.

14. A method according to claim 1, wherein said nucleic acid is produced by a method selected from the group consisting of polymerase chain reaction (PCR), primer extension, reverse transcription, DNA replication, strand displacement amplification (SDA), multiple displacement amplification (MDA), and template-independent synthesis.

15. A method according to claim 1, wherein said nucleic acid is produced by a linear isothermal amplification method comprising:

(a) hybridizing a DNA template comprising a target sequence with a composite primer, said composite primer comprising an RNA portion and a 3' DNA portion;

(b) extending the composite primer with DNA polymerase;

(c) cleaving the RNA portion of the annealed composite primer with an enzyme that cleaves RNA from an RNA/DNA hybrid such that another composite

IDT Exhibit 1020.5096

primer hybridizes to the template and repeats primer extension and strand displacement, whereby multiple copies of the target sequence are produced.

16. A method according to claim 1, wherein said nucleic acid is produced by a linear isothermal amplification method comprising:

(a) extending a first primer hybridized to a target RNA with at least one enzyme comprising RNA-dependent DNA polymerase activity, wherein the first primer is a composite primer comprising an RNA portion and a 3′ DNA portion, whereby a complex comprising a first primer extension product and the target RNA is produced;

(b) cleaving RNA in the complex of step (a) with at least one enzyme that cleaves RNA from an RNA/DNA hybrid;

(c) extending a second primer hybridized to the first primer extension product with at least one enzyme comprising DNA-dependent DNA polymerase activity and at least one enzyme comprising RNA-dependent DNA polymerase activity, whereby a second primer extension product is produced to form a complex of first and second primer extension products;

(d) cleaving RNA from the first primer in the complex of first and second primer extension products with at least one enzyme that cleaves RNA from an RNA/DNA hybrid such that a composite amplification primer hybridizes to the second primer extension product, wherein the composite amplification primer comprises an RNA portion and a 3′ DNA portion; and

(e) extending the composite amplification primer hybridized to the second primer extension product with at least one enzyme comprising DNA-dependent DNA polymerase activity; whereby said first primer extension product is displaced, RNA is cleaved from the composite amplification primer and another composite amplification primer hybridizes such that primer extension and strand displacement are repeated, and whereby multiple copies of a polynucleotide sequence complementary to the RNA sequence of interest are generated.

17. A method according to claim 1, wherein said nucleic acid is a fragment of about 10 to about 500 nucleotides in length.

IDT Exhibit 1020.5097

18. A method according to claim 27, wherein said fragment is about 10 to about 200 nucleotides in length.

19. A method according to claim 17, wherein said fragment is prepared by a method comprising:

(a) synthesizing a polynucleotide from a polynucleotide template in the presence of a non-canonical nucleotide, whereby a polynucleotide comprising the non-canonical nucleotide is generated;

(b) cleaving a base portion of the non-canonical nucleotide from the synthesized polynucleotide with an enzyme capable of cleaving the base portion of the non-canonical nucleotide, whereby an abasic site is generated;

(c) cleaving the phosphodiester backbone of the polynucleotide comprising the abasic site at or near the abasic site, whereby a polynucleotide fragment is generated.

20. A method according to claim 19, further comprising:

(d) labeling the polynucleotide at the abasic site; whereby a labeled polynucleotide fragment is generated.

21. A method according to claim 20, wherein labeling the polynucleotide at the abasic site comprises labeling with terminal deoxynucleotidyl transferase.

22. A method according to claim 20, wherein the labeled polynucleotide fragment comprises a biotin label.

23. A nucleic acid purified according to the method of claim 1.

24. A method for purifying a nucleic acid, comprising:

separating a nucleic acid-polycation, anionic substrate complex from other components of a sample or reaction mixture, wherein said complex is produced by

IDT Exhibit 1020.5098

contacting a sample or reaction mixture comprising said nucleic acid with a polycationic reagent and an poly-anionic substrate; and

eluting said nucleic acid from said nucleic acid-polycation-anionic substrate complex.

25. A kit for nucleic acid purification, comprising a polycationic reagent, an anionic substrate, and a buffer suitable for formation of a nucleic acid-polycation-anionic substrate complex.

26. A kit according to claim 25, wherein the polycationic reagent is polybrene and the anionic substrate is selected from the group consisting of carboxylated polystyrene, carboxylated beads, and carboxylated magnetic particles.

27. A kit according to claim 26, further comprising an elution reagent selected from a high ionic strength solution and an anionic reagent.

28. A kit according to claim 26, comprising instructions for use in a method for nucleic acid purification comprising:

contacting a sample or reaction mixture comprising said nucleic acid with a polycationic reagent and an anionic substrate, wherein a nucleic acid-polycation-anionic substrate complex is formed;

separating said nucleic acid-polycation-anionic substrate complex from other components of said sample or reaction mixture; and

eluting said nucleic acid from said nucleic acid-polycation-anionic substrate complex.

29. A method according to claim 1, wherein said nucleic acid is selected from the group consisting of mRNA, cDNA, and genomic DNA, synthetic RNA, and synthetic DNA.

IDT Exhibit 1020.5099

30. A method according to claim 20, wherein (d) comprises labeling the polynucleotide fragment at the abasic site with a label capable of reacting with an aldehyde residue at the abasic site.

31. A method according to claim 20, wherein the abasic site is labeled with trifluoroacetic acid salt (ARP).

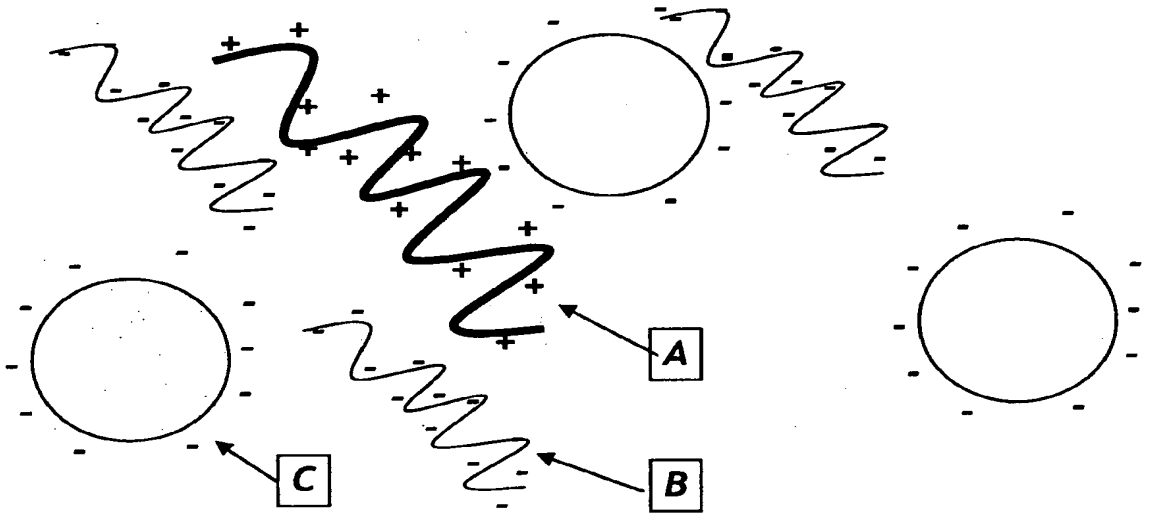
32. A method according to claim 17, wherein said nucleic acid is DNA and said nucleic acid fragment is prepared by digestion with an enzyme selected from the group consisting of DNase and a restriction endonuclease.

33. A method according to claim 17, wherein said nucleic acid fragment is prepared by chemical cleavage.

34. A method according to claim 17, wherein said nucleic acid is RNA and said nucleic acid fragment is prepared by heating the RNA at a temperature suitable to produce fragments.

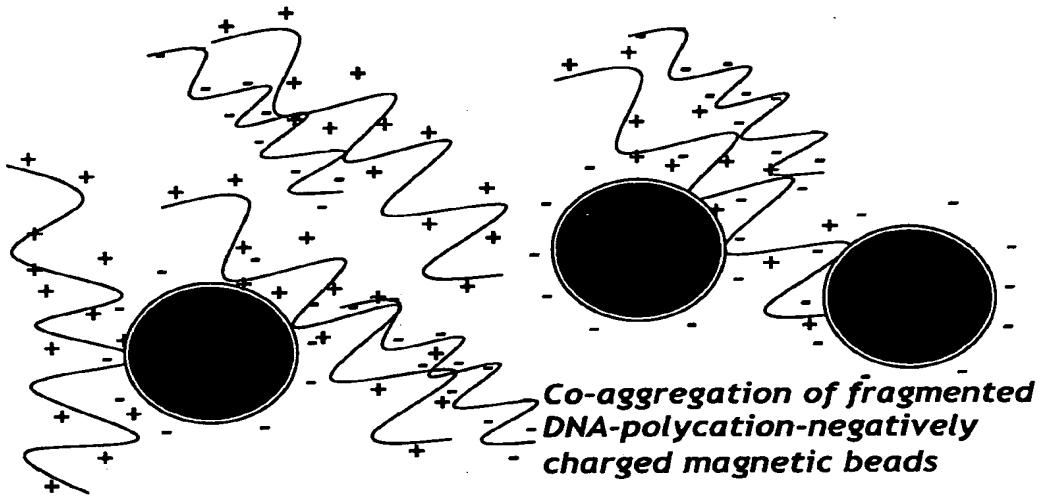
IDT Exhibit 1020.5100

FIGURE 1



IDT Exhibit 1020.5101

FIGURE 2



IDT Exhibit 1020.5102

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/011856

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/10 C12Q1/68

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N C12Q

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, WPI Data, EMBASE, BIOSIS

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	US 4 935 147 A (ULLMAN EDWIN F [US] ET AL) 19 June 1990 (1990-06-19) cited in the application the whole document	1-22, 24-34
A	WO 01/20035 A2 (NUGEN TECHNOLOGIES INC [US]; KURN NURITH [US]) 22 March 2001 (2001-03-22)	
A	EP 0 281 390 A2 (LYLE J ARNOLD JR [US]; NELSON NORMAN C [US]; REYNOLDS MARK A [US]; WAL) 7 September 1988 (1988-09-07)	
A	US 5 599 667 A (ARNOLD JR LYLE J [US] ET AL) 4 February 1997 (1997-02-04)	

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
17 October 2007	05/11/2007

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Tx. 31 651 epo nl, Fax: (+31-70) 340-3016	MARINONI, J

Form PCT/ISA/210 (second sheet) (April 2005)

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2007/011856

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 96/04404 A (MOSAIC TECHNOLOGIES INC [US]; WHITEHEAD FOR BIOMEDICAL RESEA [US]; ADA) 15 February 1996 (1996-02-15)	
A	US 5 145 784 A (COX DANIEL E [US] ET AL) 8 September 1992 (1992-09-08)	
A	EP 1 577 908 A (DU PONT [US]) 21 September 2005 (2005-09-21)	
A	US 5 279 936 A (VORPAHL JOHN [US]) 18 January 1994 (1994-01-18)	

IDT Exhibit 1020.5104

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2007/011856

Box II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This International Search Report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 23
because they relate to parts of the International Application that do not comply with the prescribed requirements to such an extent that no meaningful International Search can be carried out, specifically:
see FURTHER INFORMATION sheet PCT/ISA/210

3. ☐ Claims Nos.:
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this International Search Report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this International Search Report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this International Search Report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest.

☐ No protest accompanied the payment of additional search fees.

IDT Exhibit 1020.5105

FURTHER INFORMATION CONTINUED FROM PCT/ISA/ 210

Continuation of Box II.2

Claims Nos.: 23

Present claim 23 encompasses compounds defined only by the method used to obtain them, contrary to the requirements of clarity of Article 6 PCT, because the product-by-process type of definition does not allow the scope of the claim to be ascertained. The fact that any compound could be obtained does not overcome this objection, as the skilled person would not have knowledge beforehand as to whether it would fall within the scope claimed as virtually any nucleic acid could fall within the boundaries of the claimed subject-matter. This non-compliance with the substantive provisions is to such an extent, that no search was performed.

The applicant's attention is drawn to the fact that claims relating to inventions in respect of which no international search report has been established need not be the subject of an international preliminary examination (Rule 66.1(e) PCT). The applicant is advised that the EPO policy when acting as an International Preliminary Examining Authority is normally not to carry out a preliminary examination on matter which has not been searched. This is the case irrespective of whether or not the claims are amended following receipt of the search report or during any Chapter II procedure. If the application proceeds into the regional phase before the EPO, the applicant is reminded that a search may be carried out during examination before the EPO (see EPO Guideline C-VI, 8.5), should the problems which led to the Article 17(2) declaration be overcome.

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2007/011856

Patent document cited in search report		Publication date	Patent family member(s)		Publication date
US 4935147	A	19-06-1990	NONE		
WO 0120035	A2	22-03-2001	AT	262595 T	15-04-2004
			AU	783873 B2	15-12-2005
			AU	7483500 A	17-04-2001
			BR	0014182 A	21-05-2002
			CA	2384838 A1	22-03-2001
			CN	1373812 A	09-10-2002
			DE	60009323 D1	29-04-2004
			DE	60009323 T2	10-02-2005
			DK	1218542 T3	02-08-2004
			EP	1218542 A2	03-07-2002
			ES	2214319 T3	16-09-2004
			HK	1046021 A1	21-10-2004
			JP	3929775 B2	13-06-2007
			JP	2003509068 T	11-03-2003
			JP	2003116586 A	22-04-2003
			MX	PA02002656 A	14-10-2003
			NO	20021223 A	13-05-2002
			NZ	517121 A	28-05-2004
			PT	1218542 T	31-08-2004
EP 0281390	A2	07-09-1988	AT	107654 T	15-07-1994
			AU	1426988 A	26-09-1988
			CA	1339446 C	09-09-1997
			DE	3850273 D1	28-07-1994
			DE	3850273 T2	29-09-1994
			ES	2054797 T3	16-08-1994
			FI	885022 A	01-11-1988
			JP	1502319 T	17-08-1989
			JP	2862547 B2	03-03-1999
			PT	86881 A	30-03-1989
			WO	8806633 A1	07-09-1988
US 5599667	A	04-02-1997	NONE		
WO 9604404	A	15-02-1996	AT	216729 T	15-05-2002
			CA	2196604 A1	15-02-1996
			DE	69526511 D1	29-05-2002
			EP	0784701 A1	23-07-1997
			JP	10505492 T	02-06-1998
			US	5641658 A	24-06-1997
			US	6468751 B1	22-10-2002
			US	6090592 A	18-07-2000
			US	6060288 A	09-05-2000
US 5145784	A	08-09-1992	NONE		
EP 1577908	A	21-09-2005	EP	1577909 A1	21-09-2005
US 5279936	A	18-01-1994	US	5770388 A	23-06-1998

IDT Exhibit 1020.5107

Espacenet

Bibliographic data: WO2009112844 (A1) — 2009-09-17

METHOD OF PREPARING AN SHRNA LIBRARY

Inventor(s): TE RIELE HENRICUS PETRUS JOSEPH [NL]; WIELDERS CAMIEL LAMBERT CHRISTIAAN [NL] ± (TE RIELE, HENRICUS, PETRUS, JOSEPH, ; WIELDERS, CAMIEL, LAMBERT, CHRISTIAAN)

Applicant(s): STICHTING HET NL KANKER I [NL]; TE RIELE HENRICUS PETRUS JOSEP [NL]; WIELDERS CAMIEL LAMBERT CHRIST [NL]; DEAN JOHN PAUL [GB] ± (STICHTING HET NEDERLANDS KANKER INSTITUUT, ; TE RIELE, HENRICUS, PETRUS, JOSEPH, ; WIELDERS, CAMIEL, LAMBERT, CHRISTIAAN, ; DEAN, JOHN, PAUL)

Classification: **- international:** ***C12N15/10; C12N15/11***
- cooperative: **C12N15/1072**; **C12N15/111**; C12N2310/14; C12N2330/31

Application number: WO2009GB00684 20090313 Global Dossier

Priority number(s): GB20080004690 20080313

Abstract of WO2009112844 (A1)

There is provided a method of preparing an shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the method comprising the steps of: pre-identifying said one or more cDNA sequences using a first adaptor DNA sequence; and processing said one or more cDNA sequences using the first adaptor DNA sequence. There is further provided an shRNA library produced by such a method and uses thereof.

IDT Exhibit 1020.5108

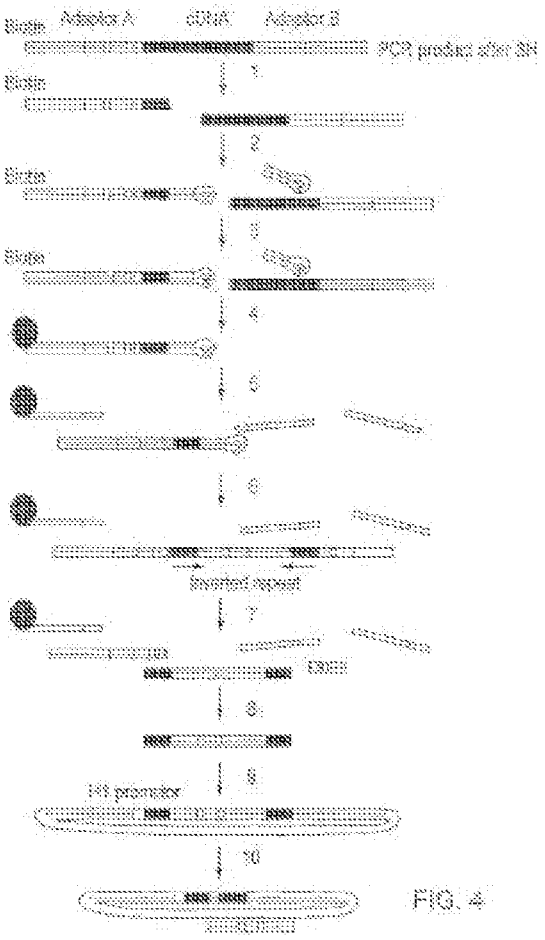


FIG. 4

IDT Exhibit 1020.5109

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 September 2009 (17.09.2009)

PCT

(10) International Publication Number
WO 2009/112844 A1

(51) International Patent Classification:
C12N 15/10 (2006.01) *C12N 15/11* (2006.01)

(21) International Application Number:
PCT/GB2009/000684

(22) International Filing Date:
13 March 2009 (13.03.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0804690.6 13 March 2008 (13.03.2008) GB

(71) Applicant *(for all designated States except US)*: **STICHTING HET NEDERLANDS KANKER INSTITUUT** [NL/NL]; Plesmanlaan 121, NL-1066 CX Amsterdam (NL).

(71) Applicant *(for LS only)*: **DEAN, John, Paul** [GB/GB]; Withers & Rogers, Goldings House, 2, Hays_Lane, London SE1 2HW (GB).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **TE RIELE, Henricus, Petrus, Joseph** [NL/NL]; Stichting Het Nederlands Kanker Instituut, Plesmanlann 121, NL-1066 CX Amsterdam (NL). **WIELDERS, Camiel, Lambert, Christiaan** [NL/NL]; Stichting Het Nederlands Kanker Instituut, Plesmanlann 121, NL-1066 CX Amsterdam (NL).

(74) Agent: **DEAN, John, Paul**; Withers & Rogers, Goldings House, 2, Hays_Lane, London SE1 2HW (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH,

[Continued on next page]

(54) Title: METHOD OF PREPARING AN SHRNA LIBRARY

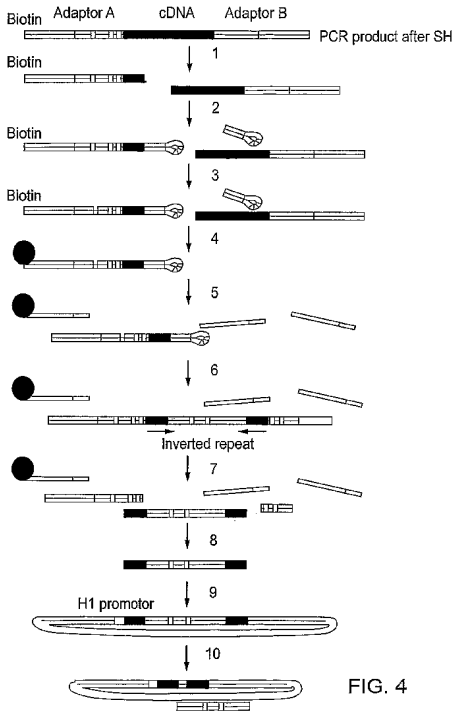


FIG. 4

(57) Abstract: There is provided a method of preparing an shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the method comprising the steps of: pre-identifying said one or more cDNA sequences using a first adaptor DNA sequence; and processing said one or more cDNA sequences using the first adaptor DNA sequence. There is further provided an shRNA library produced by such a method and uses thereof.

WO 2009/112844 A1

IDT Exhibit 1020.5110

GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

METHOD OF PREPARING AN SHRNA LIBRARY

Field Of The Invention

The invention relates to low-complexity short hairpin RNA(shRNA) libraries and to a method of producing such a library. The libraries are useful in identifying genes which are responsible for phenotypic differences existing between two cell types among the genes which are differentially expressed genes in these cell types.

Background

RNA interference (RNAi) has proved to be an important tool to knock-down gene function. Retro or lentiviral shRNA libraries targeting large parts of the human and murine genomes have been synthetically produced to study long-term effects of gene knock-down in genome-wide loss of function screens. Screening is much facilitated by a reduction of the library complexity, because fewer vectors require testing to cover the variety of library sequences, thus reducing the obligatory background. Therefore, smaller sub-libraries, for instance restricted to the deubiquitinating enzymes or the protein kinases, have been produced to reduce library complexity.

Functional screens have become a very powerful tool to identify gene function. For gain-of-function screening, cDNA libraries were expressed in cells to identify genes whose over-expression affects the behaviour of cells. In this way, stimulatory genes (including oncogenes) have been discovered (e.g., Zhu *et al.* (2005) Genetics **170** 767-77; Staudt *et al.* (2005) PLoS Genet. Oct;1(4):e55 doi:10.1371/journal.pgen.0010055; Molnar *et al*, (2006) Genetics **174** 1635-59; Mindorff *et al.* (2007) Genetics **176** 2247-63). These cDNA libraries, consisting of full-length cDNA copies of mRNA sequences, are enzymatically produced starting from RNA. Because RNA is a complex mixture, in which the number of mRNA copies of each gene varies enormously, this procedure yields very complex cDNA libraries. This means that many cDNA sequences need to be screened, in particular when the gene of interest is poorly expressed and therefore under-represented in the library. To overcome this problem, normalisation or subtractive hybridisation of cDNA preparations for gain-of-

function screening have been used to reduce the complexity of cDNA libraries (Soares *et al.* (1994) Proc Natl Acad Sci USA **91** 9228-32; Diatchenko *et al.* (1996) Proc Natl Acad Sci USA **93** 6025-30; Perrett *et al.* (1998) Gene **208** 103-15; Fischer *et al.* (2003) Biotechniques **34** 250-4; Sternberg *et al.* (2007) Methods Mol Biol. **371** 289-305; Akopyants *et al.* (1995) Proc. Natl. Acad. Sci. USA **95** 13108-13113). Normalisation aims to equalize the full-length cDNA sequences, i.e., all genes present in the library are present in equal amounts. Subtractive hybridisation is used to select cDNA sequences that are differentially expressed between cell lines, in order to reduce the complexity of the resulting cDNA libraries even further.

More recently, the advent of RNA interference (RNAi) technology, which allows suppression of gene activity, has made it possible to also perform loss-of-function screens. RNAi is conferred by short (21 nucleotides) RNA sequences, which can either be directly introduced into cells as small interfering RNAs (siRNA), or be transcribed as short hairpins (shRNA) from expression vectors to obtain long-term RNAi. Several genome-wide shRNA libraries have been produced synthetically, i.e., by inserting custom-made oligonucleotide sequences into appropriate expression vectors. Screening of such libraries has identified a series of tumour suppressor genes (Bernards *et al.* (2006) Nat Methods **3** 701-6). Synthetic shRNA libraries are of course normalized, but are also very expensive to produce. Therefore, efforts have been undertaken to produce them enzymatically starting from RNA. Again, the complexity of the RNA presents a problem similar as for cDNA libraries. Therefore, Xu *et al.* (J. Biotechnol. (2007) **128** 477-485) described a protocol to enzymatically produce shRNA libraries from normalized RNA preparations. Although in these libraries all genes are more equally represented, they are still highly complex.

One way of generating low-complexity RNAi libraries only containing genes that are differentially expressed in the cell lines under study is to perform expression profiling and then select RNAi vectors targeting differentially expressed genes from a synthetic library. However, not every laboratory has easy access to a synthetic RNAi library or the technology to perform expression profiling.

Alternatively, subtractive hybridisation may be used to pre-select cDNA sequences that are differentially expressed in cell lines under study and use these sequences to produce RNAi libraries (Zhang *et al.* (2006) Stem Cells **24** 2661-2668). However, libraries produced in this way consisted of long double-stranded RNA molecules, which are considered to confer only a mild RNAi effect, inferior to that of shRNA vectors. Further, these long double stranded RNA molecules can induce an immune response in, for example, mammalian cells. This can cause non-specific gene silencing and apoptosis which is not desired.

Summary Of The Invention

According to a first aspect of the invention, there is provided a method of preparing an shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the method comprising the steps of: pre-identifying said one or more cDNA sequences using a first adaptor DNA sequence; and processing said one or more cDNA sequences using the first adaptor DNA sequence. Preferably, the first and second cell types are phenotypically different from one another. The first and second cell types may be of different cell strains or may be of the same cell strain.

The term "shRNA library" indicates a library of vectors each comprising a short inverted repeat cDNA sequence which, when transcribed, expresses an shRNA species which can be processed by a host cell to siRNA, able to interfere with the normal process of gene expression.

The term "enriched for one or more cDNA sequences" indicates that at least one cDNA sequence, corresponding to at least a portion of a gene which is differently expressed in a first and second cell type, is present in the shRNA library more frequently than in a normalised library in which all genes are equally represented, such as a library synthesised from oligonucleotides (e.g., Bernards *et al.* (2006) Nat. Methods **3** 701-6). Preferably, the shRNA library is enriched 2-500-fold, more preferably 50-200-fold, still more preferably 100-300-fold, most preferably about 200-fold for the one or more cDNA sequences. A 200-fold enrichment would indicate that at least one cDNA sequence, corresponding to at least a portion of a gene which

is differently expressed in a first and second cell type, is present in the shRNA library 200 times more frequently than in a normalised library in which all genes are equally represented. It is within the routine abilities of the skilled person to determine such levels of enrichment in a library according to the invention.

Advantageously, the method according to the invention, including the initial step of pre-selection of putative target sequences, enables the production of very low-complexity shRNA libraries.

The term "pre-identifying said one or more cDNA sequences" indicates that genes which are differently expressed between the first and second cell types are initially selected, prior to the preparation of the shRNA library, with cDNA sequences corresponding to at least a portion of such genes being isolated.

In order to identify genetic differences between cell lines which are responsible for a selectable phenotypic difference, the library used for loss-of-function screening may be limited to the genes which are differentially expressed in these cells, because these especially are expected to contribute to the phenotypic change. For example, a low-complexity shRNA library to screen for tumour suppressor genes should primarily consist of genes whose expression is reduced in tumour cells, to test if their knock-down results in proliferation.

Therefore, the method according to the invention produces shRNA libraries for loss-of-function screening, which are strongly enriched (for example, 50-200-fold) for mRNA sequences down-regulated in the transcriptome associated with the selectable phenotype of interest compared to a control transcriptome (e.g., transformed vs. non-transformed cells). First, gene sequences are selected that are underrepresented in one transcriptome compared to another (for example, tumour suppressor genes), for example, using optimized PCR-based subtractive hybridisation. Subsequently, the selected sequences, enriched for potential target sequences, are efficiently processed into a low-complexity shRNA library. The library can then be used in a loss-of-function screen for the selectable phenotype (e.g., cell proliferation), which can identify the genes responsible for the phenotypic transition.

shRNA libraries according to the invention can be produced for loss-of-function screening in a broad range of experimental settings and species to identify genetic differences between two cell line counterparts that differ by a selectable phenotype. The shRNA libraries facilitate loss of function screening because they focus on putative target genes to reduce their complexity with respect to these genes. In addition, because the sequences which are pre-selected in the library are actually regulated during the phenotypic transition, any resulting "hits" identified by screening are likely to be relevant regulators of the phenotypic transition in the physiological context of cells.

Advantageously, the method according to the first aspect of the invention combines the steps of pre-identifying cDNA sequences and processing them into an shRNA library into one integrated procedure. This is done by using the same adaptor DNA sequence in both the first pre-identifying step and the second processing step. This reduces the complexity of the overall process and so the resulting method is simpler, easier to carry out and can be done in a relatively short period of time.

Preferably, in the method according to the first aspect of the invention, the step of pre-identifying said one or more cDNA sequences comprises using subtractive hybridisation using the first adaptor DNA sequence and a second adaptor DNA sequence. The process of subtractive hybridisation which can be used to select gene sequences that are underrepresented in one transcriptome compared to another is well known to those skilled in the art, for example, see Diatchenko *et al.* (1996) Proc. Natl. Acad Sci. USA **93** 6025-30. This method generally uses two adaptor DNA sequences which allow for PCR-based amplification of cDNA fragments that differ between a control and experimental transcriptome. The technique relies on the removal of dsDNA formed by hybridization between a control and test sample, thus reducing cDNAs of similar abundance, and retaining differentially expressed transcripts.

The two adaptor DNA sequences are generally ligated to cDNA fragments and allow the manipulation of the cDNA fragments. For example, if the adaptor DNA sequence comprises a marker, such as a stretch of adenine residues bound to Biotin, it allows

the isolation of cDNA fragments ligated to the adaptor DNA sequence. Each adaptor DNA sequence will comprise a primer binding site. This allows PCR amplification of the cDNA fragments. Preferably, Each adaptor DNA sequence has two different primer binding sites to allow nested PCR which reduces the contamination in products due to amplification of unwanted primer binding sites.

In a preferred embodiment of the invention, the pre-identifying step of the method comprises the steps of:

a. carrying out a restriction digest reaction of cDNA from the first cell type using a first restriction enzyme;
b. ligating the first adaptor DNA sequence, comprising a first primer recognition sequence and a second primer recognition sequence and a marker molecule, to DNA from step (a);
c. ligating a second adaptor DNA sequence comprising a first primer recognition sequence and a third primer recognition sequence to DNA from step (a);
d. carrying out a restriction digest reaction of cDNA from the second cell type using the first restriction enzyme;
e. mixing the denatured products of step (b) with the denatured products of step (d), optionally in the presence of a size exclusion agent, in an annealing step;
f. mixing the denatured products of step (c) with the denatured products of step (d), optionally in the presence of a size exclusion agent, in an annealing step;
g. mixing the denatured products of steps (e) and (f) in the presence of a size exclusion agent in an annealing step, wherein the concentration of the size exclusion agent is higher than the concentration of the size exclusion agent in steps (e) and (f), if present;
h. separating and isolating DNA carrying the first adaptor sequence from DNA carrying only the second adaptor sequence and/or no adaptor sequence, or separating and isolating DNA carrying the second adaptor sequence from DNA carrying only the first adaptor sequence and/or no adaptor sequence; and
i. carrying out a nested PCR reaction using the separated DNA from (h) as template DNA and including the use of a first primer which recognises the first primer recognition sequence and, subsequently, a second primer which

recognises the second primer recognition sequence and a third primer which recognises the third primer recognition sequence.

The concentration of the size exclusion agent will depend on the particular size exclusion agent used. P referably, the size exclusion agent is polyethylene glycol (PEG). Preferably, the concentration of size exclusion agent in steps (e) and (f), if present, is 1-10%. Preferably, the concentration of size exclusion agent in step (g) is 11-20%.

Having the concentration of size exclusion agent in step (g) higher than the concentration of size exclusion agent in steps (e) and (f) surprisingly causes a significant improvement in the efficiency of the pre-identifying step.

For the avoidance of doubt, not all of the steps need be carried out in the order given above; e.g., step (f) may be carried out before, or concurrently with, step (e). The term "comprises the steps of", as used throughout this specification, indicates that all steps are present but that their order need not be fixed. The skilled person would readily be able to determine a suitable alternative ordering to the steps set out above.

The first and second adaptor DNA sequences may comprise the nucleotide sequence:

GTATTACCGCACTCACTTGGAGCATCGTCCTGGCGTCTGGCTGAGGGTAGC TGGAGAGGCGATCCGACCG (SEQ ID NO. 1), which may optionally be hybridised to the sequence CGGTCGGATCGC (SEQ ID NO. 2); or

AAAAAAAAAAAAAAAAAAAGTATTACCGCACTCACTTGGACTTCTGTCAC CGTCACCGCATAGCTCATCTACGTCTTCC (SEQ ID NO. 3), which may optionally be hybridised to the sequence GGAAGACGTAGATGAGC (SEQ ID NO. 4).

The first and second adaptor sequences are preferably different to one another.

The first primer may have the nucleotide sequence:

GTATTACCGCACTCACTTGGA (SEQ ID NO. 5).

The second primer may have the nucleotide sequence:
CATCGTCCTGGCGTCTGGCT (SEQ ID NO. 6).

The third primer may have the nucleotide sequence:
CTTCTGTCACCGTCACCGCATAG (SEQ ID NO. 7).

The effect of step (i) is that cDNA carrying both the first and second adaptor sequences is amplified by the PCR reaction at an exponential level, so the PCR product is significantly dominated by cDNA carrying both adaptors.

The first restriction enzyme is preferably an enzyme having a recognition site an average of once per 250 nucleotides in a typical genomic DNA sequence, most preferably *Alu*I. An advantage of *Alu*I is that the shRNA hairpins formed start with GCT (because this is a part of the *Alu*I recognition sequence, which is AGCT). This is important for their functionality. Other suitable enzymes include: MspI, HpaII, BstUI, HinP1I, HhaI, HaeIII, PhoI, CviKI-1, TaqI, HpyCH4V, HpyCH2IV.

Preferably, the marker molecule may be any molecule which enables identification and/or separation of cDNA carrying the first adaptor sequence, for example, biotin or streptavidin. The identification of DNA carrying the first adaptor sequence, in step (h), may be carried out by use of the marker molecule, e.g., where the marker molecule is biotin, DNA carrying the marker may be separated by binding to streptavidin. Such methods are within the capabilities of the skilled person, who would be able to put them into effect without difficulty. Preferably, steps (e) and (f) of the method each continue for between 6 and 50 hours, more preferably between 40 and 50 hours, most preferably about 45 hours. In a further preference, step (g) of the method continues for about 24 hours. Preferably, the size exclusion agent (for example, PEG) is present in steps (e) and (f) at 4-8%, more preferably at about 6% and in step (g) at 15-20%, preferably at about 17.5%. The separation of DNA in step (h) may be carried out by any of the means well known to the skilled person, for

example, by gel electrophoresis, band excision and purification, or using the marker molecule present on the first adaptor, for example by affinity separation on a column.

The method may further comprise the further step of inserting at least a portion of the PCR product from (i) into an shRNA expression vector.

In the method of the invention, once differentially expressed cDNAs have been pre-identified, they can be processed to produce inverted repeat sequences encoding for shRNA molecules directed to the at least one gene which is differently expressed in the first and second cell type. This is done, at least in part, using the first adaptor DNA sequence which is ligated to the one or more cDNA sequences.

The one or more cDNA sequences can be processed using: the first adaptor DNA sequence which comprises a recognition site for a first restriction enzyme; and a third hairpin adaptor DNA sequence to form hairpin structures, allowing the formation of inverted repeats.

Preferably, in the method of the invention, the processing step comprises the steps of:

a. digesting the one or more cDNA sequences with a first restriction enzyme to produce a fragment of each of said one or more cDNA sequences attached to the first adaptor sequence;

b. ligating the product of (a) with the third hairpin adaptor DNA sequence to form hairpin structures;

c. replicating the hairpin structures of step (b) to form an inverted repeat of each of said one or more cDNA sequences;

d. removing at least a portion the duplicated first adaptor DNA sequences from the product of step (c) using a second restriction enzyme; and

e. processing the product of step (d), if necessary, to encode for a shRNA.

In step (a), the fragment of the cDNA sequence attached to the first adaptor sequence may be produced by the first restriction enzyme recognising a first recognition site in the first adaptor sequence and cleaving the cDNA attached to the first adaptor sequence at a distance from the first recognition site in the first adaptor sequence.

This would thus produce a cDNA fragment attached to the first adaptor sequence. The skilled person is well aware of restriction enzymes which cleave DNA at a distance from their recognition site, for example, *MmeI* (or another restriction enzyme which produces functional shRNA inserts, i.e., fragments which can be processed into shRNA molecules). The skilled person would have no problem in positioning the first recognition site in the first adaptor sequence in order to produce a fragment of the cDNA attached to the first adaptor sequence. The particular sequence of the first recognition site will depend on the particular restriction enzyme used and the skilled person would be able be determine this and position it accordingly. Preferably, the cDNA fragment attached to the first adaptor sequence after cleavage with the restriction enzyme is between about 10 and about 30 nucleotides in length. More preferably, the cDNA fragment is between about 15 and about 25 nucleotides in length and, most preferably, between about 18 and about 25 nucleotides in length.

In step (b), the third hairpin adaptor DNA sequence is ligated to the fragment of the cDNA sequence which is attached to the first adaptor sequence. This forms a hairpin structure. The third adaptor sequence can be any sequence which allows the formation of hairpin structures. In order to do this, the third adaptor sequence will contain two substantially complementary sequences which can bind together to form a stable hairpin structure. The two sequences do not have to be entirely complementary in order to produce a stable hairpin structure. They may be at least 90% complimentary or, more preferably, at least 95% complementary. A skilled person could readily determine a suitable sequence to give a hairpin structure. Preferably, each of the two substantially complementary sequences is between about 5 and about 50 nucleotides long, more preferably, between about 10 and about 40 nucleotides long, even more preferably, between about 20 and about 30 nucleotides long and, most preferably, about 25 nucleotides long.

In step (c), the hairpin structures are replicated to form an inverted repeat of each of said one or more cDNA sequences. This can be done in any suitable way. For example, a primer can bind to the denatured hairpin structure at a primer binding site on the first adaptor sequence. Starting from the primer, the rest of the sequence can be replicated using a polymerase, for example, the Klenow fragment of DNA

polymerase I, to produce an inverted repeat of the cDNA sequence. The primer binding site can be the same as one of the primer sites used in subtractive hybridisation or it can be different. A person skilled in the art would readily be able to determine suitable primers and binding sites.

Alternatively, the first adaptor sequence may be nicked by a third restriction enzyme which recognises a third recognition site on the first adaptor sequence. The term 'nick' means to cut only one of the strands of a double stranded DNA sequence. Nicking restriction enzymes are well known to those skilled in the art, for example, Nb.BbvC-I (or any other enzyme which produces nicks in DNA). Once the first adaptor sequence is nicked, the product can be subjected to conditions which cause the release of the nicked portion of the first adaptor sequence. A primer can then bind in the region from which the nicked portion of the first adaptor sequence was released and the hairpin structure can be replicated starting from this primer, thereby forming an inverted repeat of each the cDNA sequences.

In step (d), at least a portion of the duplicated first adaptor DNA sequences are removed from the product of step (c) using a second restriction enzyme. Preferably, the whole duplicated first adaptor sequences are removed. The second restriction enzyme recognises a second recognition site in the first adaptor sequence and cleave the first adaptor sequence from the cDNA sequence. A person skilled in the art would be able to design and position the second recognition site so that the second restriction enzyme cleaves the first adaptor sequence from the cDNA sequence. The second restriction enzyme may be *Bpm*I.

This leaves an inverted repeat of the cDNA sequence, the inverted repeat being separated by the hairpin forming structure of the third adaptor sequence. Upon transcription of this inverted repeat, a shRNA molecule will form which can interfere with gene expression.

If necessary, in step (e), the inverted repeat of the cDNA sequence can be processed to encode for a functional shRNA. For example, a portion of the third adaptor sequence which separates the inverted repeat may be removed if it is too long. This can be done

by having two restriction recognition sites in the third adaptor sequence so that a portion of the third adaptor sequence is cleaved and which can then be removed. The remaining ends of the third adaptor sequence can then be ligated together. The two restriction recognition sites may be for the same restriction enzyme or for different restriction enzymes. Preferably, they are for the same restriction enzyme. A suitable restriction enzyme is *BsgI*.

Preferably, the inverted cDNA repeat is inserted into an shRNA expression vector.

In one embodiment, step (e) can be performed after the inverted cDNA repeat has been inserted into an shRNA expression vector.

When subtractive hybridisation is used for the pre-identifying step, as discussed above, both the first and second adaptor DNA sequences will be attached to the cDNA sequences due to PCR amplification in subtractive hybridisation. Accordingly, these first and second adaptor DNA sequences need to be removed during processing of the cDNA sequences into sequences which correspond to shRNAs. The advantage of the present method is that the first adaptor DNA sequence is used both for subtractive hybridisation and for the processing of the cDNA. This means that the first adaptor sequence does not have to be removed and an alternative adaptor sequence used in the processing step. This ensures that the method is relatively simple.

Preferably, the processed cDNA sequence which encodes for a shRNA is inserted into an shRNA expression vector. Suitable expression vectors are well known to those skilled in the art.

In a slight variation to the method of the processing step described above, the processing step may comprise the steps of:

a. carrying out a restriction digest reaction of the one or more pre-identified cDNA sequences using the first restriction enzyme;

b. ligating the product of (a) with the third hairpin adaptor DNA sequence to form hairpin structures;

c. nicking the first adaptor sequence in the hairpin structure of step (b) using the third restriction enzyme;

d. isolating the product of the reaction in step (c) by binding to a marker molecule of the first adaptor sequence;

e. heating the product isolated in (d) with a primer;

f. extension of the primer by a polymerase to generate the inverted DNA sequence complementary to the product of (f); and

g. isolating at least a portion of the generated DNA product by use of the second restriction enzyme and inserting the resulting DNA into an shRNA expression vector.

The third hairpin DNA adaptor sequence may have the nucleotide sequence:

TTCAAGAGAACGCGTTGCACCGGTGCTGCACCGGTGCAGCGCGTTCTCTT GAANN (where N is A, G, C, or T) (SEQ ID NO. 8); or

TTCAAGAGAN_1N_2GCGTTGCACCGGTGCTGCACCGGTGCAGCGCn_2n_1TCTCT TGAA N_3N_4 (where N_1, N_2 , N_3, N_4, n_1 and n_2 are independently selected from A, G, C, and T, and n_1 is complementary to N_1 and n_2 complementary to N_2) (SEQ ID NO. 12).

The primer used in the processing step may have the nucleotide sequence CATCGTCCTGGCGTCTGGC (SEQ ID NO. 9).

The shRNA expression vector may be selected from, by way of non-limiting example, pRetroSuper, a retroviral or lentiviral vector, or a vector encoding inducible RNAi or microRNA primary transcripts. A final digestion with a fourth restriction enzyme, such as *Bsg*I, removes part of the hairpin adaptor and, after religation of the vector, a functional shRNA vector is formed. The skilled person would readily be able to determine alternative suitable restriction enzymes to those mentioned above. Other enzymes which may be suitable are: MspI, HpaII, BstUI, HinP1I, HhaI, HaeIII, PhoI, CviKI-1, TaqI, HpyCH4V, or HpyCH2IV.

While normalisation alone generates libraries encompassing all genes expressed in a particular tissue or cell line (Xu *et al.*, 2004), the inventors have found a method to produce libraries of very low complexity, strongly enriched for the small subset of genes which are differentially expressed in two cell types that differ by a selectable phenotype. Because the libraries of the invention focus on the few hundreds of genes that are differently regulated, screening for the actual gene that underlies the phenotypic transition becomes very easy. For the procedure, adaptors have been designed that allow subtractive hybridisation combining elements of the PCR selection method and pull down of desired sequences captured by magnetic beads. By using the same adaptors and specific restriction enzymes, selected cDNA sequences can subsequently be processed into inverted repeats that are finally introduced into shRNA-expression vectors (pSUPER). It has been found that the efficiency of subtractive hybridisation is much increased when a first round of hybridisation is performed in a low concentration (1-10% v/v) of polyethylene glycol (PEG) or another size exclusion agent, yielding a normalized number of single strands once the reaction is at equilibrium. A second hybridisation at a higher PEG concentration (11-30% v/v) shifts the equilibrium and drives further formation of hybrids. Addition of PEG enabled the optimisation of the length of the first hybridization, ultimately leading to libraries that more directly reflect the relative expression levels.

The method provides a technical solution to the problem of high cost and high complexity of current shRNA libraries, which compromise efficient screening. This is achieved by combining subtractive hybridisation and enzymatic production of shRNA libraries. To enable this, newly designed adaptors are used, which direct a more effective subtractive hybridisation procedure, followed by a new, efficient and easy procedure to process the selected sequences into shRNA libraries. The complexity of the resulting libraries is decreased by focusing on relevant sequences only, thus facilitating the subsequent screening. Such a method to produce enzymatically low complexity shRNA libraries that are enriched for differentially expressed genes has not been described or suggested in the prior art.

A second aspect of the invention relates to adaptor DNA sequences which can be used in the method described above. Accordingly, there is provided a first adaptor

nucleotide sequence for use in preparing a shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the adaptor sequence comprising:
a first primer binding site for PCR amplification of a cDNA attached thereto, for example, following subtractive hybridisation;
a first restriction recognition site for a first restriction enzyme to cause the cleavage of the cDNA attached to the first adaptor sequence into a fragment; and
a second restriction recognition site for a second restriction enzyme to cause the cleavage of the cDNA fragment from at least a portion of the first adaptor sequence.

The first adaptor sequence can be used in subtractive hybridisation to pre-identify the cDNAs with differential expression between two cell types. It is then used to process the cDNAs into an shRNA library.

It is well within the capability of a person skilled in the art to design a first primer binding site to bind a first primer to allow a cDNA attached to the first adaptor sequence to be amplified using PCR. Preferably, the first adaptor sequence comprises a second primer binding site for binding a second primer to allow nested PCR.

The first restriction recognition site is for a restriction enzyme which cuts the nucleotide sequence away from the recognition site, i.e. in the cDNA attached to the first adaptor sequence. This produces a fragment of the cDNA which is still attached to the first adaptor sequence. It is well within the capability of a person skilled in the art to select a suitable restriction enzyme, e.g. *MmeI*, and position the recognition site in the first adaptor sequence to cause cleavage of the cDNA attached to the adaptor sequence into an appropriately sized fragment.

Similarly, with regard to the second restriction recognition site, it is well within the capability of a person skilled in the art to select a suitable second restriction enzyme to recognise the second restriction recognition site and position the second recognition site in the first adaptor sequence to cause the cleavage of the cDNA fragment from at least a portion of the first adaptor sequence. For example, a suitable restriction

enzyme is *BpmI*. Preferably, the entire first adaptor sequence is cleaved from the cDNA.

Preferably, the first adaptor sequence is between about 20 and about 200 nucleotides in length. More preferably, the first adaptor sequence is between about 20 and about 120 nucleotides in length, more preferably still, the first adaptor sequence is between about 40 and about 100 nucleotides in length, even more preferably, the first adaptor sequence is between about 50 and about 90 nucleotides in length, more preferably still, the first adaptor sequence is between about 60 and about 80 nucleotides in length and, most preferably, the first adaptor sequence is about 70 nucleotides in length.

Preferably, the first restriction recognition site for the first restriction enzyme causes the cleavage of the cDNA attached to the first adaptor sequence into a fragment of between about 10 and about 30 nucleotides in length. More preferably, the cDNA fragment is between about 15 and about 25 nucleotides in length and, most preferably, between about 18 and about 25 nucleotides in length.

Preferably, the first adaptor sequence comprises a third restriction recognition site for a third restriction enzyme to introduce a nick into the first adaptor sequence. A skilled person is well aware of suitable nicking enzymes, e.g. Nb.BvCI. The nick in the first adaptor sequence provides an entry site for a polymerase, e.g. the Klenow fragment of DNA polymerase I. This allows the cDNA to be replicated into an inverted repeat when used in conjunction with the third hairpin adaptor sequence as described above.

The first adaptor sequence preferably has a third primer binding site for binding a primer to allow replication of the cDNA. When the first adaptor sequence is used in conjunction with the third hairpin adaptor sequence, the primer can be used to replicate the cDNA into an inverted repeat. When the first adaptor sequence comprises a third restriction recognition site for a third restriction enzyme to introduce a nick into the first adaptor sequence, the third primer binding site can be within the region of nucleotides which are exposed when the nicked portion of the first adaptor

sequence is released from the first adaptor sequence under suitable conditions, e.g. denaturing conditions.

The primer binding sites in the first adaptor sequence, and also in other adaptor sequences of the invention, are preferably between about 15 and about 30 nucleotides long and, more preferably, between about 18 and about 25 nucleotides long.

In one embodiment, the first adaptor sequence comprises a marker molecule. The marker molecule may be any molecule which enables identification and/or separation of the first adaptor sequence. Suitable marker molecules are well known to those skilled in the art, for example, biotin or streptavidin.

In one embodiment, the first adaptor sequence has the following sequence:
GTATTACCGCACTCACTTGGAGCATCGTCCTGGCGTCTGGCTGAGGGTAGC
TGGAGAGGCGATCCGACCG (SEQ ID NO. 1); or

The present invention also provides a second adaptor sequence which can be used in conjunction with the first adaptor sequence in subtractive hybridisation to pre-identify cDNAs with different expression between two cell types. The second adaptor sequence comprises a first primer binding site to bind a first primer for PCR amplification of a cDNA attached thereto, for example, following subtractive hybridisation. The first primer binding site of the second adaptor sequence can bind the same primer as the first primer binding site on the first adaptor sequence. Preferably, the second adaptor sequence comprises a second primer binding site for binding a second primer to allow nested PCR. Preferably, the second primer binding site of the second adaptor sequence binds a different primer than the second primer binding site of the first adaptor sequence.

In one embodiment, the second adaptor sequence comprises a marker molecule. The marker molecule may be any molecule which enables identification and/or separation of the second adaptor sequence. This can be used in subtractive hybridisation.

Preferably, the second adaptor sequence is between about 20 and about 200 nucleotides in length. More preferably, the second adaptor sequence is between about 20 and about 120 nucleotides in length, more preferably still, the second adaptor sequence is between about 40 and about 100 nucleotides in length, even more preferably, the second adaptor sequence is between about 50 and about 90 nucleotides in length, more preferably still, the second adaptor sequence is between about 60 and about 80 nucleotides in length and, most preferably, the second adaptor sequence is about 70 nucleotides in length.

In one embodiment the second adaptor sequence has the following sequence:
AAAAAAAAAAAAAAAAAAAGTATTACCGCACTCACTTGGACTTCTGTCAC
CGTCACCGCATAGCTCATCTACGTCTTCC (SEQ ID NO. 3)

The present invention also provides a third hairpin adaptor sequence. This can be any suitable sequence which allows the formation of a hairpin structure. The hairpin adaptor sequence can form hairpin structures and allows the production of inverted cDNA repeats using the method described above. In order to do this, the third adaptor sequence comprises two substantially complementary sequences which can bind together to form a stable hairpin structure. A skilled person could readily determine a suitable sequence to form a hairpin structure. As discussed earlier, the two sequences do not need to be entirely complementary. Preferably, each of the two substantially complementary sequences is between about 5 and about 50 nucleotides long, more preferably, between about 10 and about 40 nucleotides long, even more preferably, between about 20 and about 30 nucleotides long and, most preferably, about 25 nucleotides long.

Preferably, the third adaptor sequence is between about 20 and about 200 nucleotides in length. More preferably, the third adaptor sequence is between about 20 and about 120 nucleotides in length, more preferably still, the third adaptor sequence is between about 30 and about 90 nucleotides in length, even more preferably, the third adaptor sequence is between about 40 and about 80 nucleotides in length, more preferably still, the third adaptor sequence is between about 50 and about 70 nucleotides in

length and, most preferably, the third adaptor sequence is about 55 nucleotides in length.

Preferably, the third hairpin adaptor sequence comprises two restriction recognition sites for a restriction enzyme. This allows a portion of the third hairpin adaptor sequence to be removed once the cDNA inverted repeat has been produced. In the method of the invention, a suitable amount of the third hairpin adaptor can be removed to allow the formation of functional shRNA molecules as the remaining portion of the third hairpin adaptor sequence forms the hairpin section of the shRNA. Preferably, a portion of the third hairpin adaptor sequence is removed to leave between about 4 and about 15 nucleotides between the cDNA repeats and more preferably, between about 7 and about 11 nucleotides. Accordingly, a skilled person would be able to select a suitable restriction enzyme or enzymes and position the recognition sites to allow the third hairpin adaptor sequence to be processed into a suitable length for producing functional shRNAs. Preferably, the two restriction recognition sites are for the same restriction enzyme.

In one embodiment the third hairpin adaptor sequence has the following sequence: TTCAAGAGAACGCGTTGCACCGGTGCTGCACCGGTGCAGCGCGTTCTCTT GAANN (where N A, G, C, or T) (SEQ ID NO. 8); or

TTCAAGAGAN_1N_2GCGTTGCACCGGTGCTGCACCGGTGCAGCGCn_2n_1TCTCT TGAAN_3N_4 (where N_1, N_2 , N_3, N_4, n_1 and n_2 are independently selected from A, G, C, and T and n_1 is complementary to N_1 and n_2 complementary to N_2) (SEQ ID NO. 12);

According to a third aspect of the invention, there is provided a kit for preparing an shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the kit comprising a first adaptor sequence as defined above. This kit can be used in the method described above.

Preferably, the kit further comprises a second adaptor sequence as defined above for use in subtractive hybridisation. Preferably, the kit further comprises a third hairpin adaptor sequence as defined above.

Preferably, the kit further comprises suitable primers to bind to the primer binding sites on the first adaptor sequence and the second adaptor sequence.

The kit may further comprise other components, for example, chemical reagents, materials for carrying out PCR or restriction enzymes.

According to a one embodiment, there is provided a kit for use in the method according to the first aspect of the invention. The kit comprises a polynucleotide sequence comprising the sequence:
GTATTACCGCACTCACTTGGAGCATCGTCCTGGCGTCTGGCTGAGGGTAGC TGGAGAGGCGATCCGACCG (SEQ ID NO. 1);
and may comprise one or more polynucleotide sequences comprising one or more of sequences:
CGGTCGGATCGC (SEQ ID NO. 2); or
AAAAAAAAAAAAAAAAAAAGTATTACCGCACTCACTTGGACTTCTGTCAC CGTCACCGCATAGCTCATCTACGTCTTCC (SEQ ID NO. 3); or
GGAAGACGTAGATGAGC (SEQ ID NO. 4); or
GTATTACCGCACTCACTTGGA (SEQ ID NO. 5); or
CATCGTCCTGGCGTCTGGCT (SEQ ID NO. 6); or
CTTCTGTCACCGTCACCGCATAG (SEQ ID NO. 7); or
CATCGTCCTGGCGTCTGGC (SEQ ID NO. 9); or
TTCAAGAGAACGCGTTGCACCGGTGCTGCACCGGTGCAGCGCGTTCTCTT GAANN (where N is A, G, C, or T) (SEQ ID NO.8); or
TTCAAGAGAN_1N_2GCGTTGCACCGGTGCTGCACCGGTGCAGCGCn_2n_1TCTCT TGAAN_3N_4 (where N_1, N_2 , N_3, N_4, n_1 and n_2 are independently selected from A, G, C, and T and n_1 is complementary to N_1 and n_2 complementary to N_2) (SEQ ID NO. 12);
or sequences complementary to any of the above sequences.

The skilled person would readily be able to determine alternative sequences which would enable the method to be performed, subject to the requirements of, for example, the restriction enzymes, primers and marker molecules to be used.

According to a fourth aspect of the invention, there is provided an shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and a second cell type. That is, sequences from one or more genes which are expressed to the same or substantially the same level in the first and second cell types are normalised in the library, while differentially expressed genes are enriched.

The term "enriched for one or more cDNA sequences" indicates that at least one cDNA sequence, corresponding to at least a portion of a gene which is differently expressed in a first and second cell type, is present in the shRNA library more frequently than in a normalised library in which all genes are equally represented, such as a library synthesised from oligonucleotides (e.g., Bernards *et al.* (2006) Nat. Methods **3** 701-6). Preferably, the shRNA library is enriched 2-500-fold, more preferably 50-200-fold, still more preferably 100-300-fold, most preferably about 200-fold for the one or more cDNA sequences. A 200-fold enrichment would indicate that at least one cDNA sequence, corresponding to at least a portion of a gene which is differently expressed in a first and second cell type, is present in the shRNA library 200 times more frequently than in a normalised library in which all genes are equally represented.

The shRNA library is obtainable by the method of the invention. Preferably, the shRNA library is obtained by the method of the invention.

In one embodiment, a kit may comprise the shRNA library.

Advantageously, with respect to differently expressed genes, which represent potential targets of screening, such a library has a much lower complexity than prior art libraries, such that genes which differ in expression in two different cell types by only a small amount can be selected and, therefore, their effects studied.

According to a fifth aspect of the invention, there is provided a method of determining the effects of reduction in expression of a gene of interest, comprising the use of an shRNA library according to the fourth aspect of the invention.

According to a another aspect of the invention, there is provided an isolated polynucleotide sequence comprising any one of sequences:
GTATTACCGCACTCACTTGGAGCATCGTCCTGGCGTCTGGCTGAGGGTAGC TGGAGAGGCGATCCGACCG (SEQ ID NO. 1); or
CGGTCGGATCGC (SEQ ID NO. 2); or
AAAAAAAAAAAAAAAAAAAAGTATTACCGCACTCACTTGGACTTCTGTCAC CGTCACCGCATAGCTCATCTACGTCTTCC (SEQ ID NO. 3); or
GGAAGACGTAGATGAGC (SEQ ID NO. 4); or
GTATTACCGCACTCACTTGGA (SEQ ID NO. 5); or
CATCGTCCTGGCGTCTGGCT (SEQ ID NO. 6); or
CTTCTGTCACCGTCACCGCATAG (SEQ ID NO. 7); or
CATCGTCCTGGCGTCTGGC (SEQ ID NO. 9); or
TTCAAGAGAACGCGTTGCACCGGTGCTGCACCGGTGCAGCGCGTTCTCTT GAANN (where N A, G, C, or T) (SEQ ID NO. 8); or
TTCAAGAGAN_1N_2GCGTTGCACCGGTGCTGCACCGGTGCAGCGCn_2n_1TCTCT TGAAN_3N_4 (where N_1, N_2 , N_3, N_4, n_1 and n_2 are independently selected from A, G, C, and T and n_1 is complementary to N_1 and n_2 complementary to N_2) (SEQ ID NO. 12);

or sequences complementary thereto.

The skilled person would readily be able to determine alternative sequences which would enable the method to be performed, subject to the requirements of, for example, the restriction enzymes, primers and marker molecules to be used.

According to a the pre-identification step of the invention, there is provided a method of identifying at least one gene which is differently expressed in a first and a second cell type, the method comprising the steps of:

a. carrying out a restriction digest reaction of cDNA from the first cell type using a first restriction enzyme;

b. ligating a first adaptor DNA sequence, comprising a first primer recognition sequence and a second primer recognition sequence and a marker molecule, to DNA from step (a);

c. ligating a second adaptor DNA sequence comprising a first primer recognition sequence and a third primer recognition sequence to DNA from step (a);

d. carrying out a restriction digest reaction of cDNA from the second cell type using the first restriction enzyme;

e. mixing the denatured products of step (b) with denatured products of step (d), optionally in the presence of a size exclusion agent, in an annealing step;

f. mixing the denatured products of step (c) with denatured products of step (d), optionally in the presence of a size exclusion agent, in an annealing step;

g. mixing the denatured products of steps (e) and (f) in the presence of a size exclusion agent in an annealing step, wherein the concentration of the size exclusion agent is higher than the concentration of size exclusion agent in steps (e) and (f);

h. separating and isolating DNA carrying the first adaptor sequence from DNA carrying only the second adaptor sequence and/or no adaptor sequence, or separating and isolating DNA carrying the second adaptor sequence from DNA carrying only the first adaptor sequence and/or no adaptor sequence;

i. carrying out a nested PCR reaction using the separated DNA from (h) as template DNA and including the use of a first primer which recognises the first primer recognition sequence and, subsequently, a second primer which recognises the second primer recognition sequence and a third primer which recognises the third primer recognition sequence; and

j. separating and isolating the product(s) of the reaction in step (i), said product being at least a fragment of a gene which is expressed at a lower level in the second cell type than in the first cell type.

Preferably, the concentration of size exclusion agent in steps (e) and (f), if present, is 1-10%. Preferably, the concentration of size exclusion agent in step (g) is 11-30%.

The first and second adaptor DNA sequences may have the nucleotide sequence: GTATTACCGCACTCACTTGGAGCATCGTCCTGGCGTCTGGCTGAGGGTAGC TGGAGAGGCGATCCGACCG (SEQ ID NO. 1), which may optionally be hybridised to the sequence CGGTCGGATCGC (SEQ ID NO. 2); or

AAAAAAAAAAAAAAAAAAAGTATTACCGCACTCACTTGGACTTCTGTCAC CGTCACCGCATAGCTCATCTACGTCTTCC (SEQ ID NO. 3), which may optionally be hybridised to the sequence GGAAGACGTAGATGAGC (SEQ ID NO. 4)

The first and second adaptor sequences are preferably different to one another.

The first primer may have the nucleotide sequence:
GTATTACCGCACTCACTTGGA (SEQ ID NO. 5).

The second primer may have the nucleotide sequence:
CATCGTCCTGGCGTCTGGCT (SEQ ID NO. 6).

The third primer may have the nucleotide sequence:
CTTCTGTCACCGTCACCGCATAG (SEQ ID NO. 7).

Preferably, the size exclusion agent is PEG.

The first restriction enzyme is preferably an enzyme having a recognition site an average of once per 250 nucleotides in a typical DNA sequence, most preferably *Alu*I. Preferably, steps (e) and (f) of the method each continue for between 6 and 100 hours, more preferably between 30 and 50 hours, most preferably about 45 hours. In a further preference, step (g) of the method continues for about 24 hours. Preferably, the size exclusion agent (for example, PEG) is present in steps (e) and (f) at 4-8%, more preferably at about 6% and in step (g) at 15-20%, more preferably at about 17.5%.

The separation of DNA in step (h) may be carried out by any of the means well known to the skilled person, for example, by gel electrophoresis, band excision and purification, or using the marker molecule present on the first adaptor.

Brief Description Of The Figures

Embodiments of the invention will now be shown, by way of example only, with reference to the accompanying Figures 1-6 in which:

Figure 1 shows a schematic representation of the PCR based subtractive hybridisation;

Figure 2a shows normalisation and enrichment of cDNA fragments during the first hybridisation reaction.

Figure 2b shows the effect of polyethylene glycol on Hyg enrichement.

Figure 3 shows the effect of the amount of driver cDNA on the efficiency of subtractive hybridisation;

Figure 4 shows a schematic representation of the procedure to process DNA selected by subtractive hybridisation into an shRNA library using restriction sites on Adaptor A;

Figure 5 shows the difference in gene expression between a MEF cell line which grows on the surface of a culture dish cells or arrested in methyl cellulose; and

Figure 6 is a comparison of the abundance of differentially expressed genes as would be found in a completely normalised library (A), and in the shRNA library of the invention, as determined by sequencing (B); the figure shows the bins of differentially expressed genes, sequences which are not annotated as gene and genes which are not expressed in the cell lines under study.

Examples

Oligonucleotides and cDNA

All oligonucleotides were obtained from Sigma.

SH (subtractive hybridisation) Adaptor A

GTATTACCGCACTCACTTGGA*G**CATCGTCCTGGCGTCTGGCTGAGGGTAG***

CTGGAGAGGCGATCCGACCG (SEQ ID NO. 1)

(Primer 1 recognition site is bold; primer 2 recognition site is italicised)

Antisense SH Adaptor A

CGGTCGGATCGC (SEQ ID NO. 2)

SH Adaptor B

BiotinAAAAAAAAAAAAAAAAAAAA**GTATTACCGCACTCACTTGGA***CTTCT*

*GTCACCGTCACCGCATAG*CTCATCTACGTCTTCC (SEQ ID NO. 3)

(Primer 1 recognition site is bold; primer 3 recognition site is italicised)

Antisense SH Adaptor B

GGAAGACGTAGATGAGC (SEQ ID NO. 4)

SH Primer 1

GTATTACCGCACTCACTTGGA (SEQ ID NO. 5)

SH Primer 2

BiotinCATCGTCCTGGCGTCTGGCT (SEQ ID NO. 6)

SH Primer 3

CTTCTGTCACCGTCACCGCATAG (SEQ ID NO. 7)

SH Primer 4

CATCGTCCTGGCGTCTGGC (SEQ ID NO. 9)

Looped Adaptor C

TTCAAGAGAACGCGTTGCACCGGTGCTGCACCGGTGCAGCGCGTTCTCTT GAANN (where N A, G, C, or T) (SEQ ID NO. 8); or TTCAAGAGAN$_1$N$_2$GCGTTGCACCGGTGCTGCACCGGTGCAGCGCn$_2$n$_1$TCTCT TGAAN$_3$N$_4$ (where N$_1$, N$_2$, N$_3$, N$_4$, n$_1$ and n$_2$ are independently selected from A, G, C, and T and n$_1$ is complementary to N$_1$ and n$_2$ complementary to N$_2$) (SEQ ID NO. 12);

To adapt the standard pRETRO Super for insertion of inserts produced by the method of the invention, the following sequence is inserted between *Bgl*II and *Hind*III sites in the pRetroSuper vector, to allow insertion of processed fragments after BsgI digestion:

Adaptor for pRetroSuper (sense)
GATCCCCGCTGCCGTGCAGTTAACCTGCACCCGAGCTTTTTGGAAA (SEQ ID NO. 10)

Adaptor for pRetroSuper (antisense)
AGCTTTTCCAAAAAGCTCGGGTGCAGGTTAACTGCACGGCAGCGGG (SEQ ID NO. 11)

cDNA synthesis tester and driver cDNA

RNA was isolated from cells using RNeasy columns (Qiagen). cDNA was produced according to the SMART cDNA synthesis method (Clontech). cDNA was amplified using the Expand polymerase (Roche) during a maximum of 20 PCR cycles in 45 separate 50µl reactions. After PCR, the cDNA was precipitated before purification on a Microcon® YM100 column (Millipore).

Subtractive hybridisation

Adaptors A and B (Sigma) allow subtractive hybridisation and subsequent shRNA library construction. Adaptor A contains sites for library processing and B is biotinylated to allow pull down using beads. 1µg of tester cDNA preparation was *Alu*I digested and ligated in 20µl buffer 4 (New England BioLabs), supplemented with ATP (1mM), and either adaptor A or B (20µM). After incubation with *Alu*I (5U; 1

hour at 37°C), ligase (5U) was added to the reaction, which was incubated overnight in a thermocycler cycling between 30°C (10 sec) and 10°C (10 sec). After heat inactivation of the enzymes (20 min at 80°C), the samples were filtered twice to remove excess adaptors and concentrated (20ng/µl tester A and B) using a Microcon® YM 100 column.

2µg driver cDNA was digested overnight in 40 µl buffer 4 (NEB) using *Alu*I (10U). After heat inactivation of the enzyme (20 min at 80°C), the sample was concentrated (20ng/µl tester A and B) and 200ng/µl driver) using a Microcon® YM 10 column.

For subtractive hybridisation, 10ng of both the adaptor A and B ligated tester fragments were each mixed with 700ng of driver fragments. In a volume of 5µl, containing PEG (6%; MW 6/8000), NaCl (0.5mM), HEPES (100mM), and covered by mineral oil, the fragments were melted using a thermocycler (3 min 95°C) and subsequently allowed to hybridise for 45 hours at 68°C. Then, the Adaptor A and B ligated samples, and 1.25µl PEG 55% were thoroughly mixed. Hybridisation was allowed to continue for 24 hours, before the reaction was diluted by the addition of 250µl of dilution buffer (50mM NaCl, HEPES 100mM) and kept at 68°C for 10 more minutes. The reaction was removed from the thermocycler and added to 100ng of streptavidin-coated magnetic beads (Dynalbeads) in 250µl dilution buffer. After 10 minutes at room temperature (while keeping the beads in suspension), the biotinylated Adaptor B ligated fragments were pulled down from the mixture on a magnet and washed twice with dilution buffer (300µl, 5 min, 68°C). The beads were resuspended in 100µl *pfu* polymerase mix (Invitrogen) containing only dNTPs (10mM each) and heated to 75°C before the *pfu* polymerase was added to hotstart the reaction and fill in the adaptor sequences (5 min). The reaction was stopped by the addition of EDTA and the beads were recovered on a magnet. After cooling, they were washed twice with NaOH (100nM) and once in tris buffer (10mM). After recovery on a magnet, half of the beads are used as a template in a subsequent nested PCR using the *pfu* polymerase.

In a nested PCR procedure, the first PCR reaction (1000µl *pfu* PCR mix divided over 40 separate 25µl reactions) used the primer 1 and consisted of 30 cycles (15 sec 94°C;

30 sec 64°C; 90 sec 94°C). After mixing of the duplicate reactions, 2µl was used as a template in the secondary PCR (1000 µl *pfu* PCR mix divided over 40 separate 25µl reactions) using primers 2 and 3 and consisting of 16 cycles (15 sec 94°C; 30 sec 64°C; 90 sec 94°C). Primer 1 was biotinylated for further applications. After PCR, the subtracted library was precipitated before purification on a Microcon® YM100 column.

Subtractive hybridisation using isogenic Hyg+ and Hyg- cell lines.

Enrichment of differentially expressed cDNA sequences for library construction is accomplished by PCR based subtractive hybridisation, as outlined above and in Figures 1 and 2.

mRNA sequences overexpressed in one transcriptome (tester) compared to another (driver) at a ratio of t/dr are selected. First, 5' ends (but not 3' ends) of *Alu*I digested tester cDNA are ligated to either Adaptor A (step 1 of Figure 1) or B (step 2). Both Adaptor A and B ligated tester fragments are mixed separately with 75 fold excess *Alu*I digested driver cDNA (step 3), melted and allowed to hybridise. After the first hybridisation, a normalised number (n) of each fragment remains single stranded. The share of remaining single stranded fragments which carry an adaptor is proportional to their abundance in the tester (t) and driver (dr) cDNA. The equalised number n of single strands present in the two samples is mixed and, after addition of PEG (step 4), new hybrids can be formed during the second hybridisation. The number of hybrids which carry two different adaptors at their ends is then proportional to their relative expression levels in the tester and driver cDNA. Subsequently, hybrids containing the adaptor B are fished out from the mixture (step 5) and, after addition of polymerase to fill in the complementary adaptor strand, can be exponentially amplified if they carry the adaptor A at their other end (step 6). Amplification of hybrids carrying two identical adaptors is suppressed due to hairpin formation which prevents primer annealing. Hybrids which carry a single adaptor are only amplified linearly.

Two Adaptors A and B were designed to allow subtractive hybridisation and subsequent processing of selected fragments into shRNA libraries. The Adaptor A contained several restriction sites to process attached sequences into an shRNA

library. The subtractive hybridisation procedure was optimised using a pair of isogenic embryonic stem cells differing in their hyg expression. In separate reactions, the Adaptors A and B were ligated to *Alu*I digested cDNA obtained from cell line H+, which expresses the hyg gene (called tester cDNA). During a first hybridisation, the Adaptor A and -B ligated samples were allowed to anneal separately to excess cDNA obtained from cell line H-, which lacks the hyg gene (driver cDNA).

Over time, a normalized number of each of the different *Alu*I fragments present in the mixture is expected to remain single stranded due to the second order kinetics of re-hybridisation, which causes the more abundant fragments to re-anneal faster until an equilibrium concentration is reached for each different fragment. In addition, for each of the different fragments, the proportion of remaining single strands carrying either one of the adaptors A or B depends on their relative abundance in the cell line H+ and H- cDNA. Progressively more of the remaining single strands will carry an adaptor when the expression of the sequence is reduced in cell line H- relative to cell line H+.

After equilibrium was reached, the Adaptor A and B ligated samples were mixed; part of the remaining single strands were then forced to anneal during a second hybridisation by increasing the polyethylene glycol concentration to raise their effective concentration. Because the strands were first equalised, a normalised number of hybrids were formed for each of the *Alu*I fragments during the second hybridisation, including some hybrids which carry both an Adaptor A and B at their ends. Again, their number depended on the relative expression levels in the two cell lines, and was strongly enriched for sequences which are under-represented in cell line H-. Therefore, the hybrids carrying two different adaptors represent putative target sequences desired to constitute the shRNA library, which are selected in two ways.

First, after immobilising adaptor B on magnetic beads, all of the adaptor A strands, which carry the restriction sites necessary to process attached cDNA sequences into shRNA libraries, were washed away unless they were hybridised to Adaptor B strand. Subsequently, to produce a subtracted library, only fragments carrying both adaptors

A and B at their ends were exponentially amplified by PCR, using nested primers located in the two different adaptors, further restricting the resulting shRNA library to the desired sequences.

Optimisation of subtractive hybridisation procedure

To optimise subtractive hybridisation, the time required to reach equilibrium for optimal normalisation during the first hybridisation was first determined.

Adaptor A- and B-ligated tester cDNA H+ and driver cDNA H- were mixed 1:35, and the first hybridisation was allowed to proceed between 6 and 50 hours, before the two Adaptor A and B ligated samples were mixed and the PEG concentration was raised from 6% to 22%. After the second hybridisation (24 hour), the subtracted library was amplified by PCR, and the presence of a panel of 10 genes, including the differently expressed hyg gene and abundant genes (such as actin, tubulin, E cadherin, and msh6) among the selected cDNA fragments was measured using dotblots. The results are shown in Figure 2a. When the first hybridisation was allowed to proceed for 45 hours, those genes expressed equally in the two cell lines H+ and H- were reduced to normalised low levels in the subtracted library. Even the levels of the abundantly expressed sequences were effectively reduced. In contrast, the differentially expressed hyg gene was strongly enriched and exceeded the other sequences approximately 200 fold in the subtracted library. The addition of PEG during the second hybridisation increases the efficiency of the subtractive hybridisation (Fig. 2b).

In a second experiment, the effect of the amount of driver cDNA added in excess on the efficiency of subtractive hybridisation was determined.

The procedure was followed using 15, 30, 45, or 60 fold excess of driver (H-) cDNA compared to tester (H+) cDNA for subtractive hybridisation. As expected, the efficiency of subtractive hybridisation was increased when more driver cDNA was added to the reaction (Figure 3). When the first hybridisation was allowed to proceed for only 24 hours, normalisation and enrichment rapidly improved when increasing amounts of driver cDNA are added to the reaction.

Construction of shRNA libraries

Restriction sites on the Adaptor A were used to process the selected sequences into inverted repeats, which were ligated into a pRetroSuper vector. 10 µg of the amplified subtracted library was processed in 100µl buffer 4 (New England BioLabs), containing SAM (1mM) and the Looped Adaptor C. One hour after addition of *Mme*I (5U), ligase was added overnight to form hairpins by Looped Adaptor C ligation. After inactivation of these enzymes (80°C 15 min), Nb.BbvC I (5U) was added for 1 hour to nick the hairpin. On ice, streptavidin-coated magnetic beads (100ng) were used to pull down the biotinylated nicked hairpin. The supernatant was then replaced by 100 µl buffer 2 (New England BioLabs) containing primer 4 (100nM) and dNTP's (10mM each) the sample was heated (70°C 20 min) to inactivate the enzyme and release the nicked hairpins from the beads. Exo$^-$ Klenow polymerase (10U) is added to elongate the primer 4 (1 hour at 37°C) and form an inverted repeat. *Bpm*I (10U) digestion (1 hour at 37°C) removed the duplicated adaptor A ends, generating CT overhangs on both sides. The sample was separated on a 4.5% agarose gel and the 91bp inverted repeat was isolated from gel by freeze squeezing and phenol extracted. After ligation into a modified pRetroSuper vector, the plasmids were used to transform bacteria. At least 40000 individual small colonies were pooled and a maxiprep (Genomed) was performed to isolate plasmids. 1µg was digested with *Bsg*I to remove an obsolete part of the Looped Adaptor C, and the linearised 6834bp vector fragment isolated from gel (Qiagen). After religation of the vector, the plasmids were used to transform bacteria. The plasmids from 40000 individual small colonies were pooled and a maxiprep (Genomed) was performed. An aliquot of the maxiprep was used to transform *E. coli* in order to isolate and sequence individual plasmids. Sequenced inserts were compared to annotated sequences (Genbank). The library of pRETRO SUPER plasmids was calcium transfected into Phoenix cells to produce viruses.

The steps to process the selected cDNA sequences into an shRNA library are shown in Figure 4 and are outlined in Table 1:

1	*Mme*I digestion
2	Ligation with looped adaptor

3	Nicking by Nb.BbvCI enzyme
4	Immobilisation on beads and washing
5	Heating and replacement by primer
6	Replication, Klenow primed by nick
7	*Bpm*I digestion
8	Separation and isolation from gel
9	Ligation into p(RETRO)Super
10	*Bsg*I digestion and religation into p(RETRO)Super

Table 1: outline of experimental steps shown in Figure 4

Explaining now in more detail, digestion with *Mme*I, which cuts away from its recognition site in adaptor A, leaves around 18 base pairs of cDNA attached to the adaptor A. This sequence will constitute the sequence which is expressed in shRNA in the shRNA vector. The cDNA end is ligated to the looped adaptor C, forming a hairpin structure. The looped adaptor C carries a nine nucleotide sequence which ultimately ends up in between the inverted cDNA repeats in the final shRNA vector. In the hairpin consisting of adaptor A, cDNA and adaptor C, a nick is created in adaptor A using Nb.BbvCI. Since a biotinylated primer was used during PCR amplification of the subtracted library, the nicked hairpins can be immobilised on streptavidin coated beads,. Due to the nick, part of the hairpin is released from the bead by heating, leaving part of adaptor A attached to the bead. Primer 4, which binds to part of the exposed sequence of adaptor A, is added to restore the nicked hairpin. Using the nick as an entry site for DNA synthesis by the Klenow fragment of DNA polymerase I, a complementary strand is synthesized, starting from primer 4, to form a double stranded inverted repeat. *Bpm*I, which cuts 16 bp away from its restriction recognition site present in adaptor A, is used to remove the now duplicated adaptor A ends. The resulting inverted repeats of cDNA and adaptor C sequence are purified by gel electrophoresis. Subsequently, they are ligated behind an appropriate promoter sequence, preferably the H1 promoter, in a vector of choice, preferably pRETRO-SUPER. Finally, most of the adaptor C sequence is removed using *Bsg*I, which recognizes two restriction sites in adaptor C, to yield a pRETRO-Super shRNA vector. In the vector, inverted cDNA sequences are separated by the nine nucleotide sequence originally present in adaptor C.

The skilled person would readily be able to determine suitable alternative adaptors, primers, vectors and/or restriction enzymes for use in the method of the invention.

An shRNA library was produced using a subtracted library generated from cell line H+ and H- using 60 fold excess of cell line H- as a driver. To validate the resulting shRNA library, the inserted sequences in 88 shRNA plasmids were verified by sequencing. 81 of the sequences corresponded to inverted repeats capable of expressing genuine shRNA molecules. The inserted sequences matched perfectly to mRNA sequences found in GenBank. They all followed *Alu*I sites and were distributed over the full length of the corresponding mRNA. It was found that one of the 88 vectors targeted the hyg gene, the result of efficient enrichment by subtractive hybridization.

Validation of subtracted libraries.

The composition of the subtracted libraries was determined by microarray, dotblot analysis or sequencing. Microarrays were performed by the Netherlands Cancer Institute microarray facility. For dotblots, a panel of 15 genes was spotted onto a membrane and hybridised with radiolabelled subtracted libraries. The abundance of the panel genes in the libraries was quantified by phosphoimaging (Figures 2 and 3). Sequencing was performed at the Netherlands Cancer Institute sequencing facility.

Production of shRNA library from cells either grown on a culture dish or arrested in methyl cellulose.

To test the method of the invention, an shRNA library was produced, which was enriched for tumour suppressor sequences that are induced when cells arrest in methyl cellulose. Subsequently, a loss of function screen for colony growth in methyl cellulose was performed. For microarray analysis and library construction, cDNA was produced from a mouse embryonic fibroblast (MEF) cell line expressing RAS^{V12} and lacking the retinoblastoma proteins Rb and p107, which was either grown on a culture dish or arrested in methyl cellulose to induce tumor suppressor genes. Subtractive hybridization was performed using cDNA derived from the arrested cells as the tester

and from the growing cells as a driver to enrich for gene sequences induced by methyl cellulose.

Microarray analysis of cells residing in methyl cellulose revealed numerous differences in gene expression, including many genes which were induced in the arrested cells, but not in the growing cells. Figure 5 shows the expression profile of the growing and arrested cells as determined by microarray analysis. Black dots represent genes expressed equally in the two culture situations. Grey dots represent genes either overexpressed (negative M value) or repressed (positive M value) by culturing the cells in methyl cellulose. The 21331 expressed genes were divided into 5 bins: more than 4-fold induction (47 genes), 4-3 fold (83 genes), 3-2 fold (425 genes), 1-2 fold (10490 genes) or reduced expression (10440 genes). In Figure 6, the representation of the different bins as would be found in a normalised library containing all 25000 annotated genes (i.e., genes that encode a protein) (A) is compared to the representation of the bins in the library made according to the invention (B). The genes induced mostly by methyl cellulose (>4-fold induction) were enriched up to 51 fold.

Next, untransformed RAS^{V12} expressing Rb/p107 -/- cells either mock infected or infected with the shRNA library of the invention were inoculated in methyl cellulose. Although RAS^{V12} expressing Rb/p107 -/- cells normally arrest under these conditions, expression of the library of the invention caused a small number of colonies to form, indicating that this system can be used to screen for tumor suppressor sequences. Sequencing of vectors indicated that they target both coding and non-coding genes (Table 2).

Table 2

Target	Frequency	Function	Reference
MAPKAP K3	12	p38 pathway, regulation BMI1 and p19	(Voncken, Niessen et al. 2005)
Hspb1	1	p38 pathway, growth suppression and invasion	(Xu, Chen et al. 2006; Matsushima-Nishiwaki, Takai et al. 2008)
TNFa induced protein	1		
Dishevelled 3	1	*Wnt* signaling	(Polakis 2000)
Sca1 (Ly6a, ataxin)	2	Stem cell factor regulating proliferation and differentiation	(Epting, King et al. 2008)
Dppa5	2	Stem cell factor implicated in pluripotency	(Kim, Suh et al. 2005)
Ptrf (cavin)	1	Formation caveoli through regulation tumor suppressor caveolin-1	(Williams, Hassan et al. 2005; Williams and Lisanti 2005; Hill, Bastiani et al. 2008)
Enam/eIF-4G 3	6	unknown	
Non coding sequence on chromosome 7	3	unknown	
Non coding sequence on chromosome 8	1	unknown	

Conclusions

Expression profiling using microarrays may reveal many genes which are differentially expressed between two different cellular phenotypes. However, it usually fails to identify the genes responsible to bring about this phenotypic change, because it is impossible to distinguish between cause and effect. To identify the genes responsible for a phenotypic transition, functional genetic screens may be performed. In recent years, loss of function screening using viral RNAi libraries has proved to be a powerful method to study gene function. Loss of function screens are facilitated by reducing the complexity of the RNAi library. Since fewer cells require testing to cover the variety of library sequences, this benefits the ratio between truly positive "hits" and obligatory background. In addition, it becomes easier to perform screens in a single well format. Finally, when the complexity of the library becomes low enough, it may be possible to pick up interactions between genes in screens.

To reduce their complexity, RNAi libraries for loss-of-function screening should ideally be limited to genes which are suspected to suppress the particular selectable phenotype which is the focus of the screen, such as the genes which are actually differentially expressed during the phenotypic change.

Although post-translational modifications of proteins may play an important role to bring about a phenotypic change, it may be beneficial to use shRNA libraries as

disclosed here, if there are accompanying changes in expression patterns, for two reasons. First, the subpopulation of genes which are regulated during a phenotypic transition are strongly suspected to play a underlying role. At the expense of disregarding others, these candidate genes are strongly enriched to lower the complexity of libraries of the invention. Second and equally important, because the sequences selected in the library were regulated during the phenotypical transition, they actually play a role in the physiological context of cells.

Using the method of the invention, more shRNA vectors can be generated per gene compared to conventional synthetic shRNA libraries, depending on the number of *Alu*I sites present in a gene, which averages 20. This provides an internal control for off-target effects and increases the chance to come across vectors that give the right degree of functionality. In addition, libraries may be produced for every organism. In addition to pRetroSuper, other shRNA delivering vector backbones, including retroviral or lentiviral vectors, vectors encoding inducible RNAi or microRNA primary transcripts, can be used with only a minor adaptation to meet experimental demands.

An important advantage is the simplicity of the method. After isolation of RNA and the production of cDNA, it requires only three weeks to perform the subtractive hybridisation and to produce tailor-made libraries.

Claims

1. A method of preparing a shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the method comprising the steps of: pre-identifying said one or more cDNA sequences using a first adaptor DNA sequence; and processing said one or more cDNA sequences using the first adaptor DNA sequence.

2. A method according to claim 1 wherein the first and second cell types are phenotypically different from one another.

3. A method according to claim 1 or 2 wherein the first and second cell types are different cell strains.

4. A method according to claim 1 or 2 wherein the first and second cell types are the same cell strain.

5. A method according to any preceding claim, wherein the step of pre-identifying said one or more cDNA sequences comprises using subtractive hybridisation using the first adaptor DNA sequence and a second adaptor DNA sequence.

6. A method according to any preceding claim, wherein said one or more cDNA sequences are processed to produce inverted repeat sequences encoding for shRNA molecules directed to the at least one gene which is differently expressed in the first and second cell type.

7. A method according to any preceding claim, wherein said one or more cDNA sequences are processed using: the first adaptor DNA sequence which comprises a recognition site for a first restriction enzyme; and a third adaptor DNA sequence to form hairpin structures.

8. A method according to any one of preceding claim wherein the step of pre-identifying each one or more cDNA sequences comprises the steps of:
   a. carrying out a restriction digest reaction of cDNA from the first cell type using a first restriction enzyme;
   b. ligating a first adaptor DNA sequence, comprising a first primer recognition sequence and a second primer recognition sequence and a marker molecule, to DNA from step (a);
   c. ligating a second adaptor DNA sequence comprising a first primer recognition sequence and a third primer recognition sequence to DNA from step (a);
   d. carrying out a restriction digest reaction of cDNA from the second cell type using the first restriction enzyme;
   e. mixing the denatured products of step (b) with denatured products of step (d), optionally in the presence of a size exclusion agent, in an annealing step;
   f. mixing the denatured products of step (c) with denatured products of step (d), optionally in the presence of a size exclusion agent, in an annealing step;
   g. mixing the denatured products of steps (e) and (f) in the presence of a size exclusion agent in an annealing step, wherein the concentration of the size exclusion agent is higher than the concentration of size exclusion agent in steps (e) and (f), if present;
   h. separating and isolating DNA carrying the first adaptor sequence from DNA carrying only the second adaptor sequence or no adaptor sequence, or separating and isolating DNA carrying the second adaptor sequence from DNA carrying only the first adaptor sequence or no adaptor sequence;
   i. carrying out a nested PCR reaction using the separated DNA from (h) as template DNA and including the use of a first primer which recognises the first primer recognition sequence and, subsequently, a second primer which recognises the second primer recognition sequence and a third primer which recognises the third primer recognition sequence.

9. A method according to claim 8 wherein the marker molecule is biotin or streptavidin.

10. A method according to claim 8 or 9 comprising the further step of inserting at least a portion of the PCR product from (i) into an shRNA expression vector.

11. A method according to any one of claims 1 to 9, wherein the processing step comprises the steps of:

a. digesting the one or more cDNA sequences with a first restriction enzyme to produce a fragment of each of said one or more cDNA sequences attached to the first adaptor sequence;

b. ligating the product of (a) with the third hairpin adaptor DNA sequence to form hairpin structures;

c. replicating the hairpin structures of step (b) to form an inverted repeat of each of said one or more cDNA sequences;

d. removing at least a portion of the duplicated first adaptor DNA sequences from the product of step (c) using a second restriction enzyme; and

e. processing the product of step (d), if necessary, to encode for a shRNA.

12. A method according to any one of claims 1 to 9, wherein the processing step comprises the steps of:

a. carrying out a restriction digest reaction of the one or more pre-identified cDNA sequences using a second restriction enzyme;

b. ligating the product of (a) with a third adaptor sequence, comprising a recognition site for a third restriction enzyme, to form hairpin structures;

c. nicking the hairpin structure of step (b) using a third restriction enzyme;

d. isolating the product of the reaction in step (c) by binding to the marker molecule;

e. heating the product isolated in (d) with a fourth primer;

f. extension of the fourth primer by a polymerase to generate the inverted DNA sequence complementary to the product of (e); and

g. isolating at least a portion of the generated DNA product by use of a fourth restriction enzyme and inserting the resulting DNA into an shRNA expression vector.

13. A first adaptor nucleotide sequence for use in preparing a shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the adaptor sequence comprising:

a first primer binding site for PCR amplification of a cDNA attached thereto;
a first restriction recognition site for a first restriction enzyme to cause the cleavage of the cDNA attached to the first adaptor sequence into a fragment; and
a second restriction recognition site for a second restriction enzyme to cause the cleavage of the cDNA fragment from at least a portion of the first adaptor sequence.

14. The adaptor sequence of claim 13, wherein the first restriction recognition site for a first restriction enzyme causes the cleavage of the cDNA attached to the first adaptor sequence into a fragment of between about 10 and about 30 nucleotides in length.

15. The adaptor sequence of claim 13 or claim 14, further comprising a third restriction recognition site for a third restriction enzyme to introduce a nick into the first adaptor sequence.

16. The adaptor sequence of any one of claims 13 to 15, further comprising a third primer binding site for binding a primer to allow replication of the cDNA.

17. A kit for preparing an shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and second cell type, the kit comprising a first adaptor sequence according to any one of claims 13 to 16.

18. The kit of claim 17, further comprising a second adaptor sequence for use in subtractive hybridisation.

19. The kit of claim 17 or claim 18, further comprising a third hairpin adaptor sequence.

20. An shRNA library enriched for one or more cDNA sequences corresponding to at least a portion of at least one gene which is differently expressed in a first and a second cell type, wherein the shRNA library is obtainable by the method of any of claims 1 to 12.

21. A method of determining the effects of reduction in expression of a gene of interest, comprising the use of an shRNA library according to claim 20.

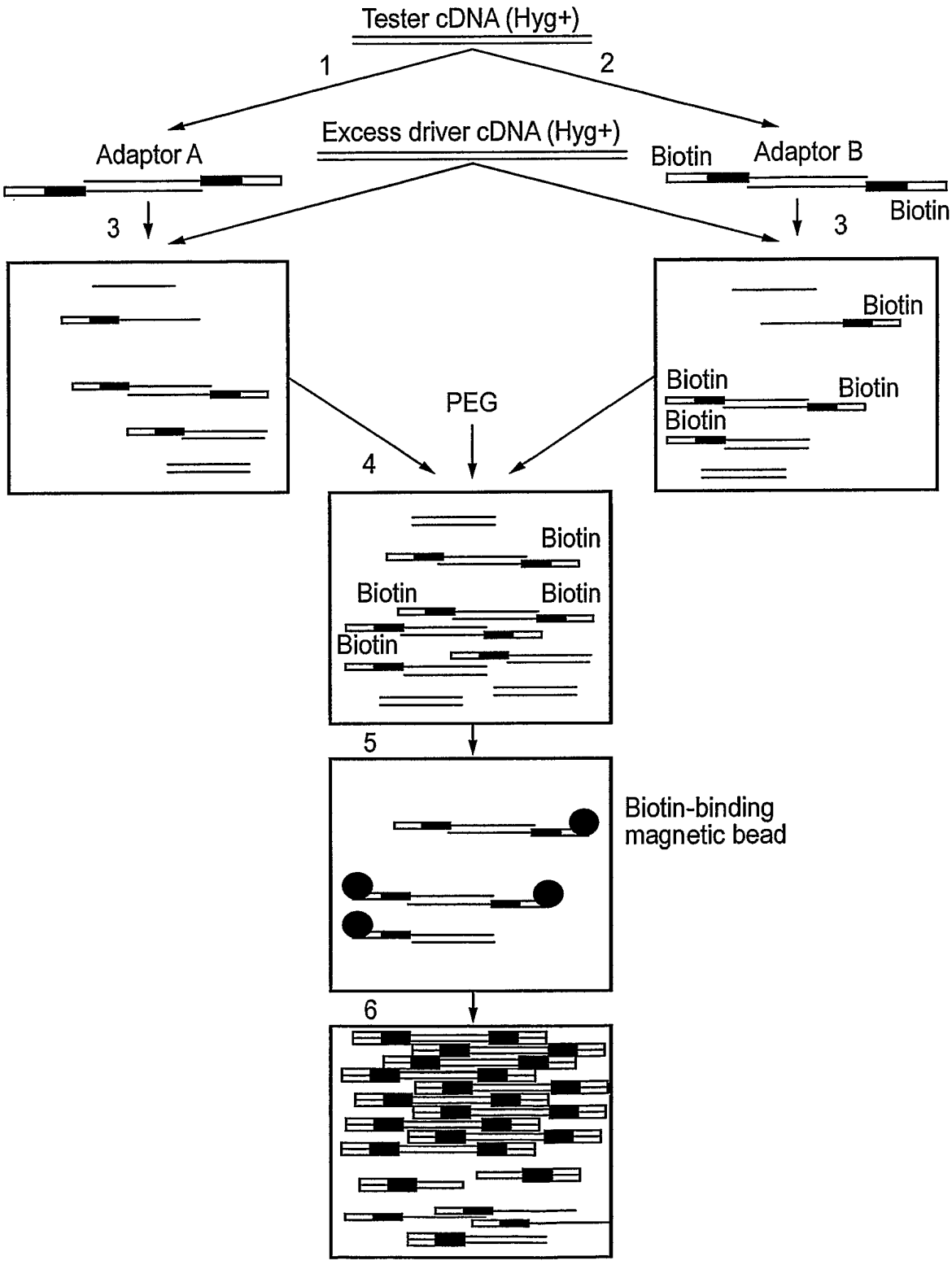

Tester cDNA (Hyg+)
1
2
Excess driver cDNA (Hyg+)
Adaptor A
Biotin
Adaptor B
Biotin
3
3
Biotin
Biotin
Biotin
PEG
4
Biotin
Biotin
Biotin
Biotin
5
Biotin-binding
magnetic bead
6

FIG. 1

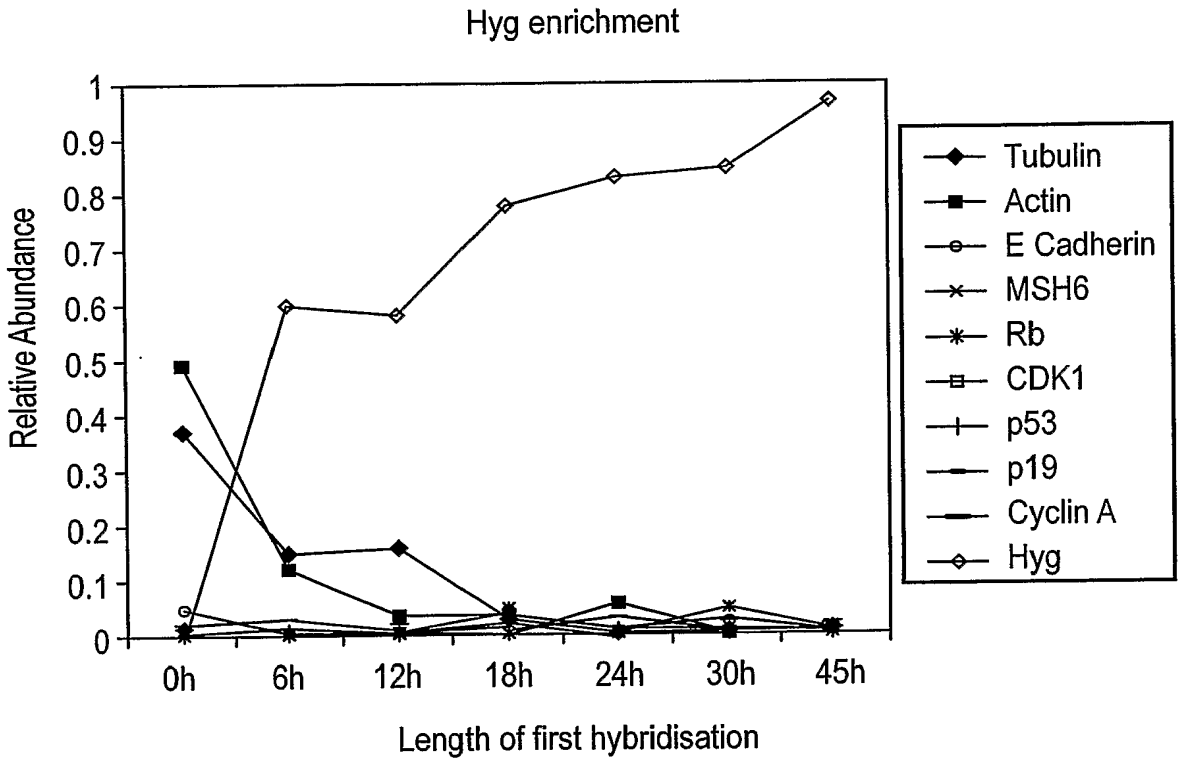

Hyg enrichment
Relative Abundance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0h
6h
12h
18h
24h
30h
45h
Length of first hybridisation
Tubulin
Actin
E Cadherin
MSH6
Rb
CDK1
p53
p19
Cyclin A
Hyg

FIG. 2A

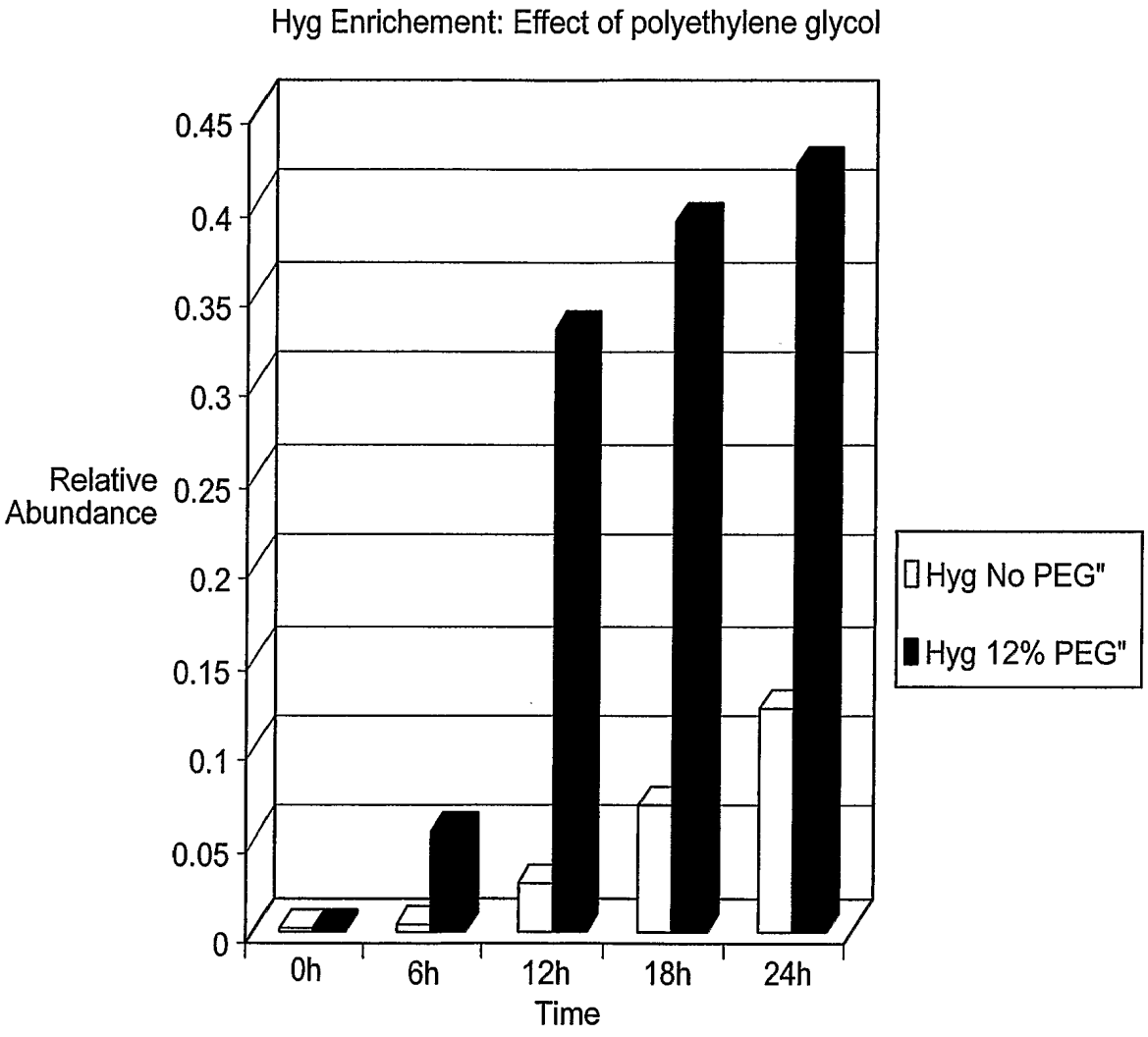

Hyg Enrichement: Effect of polyethylene glycol
Relative Abundance
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0h
6h
12h
18h
24h
Time
Hyg No PEG"
Hyg 12% PEG"

FIG. 2B

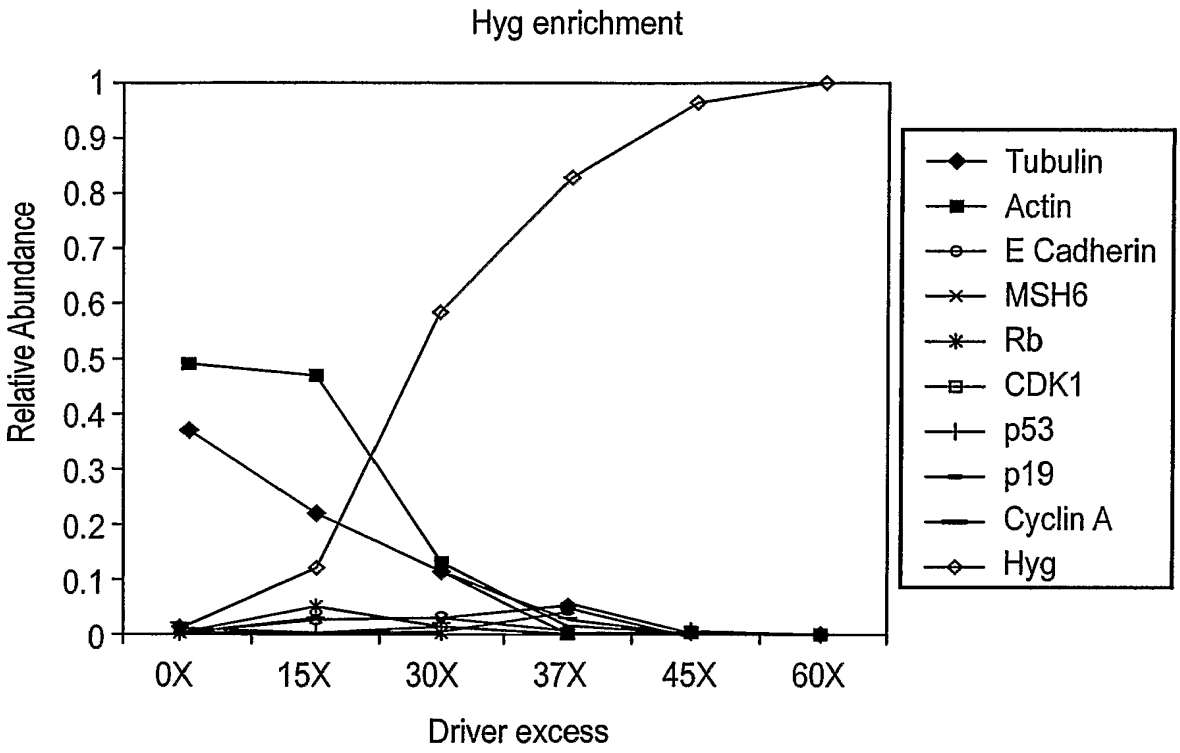

Hyg enrichment
Relative Abundance
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0X
15X
30X
37X
45X
60X
Driver excess
Tubulin
Actin
E Cadherin
MSH6
Rb
CDK1
p53
p19
Cyclin A
Hyg

FIG. 3

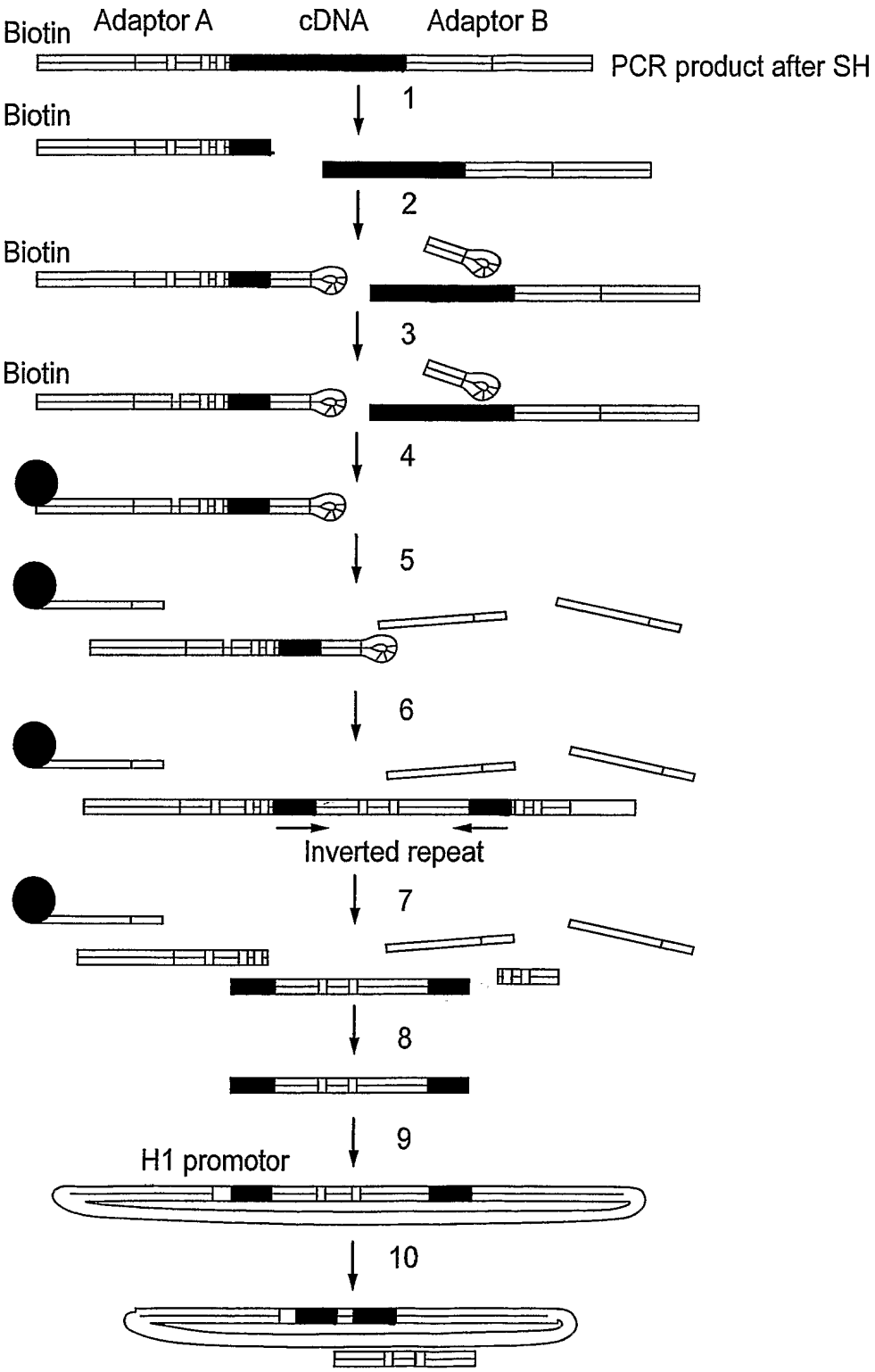

Biotin
Adaptor A
cDNA
Adaptor B
PCR product after SH
1
Biotin
2
Biotin
3
Biotin
4
5
6
Inverted repeat
7
8
9
H1 promotor
10

FIG. 4

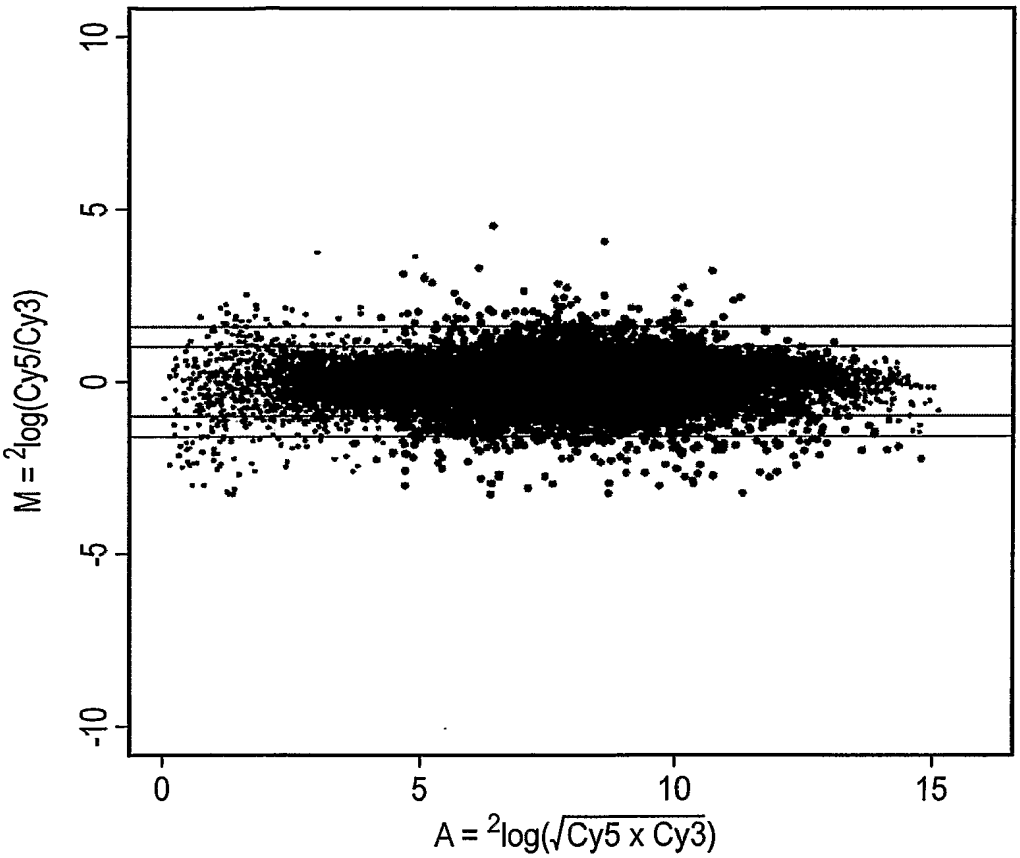


FIG. 5

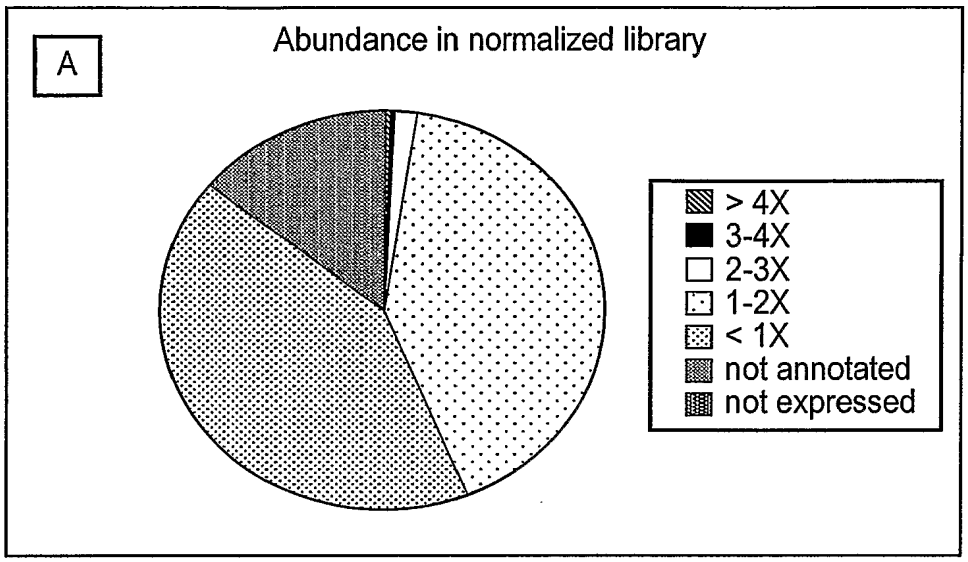

A
Abundance in normalized library
> 4X
3-4X
2-3X
1-2X
< 1X
not annotated
not expressed

FIG. 6A

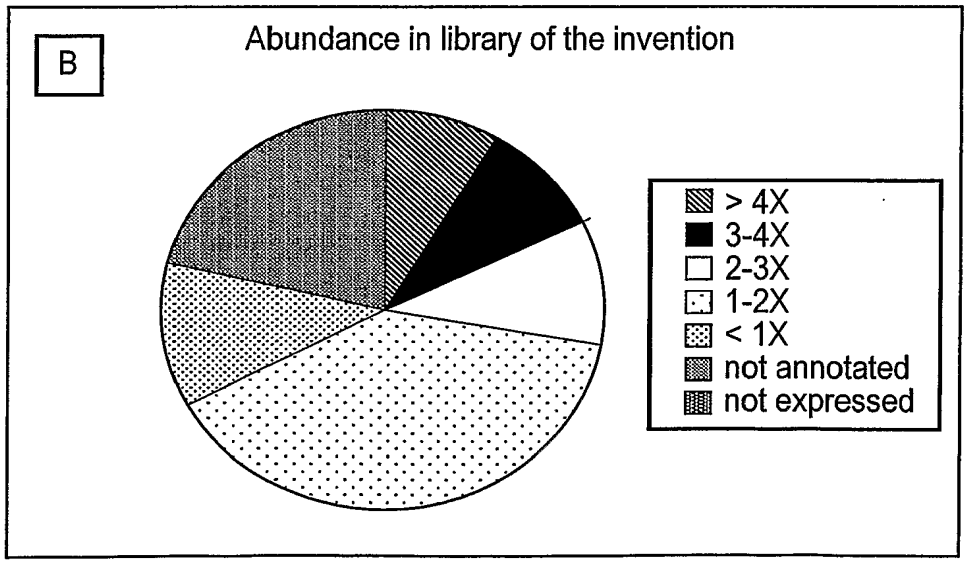

B
Abundance in library of the invention
> 4X
3-4X
2-3X
1-2X
< 1X
not annotated
not expressed

FIG. 6B

INTERNATIONAL SEARCH REPORT

International application No
PCT/GB2009/000684

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/10 C12N15/11

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, BIOSIS, EMBASE, WPI Data, Sequence Search

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	XU ET AL: "Construction of equalized short hairpin RNA library from human brain cDNA" JOURNAL OF BIOTECHNOLOGY, ELSEVIER SCIENCE PUBLISHERS, AMSTERDAM, NL, vol. 128, no. 3, 24 January 2007 (2007-01-24), pages 477-485, XP005856745 ISSN: 0168-1656 cited in the application the whole document --- -/--	1-21

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
24 June 2009	14/07/2009

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL - 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Ferreira, Roger

Form PCT/ISA/210 (second sheet) (April 2005)

IDT Exhibit 1020.5161

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	ZHANG JUN-ZHENG ET AL: "Screening for genes essential for mouse embryonic stem cell self-renewal using a subtractive RNA interference library." STEM CELLS (DAYTON, OHIO) DEC 2006, vol. 24, no. 12, December 2006 (2006-12), pages 2661-2668, XP002533583 ISSN: 1066-5099 cited in the application the whole document	1-21
X	WO 2005/023991 A (GEN HOSPITAL CORP [US]; MCCAMPBELL ALEX [US]; KAZANTSEV ALEKSEY G [US]) 17 March 2005 (2005-03-17) abstract page 2, line 27 - page 9, line 32 page 12, line 27 - page 17, line 34 page 24, line 7 - line 34 claims 1-54 figures 1-4 examples 1-8	1-21
X	WO 03/033673 A (AGY THERAPEUTICS INC [US]; MELCHER THORSTEN [US]; MCFARLAND K C [US];) 24 April 2003 (2003-04-24) abstract page 2, line 1 - page 7, line 8 page 18, line 1 - page 22, line 11 page 31, line 5 - page 36, line 22 claims 1-51 figures 1-10 examples 1-7	1-21
A	HANAZAWA M ET AL: "Use of cDNA subtraction and RNA interference screens in combination reveals genes required for germ-line development in Caenorhabditis elegans." PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF THE UNITED STATES OF AMERICA 17 JUL 2001, vol. 98, no. 15, 17 July 2001 (2001-07-17), pages 8686-8691, XP002533584 ISSN: 0027-8424 the whole document	1-21

-/--

IDT Exhibit 1020.5162

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	SILVA J M ET AL: "Second-generation shRNA libraries covering the mouse and human genomes" NATURE GENETICS, NATURE PUBLISHING GROUP, NEW YORK, US, vol. 37, no. 11, 2 October 2005 (2005-10-02), pages 1281-1288, XP002399751 ISSN: 1061-4036 the whole document	1-21
A	BERNARDS RENÉ ET AL: "shRNA libraries and their use in cancer genetics." NATURE METHODS SEP 2006, vol. 3, no. 9, September 2006 (2006-09), pages 701-706, XP002533585 ISSN: 1548-7091 cited in the application the whole document	1-21
A	BERNS K ET AL: "A large-scale RNAi screen in human cells identifies new components of the p53 pathway" NATURE, NATURE PUBLISHING GROUP, LONDON, UK, vol. 428, 25 March 2004 (2004-03-25), pages 431-437, XP003002475 ISSN: 0028-0836 the whole document	1-21
A	ZHUMABAYEVA B ET AL: "Generation of full-length cDNA libraries enriched for differentially expressed genes for functional genomics." BIOTECHNIQUES MAR 2001, vol. 30, no. 3, March 2001 (2001-03), pages 512-516 , 518, XP002533586 ISSN: 0736-6205 the whole document	1-21
A	DIATCHENKO L ET AL: "SUPPRESSION SUBTRACTIVE HYBRIDIZATION: A VERSATILE METHOD FOR IDENTIFYING DIFFERENTIALLY EXPRESSED GENES" METHODS IN ENZYMOLOGY, ACADEMIC PRESS INC, SAN DIEGO, CA, US, vol. 303, 1 January 1999 (1999-01-01), pages 349-380, XP009016828 ISSN: 0076-6879 the whole document	1-21

-/--

IDT Exhibit 1020.5163

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DIATCHENKO L ET AL: "SUPPRESSION SUBTRACTIVE HYBRIDIZATION: A METHOD FOR GENERATING DIFFERENTIALLY REGULATED OR TISSUE-SPECIFIC CDNA PROBES AND LIBRARIES" PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES OF USA, NATIONAL ACADEMY OF SCIENCE, WASHINGTON, DC.; US, vol. 93, 1 June 1996 (1996-06-01), pages 6025-6030, XP002911922 ISSN: 0027-8424 cited in the application the whole document	1-21
A	PACCHIONI B ET AL: "Semi-multiplex PCR technique for screening of abundant transcripts during systematic sequencing of cDNA libraries." BIOTECHNIQUES OCT 1996, vol. 21, no. 4, October 1996 (1996-10), pages 644-646 , 648, XP002533587 ISSN: 0736-6205 the whole document	1-21
P,X	ZHOU DEMIN ET AL: "Generation of shRNA pool library: a revision of the biological technique from the viewpoint of chemistry." CHEMBIOCHEM : A EUROPEAN JOURNAL OF CHEMICAL BIOLOGY 16 JUN 2008, vol. 9, no. 9, 16 June 2008 (2008-06-16), pages 1365-1367, XP002533588 ISSN: 1439-7633 the whole document	1-21

IDT Exhibit 1020.5164

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/GB2009/000684

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2005023991 A	17-03-2005	NONE	
WO 03033673 A	24-04-2003	CA 2461171 A1	24-04-2003
		EP 1444362 A2	11-08-2004

IDT Exhibit 1020.5165

Espacenet

Bibliographic data: WO2010063711 (A1) — 2010-06-10

METHOD FOR INTRODUCING DIVERSITY INTO AND ASSEMBLY OF POLYNUCLEOTIDE SEQUENCES

Inventor(s): MONACI PAOLO [IT]; BENDTSEN CLAUS [IT]; LAHM ARMIN [IT] ± (MONACI, PAOLO, ; BENDTSEN, CLAUS, ; LAHM, ARMIN)

Applicant(s): ANGELETTI P IST RICHERCHE BIO [IT]; MONACI PAOLO [IT]; BENDTSEN CLAUS [IT]; LAHM ARMIN [IT] ± (ISTITUTO DI RICERCHE DI BIOLOGIA MOLECOLARE P. ANGELETTI S.P.A, ; MONACI, PAOLO, ; BENDTSEN, CLAUS, ; LAHM, ARMIN)

Classification: - international: ***C12N15/10;*** *C07K16/00*
- cooperative: **C12N15/1027; C12N15/66**

Application number: WO2009EP66135 20091201 Global Dossier

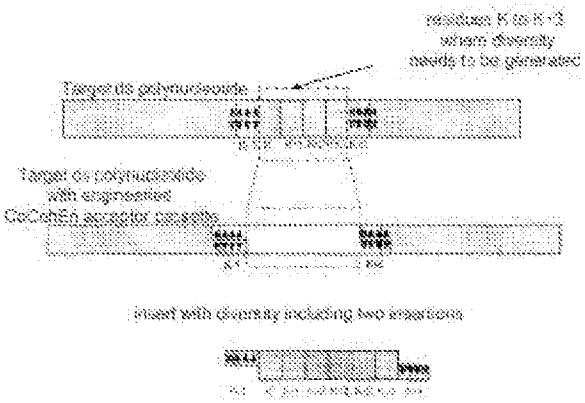
Priority number(s): GB20080022107 20081204

Also published as: GB2465986 (A)

Abstract of WO2010063711 (A1)

The invention provides a method for introducing a unique DNA sequence or a mixture of different DNA sequences within regions of selectable size and location of a target DNA vector by generating and using compatible cohesive ends. Specifically the invention provides a methodology for introducing a population of variable DNA sequences within regions of selectable size and location (e.g. segments containing complementarity determining regions in the genes coding for immunoglobulins) by generating and using compatible cohesive ends.

Figure 4



IDT Exhibit 1020.5166

IDT Exhibit 1020.5167

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 June 2010 (10.06.2010)

PCT

(10) International Publication Number
WO 2010/063711 A1

(51) International Patent Classification:
C12N 15/10 (2006.01) *C07K 16/00* (2006.01)

(21) International Application Number:
PCT/EP2009/066135

(22) International Filing Date:
1 December 2009 (01.12.2009)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
0822107.9 4 December 2008 (04.12.2008) GB

(71) Applicant *(for all designated States except US)*: **ISTITUTO DI RICERCHE DI BIOLOGIA MOLECOLARE P. ANGELETTI S.P.A.** [IT/IT]; Via Pontina KM 30.600, I-00040 Pomezia (Rome) (IT).

(72) Inventors; and

(75) Inventors/Applicants *(for US only)*: **MONACI, Paolo** [IT/IT]; IRBM, Via Pontina KM 30.600, Pomezia, I-00040 Rome (IT). **BENDTSEN, Claus** [DK/IT]; IRBM, Via Pontina KM 30.600, Pomezia, I-00040 Rome (IT). **LAHM, Armin** [DE/IT]; IRBM, Via Pontina KM 30.600, Pomezia, I-00040 Rome (IT).

(74) Agent: **BUCHAN, Gavin**; Hertford Road, Hoddesdon, Hertfordshire EN11 9BU (GB).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Declarations under Rule 4.17:

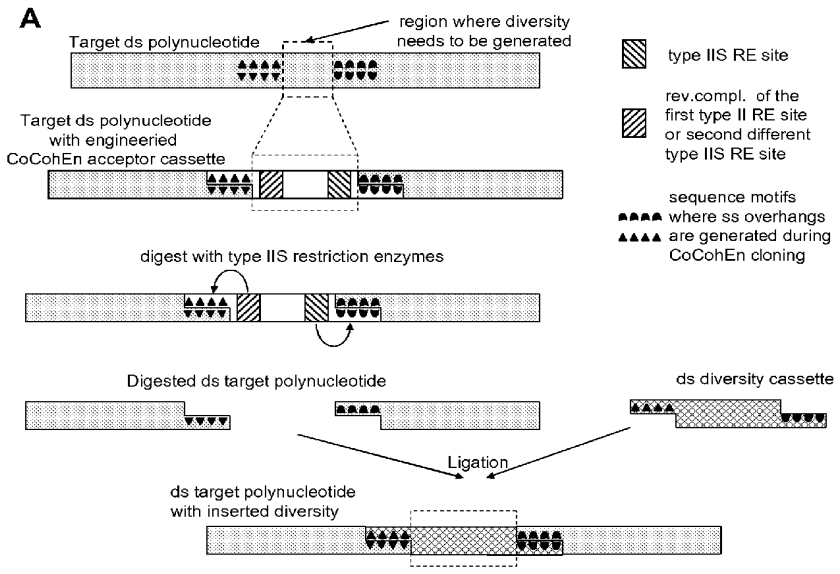
— *as to applicant's entitlement to apply for and be granted a patent (Rule 4.17(ii))*

— *as to the applicant's entitlement to claim the priority of the earlier application (Rule 4.17(iii))*

[Continued on next page]

(54) **Title**: METHOD FOR INTRODUCING DIVERSITY INTO AND ASSEMBLY OF POLYNUCLEOTIDE SEQUENCES

Figure 2



(57) **Abstract**: The invention provides a method for introducing a unique DNA sequence or a mixture of different DNA sequences within regions of selectable size and location of a target DNA vector by generating and using compatible cohesive ends. Specifically the invention provides a methodology for introducing a population of variable DNA sequences within regions of selectable size and location (e.g. segments containing complementarity determining regions in the genes coding for Immunoglobulins) by generating and using compatible cohesive ends.

WO 2010/063711 A1

Published:

— *with international search report (Art. 21(3))*

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

IDT Exhibit 1020.5169

METHOD FOR INTRODUCING DIVERSITY INTO AND ASSEMBLY OF POLYNUCLEOTIDE SEQUENCES

BACKGROUND OF THE INVENTION

Generating diversity

Methods to create diversity in specific regions of a nucleic acid sequence encoding a target protein can be grouped in two classes, depending on whether the process is performed *in vivo* or *in vitro*. A first class includes bacterial or other microorganism strains with an inherently increased frequency of mutation. Though their use is straightforward, the rate of mutagenesis is rather low. In addition, mutations are introduced in the bacterial chromosome as well as in the entire plasmid DNA containing the target sequence. As a consequence, target sequence mutants affecting viability of the bacteria are eliminated during the process.

A second class, by far the most popular, comprises enzymatic or synthetic processes to generate sequence diversity. Several PCR-based protocols have been reported. Error-prone polymerase chain reaction (PCR) uses low-fidelity polymerization conditions to introduce point mutations randomly within a nucleic acid sequence. This method can be used to mutagenize one or more nucleic acid sequence fragments, whose sequence can be unknown. However, the mutagenesis rate is too slow to allow the introduction of focused changes that are required for continuous sequence evolution. On the other hand, increasing the number of PCR cycles results in the accumulation of harmful or neutral mutations. In particular, for therapeutic proteins an undesired side effect of this approach can be the generation of point mutations that can increase the immunogenicity of the protein. Several variant protocols have been reported where variation in buffer composition alters the mutagenesis rate. Similarly, incorporating nucleotide derivatives can affect the specificity and the efficiency of the process.

DNA shuffling, also known as "sexual PCR", is an alternative method to generate diversity. It is based on in vitro homologous recombination of pools of selected mutant genes by random fragmentation and PCR reassembly. Iterative homologous recombination makes DNA shuffling a powerful protocol, for example for selecting protein mutants with increased enzymatic activity.

An alternative to enzyme-mediated random mutagenesis acting on the DNA sequence is *in vitro* synthesis of oligonucleotides which, in principle, allows the amino acid composition and relative frequency at each position in the target sequence to be defined. A number of protocols have been reported where only a subset of desired codon triplets are being used, eliminating the redundancy and occurrence of stop codons (Virnekäs *et al.* 1994; Kayushin et al. 1996; Neuner et al., 1998). Whilst allowing better control of the diversity introduced, the respective protocols can become relatively complex if the target regions to be modified are not contiguous.

IDT Exhibit 1020.5170

Introducing diversity at specific sites

Several different strategies have been adopted to assemble and clone the diversity generated through the above methods. The simplest approach is direct cloning of a mutagenized sequence into the nucleic acid sequence coding for the target protein exploiting the use of unique restriction sites compatible with the amino acid sequence of the target protein. This implies that the required restriction site(s) are already present or are engineered into the DNA sequence encoding for the target protein without altering its amino acid sequence. When this requirement is fulfilled, this approach is straightforward. In most cases two different restriction enzymes are used which generate incompatible restriction ends. This reduces the frequency of intra-molecular ligation events of the vector in the absence of the mutagenized insert. In addition it also prevents multiple DNA fragments (vector and/or insert) to polymerize and ligate together.

An effective example of this approach is the modular cloning strategy developed and implemented in WO97/08320 to introduce sequence diversity into the complementarity determining regions (CDR) of the *HuCAL Gold*™ synthetic phage-displayed antibody libraries. Unique restriction sites were engineered in the region adjacent to each of the six CDRs present in the seven light and seven heavy chain variable domain consensus sequences included in the *HuCAL Gold*™ library. These restriction sites are located in the region flanking the target antibody variable chain and were chosen to preserve the amino acid sequence of the polypeptide.

Other methods described in the literature to generate diversity using specific restriction enzymes include:

i) ONCL-DYAX WO 01/79481-A2 and Nucleic Acids Research 2005 33(9): e81. A method is described where specificity of cloning is achieved by generating local double-stranded (ds) DNA regions on a single-stranded (ss) DNA. Initially a ss oligonucleotide is synthesized which: i) hybridizes to a specific complementary sequence in the ssDNA template (and where cleavage is desired); and ii) includes a sequence that together with its complement in the ssDNA template forms a restriction endonuclease recognition site. Preferably, the restriction endonuclease recognition site is that of a type IIS restriction endonuclease whose cleavage site is located at a known distance from its recognition site. Upon incubation, the restriction enzyme binds to the dsDNA region and cleaves at the desired location within the dsDNA generated by the ssDNA template and ssDNA oligonucleotide.

ii) WO2004/013328 (COSMIX Molecular Biologicals) describes the recombination of heterologous polynucleotides comprising variant groups and/or regions wherein recombination is performed within one group of polynucleotides and/or between defined regions of polynucleotides by using different restriction sites in different groups and/or

IDT Exhibit 1020.5171

regions of polynucleotides. The overall result is therefore the generation, for each heavy chain subclass, of new heavy/light chain combinations, where only the heavy chain CDR3 region has been exchanged. The method therefore allows to considerably enlarge the complexity of a pre-existing library of antibody Fab fragments by generating new variants where the heavy chain CDR3 region of a particular heavy/light chain combination has been substituted by heavy chain CDR3 regions present in heavy/light chain combinations of the same heavy chain subclass.

iii) WO2001/000816 (Complete Genomics) describes the generation of dsDNA fragments with unique single stranded regions and then mediating the binding between a first and second nucleic acid molecule. In this method, restriction nucleases are used that form non-identical overhangs, e. g. type IIP or IIS restriction endonucleases. This method of mediation has particular applications for effectively identifying and selecting a first nucleic acid molecule fragment and then mediating its binding to a second nucleic acid molecule where this was not previously possible. The intended use is aimed at inserting, without too much explicit control, an insert sequence into a vector without, however, removing the type II restriction site. The method requires the use of adapter sequences and will in general introduce additional segments in addition to the primary insert sequence. This is not useful for generating diversity in a specific region of a polypeptide as the reading frame of the polypeptide will be enlarged and modified and in general also disrupted.

Ligation–PCR is an alternative approach. In this method partially overlapping DNA fragments which include the desired sequence diversity to be introduced at one or more locations in the target peptide are assembled and amplified by PCR. Several amplification rounds are required to efficiently combine the different diversity regions. During this process the imperfect pairing between regions containing nearly complementary diversity creates a number of undesired mutations, including deletions and inversions, which frequently generate out-of-frame mutations. These can be effectively removed by fusing a beta-lactamase coding sequence downstream of the diversity cassette

Another PCR-based approach uses an inverse-PCR reaction of the entire plasmid with the 5'-region of one or both PCR primers contains a diversified region. The ends of these amplification products are then trimmed by restriction enzymes to generate compatible ends, intra-molecularly ligated and amplified by transformation into bacteria. However, the efficiency of this approach decreases with the length of the plasmid to be amplified. Furthermore mutations are expected to occur throughout the plasmid though they will be readily detected only when leading to frame shifts or stop codons.

These examples clearly identify the need for a novel strategy enabling a direct cloning of a pool of DNA fragments anywhere into the DNA sequence of a target protein. The same approach

IDT Exhibit 1020.5172

should be applicable to any DNA sequence and at the same time preserving the amino acid sequence of the target protein outside the region of interest.

TECHNICAL PROBLEM

Often the goal of research is the identification of a protein derivative with altered properties compared to the wild type (wt) sequence (e.g. binding to a target receptor molecule). In many cases these amino acid changes need to be confined to one or more well defined segments of the target protein. For example, in the case of antibodies, changes are directed to the CDR loops. This task is usually accomplished by generating in vitro a large collection of variants of the target protein DNA sequence. In order to preserve the correct reading frame coded by the DNA sequence, the wt sequence segment has to be replaced precisely by the corresponding mutated sequence segment, the "diversity cassette". In addition, the process has to be capable of efficiently generating large numbers of mutant proteins. Cloning by means of restriction sites is by far the most efficient and reliable approach. However, in many cases, it is difficult or even impossible to identify existing or to introduce new unique restriction sites adjacent to the target segment without also altering the protein sequence outside the region of interest.

The technical problem solved by the present invention is the introduction of potentially very high sequence diversity into a defined segment of a DNA sequence which does not contain useful restriction sites. In addition for a protein-coding DNA sequence, cloning should not alter the amino acid sequence of the protein outside the region of interest.

SUMMARY OF THE INVENTION

We describe a solution to the above problem by using the compatible cohesive ends (CoCohEn) approach described below. CoCohEn is a generally applicable method based on the cloning of a single sequence or the cloning of a diversity cassette into a defined region of a target double-stranded (ds) DNA sequence (Figure 1) by means of compatible cohesive ends. The cohesive ends are, at least in the target vector, generated using restriction enzymes that recognize a defined sequence of the dsDNA and then cleave at a defined distance outside their recognition site. Engineering a vector as illustrated in Figure 2 allows the use of certain restriction enzymes (the same or two different enzymes) to generate the desired compatible cohesive ends in the target region of the dsDNA. The CoCohEn process only transiently introduces an additional sequence segment (the acceptor cassette) within the target region. Indeed, the additional segment is subsequently eliminated upon introduction of the diversity segment. The restriction enzymes used may be, but are not limited to, type IIS or type IIB restriction enzymes that cleave outside their recognition site. The additional segment containing the restriction sites is introduced at the desired region of the target sequence (or target sequences in cases of libraries of sequences). This

IDT Exhibit 1020.5173

is done in order to generate, upon digestion with the corresponding enzyme, compatible cohesive ends which unambiguously anneal to the compatible (complementary) cohesive ends in the region flanking the diversity cassette. Design of the acceptor cassette does require that the restriction sites at the 5'- and 3'-ends are correctly positioned and orientated and that these sites are not present within the rest of the vector. The sequence composition for the rest of the acceptor cassette can be chosen as one sees fit, for example by introducing a unique un-related restriction site for counter-selection by digestion of the un-digested vector. Unlike as described in WO2001/000816, no adapter is required in CoCohEn and no material is introduced in addition to what is desired, thereby reducing the risk of mutations.

Thus, the present invention provides a process for introducing a unique DNA sequence, or a mixture of different DNA sequences, in a modular fashion at specific site(s) in a vector by generating and using compatible cohesive ends (Figure 2 and 3). More specifically, the invention provides a protocol, CoCohEn, for introducing a population of variable DNA sequences at target sites (e.g. complementarity determining regions) in polynucleotides coding, for example, for the immunoglobulin variable domains. The disclosed CoCohEn method can be used to introduce a dsDNA fragment with a unique sequence, or a mixture of dsDNA fragments, with different sequence and/or length, into an expression vector or, more generally, into any dsDNA segment. In particular, the CoCohEn protocol can be used to create antibody libraries that are suitable either for antibody lead discovery or for optimization (maturation or humanization) of existing antibodies.

As shown herein, the CoCohEn protocol can be used to replace specific regions of a polynucleotide sequence, for example an antibody CDR that is targeted for diversification. More specifically, the protocol can be used to diversify any of the three CDR regions present in an antibody heavy or light chain polypeptide by introducing one or more cassettes encoding diversified polynucleotide sequences into specific locations of a recombinant vector that has been engineered appropriately to accept diversity inserts with compatible cohesive ends.

The disclosed method provides an efficient and simple cloning procedure that is generally applicable to the cloning of a sequence or a group of sequences, into a vector at one or more desired target position(s). Generally speaking, the disclosed method uses restriction enzymes which cut at a defined distance from their recognition site, with the distance being restriction enzyme specific and independent of the nucleotide sequence present at the site of restriction and independently of the sequence present between the restriction enzyme recognition site and the site of the cut. The nucleotide sequence of the resulting overhangs will thus depend on the sequence flanking the recognition site and is appropriately chosen in order to vary from one site to another thereby generating site specific cohesive ends with distinct single stranded overhangs.

IDT Exhibit 1020.5174

In general, an acceptor cassette containing restriction enzyme (RE) recognition motifs is inserted into the vector in order to place the cohesive ends which are generated upon digestion adjacent to the region where diversity needs to be introduced. For example, if diversity needs to be introduced in a region of a protein comprising residues K to K+3, the acceptor cassette will in general be positioned in the vector in such a way that cohesive ends are generated around residue K-1 or K-2 and around K+4 or K+5 (Figure 4). Start and end of the cohesive ends do not need to coincide with a specific nucleotide of the residue's codon triplet since appropriate design of the diversity cassette ensures that the correct peptide sequence and correct reading frame that needs to be maintained is restored after insertion of the diversity cassette.

A diversity cassette, containing a collection of dsDNA fragments harboring diversity in a defined region flanked both at their 5'- and 3'-ends by single-stranded overhangs (cohesive ends) complementary to those present in the acceptor cassette described above is generated and ligated into the digested vector.

Methods to generate these compatible cohesive ends in the diversity cassette include but are not limited to introducing type II RE recognition motifs at both the 5'- and the 3'-end of the diversity cassette. When using such motifs, these are localized and oriented to generate, after digestion of the dsDNA, single-stranded overhangs complementary to those generated by the digestion of the acceptor cassette as previously described. The type II RE recognition motifs present within the 5' and 3' flanking region might be recognized by the same type II RE or represent distinct motifs recognized by two different type II REs.

In summary the invention is a method of introducing diversity at a desired point into one or more target polynucleotide sequences and assembling a library of double-stranded DNA sequences, wherein said method comprises the steps of:

a) providing one or more double-stranded polynucleotide sequences each of which is engineered to contain at least one recognition motif for a type II restriction enzyme that is not native to that polynucleotide sequence in a region common to all the polynucleotide sequences targeted for diversification,

b) digesting the polynucleotide sequences of step (a) with a type II restriction enzyme that recognizes the recognition motif and, if necessary, one or more further type II restriction enzymes in the same region of the polynucleotides such that a portion of the polynucleotide sequences can be excised to give polynucleotide sequences with overhanging single stranded nucleotide ends of two to seven nucleotides,

c) preparing a collection of dsDNA variant fragments having at their 5' and 3' ends overhanging single stranded nucleotide ends that are complementary to the ones generated in step (b), and

IDT Exhibit 1020.5175

d) ligating the dsDNA variant fragments of step (c) into the restriction enzyme digested polynucleotide sequences of step (b) to give a diverse library of ds DNA sequences.

DESCRIPTION OF THE DRAWINGS

Figure 1: Application of the CoCohEn method allows inserting diversity in the region of dsDNA between sites X and Y. The method has the advantage that the location of sites X and Y can be chosen as desired.

Figure 2 A): Generation of vector applicable to CoCohEn cloning. A cassette containing type IIS restriction enzyme recognition sites is inserted into the region of the polynucleotide sequence where diversity needs to be inserted. Recognition sites are placed within the cloning cassette in such a way that the digestion with the type IIS restriction enzymes generates single-strand overhangs flanking the region where diversity needs to be introduced. One or two different type IIS restriction enzymes can be used. A suitable ds diversity cassette containing complementary single-strand overhangs is then ligated into the digested vector.

Figure 2 B): CoCohEn cloning using a cloning cassette containing a type IIB restriction enzyme recognition site. The recognition site is placed in such a way that single-strand overhangs flanking the region where diversity needs to be introduced are generated on digestion. A suitable ds diversity cassette containing complementary single-strand overhangs is then ligated into the digested vector.

Figure 3 A): Generation of a ds diversity cassette for CoCohEn cloning. Flanking the diversity segment are two regions containing type IIS RE recognition sites positioned at the right distance and in the right orientation. One or two different type IIS restriction enzymes can be used. On digestion single-strand overhangs complementary to those present in the digested target vector are generated.

Figure 3 B): Same as Figure 3 A) but using instead a type IIB RE site appropriately placed and oriented in the flanking regions. Suitable cassettes can also be prepared using a combination of type IIS and type IIB restriction enzymes. Also, if compatible with the composition of the desired diversity, the restriction sites can also be located within the diversity region.

Figure 3 C): Generation of diversity cassettes using a suitable combination of restriction enzymes sites compatible with the generation of single-strand overhangs complementary to those present in the digested target vector. Also, cassettes might be generated using a combination of approaches A, B and C. D) Generation of diversity inserts by annealing of complementary single-stranded oligonucleotides. Following annealing single-strand overhangs complementary to those present in the digested target vector are present.

Figure 4: Example illustrating how diversity is generated in a region covering residues K to K+3 of a peptide sequence, including insertion of two additional residues. In the target

IDT Exhibit 1020.5176

polynucleotide (vector) the cohesive ends are positioned around residues K-1, K-2 and K+4, K+5. On insertion of the appropriately designed diversity cassette the correct peptide sequence is restored. The precise positioning of the cohesive ends (acceptor cassette) in the target polynucleotide vector can be chosen as appropriate because the design of the diversity cassette can be adjusted accordingly. Also, co-incidence of the start and end of the cohesive ends with the start or the end of a residue codon triplet is not required.

Figure 5: Variable domain of the light chain of the vector 3.1_5B. BpiI and KpnI sites are indicated.

Figure 6: Forward (Primer 3.1_Lfor) and reverse (Primer3.1_Lrev) primer used to generate the acceptor cassette. BpiI and KpnI sites are indicated

Figure 7: dsDNA VLes construct. BpiI and KpnI sites are indicated as are the Esp3I site

Figure 8: Acceptor cassette at the light chain CDR3 region of plasmid p3.1_Les obtained by ligation of the BpiI and KpnI digested dsDNA_VLes construct into the BpiI and KpnI digested 3.1_5B vector. Esp3I sites of the acceptor cassette are indicated.

Figure 9: Composition of the degenerate oligonucleotides CDR3-L Oligo-S, Oligo-S2 and the reverse primer CDR3-L-S12. The Esp3I sites present in the oligonucleotide sequences are underlined.

Figure 10: PCR product dsDNA_Vlib obtained by elongating oligonucleotides CDR3-L-S1 and CDR3-L-S2 with the reverse primer CDR3-L-S12. The Esp3I sites present in the oligonucleotide sequences are underlined.

Figure 11: Cloning of the dsDNA_Vlib_Insert insert into the Esp3I digested p3.1_Les vector. The acceptor cassette in p3.1_Les is replaced by the insert dsDNA_Vlib_Insert. Amino acid diversity present at the desired site in the resulting ligation product p3.1_Vlib is indicated.

Figure 12: Observed and expected amino acid residue frequency in p3.1_Vlib.

DETAILED DESCRIPTION OF THE INVENTION

The present invention provides: A method of introducing diversity at a desired point into one or more target polynucleotide sequences and assembling a library of double-stranded DNA sequences, wherein said method comprises the steps of:

a) providing one or more double-stranded polynucleotide sequences each of which is engineered to contain at least one recognition motif for a type II restriction enzyme that is not native to that polynucleotide sequence in a region common to all the polynucleotide sequences targeted for diversification,

b) digesting the polynucleotide sequences of step (a) with a type II restriction enzyme that recognizes the recognition motif and, if necessary, one or more further type II restriction enzymes in the same region of the polynucleotides such that a portion of the

IDT Exhibit 1020.5177

polynucleotide sequences can be excised to give polynucleotide sequences with overhanging single stranded nucleotide ends of two to seven nucleotides,

c) preparing a collection of dsDNA variant fragments having at their 5' and 3' ends overhanging single stranded nucleotide ends that are complementary to the ones generated in step (b), and

d) ligating the dsDNA variant fragments of step (c) into the restriction enzyme digested polynucleotide sequences of step (b) to give a diverse library of ds DNA sequences.

Recognition motifs for any type II restriction enzymes might be used provided that, on digestion with the appropriate type II restriction enzymes, overhanging single-stranded ends are generated which are complementary to the cohesive ends generated in the target polynucleotide sequences on digestion of the acceptor cassette. However, the use of at least one type IIB or type IIS restriction enzyme is preferred for use in the methods of the present invention. If a type IIS restriction enzyme is used in the methods of the present invention, then preferably there are two recognition motifs for restrictions enzymes in the polynucleotide sequence, at least one of which recognition motifs is not native to that polynucleotide sequence, and preferably there are at least two recognition motifs for type IIS restriction enzyme that are not native to that polynucleotide sequence. However, if a type IIB restriction enzyme is used in the methods of the present invention, then the use of only one restriction enzyme is necessary together with a recognition motif for that restrictions enzyme.

Suitably the ds polynucleotide encodes a polypeptide, for example a polypeptide which is an antibody variable heavy or light chain.

The dsDNA variant fragments to be inserted in the polynucleotide sequence may be generated by digestion of ds DNA fragments comprising within their 5' and 3' segment type II restriction enzymes recognition sites. Thus, the dsDNA variant fragments to be inserted in the polynucleotide sequence may be generated by digestion of ds DNA fragments comprising within their variable regions one or two IIB restriction enzymes recognition sites that generate, on digestion with the appropriate type IIB restriction enzymes, cohesive ends which are complementary to the overhanging ends generated by digestion of the polynucleotide sequence. The dsDNA variant fragments to be inserted in the polynucleotide sequence may be generated by starting from a collection of single-stranded oligonucleotides containing the desired variability; a reverse primer being annealed to the 3' end of each of the collection of the single-stranded oligonucleotides followed by double-strand elongation with an appropriate DNA polymerase to give the ds DNA variant fragments. Oligonucleotides of the collection contain a 5' segment complementary to the respective cohesive end generated at the 5' end of the acceptor cassette. The reverse primer contains a 5' segment complementary to the respective cohesive end generated at the 3' end of the acceptor cassette. After annealing the reverse primer to the collection of

IDT Exhibit 1020.5178

single-stranded oligonucleotides the 5' segments of the oligonuclotides and the reverse primer representing the complement of the respective cohesive ends of the acceptor cassette are protected by addition of appropriate short single-stranded RNA molecules before initiation of the polymerase reaction. After completion of the polymerase reaction the RNA components of the resulting double-strand are digested by RNAse thus generating dsDNA variant fragments with cohesive ends complementary to the cohesive ends generated in the target polynucleotide sequences on digestion of the acceptor cassette.

Alternatively, the dsDNA variant fragments to be inserted in the polynucleotide sequence may be generated by annealing two complementary collections of oligonucleotides where sequences within each collection of oligonucleotides contains a 5' segment remaining single-stranded after annealing and the 5' segment being complementary to the cohesive ends generated in the target polynucleotide sequences on digestion.

The dsDNA variant fragments with two cohesive ends following digestion or annealing are inserted into a library of double-stranded polynucleotide sequences with appropriate complementary cohesive ends generated by PCR amplification of a library of double-stranded polynucleotide sequences with appropriate PCR primers.

The collection of dsDNA variants may also be prepared using single-stranded oligonucleotides generated by Dimer-based codon synthesis, a reverse primer being annealed to the 3' end of the single-stranded oligonucleotides so prepared followed by double-strand elongation using an appropriate DNA polymerase.

The collection of dsDNA variants may be prepared starting with single-stranded oligonucleotides; a reverse primer is annealed to the 3' end of the single-stranded oligonucleotides followed by double-strand elongation using an appropriate DNA polymerase.

The method described above wherein digestion by the restriction enzymes results in overhangs of 2 to 5 bases on the sense strand and/or 2 to 6 bases on the reverse strand, suitably overhangs of 2 to 4 bases on the sense strand and/or 2 to 4 bases on the reverse strand, for example overhangs of 4 bases.

In a particular embodiment, the polynucleotide sequence encoding an antibody variable chain, and the region targeted for diversification is one of the three complementarity determining regions (i.e., CDR1, CDR2 or CDR3). More specifically, in a particular embodiment the disclosed method is used to introduce diversity into the region of a polynucleotide sequence encoding the heavy or light chain CDR3 region. A skilled artisan will readily appreciate that the disclosed methodology can be used to introduce diversity into any one of the three CDR regions present in a parental antibody variable heavy (VH) or variable light (VL) chain.

In many cases, diversity needs to be introduced at various sites in a protein (e.g. the CDR regions of an antibody). Using the CoCohEn approach, we inserted the diversity cassettes by

IDT Exhibit 1020.5179

cloning one cassette per round of cloning. The reason for following this strategy stems from the much lower efficiency of the ligation step when three or more different fragments are present during ligation as compared to the case involving only two fragments. Attempts to first ligate two insert segments and then ligate the product into the vector rely on the fidelity of the ligation between the two insert fragments and on the reduced efficiency as the size of the ligated fragment increases. In addition, it is important to remember that the complexity generated during each cloning step must be conserved since the material produced by a cloning step becomes the template into which diversity needs to be introduced during subsequent cloning step. For example, after having inserted a cassette of diversity C in a first step, the population thus generated has to be submitted to the next cloning step with a number of clones $N > 10C$ to ensure the presence of the majority of the clones generated during the first step. As a consequence, multiple diversity cassettes are usually introduced in order of increasing complexity.

As used herein the term "complementary" refers to specific base recognition via for example base-base complementary. However, complementary as referred to herein includes also any Watson-Crick base-pairing of nucleoside analogs, e.g. deoxyinosine, which are capable of specific hybridization to the C base and other analogs which result in such specific hybridization, e.g. PNA, DNA and their analogs. Complementary of one single stranded region to another is considered to be sufficient when, under the conditions used, specific binding is achieved. Thus in the case of long single stranded regions including some lack of base-base specificity, e.g. presence of a mismatch, may be tolerated (e.g. if one base in a series of 10 bases is not complementary). Single-strand regions containing such small mismatches which do not affect the ultimate binding and ligation of the complementary single stranded region are considered to be complementary for the purposes of this invention. The complementary single stranded regions may retain portions, after binding, which remain single stranded, e.g. when single-strand overhangs of different length are used. In these cases, prior to ligation, missing bases may be filled in e.g. using Klenow fragment or other appropriate techniques as necessary.

TERMS

Ab: Antibody: an immunoglobulin. The term also covers any protein having a binding domain which is homologous to an immunoglobulin. A few examples of antibodies within this definition are the Fab, $F(ab^1)_2$,scFv, Fv, dAb and Fd fragments.

Antisense strand: The lower strand of ds DNA as usually written. In the antisense strand, 3'-TAC-5' would correspond to a Met codon in the sense strand.

IDT Exhibit 1020.5180

Acceptor cassette: A region of a polynucleotide sequences targeted for diversification that has been engineered in a way to comprise at least two restriction enzyme recognition motifs. The restriction enzyme recognition motifs are placed such that single stranded overhangs are generated upon digestion at the boundaries of the acceptor cassette. The latter is eliminated upon digestion.

CDR: Complementarity determining region.

CoCohEn: Compatible Cohesive Ends.

Complementary: Refers to specific base recognition via for example base-base complementary.

DNA : Deoxyribonucleic Acid

ds: double-stranded

Forward primer: A "forward" primer is complementary to a part of the antisense strand and primes for synthesis of a new sense-strand molecule.

Forward strand: The upper strand of ds DNA as usually written. In the forward strand, 5'-ATG-3' codes for Met.

Library: A collection of ds DNA sequences.

PCR: Polymerase chain reaction.

PNA: Peptide nucleic Acid.

Primer: A primer is an oligonucleotide complementary to a part of the antisense or sense strand and primes for synthesis of a new sense-or anti-sense strand molecule.

Reverse primer: A "reverse" primer is complementary to a part of the sense strand and primes for synthesis of a new antisense-strand molecule.

Reverse strand: The lower strand of ds DNA as usually written. In the reverse strand, 3'-TAC-5' would correspond to a Met codon in the sense strand.

IDT Exhibit 1020.5181

RE: Restriction enzyme.

Overhang: A segment of single-stranded DNA generated by digesting double-stranded DNA with restriction enzymes.

Sense strand: The upper strand of ds DNA as usually written. In the sense strand, 5'-ATG-3' codes for Met.

ss: single-stranded.

VH: A variable domain of an immunoglobulin heavy chain.

VL: A variable domain of an immunoglobulin light chain.

wt: wild type.

Type IIS and IIB restriction enzymes

Restriction enzymes (REs) are endonucleases which recognize a specific, generally consecutive, sequence of bases in a double-stranded (ds) DNA molecule, known as the recognition motif. REs are traditionally classified into three types on the basis of their subunit composition, cleavage position, sequence-specificity and cofactor-requirements:

Type I restriction enzymes bind to the recognition site and then cut at a varying distance from the recognition sequences.

Type II restriction enzymes bind at a recognition site and then cleave the DNA molecule at a site within to the recognition site DNA or at a defined distance outside but close to the recognition site DNA.

Type III enzymes cleave outside of their recognition sequences and require two such sequences to be present in opposite orientations within the same DNA molecule in order to accomplish cleavage. They rarely give complete digests.

Cloning methods have used restriction enzymes to generate nucleic acid fragments suitable for insertion into vectors and to prepare vectors to accept a nucleic acid sequence. Typically, restriction enzymes are used to cleave vectors to produce complementary terminal sequences relative to the terminal sequences present in the nucleic acid sequences which are being introduced into the vector. Most commonly, restriction enzymes cut palindromic sequences which are characterized by generating identical overhangs. The subsequent ligation of nucleic acid

IDT Exhibit 1020.5182

fragments that have been cut with the same restriction endonucleases as the vector results in the formation of a new recombinant nucleic acid sequence.

However, cloning strategies which rely on the use of one endonuclease that produce identical overhangs are relatively inefficient due to the formation of ligation products containing multiple fragments and to the inability to control the orientation of the segment to be inserted in these cases, thus, additional steps are required to separate and recover the desired ligation product.

The property of type II restriction enzymes to bind a single, specific recognition sequence, and cleave the DNA molecule at a specific site has promoted their extensive use in recombinant DNA technology type. Type II REs can be grouped into various subclasses according to the following properties: recognition of symmetric (EcoRI, GAATTC) or asymmetric (e.g., BbvCI, CCTCAGC) DNA sequences, binding to contiguous sequences in which the two half-sites of the recognized sequence are adjacent (e.g. EcoRI, GAATTC) or binding to discontinuous sequences in which the half-sites are separated (e.g. BglI, GCCNNNNNGGC), cleavage of DNA within or outside their recognition sequence (i.e. HindIII or FokI), the type of cleavage (blunt, 3' or -5'-protruding), and the distance of the cleavage site from the recognition site. Type IIP REs, which include most of the type II REs, recognize palindromic DNA sequences and cleave both DNA strands within the target at symmetrical positions. Type IIA REs recognize asymmetric DNA sequences and cleave within or outside of the recognition sequence. Type IIS REs are those type IIA REs that bind an asymmetric recognition sequences and cleave at least one strand of the DNA duplex outside of the recognition sequence. Type IIB REs cleave on both sides of the recognition sequence.

SubType	Recognition sequence	Cleavage site	Examples			
A	Asymmetric	Inside or outside	FokI	GGATC(9/13)	AciI	C↓CG↑C
S	Asymmetric	Outside	FokI	GGATC(9/13)	MmeI	TCCRAC(20/18)
P	Symmetric	Symmetric inside	EcoRI	G↓AATT↑C	PpuMI	RG↓GWC↑CY
B		Outside on both sides	BcgI	(10/12)CGANNNNNNTGC(12/10)		

Type IIS endonucleases exhibit no specificity with respect to the sequence that is cut and they can therefore generate single-strand overhangs with any base composition. Depending on the enzyme, cleavage with IIS enzymes results in overhangs of various lengths, e.g. from −5 to +6 bases in length (negative numbers indicate overhangs on the sense strand, positive numbers overhangs on the reverse strand). Preferably for performing the method of the invention, enzymes are chosen which generate 3-6, e.g. 4 base pair overhangs. Suitable restriction enzymes for use in

IDT Exhibit 1020.5183

the invention include, but are not limited to enzymes which produce 4 base overhangs at the 3′ end of the recognition site: BstXI; 5 base overhang at the 3′ end: AloI, BaeI, BplI, Bsp24I; 6 base overhang at the 3′ end: CjeI, CjePI, HaeIV; 4 base overhang at the 5′ end: AceIII, Acc36I, Alw26I, AlwXI, Bbr7I, BbsI, BbvI, BbvII, Bvb16II, Bli736I, BpiI, BpuAI, BsaI, Bsc91I, BseKI, BseXI, BsmAI, BsmBI, BsmFI, Bso31I, Bsp423I, BspBS31I, BspIS4I, BspLU11III, BspMI, BspST5I, BspTS514I, Bst12I, Bst71I, BstBS32I, BstGZ53I, BstTS5I, BstOZ616I, BstPZ418I, Eco31I, EcoA41, EcoO44I, Esp3I, FokI, PhaI, SfaNI, Sth132I, StsI; and 5 base overhang at the 5′ end: HgaI.

Type IIB endonucleases exhibit no specificity with respect to the sequence that is cut and they can therefore generate single-strand overhangs with any base composition. Depending on the enzyme, cleavage with IIB enzymes results in overhangs of various lengths, e.g. from −5 to +6 bases in length (negative numbers indicate overhangs on the sense strand, positive numbers overhangs on the reverse strand). Preferably for performing the method of the invention, enzymes are chosen which generate 3-6, e.g. 4 base pair overhangs. Suitable restriction enzymes for use in the invention include, but are not limited to FalI, BcgI, BpII, AlfI, BdaI, BaeI, BsaXI, PpiI.

EXAMPLE

The following example serves to exemplify the preparation of the introduction of diversity into nucleosides as claimed in the present invention. This should not, however, be construed as forming the only genus that is considered as the invention.

INTRODUCING DIVERSITY AT SPECIFIC SITE BY THE COMPLEMENTARY COHESIVE ENDS PROTOCOL (CoCohEn)

We evaluated the efficiency of the CoCohEn protocol by cloning a variable sequence in the CDR3-L region of an antibody fragment. The p3.1_5B vector selected from the PDL1 mini library (obtained from P. Luo, Abmaxis) expresses the 3.1_5B Fab. The VL sequence from the p3.1_5B vector is reported in Figure 5.

A PCR elongation reaction mix was assembled containing 10 µg of primer 3.1_Lfor and 3.1_Lrev (Table I), 2U Phusion DNA polymerase in 1x Phusion buffer (NEB) in a final volume of 50 µL. Reaction was thermal-cycled according to the following profile: 95°C 30", 72°C 1'. The dsDNA VLes construct thus obtained contains two Esp3i sites arranged as reported in Figure 7.

The dsDNA_VLes was digested with BpiI and KpnI restriction enzymes, purified using Qiaquick PCR kit and ligated overnight at 16° to the 3.1_5B vector digested with the same enzymes and 5'-ends dephosphorylated with Calf Intestinal Phosphatase (Boehringer). Ligation reaction was transformed in Top10 electro-competent cells (Invitrogen), plated on 100 µg/mL

IDT Exhibit 1020.5184

Amp, 2% glucose, 2xTY (2xTYAG) and grown overnight at 30°C. The resulting plasmid is referred to as p3.1_Les (Figure 8).

Design of degenerated oligonucleotides.

CDR3-L.S1 and CDR3-L.S2 oligonucleotides, together with primer CDR3-L.S12, were designed and synthesized according to the sequence reported in Figure 9. The synthesis of the oligos CDR3-L.S1 and CDR3-L.S2 was performed incorporating an equimolar ratio of two or more dNTPs at 7 positions which creates 128 variants (2x(2^6)). CDR3-L.S1 and CDR3-L.S2 oligonucleotides and primer CDR3-L.S12 were synthesized 5'-biotinylated.

Processing of degenerated oligonucleotides.

The CDR3-L.S12 primer was annealed to its complementary sequence in oligonucleotides CDR3-L-S1 and CDR3-L-S2 and elongated. A reaction mix was assembled containing 600 pmoles of each oligo, 0.25mM dNTPs, 1x PCR buffer, 2 U Fusion DNA polymerase and incubated at 95°C for 30", 43°C for 1', 72°C for 30". The PCR product is referred to as dsDNA_Vlib (Figure 10).

Introducing sequence diversity for CDR3-L in 3.1_Les.

The dsDNA_Vlib product was digested with Esp3I, loaded onto a Streptavidin-agarose column. The flow-through was analyzed on a 2.5% agarose gel to confirm that digestion was effectively accomplished. Digestion removed the flanking regions and generated dsDNA_Vlib_Insert with 4 bp long 5'-overhanging ends at both extremes (Figure 11). At the same time the vector 3.1_Les was digested with Esp3i restriction enzyme, creating two 4bp 5'-overhanging ends, compatible with those present in dsDNA_Vlib_Insert. Vector and insert were ligated overnight, transformed into electro-competent Top10 bacterial cells, plated onto 2xTYAG plates and grown overnight at 30°C. The cloning strategy adopted is schematically illustrated in Figure 11.

Analysis of the clones

Ninety-five clones derived from the ligation product p3.1_Vlib were sequenced with the following results:

- 3 sequences retained the original Esp3i cloning sites
- the remaining clones displayed the expected inserted variable sequence

Figure 12 summarizes the data and compares the observed against the expected frequency of residues at each position.

In some cases the observed frequency of a variant sensibly differs from the expected value. The results of similar experiments where dsDNA_Vlib was cloned blunt into a different plasmid yielded similar results. These data indicate that the incorporation of the nucleotides in the synthesis process is responsible for these variations.

IDT Exhibit 1020.5185

WHAT IS CLAIMED IS:

1. A method of introducing diversity at a desired point into one or more target polynucleotide sequences and assembling a library of double-stranded DNA sequences, wherein said method comprises the steps of:

a) providing one or more double-stranded polynucleotide sequences each of which is engineered to contain at least one recognition motif for a type II restriction enzyme that is not native to that polynucleotide sequence in a region common to all the polynucleotide sequences targeted for diversification,

b) digesting the polynucleotide sequences of step (a) with a type II restriction enzyme that recognizes the recognition motif and, if necessary, one or more further type II restriction enzymes in the same region of the polynucleotides such that a portion of the polynucleotide sequences can be excised to give polynucleotide sequences with overhanging single stranded nucleotide ends of two to seven nucleotides,

c) preparing a collection of dsDNA variant fragments having at their 5' and 3' ends overhanging single stranded nucleotide ends that are complementary to the ones generated in step (b), and

d) ligating the dsDNA variant fragments of step (c) into the restriction enzyme digested polynucleotide sequences of step (b) to give a diverse library of ds DNA sequences.

2. A method according to claim 1 wherein at least one of the recognition motifs is a recognition motif for a type IIB restriction enzyme that is not native to that polynucleotide sequence.

3. A method according to claim 1 wherein at least one of the recognition motifs is a recognition motif for a type IIS restriction enzyme that is not native to that polynucleotide sequence.

4. A method according to any one of claims 1 to 3 wherein the ds polynucleotide encodes a polypeptide.

5. A method according to claim 4 wherein the polypeptide is an antibody heavy or light chain variable domain.

IDT Exhibit 1020.5186

6. The method according to claim 5 wherein the diversity introduced corresponds to a complementarity determining region (CDR) of the polynucleotide sequences encoding an antibody light or heavy chain variable domain.

7. The method according to claim 5 wherein the complementarity determining region (CDR) is a CDR3 of the polynucleotide sequences encoding an antibody light or heavy chain variable domain.

8. The method of any previous claim wherein the dsDNA variant fragments to be inserted in the target polynucleotide sequences are generated by digestion of ds DNA fragments comprising within their 5' and 3' segment type II restriction enzymes recognition sites.

9. The method of any previous claim wherein the dsDNA variant fragments to be inserted in the target polynucleotide sequences are generated by digestion of ds DNA fragments comprising within their variable regions one or two IIB restriction enzymes recognition sites that generate, on digestion with the appropriate type IIB restriction enzymes, overhanging single stranded nucleotide ends which are complementary to the overhanging ends generated by digestion of the target polynucleotide sequences.

10. The method of claims 1 to 7 wherein the dsDNA variant fragments to be inserted in the target polynucleotide sequences are generated by starting from a collection of single-stranded oligonucleotides containing the desired variability; a reverse primer is annealed to the 3' end of each of the collection of the single-stranded oligonucleotides followed by double-strand elongation with an appropriate DNA polymerase to give the ds DNA variant fragments.

11. The method of claims 1 to 7 wherein the dsDNA variant fragments to be inserted in the target polynucleotide sequences are generated by annealing two complementary collections of oligonucleotides where sequences within each collection of oligonucleotides contains a 5' segment remaining single-stranded after annealing and the 5' segment being complementary to the overhanging ends generated in the target polynucleotide sequences on digestion.

12. The method of claims 1 to 7 wherein the collection of dsDNA variants is prepared using single-stranded oligonucleotides generated by Dimer-based codon synthesis, a reverse primer being annealed to the 3' end of the single-stranded oligonucleotides followed by double-strand elongation using an appropriate DNA polymerase.

IDT Exhibit 1020.5187

13. The method of claims 1 to 7 wherein the collection of dsDNA variants is prepared starting with single-stranded oligonucleotides; a reverse primer is annealed to the 3' end of the single-stranded oligonucleotides followed by double-strand elongation using an appropriate DNA polymerase.

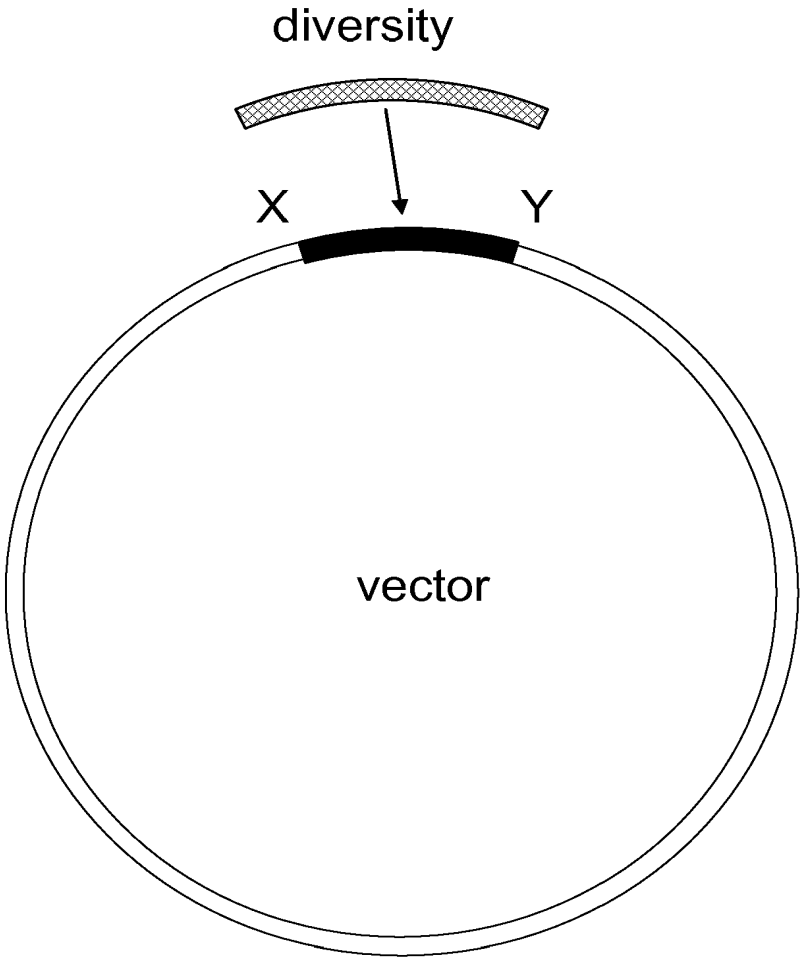
14. The method of claims 1 to 7 wherein the dsDNA variant fragments with two overhanging ends following digestion or annealing are inserted into a library of double-stranded polynucleotide sequences with appropriate complementary overhanging ends generated by PCR amplification of a library of double-stranded polynucleotide sequences with appropriate PCR primers.

15. The method of any previous claim wherein the cleavage by the restriction enzymes results in overhangs of up to 5 bases on the sense strand and/or up to 6 bases on the reverse strand.

16. The method of any previous claim wherein the cleavage by the restriction enzymes results in overhangs of up to 4 bases on the sense strand and/or up to 4 bases on the reverse strand.

IDT Exhibit 1020.5188

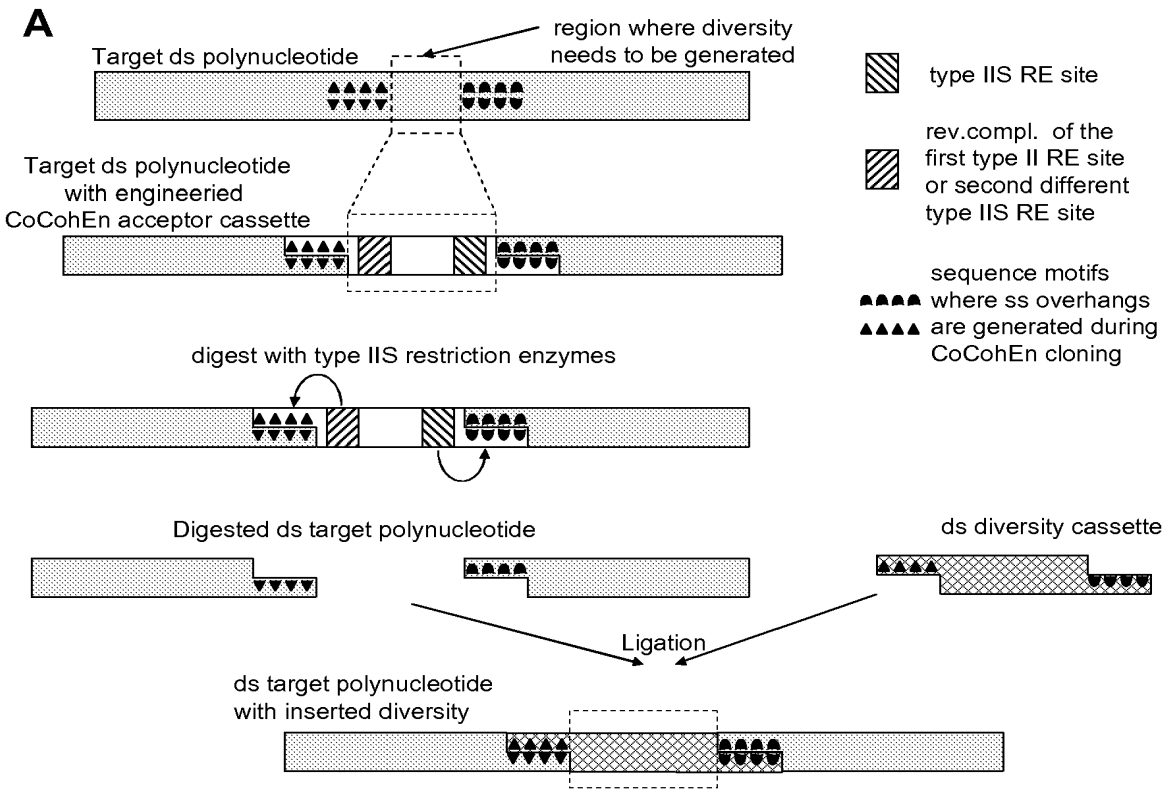
Figure 1



5

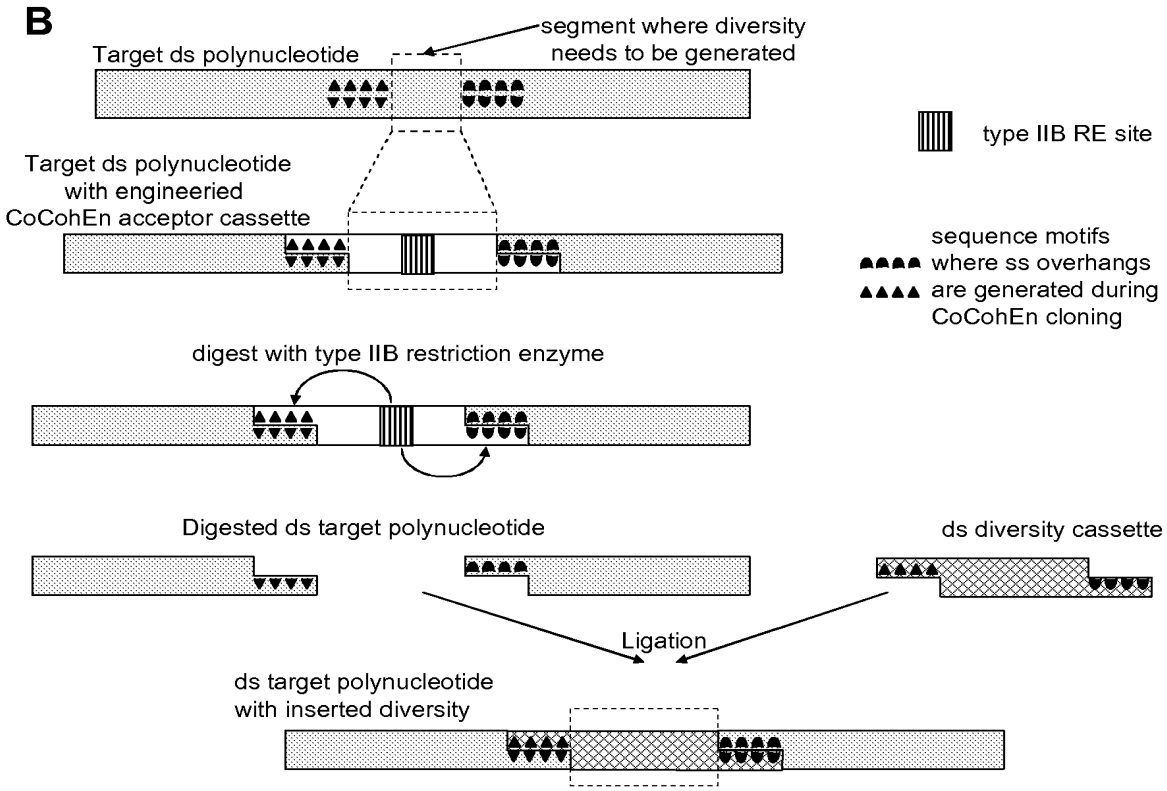
IDT Exhibit 1020.5189

Figure 2



IDT Exhibit 1020.5190

Figure 2 (cont.)



IDT Exhibit 1020.5191

Figure 3

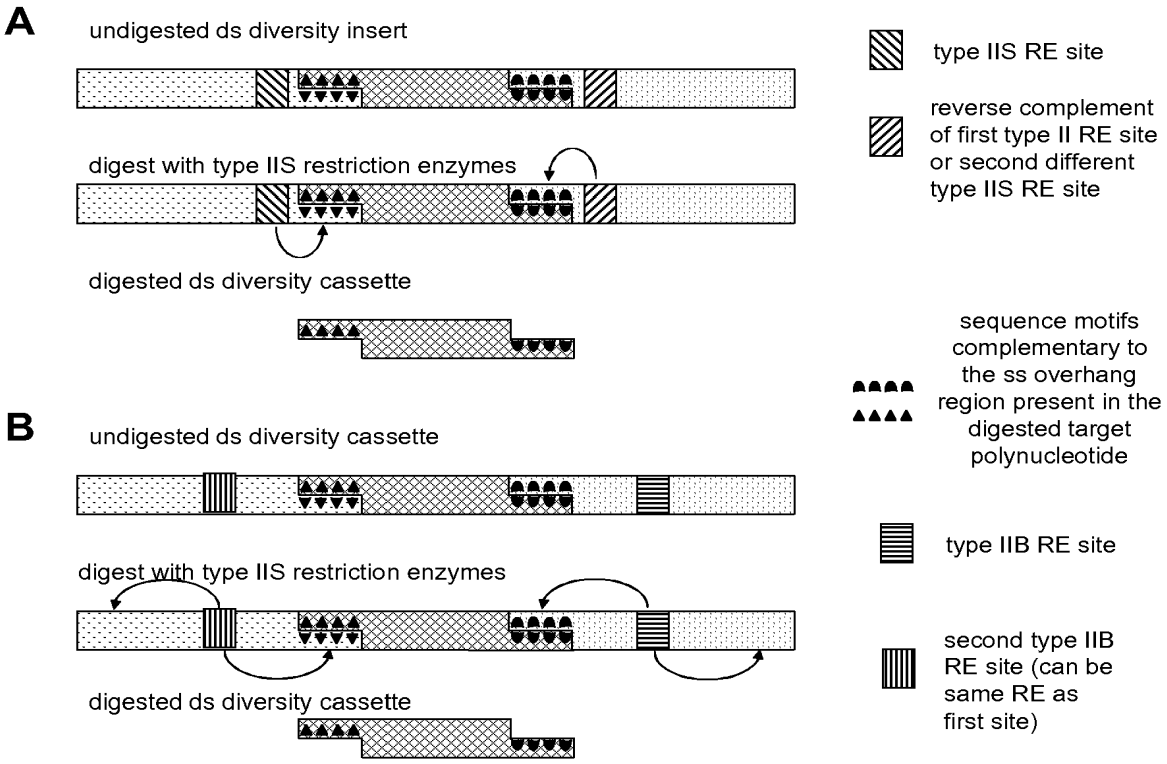
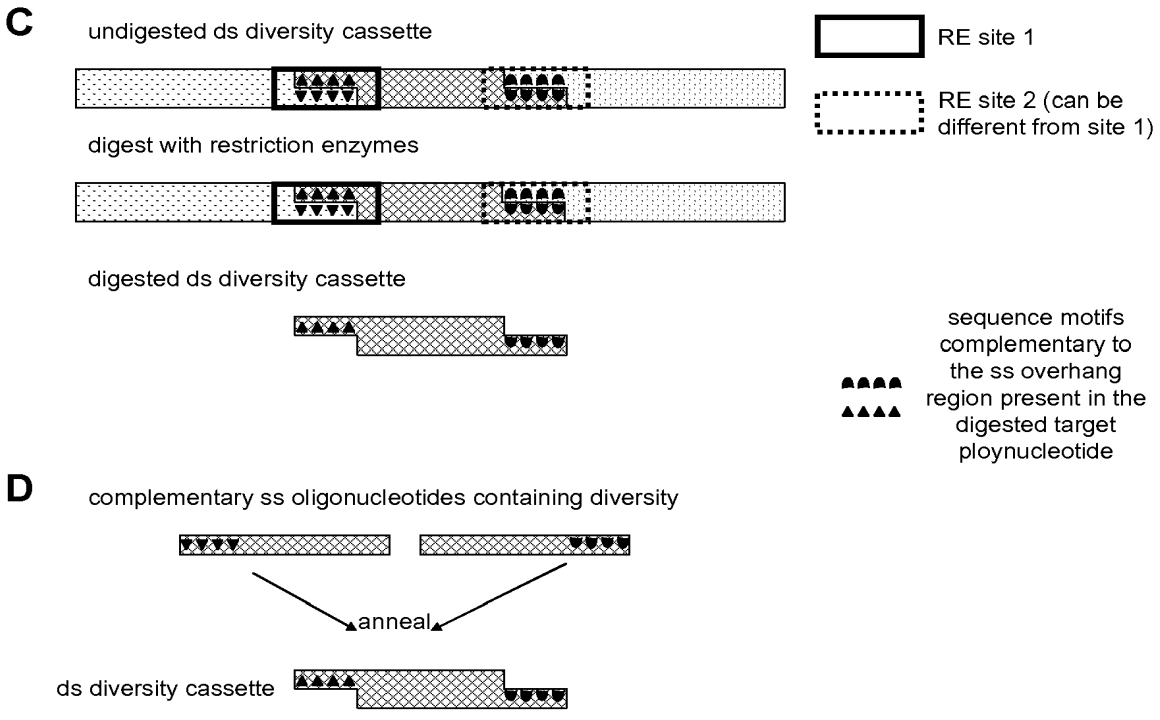


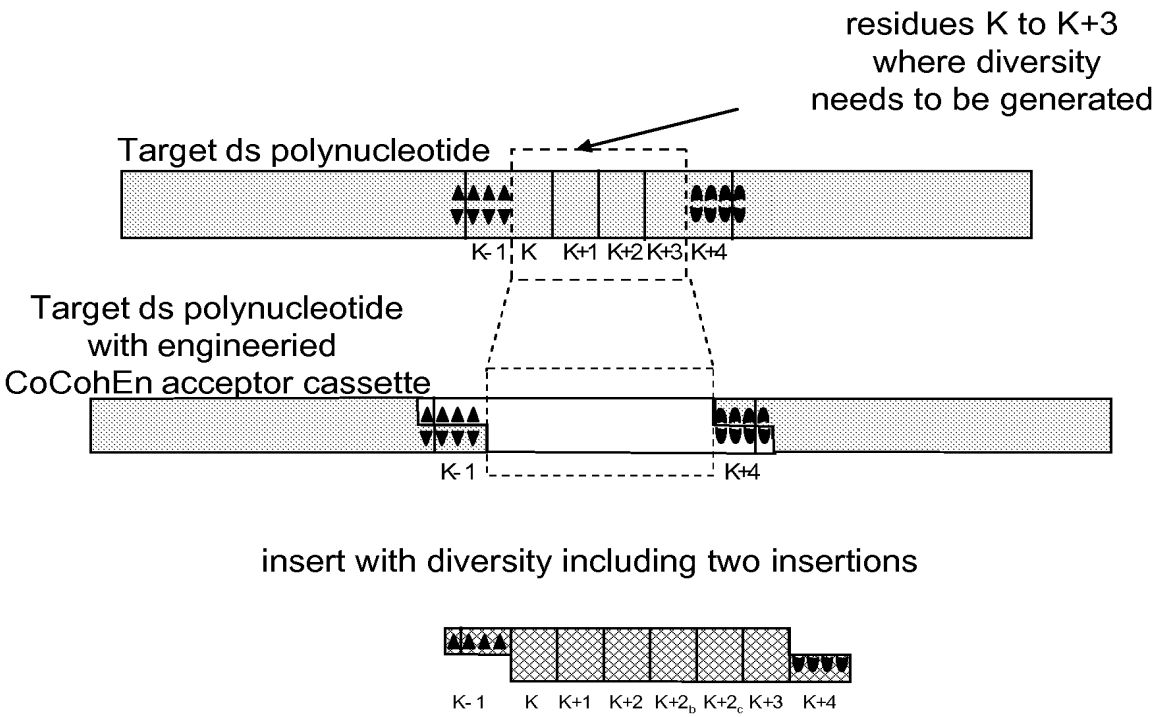
Figure 3 (cont.)



5

IDT Exhibit 1020.5192

Figure 4



IDT Exhibit 1020.5193

Figure 5



Figure 5: Variable domain of the light chain of the vector 3.1_5B. BpiI and KpnI sites are indicated.

Figure 6

Primer 3.1_Lfor 5'-GAGGAAGACTTCGCCACCTACTACTGCCAGCTGAGACGGACG-3'

Primer 3.1_Lrev 5'-CGTGGTACCACCACCGAATTGAGACGGACGTCCGTCTCAGCTGGC-3'

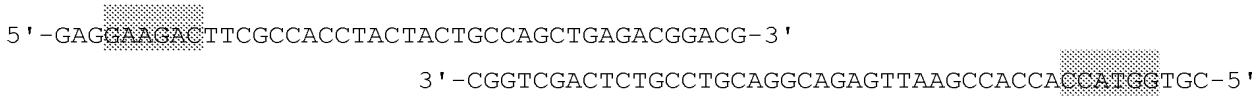


Figure 6: Forward (Primer 3.1_Lfor) and reverse (Primer3.1_Lrev) primer used to generate the acceptor cassette. BpiI and KpnI sites are indicated

IDT Exhibit 1020.5194

Figure 7

dsDNA VLes



Figure 7: dsDNA VLes construct. BpiI and KpnI sites are indicated as are the Esp3I site

Figure 8



Figure 8: Acceptor cassette at the light chain CDR3 region of plasmid p3.1_Les obtained by ligation of the BpiI and KpnI digested dsDNA_VLes construct into the BpiI and KpnI digested 3.1_5B vector. Esp3I sites of the acceptor cassette are indicated.

IDT Exhibit 1020.5195

Figure 9:

Degenerate CDR3-L-S1:

```
                          N  N
                          G  G
                          S  S  S     L
                    Q  Q  Y  D  D  W  P  W  T
5'bio-GAGCGTCTCACAGCAGTACRRTRRTTSGCCATKGACATTCGGAGACGACG
```

Degenerate CDR3-L-S2:

```
                          N  N
                          G  G
                          S  S  S     F
                    Q  Q  Y  D  D  W  P  Y  T
5'bio-GAGCGTCTCACAGCAGTACRRTRRTTSGCCATWCACATTCGGAGACGACG
```

Primer CDR3-L-S12: 5'bio-CGTCGTCTCCGAATG

Nomenclature degenerate bases:

R	A or G
Y	T or C
W	A or T
S	C or G
K	G or T
M	A or C

Figure 9: Composition of the degenerate oligonucleotides CDR3-L Oligo-S, Oligo-S2 and the reverse primer CDR3-L-S12. The Esp3I sites present in the oligonucleotide sequences are underlined.

RECTIFIED SHEET (RULE 91) ISA/EP

IDT Exhibit 1020.5196

Figure 10:

PCR product dsDNA_Vlib

5'bio-GAG*CGTCTC*ACAGCAGTACRRTRRTTSGCCATKGACATTCG*GAGACG*ACG
CTC*GCAGAG*TGTCGTCATGYYAYYAAWCGGTAMCTGTAAGC*CTCTGC*TGC-bio5'

5'bio-GAG*CGTCTC*ACAGCAGTACRRTRRTTSGCCATWCACATTCG*GAGACG*ACG
CTC*GCAGAG*TGTCGTCATGYYAYYAAWCGGTASGTGTAAGC*CTCTGC*TGC-bio5'

Figure 10: PCR product dsDNA_Vlib obtained by elongating oligonucleotides CDR3-L-S1 and CDR3-L-S2 with the reverse primer CDR3-L-S12. The Esp3I sites present in the oligonucleotide sequences are underlined.

IDT Exhibit 1020.5197

Figure 11:

Region of the acceptor cassette in vector p3.1_Les after Esp3I digestion:

```
     A  T  Y  Y  C                                        F  G  Q  G
TCGCCACCTACTACTGC     CAGCTGAGACGATATCGTCTCA     ATTCGGTGGTGGTA
AGCGCCACATGATGACGGTCG     ACTCTGCTATAGCAGAGTTAAG     CCACCACCAT
```

dsDNA_Vlib_Insert (dsDNA_Vlib after Esp3I digestion):

```
          5'-CAGCAGTACRRTRRTTSGCCATKGAC
                TCATGYYAYYAAWCGGTAMCTGTAAG-5'

          5'-CAGCAGTACRRTRRTTSGCCATWGAC
                TCATGYYAYYAAWCGGTASCTGTAAG-5'
```

dsDNA_Vlib_Insert ligated into 3.1_Les CDR3-L cloning site:

```
                          N  N
                          G  G
                          S  S  S     L
   A  T  Y  Y  C  Q  Q  Y  D  D  W  P  W  T  F  G  Q  G
TCGCCACCTACTACTGCCAGCAGTACRRTRRTTSGCCATKGACATTCGGTGGTGGTA
AGCGCCACATGATGACGGTCGTCATGYYAYYAAWCGGTAMCTGTAAGCCACCACCAT

                          N  N
                          G  G
                          S  S  S     F
   A  T  Y  Y  C  Q  Q  Y  D  D  W  P  Y  T  F  G  Q  G
TCGCCACCTACTACTGCCAGCAGTACRRTRRTTSGCCATWGACATTCGGTGGTGGTA
AGCGCCACATGATGACGGTCGTCATGYYAYYAAWCGGTASCTGTAAGCCACCACCAT
```

Figure 11: Cloning of the dsDNA_Vlib_Insert insert into the Esp3I digested p3.1_Les vector. The acceptor cassette in p3.1_Les is replaced by the insert dsDNA_Vlib_Insert. Amino acid diversity present at the desired site in the resulting ligation product p3.1_Vlib is indicated.

IDT Exhibit 1020.5198

Figure 12:

Position	Residue	Expected (%)	Observed (%)
92	N	25	4
	S	25	13
	D	25	20
	G	25	64
93	N	25	7
	S	25	13
	D	25	28
	G	25	52
94	S	50	57
	W	50	43
96	L	25	3
	Y	25	35
	F	25	49
	W	25	21

Figure 12: Observed and expected amino acid residue frequency in p3.1_Vlib.

IDT Exhibit 1020.5199

INTERNATIONAL SEARCH REPORT

International application No
PCT/EP2009/066135

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/10
ADD. C07K16/00

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12N

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched

Electronic data base consulted during the international search (name of data base and, where practical, search terms used)
EPO-Internal, EMBASE, BIOSIS, WPI Data, PAJ

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2008/095927 A1 (UNIV MARBURG PHILIPPS [DE]; SELMER THORSTEN [DE]; PINKENBURG OLAF [DE]) 14 August 2008 (2008-08-14)	1,3-8, 10-16
Y	page 53; claim 58; figures 1, 2, 5	2,9
X	BERGER S L ET AL: "PHOENIX MUTAGENESIS: ONE-STEP REASSEMBLY OF MULTIPLY CLEAVED PLASMIDS WITH MIXTURES OF MUTANT AND WILD-TYPE FRAGMENTS" ANALYTICAL BIOCHEMISTRY, ACADEMIC PRESS INC, NEW YORK LNKD- DOI:10.1006/ABIO.1993.1540, vol. 214, 1 January 1993 (1993-01-01), pages 571-579, XP002043107 ISSN: 0003-2697	1,3-8, 10-16
Y	the whole document	2,9

-/--

[X] Further documents are listed in the continuation of Box C.

[X] See patent family annex.

* Special categories of cited documents :

"A" document defining the general state of the art which is not considered to be of particular relevance

"E" earlier document but published on or after the international filing date

"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)

"O" document referring to an oral disclosure, use, exhibition or other means

"P" document published prior to the international filing date but later than the priority date claimed

"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention

"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone

"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art.

"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
5 May 2010	17/05/2010

Name and mailing address of the ISA/	Authorized officer
European Patent Office, P.B. 5818 Patentlaan 2 NL – 2280 HV Rijswijk Tel. (+31-70) 340-2040, Fax: (+31-70) 340-3016	Weinberg, Suzanna

2

Form PCT/ISA/210 (second sheet) (April 2005)

IDT Exhibit 1020.5200

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	MARSHALL JACQUELINE J T ET AL: "Restriction endonucleases that bridge and excise two recognition sites from DNA." JOURNAL OF MOLECULAR BIOLOGY 23 MAR 2007 LNKD- PUBMED:17266985, vol. 367, no. 2, 23 March 2007 (2007-03-23), pages 419-431, XP002580023 ISSN: 0022-2836 abstract	2,9
A	DE 103 37 407 A1 (TRANSMIT TECHNOLOGIETRANSFER [DE]) 10 March 2005 (2005-03-10) paragraph [0019]; figures 1,2	1-16

2

IDT Exhibit 1020.5201

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/EP2009/066135

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2008095927 A1	14-08-2008	AU 2008212907 A1	14-08-2008
		EP 2115144 A1	11-11-2009
DE 10337407 A1	10-03-2005	NONE	

Form PCT/ISA/210 (patent family annex) (April 2005)

IDT Exhibit 1020.5202

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
17 March 2011 (17.03.2011)

PCT

(10) International Publication Number
WO 2011/032040 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01)

(21) International Application Number:
PCT/US2010/048526

(22) International Filing Date:
10 September 2010 (10.09.2010)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/241,378 10 September 2009 (10.09.2009) US

(71) Applicant *(for all designated States except US)*: **CENTRILLION TECHNOLOGY HOLDING CORPORATION** [KZ/KZ]; Maples Corporate Services Limited, PO Box 309, Ugland House, Grand Cayman, KY1-1104 (KY).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **ZHOU, Wei** [US/US]; 1520 Oak Avenue, Los Altos, CA 94024 (US). **MEI, Rui** [US/US]; 234 Rodonovan Dr., Santa Clara, CA 95051 (US). **MATSUZAKI, Hajime** [JP/US]; 10471 Davison Ave., Cupertino, CA 95014 (US).

(74) Agents: **HARBURGER, David** et al.; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, CA 94304-1050 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PE, PG, PH, PL, PT, RO, RS, RU, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, MD, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

WO 2011/032040 A1

(54) **Title**: METHODS OF TARGETED SEQUENCING

(57) **Abstract**: The present invention provides methods for targeted sequencing of polynucleotide. In one aspect, the present invention provides a method of sequencing a target polynucleotide with fewer probes. In another aspect, the present invention provides a method of sequencing a target polynucleotide with longer reads. Locus-specific, ligation-assisted sequencing/ genotyping method and ligation-captured sequencing method are also provided in the present invention. The methods of the present invention allow low-cost, high-throughput and accurate sequencing of nucleic acids.

METHODS OF TARGETED SEQUENCING

RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application No. 61/241,378, filed September 10, 2009, which is incorporated by reference herein in its entirety.

BACKGROUND OF THE INVENTION

[0002] Large-scale multiplex analysis of highly polymorphic loci is needed for practical identification of individuals, e.g., for paternity testing and in forensic science (Reynolds et al., Anal. Chem., 63:2-15 (1991)), for organ-transplant donor-recipient matching (Buyse et al., Tissue Antigens, 41:1-14 (1993) and Gyllensten et al., PCR Meth. Appl, 1:91-98 (1991)), for genetic disease diagnosis, prognosis, and pre-natal counseling (Chamberlain et al., Nucleic Acids Res., 16:11141-11156 (1988) and L. C. Tsui, Human Mutat., 1:197-203 (1992)), and the study of drug metabolism and oncogenic mutations (Hollstein et al., Science, 253:49-53 (1991)). In addition, the cost-effectiveness of infectious disease diagnosis by nucleic acid analysis varies directly with the multiplex scale in panel testing. Many of these applications depend on the discrimination of single-base differences at a multiplicity of sometimes closely spaced loci.

[0003] A variety of DNA hybridization techniques are available for detecting the presence of one or more selected polynucleotide sequences in a sample containing a large number of sequence regions. In a simple method, which relies on fragment capture and labeling, a fragment containing a selected sequence is captured by hybridization to an immobilized probe. The captured fragment can be labeled by hybridization to a second probe which contains a detectable reporter moiety.

[0004] Another widely used method is Southern blotting. In this method, a mixture of DNA fragments in a sample is fractionated by gel electrophoresis, and then fixed on a nitrocellulose filter. By reacting the filter with one or more labeled probes under hybridization conditions, the presence of bands containing the probe sequences can be identified. The method is especially useful for identifying fragments in a restriction-enzyme DNA digest which contains a given probe sequence and for analyzing restriction-fragment length polymorphisms ("RFLPs").

[0005] Another approach to detecting the presence of a given sequence or sequences in a polynucleotide sample involves selective amplification of the sequence(s) by polymerase chain reaction. U.S. Pat. No. 4,683,202 to Mullis, et al. and R. K. Saiki, et al., Science 230:1350 (1985). In this method, primers complementary to opposite end portions of the selected sequence(s) are used to promote, in conjunction with thermal cycling, successive rounds of primer-initiated replication. The amplified sequence(s) may be readily identified by a variety of techniques. This approach is particularly useful for detecting the presence of low-copy sequences in a polynucleotide-containing sample, e.g., for detecting pathogen sequences in a body-fluid sample.

[0006] More recently, methods of identifying known target sequences by probe ligation methods have been reported. U.S. Pat. No. 4,883,750 to N. M. Whiteley, et al., D. Y. Wu, et al., Genomics 4:560 (1989), U. Landegren, et al., Science 241:1077 (1988), and E. Winn-Deen, et al., Clin. Chem. 37:1522 (1991). In one approach, known as oligonucleotide ligation assay ("OLA"), two probes or probe elements which span a target region of interest are hybridized to the target region. Where the probe elements basepair with adjacent target bases, the confronting ends of the probe elements can be joined by ligation, e.g., by treatment with ligase. The ligated probe element is then assayed, evidencing the presence of the target sequence.

[0007] In a modification of this approach, the ligated probe elements act as a template for a pair of complementary probe elements. With continued cycles of denaturation, hybridization, and ligation in the presence of pairs of probe elements, the target sequence is amplified linearly, allowing very small amounts of target sequence to be detected and/or

IDT Exhibit 1020.5204

amplified. This approach is referred to as ligase detection reaction. When two complementary pairs of probe elements are utilized, the process is referred to as the ligase chain reaction which achieves exponential amplification of target sequences. F. Barany, "Genetic Disease Detection and DNA Amplification Using Cloned Thermostable Ligase," Proc. Nat'l Acad. Sci. USA, 88:189-93 (1991) and F. Barany, "The Ligase Chain Reaction (LCR) in a PCR World," PCR Methods and Applications, 1:5-16 (1991).

[0008] Another scheme for multiplex detection of nucleic acid sequence differences is disclosed in U.S. Pat. No. 5,470,705 to Grossman et. al. where sequence-specific probes, having a detectable label and a distinctive ratio of charge/translational frictional drag, can be hybridized to a target and ligated together. This technique was used in Grossman, et. al., "High-density Multiplex Detection of Nucleic Acid Sequences: Oligonucleotide Ligation Assay and Sequence-coded Separation," Nucl. Acids Res. 22(21):4527-34 (1994) for the large scale multiplex analysis of the cystic fibrosis transmembrane regulator gene.

[0009] Jou, et. al., "Deletion Detection in Dystrophia Gene by Multiplex Gap Ligase Chain Reaction and Immunochromatographic Strip Technology," Human Mutation 5:86-93 (1995) relates to the use of a so called "gap ligase chain reaction" process to amplify simultaneously selected regions of multiple exons with the amplified products being read on an immunochromatographic strip having antibodies specific to the different haptens on the probes for each exon.

[0010] Solid-phase hybridization assays require multiple liquid-handling steps, and some incubation and wash temperatures must be carefully controlled to keep the stringency needed for single-nucleotide mismatch discrimination. Multiplexing of this approach has proven difficult as optimal hybridization conditions vary greatly among probe sequences.

[0011] Ligation of allele-specific probes generally has used solid-phase capture (U. Landegren et al., Science, 241:1077-1080 (1988); Nickerson et al., Proc. Natl. Acad. Sci. USA, 87:8923-8927 (1990)) or size-dependent separation (D. Y. Wu, et al., Genomics, 4:560-569 (1989) and F. Barany, Proc. Natl. Acad. Sci, 88:189-193 (1991)) to resolve the allelic signals, the latter method being limited in multiplex scale by the narrow size range of ligation probes. Further, in a multiplex format, the ligase detection reaction alone cannot make enough product to detect and quantify small amounts of target sequences. The gap ligase chain reaction process requires an additional step--polymerase extension. The use of probes with distinctive ratios of charge/translational frictional drag for a more complex multiplex will either require longer electrophoresis times or the use of an alternate form of detection.

[0012] There is a great need for a rapid, high-throughput, and low-cost assay format that allows targeted sequencing with high specificity. Most sequence methods do not distinguish between the multiple copies of DNA that organisms may have. For example, human genome contains DNA sequences of both maternal and paternal origin. Therefore, polymorphisms may exist at loci and provide multiple different readings at the same locus during standard sequencing methods, complicating the sequencing process. The present invention permits reading one copy of DNA at regions containing variations, such as single nucleotide polymorphisms (SNPs).

SUMMARY OF THE INVENTION

[0013] In one aspect, the present invention provides a locus-specific sequencing method which utilizes genotyping steps to attain enhanced specificity. In one embodiment, the present invention provides a method of sequencing comprising forming a hybridization complex, which comprises a first set of capture probes, a target polynucleotide, and a first solution probe. At least one of the first capture probes binds a first region of the target polynucleotide. The first solution probe binds a second region of the target polynucleotide. The first solution probe has a base at the 3' end complementary to a first loci site of the target polynucleotide. Therefore, when hybridization occurs at least one of the first set of capture probes and the first solution probe may be ligated. Additionally, the first solution probe comprises a

IDT Exhibit 1020.5205

cleavable first cap. This method includes ligating at least one of the first set of capture probes and the first solution probe, adding a second cap to those capture probes of the first set of capture probes not ligated to the first solution probe, removing the first cap, and then sequencing the target polynucleotide.

[0014] In some embodiments, the methof further comprises forming a hybridization complex comprising a second solution probe that hybridizes to the target polynucleotide, wherein the second solution probe binds to a third region of the target polynucleotide, wherein the second solution probe has a base complementary to a second loci site of the target polynucleotide. The second solution probe may be ligated to the first solution probe. The method includes forming a hybridization complex comprising a third solution probe that hybridizes to the target polynucleotide. The third solution probe binds to a fourth region of the target polynucleotide, has a base complementary to a third loci site of the target polynucleotide and comprises a universal primer sequence. This aspect of the invention also involves ligating the third solution probe to the second solution probe, removing the target polynucleotide, and sequencing the target polynucleotide with a primer that hybridizes the universal primer sequence of the third solution probe.

[0015] In some embodiments, the sequencing comprises placing the target-probe hybridization complex under nucleic acid synthesis conditions in the presence of free nucleotides to allow extension of the probe that is complementary to the target polynucleotide.

[0016] In some embodiments, the method further comprises removing non-specific hybridization.

[0017] In some embodiments, the first cap comprises a capping dye and the first solution probe comprises a cleavage site. In one aspect the second cap comprises a protection group.

[0018] In some embodiments, the first set of capture probes is attached to a flat surface or a bead. In one aspect, the flat surface is a flow cell. In some embodiments, the first set of capture probes is synthesized or spotted on said flat surface. In some embodiments, the first set of capture probes is spotted randomly at a controlled density. In some embodiments, the first set of capture probes is spotted at known locations on the flat surface.

[0019] In some embodiments, the first set of capture probes comprises polynucleotides. In some embodiments, the each capture probe of said first set of capture probes contains 50-150 nucleotides.

[0020] In some embodiments, the first, second, or third solution probe comprises a polynucleotide. In one aspect, the first, second, or third solution probe contains 5 to 50 nucleotides. In some embodiments, the target polynucleotide comprises a single nucleotide polymorphism (SNP), an insertion or deletion, or a copy number variation.

[0021] In some embodiments, the target polynucleotide is removed from the hybridization duplex.

[0022] In some embodiments, the first set of capture probes comprises allele-specific capture probes for a locus. In some embodiments, the allele-specific capture probes are to known alleles of the first loci.

[0023] In some embodiments, a nucleotide reversible terminator is added to the hybridization complex as a first cap. In some embodiments, the nucleotide reversible terminator contains an amino-2-hydroxypropyl group.

[0024] In some embodiments, at least 50% of a genome is sequenced using no more than 500 million probes. In another aspect, 99.5% of the genome is sequenced. In some embodiments, the nucleic acid sequence of the first, second or third solution probe is known. In some embodiments, the number of total probes used is about 5 million to 500 million. In another aspect, the number of duplicate probes is less and 5. In some embodiments, multiple loci in the genome are sequenced. In some embodiments, the sequencing avoids SNP with minor allele frequency of greater than 0.05 at the extreme 3' end of the first, second, third or fourth solution probe or first or second capture probe sequences. In some embodiments, sequencing said target polynucleotide comprises generating reads that are greater than 5-50 bases.

[0025] In some embodiments, the first, second, third or fourth solution probe is labeled. In some embodiments, the first, second, third or fourth solution probe has at least one base that is degenerate. In some embodiments, sequencing

IDT Exhibit 1020.5206

information of the target polynucleotide or genome is obtained via a computer system. In some embodiments, the computer system comprises an algorithm for detecting, processing, analyzing and/or assembling nucleotide sequence information.

INCORPORATION BY REFERENCE

[0026] All publications and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0027] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following detailed description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0028] **FIGURE 1** shows the representative steps of ligation assisted sequencing.

[0029] **FIGURE 2** shows the representative steps of ligation captured sequencing. FIG. 2A shows the hybridization, first ligation, capping and cleaving labels steps of the ligation captured sequencing; FIG. 2B shows the washing and subsequent ligation steps of the ligation captured sequencing; and FIG. 2C shows the primer annealing and base extension steps of the ligation captured sequencing.

[0030] **FIGURE 3** shows an example of probe design for targeted sequencing. The figure shows a 10bp locus of interest within genome 50-mer probes immobilized on surface at 5' end with forward and reverse orientation probes.

[0031] **FIGURE 4** shows an example of probe design for targeted sequencing. The probe should avoid HapMap single-nucleotide polymorphism (SNP) with MAP > 0.05 at 3' end of probe sequences.

[0032] **FIGURE 5** shows an example of probe design for targeted sequencing and expected distribution of read depth. Arrows represent sequence reads.

[0033] **FIGURE 6** shows a schematic on reading heterozygous insertions/deletions. The top panel shows "in situ" array feature in which the probes have both maternal and paternal chromosomes hybridized. The bottom panel shows that on a "lawn" flow cell, each probe will have either maternal or paternal chromosome hybridized.

[0034] **FIGURE 7** shows that nucleic acid sequence information can be obtained, processed, analyzed and/or assembled via a computer system.

[0035] **FIGURE 8** shows probe design sequences for five example loci from the PharmaADME panel (PharmaADME.org). Each locus is represented by sets of five overlapping probe sequences in both the forward and reverse orientations. The five loci include examples of a deletion, insertion, copy number, and two SNP variants. The probe sequences for the copy number example at CYP2A6 were modified from a primer sequence described in Fukami et al 2006 (*Pharmacogenomics Journal* 6: 401-412).

[0036] **FIGURE 9** shows results from sequencing reactions on a human SNP chip.

DETAILED DESCRIPTION OF THE INVENTION

[0037] Unless defined otherwise, all technical and scientific terms used herein have the same meaning as commonly understood by one of ordinary skill in the art to which this invention belongs. Although any methods and materials similar or equivalent to those described herein can also be used in the practice or testing of the present invention, representative illustrative methods and materials are now described.

IDT Exhibit 1020.5207

[0038] Such conventional techniques and descriptions can be found in standard laboratory manuals such as Genome Analysis: A Laboratory Manual Series (VoIs. I-IV), Using Antibodies: A Laboratory Manual, Cells: A Laboratory Manual, PCR Primer: A Laboratory Manual, and Molecular Cloning: A Laboratory Manual (all from Cold Spring Harbor Laboratory Press); Stryer, L. (1995) Biochemistry (4th Ed.) Freeman, New York; Gait, "Oligonucleotide Synthesis: A Practical Approach" 1984, IRL Press, London, Nelson and Cox (2000), Lehninger, (2004) Principles of Biochemistry 4th Ed., W. H. Freeman Pub., New York, N.Y. and Berg et al. (2006) Biochemistry, 6th Ed., W. H. Freeman Pub., New York, N.Y., all of which are herein incorporated in their entirety by reference for all purposes.

[0039] The present invention provides a locus-specific sequencing method which utilizes genotyping steps to attain enhanced specificity.

[0040] In one aspect, the present invention provides a method of sequencing comprising: (1) forming a hybridization complex, which comprises a first set of capture probes, a target polynucleotide, and a first solution probe, at least one of the first capture probes binds a first region of the target polynucleotide, the first solution probe binds a second region of the target polynucleotide, the first solution probe has a base at the 3' end complementary to a first loci site of the target polynucleotide so that when hybridization occurs at least one of the first set of capture probes and the first solution probe may be ligated; (2) ligating at least one of the first set of capture probes and the first solution probe; (3) adding a second cap to those capture probes of the first set of capture probes not ligated to the first solution probe; (4) optionally removing the first cap; and (5) sequencing the target polynucleotide. The first solution probe preferably comprises a cleavable first cap.

[0041] In another aspect, the present invention provides a method for sequencing a genome or at least 50% of a genome using no more than 5 million, 10 million, 50 million, 100 million, 150 million, 200 million, 500 million capture probes. In some embodiments, at least 50%, 85%, 95%, 99.0%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9% or more of the genome is sequenced. In one embodiment, the genome is a human genome. In some embodiments, the capture probes used in the subject methods are loci-specific and allele specific. The loci can contain a single nucleotide polymorphism (SNP), an insertion or deletion, or a copy number variation. For example, the loci can contain a drug metabolizing enzyme and transporter gene. In some embodiments, the nucleic acid sequences of the capture probes are known. In some embodiments, the capture probes can have a sequence located next to a loci of interest or proximate to the loci at no more than 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 100 bases distance.

[0042] In some embodiments, the present invention provides a method of performing a genomic sequencing reaction using approximately 0, 1, 2, 3, 4, 5, or more duplicated capture probes. In some embodiments, the same locus may be covered by multiple different capture probes and each locus will generally have at least one, but more likely at least two capture probes for biallelic feature such as a SNP. In one example, the number of duplicate capture probes is 0. In some embodiments, multiple loci in the genome are sequenced. In some embodiments, multiple alleles in the genome are sequenced. The sequencing may avoid SNP with minor allele frequency of > 0.05 at extreme 3' end of probe sequences. In some embodiments, the capture probe contains 10-10,000 nucleotides, for example, a capture probe can be 50 nucleotides. The capture probes can be located at known geographic locations or random geographic locations at a controlled density on a substrate, such as a solid phase support (e.g. a flat surface or a bead).

[0043] In another aspect, the present invention also provides a method for sequencing a target polynucleotide by generating reads that are greater 5, 10, 15, 20, 30, 40, 50, 60, 74, 100, 150 bases. In some embodiments, a lawn of randomly placed capture probes is used, wherein each capture probe generates a read greater than 5, 10, 15, 20, 30, 40, 50, 60, 74, 100, 150 bases. In some embodiments, the read is greater than 10, 11, 12, 13, 14, 15, 16, 17, 18, 19 or 20 bases. Loci-specific and/or allele-specific capture probes can be used. These capture probes can be located at known or

IDT Exhibit 1020.5208

random geographic locations on a substrate. In some embodiments, the target polynucleotide is genomic DNA, for example, human genomic DNA. Multiple loci and/or alleles in the genomic DNA can be sequenced.

[0044] In some embodiments, the target polynucleotide region contains a single nucleotide polymorphism (SNP), an insertion or deletion, or a copy number variation. The target polynucleotide region may be a loci or allele. For example, the alleles or loci may contain a drug metabolizing enzyme and transporter gene.

[0045] In one embodiment, the methods of the present invention provide for selecting a subset of SNPs located to a region of interest. In one aspect, methods of the present invention can be used to sequence an entire genome, part of a genome, one chromosome, or part of a chromosome, permitting chromosome specific sequencing. 1, 2, 3, 4 or more capture probes may be used to capture all possible alleles at known loci.

[0046] In some embodiments, the methods of the present invention comprise the steps of: a) immobilizing a set of loci-specific probes (also referred to as "capture probes") on a substrate, wherein the probes have defined sequences; b) hybridizing a sample containing a target polynucleotide to the set of loci-specific probes; c) ligating a first solution probe to the loci-specific probe, wherein the solution probe carries a cleavage site (near the 3'end) and hybridizes to the target polynucleotide; d) providing a nucleotide reversible terminator to the hybridization duplexes resulted from steps b) and c), wherein the nucleotide reversible terminator prevents base extension of complexes formed by nonspecific hybridization; e) cleaving the nucleotide reversible terminator at the cleavage site present on the first solution probe; and f) placing the target-probe hybridization duplexes under nucleic acid synthesis conditions in the presence of free nucleotides to allow extension of the probe that is complementary to the target polynucleotide. In some embodiments, the method further comprises obtaining sequence information of the target polynucleotide via a computer system. In some embodiments, the target polynucleotide is genomic DNA. In some embodiments, the first solution probe is labeled and have at least one base that is degenerate. In one embodiment, the first solution probe is a 9-mer, i.e. contains 9 nucleotides. The subject method provides higher hybridization specificity of at least 95%, 96%, 97%, 98%, 99%, 99.5% or more. In some embodiments, the loci-specific first solution probes and/or capture probes are allele-specific. In some embodiments, 1, 2, 3, 4 or more different solution probes are used.

[0047] In some embodiments, the methods of the present invention comprise the steps of: a) immobilizing a set of loci-specific probes on a substrate, wherein the probes have defined sequences; b) hybridizing a sample containing a target polynucleotide to the set of loci-specific probes which bind to loci sites; c) ligating a first solution probe (5'phosphate) to the first loci-specific probe, wherein the solution probe carries a cleavage site (near 3'end) and hybridizes to the target polynucleotide; d) providing a nucleotide reversible terminator to the hybridization duplexes resulted from steps b) and c), wherein the capping step blocks base extension; e) cleaving the nucleotide reversible terminator at the cleavage site present on the solution probe to allow base extension; f) ligating a second solution probe to the first solution probe such that the second solution probe hybridizes to the target polynucleotide; g) washing the hybridization duplexes from step f) to remove nonspecific hybridization between the second solution probe and the target polynucleotide; h) ligating a third set of solution probes carrying a common primer at 3' end to the previously ligated solution probe such that the solution probe carrying the primer hybridizes to the target polynucleotide; i) washing the hybridization duplexes from step h) to remove the target polynucleotide from the hybridization duplexes; j) adding primers that are complementary to the common primer at 3' end of the solution probes to allow primer annealing; and k) placing the single strand probe polynucleotide under nucleic acid synthesis conditions in the presence of free nucleotides to allow base extension from the annealed primer at the 3' end based on nucleic acid sequence of the probe strand that is complementary to the target polynucleotide. In some embodiments, the method further comprises obtaining sequence information of the target polynucleotide via a computer system. In some embodiments, the target polynucleotide is genomic DNA. In some

IDT Exhibit 1020.5209

embodiments, the solution probes are labeled and have at least one base that is degenerate. In one embodiment, the solution probe is a 9-mer, i.e. contains 9 nucleotides. In some embodiments, the washing condition is sufficiently stringent such that the solution probes that have nonspecifically hybridized to the target polynucleotide can be washed off. In other embodiments, the washing condition is sufficiently stringent such that the target polynucleotide can be washed off from the hybridization duplex. The cycle of ligation between the solution probe and the target polynucleotide followed by stringent washing to remove nonspecific ligation can be repeated more than one time. The subject ligation-captured sequencing method provides higher hybridization specificity. The specificity can be at least 95%, 96%, 97%, 98%, 99%, 99.5% or more. In some embodiments, the loci-specific probes or capture probes are allele-specific.

I. Ligation Assisted Sequencing

[0048] One aspect of the present invention relates to a method for determining or identifying nucleotide sequence via ligation assisted sequencing. This method generally comprises hybridization between an allele-specific, sequence-defined probe and a target polynucleotide, ligation between a first solution probe comprising a cleavable cap and the target polynucleotide, capping of the 3' end of capture probes not ligated to the first solution probe, removing the 3' end cap on the first solution probe to prevent blocking of base extension of the specific hybridization product, and enzymatic reaction to allow base extension complementary to the target polynucleotide by for example, polymerase chain reaction.

Target Preparation

[0049] In one aspect, the present invention provides a method for sequencing a genome or at least 50%, 85%, 95% or 99% of the genome. In some embodiments, the target polynucleotide is genomic DNA or a part of the genomic DNA. While some embodiments are for sequencing whole genome at more than 50% coverage, these embodiments and some other embodiments are also suitable for sequencing a targeted region such as genomic regions relating to drug metabolism. In one example, the target polynucleotide is human genomic DNA. The subject methods can sequence at least 50%, 85%, 90%, 95%, 96%, 97%, 98%, 99%, 99.1%, 99.2%, 99.3%, 99.4%, 99.5%, 99.6%, 99.7%, 99.8%, 99.9% or more of the genome, for example, human genome. The number of probes used for sequencing a genome using the subject methods of the present invention is fewer as compared to other genome sequencing techniques currently available. In some embodiments, no more than 200 million probes are used. For example, 200 million probes can be used to sequence at least 95% of the human genome. In some embodiments, the probes are loci- and/or allele-specific. In some embodiments, the sequences of the probes are known or defined. The design of the probes of the present invention is described herein below.

[0050] "Nucleic acid" or "oligonucleotide" or "polynucleotide" or grammatical equivalents typically refer to at least two nucleotides covalently linked together. A nucleic acid of the present invention will generally contain phosphodiester bonds, although in some cases, as outlined below (for example in the construction of primers and probes such as label probes), nucleic acid analogs are included that may have alternate backbones, comprising, for example, phosphoramide (Beaucage et al., Tetrahedron 49(10):1925 (1993) and references therein; Letsinger, J. Org. Chem. 35:3800 (1970); Sprinzl et al., Eur. J. Biochem. 81:579 (1977); Letsinger et al., Nucl. Acids Res. 14:3487 (1986); Sawai et al, Chem. Lett. 805 (1984), Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); and Pauwels et al., Chemica Scripta 26:141 91986)), phosphorothioate (Mag et al., Nucleic Acids Res. 19:1437 (1991); and U.S. Pat. No. 5,644,048), phosphorodithioate (Briu et al., J. Am. Chem. Soc. 111:2321 (1989), O-methylphophoroamidite linkages (see Eckstein, Oligonucleotides and Analogues: A Practical Approach, Oxford University Press), and peptide nucleic acid (also referred to herein as "PNA") backbones and linkages (see Egholm, J. Am. Chem. Soc. 114:1895 (1992); Meier et al., Chem. Int. Ed. Engl. 31:1008 (1992); Nielsen, Nature, 365:566 (1993); Carlsson et al., Nature 380:207 (1996), all of which are incorporated by

IDT Exhibit 1020.5210

reference). Other analog nucleic acids include those with bicyclic structures including locked nucleic acids (also referred to herein as "LNA"), Koshkin et al., J. Am. Chem. Soc. 120.13252 3 (1998); positive backbones (Denpcy et al., Proc. Natl. Acad. Sci. USA 92:6097 (1995); non- ionic backbones (U.S. Pat. Nos. 5,386,023, 5,637,684, 5,602,240, 5,216,141 and 4,469,863; Kiedrowshi et al., Angew. Chem. Intl. Ed. English 30:423 (1991); Letsinger et al., J. Am. Chem. Soc. 110:4470 (1988); Letsinger et al., Nucleoside & Nucleotide 13:1597 (1994); Chapters 2 and 3, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook; Mesmaeker et al., Bioorganic & Medicinal Chem. Lett. 4:395 (1994); Jeffs et al., J. Biomolecular NMR 34:17 (1994); Tetrahedron Lett. 37:743 (1996)) and non-ribose backbones, including those described in U.S. Pat. Nos. 5.235,033 and 5,034,506, and Chapters 6 and 7, ASC Symposium Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Cook. Nucleic acids containing one or more carbocyclic sugars are also included within the definition of nucleic acids (see Jenkins et al., Chem. Soc. Rev. (1995) pp 169 176). Several nucleic acid analogs are described in Rawls, C & E News Jun. 2, 1997 page 35. "Locked nucleic acids" are also included within the definition of nucleic acid analogs. LNAs are a class of nucleic acid analogues in which the ribose ring is "locked" by a methylene bridge connecting the 2'-0 atom with the 4'-C atom, All of these references are hereby expressly incorporated by reference. These modifications of the ribose-phosphate backbone may be done to increase the stability and half-life of such molecules in physiological environments. For example, PNA:DNA and LNA-DNA hybrids can exhibit higher stability and thus may be used in some embodiments. The target nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. Depending on the application, the nucleic acids may be DNA (including genomic and cDNA), RNA (including mRNA and rRNA) or a hybrid, where the nucleic acid contains any combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xathanine hypoxathanine, isocytosine, isoguanine, etc.

[0051] The term "target polynucleotide" typically refers to a nucleic acid of interest. In one aspect, target nucleic acids of the invention are genomic nucleic acids. DNA derived from the genetic material in the chromosomes of a particular organism is genomic DNA. A genomic library is a collection of clones made from a set of randomly generated overlapping DNA fragments representing the entire genome of an organism. Target nucleic acids include naturally occurring or genetically altered or synthetically prepared nucleic acids (such as genomic DNA from a mammalian disease model). Target nucleic acids can be obtained from virtually any source and can be prepared using methods known in the art. For example, target nucleic acids can be directly isolated without amplification, isolated by amplification using methods known in the art, including without limitation polymerase chain reaction (PCR), whole genome amplification (WGA), multiple displacement amplification (MDA), rolling circle amplification (RCA), rolling circle amplification (RCR) and other amplification methodologies. Target nucleic acids may also be obtained through cloning, including cloning into vehicles such as plasmids, yeast, and bacterial artificial chromosomes.

[0052] In some embodiments, the methods of the present invention involve no amplification of the target polynucleotide, for example, genomic DNA. In this case, a few micrograms of genomic DNA will have about million copies of genomic DNA per microliter.

[0053] In some embodiments, the methods of the present invention comprise capture of target polynucleotide. The target polynucleotide may be from a known region of the genome. In one embodiment, oligo probes can be immobilized on beads and these oligo beads which are inexpensive and reusable can be used to capture the target genomic polynucleotide. In another embodiment, Nimblegen microarrays are used to capture target polynucleotide. NimbleGen arrays are particularly suited for chromatin immunoprecipitation--microarray hybridizations (ChIP-chip) which use whole genome arrays, whole genome promoter specific arrays, or custom designed arrays; and comparative genomic

IDT Exhibit 1020.5211

hybridization (CGH) which uses whole human genome tiled arrays or custom designed region specific arrays. In another embodiment, Agilent RNA oligos can be used to capture target polynucleotide of the present invention.

[0054] In some embodiments, target polynucleotide is amplified using standard amplification methods known in the relevant art. In one embodiment, the target polynucleotide is prepared by whole genome amplification (WGA). The most recently developed WGA methods include the ligation-mediated PCR (LMP), the T7-based linear amplification of DNA (TLAD) and the multiple displacement amplification (MDA). LMP is a method that uses endonuclease or chemical cleavage to fragment the gDNA sample and uses linkers and primers for its amplification. It was adapted for the WGA of small quantities of gDNA and single cells (Klein *et al.*, 1999; Tanabe *et al.*, 2003). Rubicon Genomics commercializes different kits (Omniplex) that allow for the amplification of RNA, DNA and methylated DNA sequences. The main advantages are that the method is able to amplify degraded DNA, and allows for different variations and that all steps are performed in the same tube. The main disadvantages are that it only amplifies a representation of the genome and it generates fragments only up to 2 kb. It has been reported to provide better CGH results than DOP-PCR (Pirker *et al.*, 2004) but worse genotyping results than MDA (Bergen *et al.*, 2005a). TLAD is a variant on the protocol originally designed by Phillips and Eberwine to amplify mRNA (Phillips and Eberwine, 1996) that has been adapted for WGA (Liu *et al.*, 2003). It uses Alu I restriction endonuclease digestion and a terminal transferase to add a polyT tail on the 3' terminus. A primer is then used with a 5' T7 promoter and a 3' polyA tract, and Taq polymerase is used to synthesise the second strand. Then the sample is submitted to *in vitro* transcription reaction and posterior reverse transcription. The major advantage is that TLAD does not introduce sequence and length-dependent biases. Multiple displacement amplification (MDA) is a non-PCR-based isothermal method based on the annealing of random hexamers to denatured DNA, followed by strand-displacement synthesis at constant temperature (Blanco *et al.*, 1989). It has been applied to small genomic DNA samples, leading to the synthesis of high molecular weight DNA with limited sequence representation bias (Lizardi *et al.*, 1998; Dean *et al.*, 2002). As DNA is synthesized by strand displacement, a gradually increasing number of priming events occur, forming a network of hyper-branched DNA structures. The reaction can be catalyzed by the Phi29 DNA polymerase or by the large fragment of the Bst DNA polymerase. The Phi29 DNA polymerase possesses a proofreading activity resulting in error rates 100 times lower than the Taq polymerase (Eckert and Kunkel, 1991; Esteban *et al.*, 1993). Recently, it has been shown that MDA, when used on genomic DNA sequences with high variability, results in a loss of heterozygosity (Murthy *et al.*, 2005). The technology has been shown to be very sensitive and can amplify from single cells (Hellani et al., 2004, Handyside et al., 2005) and single bacteria (Raghunathan et al., 2005). Any of the herein disclosed method can be used in the methods of the present invention.

[0055] In another embodiment, the target polynucleotide is prepared by whole genome sampling assay (WGSA). The WGSA reduces the complexity of a nucleic acid sample by amplifying a subset of the fragments in the sample. A nucleic acid sample is fragmented with one or more restriction enzymes and an adapter is ligated to both ends of the fragments. A primer that is complementary to the adapter sequence is used to amplify the fragments using PCR. During PCR fragments of a selected size range are selectively amplified. The size range may be, for example, 400-800 or 400 to 2000 base pairs. Fragments that are outside the selected size range are not efficiently amplified. The fragments that are amplified by WGSA may be predicted by in silico digestion and an array may be designed to genotype SNPs that are predicted to be amplified. Genotyping may be done by allele specific hybridization with probes that are perfectly complementary to individual alleles of a SNP. A set of probes that are complementary to the region surrounding each SNP may be present on the array. Perfect match probes are complementary to the target over the entire length of the probe. Mismatch probes are identical to PM probes except for a single mismatch base. The mismatch position is typically the central position. WGSA is disclosed in Kennedy et al. (2003), *Nat Biotechnol* , Vol., pp.1233-1237, and U.S. patent

IDT Exhibit 1020.5212

application Ser. Nos. 09/920,492, 09/904,039, 10/681,773, 10/316,517, 10/442,021, 10/463,991, 10/316, 629, and 10/264,945 and U.S. Pat. No. 6,361,947, all of which are herein incorporated by reference. WGSA can simultaneously genotype more than 10,000 SNPs in parallel by allele-specific hybridization to perfect match (PM) and mismatch (MM) probes synthesized on an array. WGSA may not be able to assay the entire panels of loci.

[0056] In another embodiment, the target polynucleotide is prepared by long-range PCR. Long range PCR allows the amplification of PCR products, which are much larger than those achieved with conventional *Taq* polymerases. Up to 27 kb fragments are possible from good quality genomic DNA, although 10 - 20 kb fragments are routinely achievable, given the appropriate conditions. The method relies on a mixture of thermostable DNA polymerases, usually *Taq* DNA polymerase for high processivity (i.e. 5'-3' polymerase activity) and another DNA polymerase with 3'-5' proofreading abilities (usually *Pwo*). This combination of features allows longer primer extension than can be achieved with *Taq* alone. This method for detection of the FVIII gene intron 22 inversion (Liu *et al*, 1998) removes the requirement for Southern Blotting. Results can be obtained within 24 hours. Modifications from standard long range PCR protocols include the addition of DMSO and incorporation of deaza GTP to enable read through of a high GC content region upstream of the FVIII gene. The method relies on overlapping PCR to generate a constant band, which appears in all template DNA's. This band acts as a control to show that the reaction has worked efficiently. The largest amplification product seen using this method is 12 kb, well within the range of the enzyme mix utilized. Long-range PCR may be costly to assay singleton loci that are distant from groups of clustered loci.

[0057] In another embodiment, the target polynucleotide is prepared by locus-specific multiplex PCR. Multiplex locus specific amplification can be used to amplify a plurality of pre-selected target sequences from a complex background of nucleic acids. The targets are selected for amplification using splint oligonucleotides that are used to modify the ends of the fragments. The fragments have known end sequences and the splints are designed to be complementary to the ends. The splint can bring the ends of the fragment together and the ends are joined to form a circle. The splint can also be used to add a common priming site to the ends of the target fragments. Specific loci are amplified and can be subsequently analyzed. Locus-specific multiplex PCR can be very costly.

[0058] Other suitable amplification methods include but are not limited to the ligase chain reaction (LCR) (e.g., Wu and Wallace, Genomics 4, 560 (1989), Landegren et al., Science 241, 1077 (1988) and Barringer et al. Gene 89:117 (1990)), transcription amplification (Kwoh et al., Proc. Natl. Acad. Sci. USA 86, 1173 (1989) and WO88/10315), self-sustained sequence replication (Guatelli et al., Proc. Nat. Acad. Sci. USA, 87, 1874 (1990) and WO90/06995), selective amplification of target polynucleotide sequences (U.S. Pat. No. 6,410,276), consensus sequence primed polymerase chain reaction (CP-PCR) (U.S. Pat. No. 4,437,975), arbitrarily primed polymerase chain reaction (AP-PCR) (U.S. Pat. Nos. 5,413,909, 5,861,245) and nucleic acid based sequence amplification (NABSA). (See, U.S. Pat. Nos. 5,409,818, 5,554,517, and 6,063,603 each of which is incorporated herein by reference). Other amplification methods that may be used are described in, U.S. Pat. Nos. 5,242,794, 5,494,810, 4,988,617 and in U.S. Ser. No. 09/854,317, each of which is incorporated herein by reference. Additional methods of sample preparation and techniques for reducing the complexity of a nucleic sample are described in Dong et al., Genome Research 11, 1418 (2001), in U.S. Pat. Nos. 6,361,947, 6,391,592 and U.S. Ser. Nos. 09/916,135, 09/920,491, 09/910,292, and 10/013,598.

[0059] Naturally-existing targets can be assayed directly in cell lysates, in nucleic acid extracts, or after partial purification of fractions of nucleic acids so that they are enriched in targets of interest. In one example, the target polynucleotide is human genomic DNA. The polynucleotide target to be detected can be unmodified or modified. Useful modifications include, without limitation, radioactive and fluorescent labels as well as anchor ligands such as biotin or digoxigenin. The modification(s) can be placed internally or at either the 5' or 3' end of the targets. Target modification

IDT Exhibit 1020.5213

can be carried out post- synthetically, ether by chemical or enzymatic reaction such as ligation or polymerase- assisted extension. Alternatively, the internal labels and anchor ligands can be incorporated into an amplified target or its complement directly during enzymatic polymerization reactions using small amounts of modified NTPs as substrates.

[0060] The target polynucleotide can be isolated from a subject. The subject is not limited to a human being but may also be other organisms including but not limited to mammals, plants, bacteria, virus or fungi. In one example, the target polynucleotide is genomic DNA extracted from a human.

Methods of laying probes on surface

[0061] In some embodiments, the methods of the invention comprise oligonucleotide probes, generally stretched on a substrate. The terms "substrate" or "solid support" or other grammatical equivalents as used herein typically refer to any material that is modified to allow "stretching" of nucleic acid molecules as described herein. In general, the substrate contains discrete individual sites (for example, nanochannels, flow cells, or lines) appropriate for the attachment or association of decorated nucleic acid molecules to form stretched nucleic acids and is amenable to at least one detection method. As will be appreciated by those in the art, the number of possible substrates is very large. Possible substrates include, but are not limited to, glass and modified or functionalized glass, plastics (including acrylics, polystyrene and copolymers of styrene and other materials, polypropylene, polyethylene, polybutylene, polyurethanes, TeflonJ, etc.), polysaccharides, nylon or nitrocellulose, resins, silica or silica-based materials including silicon and modified silicon, carbon, metals, inorganic glasses, plastics, optical fiber bundles, and a variety of other polymers. In general, the substrates allow optical detection and do not themselves appreciably fluoresces.

[0062] Substrates of the invention can be configured to have any convenient geometry or combination of structural features. The substrates can be either rigid or flexible and can be either optically transparent or optically opaque, or have combinations of these surfaces. The substrates can also be electrical insulators, conductors or semiconductors. Further the substrates can be substantially impermeable to liquids, vapors and/or gases or, alternatively, the substrates can be substantially permeable to one or more of these classes of materials. In general, the substrates fall into two different classes: substrates comprising particular geometries such as nanochannels or nanopores, as more fully discussed below, or those that have surface characteristics to allow the stretching of decorated nucleic acids, such as the use of linear patterns of surface chemistries.

[0063] In one aspect of the invention, substrates of the invention comprise nanostructures or cells. Such structures can include without limitation nanopillars, nanopores and nanochannels. In many exemplary aspects, substrates of the invention comprise nanochannels. Such substrates are known in the art. For example, U.S. Patent Nos. 7,217,562; 6,685,841; 6,518,189; 6,440,662; 6,214,246 describe nanostructures, including nanochannels, of use in accordance with the present invention. These patents are hereby incorporated by reference in their entirety. Generally, in these nanochannel substrates, there is a reservoir into which the oligonucleotide probes are placed, which are then moved into nanochannels, a single molecule of oligonucleotide probe per nanochannel, to form the stretched nucleic acids, followed by detection of the order, and optionally, the distance between the labels of the incorporated probes.

[0064] In some embodiments, the substrates comprise cells that are generally 1-2 millimeters thick. In one example, the substrate, e.g. a slide can be about 10 centimeters long. Another embodiment of nanostructures that finds use in the present invention are substrates comprising nanopores. Nanopore devices can provide single-molecule detection of molecules driven electrophoretically in solution through a nano-scale pore, and the sequence of nucleotides can be detected by the sequence of signals generated as each nucleotide passes through the pore. Such nanopores and methods of sequencing using nanopores are known in the art and discussed in for example, Branton et al., (2008), Nature, 26(10):1 146-53 and in U.S. Patent Nos. 6,673,615; 7,258,838; 7,238,485; 7,189,503;6,627,067; 6,464,842; 6,267,872 and U.S.

Patent Application Nos. 20080248561; 20080171316, 20080102504, each of which is herein incorporated by reference in its entirety for all purposes, and in particular for the figures, legends and accompanying text describing the compositions, methods of using the compositions and methods of making the compositions.

[0065] In some embodiments, the oligonucleotide probe is immobilized on a solid support before binding to the target polynucleotide. In one embodiment, the 5' end of an oligonucleotide probe of the present invention is attached to a solid surface or substrate. Oligonucleotide can be immobilized by various methods known in the art including, (without limitation) covalent cross-linking to a surface (e.g., photochemically or chemically), non-covalent attachment to the surface through the interaction of an anchor ligand with a corresponding receptor protein (e.g. biotin- streptavidin or digoxigenin-anti-digoxigenin antibody), or through hybridization to an anchor nucleic acid or nucleic acid analog. The anchor nucleic acid or nucleic acid analog have sufficient complementarity to the target (i.e., their formed duplex has sufficiently high Tm) that the anchor-target-probe complex will survive stringent washing to remove unbound targets and probes, but they do not overlap with the target site that is complementary to the probe antisense sequence.

[0066] The solid substrate can be made of any material to which the molecules can be bound, either directly or indirectly. Examples of suitable solid substrates include flat glass, quartz, silicon wafers, mica, ceramics and organic polymers such as plastics, including polystyrene and polymethacrylate. The surface can be configured to act as an electrode or a thermally conductive substrate (which enhances the hybridization or discrimination process). For example, micro and sub-micro electrodes can be formed on the surface of a suitable substrate using lithographic techniques. Smaller nanoelectrodes can be made by electron beam writing/lithography. Electrodes can also be made using conducting polymers which can be pattern a substrate by ink-jet printing devices by soft lithography or be applied homogenously by wet chemistry. TnO_2 coated glass substrates are available. Electrodes can be provided at a density such that each immobilized molecule has its own electrode or at a higher density such that groups of molecules or elements are connected to an individual electrode. Alternatively, one electrode may be provided as a layer below the surface of the array which forms a single electrode. The solid substrate may optionally be interfaced with a permeation layer or a buffer layer. It is also possible to use semi-permeable membranes such as nitrocellulose or nylon membranes, which are widely available. The semi-permeable membranes can be mounted on a more robust solid surface such as glass. The surface layer may comprise a sol-gel. The surfaces may optionally be coated with a layer of metal, such as gold, platinum or other transition metal. A particular example of a suitable solid substrate is the commercially available SPR BIACore™ chip (Pharmacia Biosensors). Heaton et al., 2001 (PNAS 98:3701-3704) have applied an electrostatic field to an SPR surface and used the electric field to control hybridization.

[0067] The solid substrate is generally a material having a rigid or semi-rigid surface. In some embodiments, at least one surface of the substrate is substantially flat, although in some embodiments it may be desirable to physically separate discrete elements with, for example, raised regions or etched trenches. For example, the solid substrate may comprise nanovials—small cavities in a flat surface e.g. 10 μm in diameter and 10 μm deep. This is particularly useful for cleaving molecules from a surface and performing assays or other processes such as amplification in them. The solution phase reaction is more efficient than the solid phase reaction, whilst the results remains spatially addressable, which is advantageous. Other formats include but are not limited to synthetic or natural beads, membranes or filters, slides including microarray slides, microtiter plates, microcapillaries, and microcentrifuge tubes.

[0068] In some embodiments, the loci-specific oligo probes are coated or attached onto beads for capturing genomic DNA. The oligo probes can be directed against large regions on genomic DNA that include multiple loci of interest. For example, many ADME (absorption, distribution, metabolism, and excretion) markers are on about 200 genes. Hybridization between loci-specific oligo probes and target polynucleotide can be carried out on beads in columns at a

IDT Exhibit 1020.5215

controlled temperature and salt concentration. The hybridization products can be eluted from the beads with moderate pressure.

[0069] The use of a solid support with an array of capture oligonucleotides is fully disclosed in U.S. Patent Application Ser. No. 60/011,359, which is hereby incorporated by reference. When using such arrays, the oligonucleotide primers or probes used in the above-described coupled PCR and LDR phases, respectively, have an addressable array-specific portion. After the LDR or PCR phases are completed, the addressable array-specific portions for the products of such processes remain single stranded and are caused to hybridize to the capture oligonucleotides during a capture phase. C. Newton, et al., "The Production of PCR Products With 5' Single-Stranded Tails Using Primers That Incorporate Novel Phosphoramidite Intermediates," Nucl. Acids Res. 21(5): 1155-62 (1993), which is herein incorporated by reference.

[0070] During the capture phase of the process, the mixture can be contacted with the solid support at a temperature of 45-90°C and for a time period of up to 60 minutes. Hybridizations may be accelerated by adding cations, volume exclusion or chaotropic agents. When an array consists of dozens to hundreds of addresses, it is important that the correct ligation product sequences have an opportunity to hybridize to the appropriate address. This may be achieved by the thermal motion of oligonucleotides at the high temperatures used, by mechanical movement of the fluid in contact with the array surface, or by moving the oligonucleotides across the array by electric fields. After hybridization, the array is washed sequentially with a low stringency wash buffer and then a high stringency wash buffer.

[0071] Loading of nucleic acids onto these substrates can be modulated and/or controlled by the flow and/or electrical forces, including diffusion forces and surface forces exerted by areas of differential charge and/or hydrophobicity. The number of nucleic acids applied to the substrate (i.e., with a loading buffer or other solution) can be adjusted to assure maximal occupancy of the linear features with non-overlapping nucleic acid molecules and thus minimize the number of empty linear features on the substrate. In an exemplary embodiment, at least 50% of the linear features of a substrate are occupied by at least one nucleic acid molecule. In a further embodiment, at least 60%, 70%, 80%, 90%, and 95% of the linear features are occupied by one or more nucleic acids.

[0072] Two exemplary approaches of laying probes are disclosed herein below for illustrative purposes. The first approach is “In Situ” oligonucleotide synthesis in which the probes are in known geographic locations in the X-Y coordinate plane. In one embodiment, the oligonucleotide probe is synthesized on the surface. Examples of technologies that allow on-surface oligo synthesis include but are not limited to photolithography and ink jet. In another embodiment, the pre-synthesized oligonucleotide probes are spotted onto the surface. Various microarray protocols, for example, protocol for Agilent inkjet-deposited pre-synthesized oligo arrays are known to one skilled in the art.

[0073] Polymers such as nucleic acids or polypeptides can be synthesized in situ using photolithography and other masking techniques whereby molecules are synthesized in a step-wise manner with incorporation of monomers at particular positions being controlled by means of masking techniques and photolabile reactants. For example, U.S. Pat. No. 5,837,832 describes a method for producing DNA arrays immobilized to silicon substrates based on very large scale integration technology. In particular, U.S. Pat. No. 5,837,832 describes a strategy called “tiling” to synthesize specific sets of probes at spatially-defined locations on a substrate. U.S. Pat. No. 5,837,832 also provides references for earlier techniques that can also be used. Light directed synthesis can also be carried out by using a Digital Light Micromirror chip (Texas Instruments) as described (Singh-Gasson et al., (1999) Nature Biotechnology 17:974-978). Instead of using photo-deprotecting groups which are directly processed by light, conventional deptotecting groups such as dimethoxy trityl can be employed with light directed methods where for example a photoacid is generated in a spatially addressable way which selectively deprotects the DNA monomers (McGall et al PNAS 1996 93: 1355-13560; Gao et al J. Am. Chem

IDT Exhibit 1020.5216

Soc. 1998 120: 12698-12699). Electrochemical generation of acid is another means that can be used in the subject methods of the present invention.

[0074] The "in situ" arrays can have about 1,000 to 100,000,000 array probes (features). In one embodiment, the "in situ" array carries approximately 200,000,000 probes.

[0075] The second approach of laying probes on a surface is "Lawn" approach in which a finite set of probes, for example, pre-synthesized oligos, are laid down randomly at controlled density on a surface or flow cell. In this case, single probe molecules should be sufficiently separated to allow detection of independent events. In some embodiments, the "lawn" surface has at least one hundred of independent redundant reads off each unique probe sequence. In some embodiments, Helicos molecular barcoding techniques and single molecule sequencing can be used in the subject methods (Pushkarev D, Neff NF, Quake SR. Nat Biotechnol. Aug 2009). The HeliScope™ Single Molecule Sequencer is a genetic analyzer for direct DNA measurement utilizing Helicos True Single Molecule Sequencing (tSMS)™ technology. As a DNA microscope, the HeliScope instrument performs tSMS chemistry and captures images to observe sequencing-by-synthesis reactions for billions of individual DNA molecules in parallel. Multiplexed barcoded samples run on the HeliScope™ Single Molecule Sequencer saves money and boosts research productivity by increasing the number of analyzable samples per run. In some embodiments, barcoding used in the subject methods allows loading of 5 identifiable individual samples per channel on a 50-channel flow cell. In one embodiment, it allows 250 individuals, i.e. barcoded DNA samples to be analyzed per flow cell. One can sequence 2.5 megabases from five different samples per channel at 15X coverage for less than $72 per sample. Samples can be prepared simply without amplification. Oligonucleotides can be ligated to the DNA in the samples followed by barcode reading.

[0076] Molecules that can be immobilized in the array include nucleic acids such as DNA and analogues and derivatives thereof, such as PNA. Nucleic acids can be obtained from any source, for example genomic DNA or cDNA or synthesized using known techniques such as step-wise synthesis. Nucleic acids can be single or double stranded. DNA nanostructures or other supramolecular structures can also be immobilized. Other molecules include but are not limited to compounds joined by amide linkages such as peptides, oligopeptides, polypeptides, proteins or complexes containing the same; defined chemical entities, such as organic molecules; conjugated polymers and carbohydrates or combinatorial libraries thereof.

[0077] Molecules can be labeled to enable interrogation using various methods. Suitable labels include: optically active dyes, such as fluorescent dyes; nanoparticles such as fluorospheres and quantum dots, rods or nanobars; and surface plasmon resonant particles (PRPs) or resonance light scattering particles (RLSs)—particles of silver or gold that scatter light (the size and shape of PRP/RLS particles determines the wavelength of scattered light). See Schultz et al., 2000, PNAS 97: 996-1001; Yguerabide, J. and Yguerabide E., 1998, Anal Biochem 262: 137-156.

Probe Design

[0078] In some embodiments of the hybridization phase, a loci-specific oligonucleotide probe is provided ("capture probe"). The nucleotide sequence of the probe is known. In some embodiments, the probe is specific to a locus of interest. The loci of interest include but are not limited to single nucleotide polymorphism (SNP), adjacent SNPs, insertion or deletions, for example, up to 10 base-pair (bp), and copy number variation (CNV), for example, integer copy calls.

[0079] In some embodiments, the locus of interest is a SNP including tri-allelic and quad-allelic. A single-nucleotide polymorphism (SNP) is a DNA sequence variation occurring when a single nucleotide — A, T, C, or G — in the genome (or other shared sequence) differs between members of a species (or between paired chromosomes in an individual). For example, two sequenced DNA fragments from different individuals, AAGCCTA to AAGCTTA, contain a difference in a

IDT Exhibit 1020.5217

single nucleotide. In this case there are two alleles: C and T. Within a population, SNPs can be assigned a minor allele frequency — the lowest allele frequency at a locus that is observed in a particular population. This is simply the lesser of the two allele frequencies for single-nucleotide polymorphisms. There are variations between human populations, so a SNP allele that is common in one geographical or ethnic group may be much rarer in another.

[0080] Single nucleotide may be changed (substitution), removed (deletions) or added (insertion) to polynucleotide sequence. Ins/del SNP may shift translational frame. Single-nucleotide polymorphisms may fall within coding sequences of genes, non-coding regions of genes, or in the intergenic regions between genes. SNPs within a coding sequence will not necessarily change the amino acid sequence of the protein that is produced, due to degeneracy of the genetic code. A SNP in which both forms lead to the same polypeptide sequence is termed *synonymous*— if a different polypeptide sequence is produced they are *nonsynonymous*. A nonsynonymous change may either be missense or nonsense, where a missense change results in a different amino acid, while a nonsense change results in a premature stop codon. SNPs that are not in protein-coding regions may still have consequences for gene splicing, transcription factor binding, or the sequence of non-coding RNA.

[0081] Variations in the DNA sequences of humans can affect how humans develop diseases and respond to pathogens, chemicals, drugs, vaccines, and other agents. SNPs are also thought to be key enablers in realizing the concept of personalized medicine. Another great importance of SNPs in biomedical research is for comparing regions of the genome between cohorts (such as with matched cohorts with and without a disease). The study of single-nucleotide polymorphisms is also important in crop and livestock breeding programs (genotyping). In one aspect, the present invention provides methods used to identify SNPs.

[0082] In some embodiments, the loci of interest include CNVs. A copy number variation (CNV) is a segment of DNA in which copy-number differences have been found by comparison of two or more genomes. The segment may range from one kilobase to several megabases in size. Humans (being diploid) ordinarily have two copies of each autosomal region, one per chromosome. This may vary for particular genetic regions due to deletion or duplication. CNVs may either be inherited or caused by *de novo* mutation. A recently proposed mechanism for the cause of some CNVs is fork stalling and template switching, a replication misstep. Therefore, CNVs can be caused by genomic rearrangements such as deletions, duplications, inversions, and translocations. Low copy repeats (LCRs) which are region specific repeat sequences are susceptible to such genomic rearrangements resulting in CNVs. Factors such as size, orientation, percentage similarity and the distance between the copies renders them susceptible (Lee JA, Carvalho CM, Lupski JR. (2007) *Cell* **131** (7): 1235–47; Lee J.A., Lupski J.R. (2006) Neuron, 52 (1), pp. 103-121).

[0083] Copy number variation can be discovered by cytogenetic techniques such as fluorescent in situ hybridization, comparative genomic hybridization, array comparative genomic hybridization, and by virtual karyotyping with SNP arrays.

[0084] In one embodiment, the present invention provides methods for identifying and sequencing CNVs. It is estimated that approximately 0.4% of the genomes of unrelated people typically differ with respect to copy number (Kidd JM, Cooper GM, Donahue WF, *et al.* (May 2008). *Nature* **453** (7191): 56–64). Like other types of genetic variation, some CNVs have been associated with susceptibility or resistance to disease. Gene copy number can be elevated in cancer cells. For instance, the EGFR copy number can be higher than normal in non-small cell lung cancer (Cappuzzo F, Hirsch, et al. (2005 *Journal of the National Cancer Institute* **97**: 643–655). In addition, a higher copy number of CCL3L1 has been associated with lower susceptibility to human HIV infection (Gonzalez, E. et al. (2005). *Science* **307**: 1434–1440), and a low copy number of FCGR3B (the CD16 cell surface immunoglobulin receptor) can increase susceptibility to systemic lupus erythematosus and similar inflammatory autoimmune disorders (Aitman T. J. et al. (2006). *Nature* **439**:

IDT Exhibit 1020.5218

851–855). Copy number variation has also been associated with autism, schizophrenia, and idiopathic learning disability (Cook EH, Scherer SW (2008). *Nature* **455** (7215): 919–23). In humans, CNVs encompass more DNA than SNPs. CNVs can be limited to a single gene or include a contiguous set of genes. CNVs can result in having either too many or too few of the dosage sensitive genes, which may be responsible for a substantial amount of human phenotypic variability, complex behavioral traits and disease susceptibility. In certain cases, such as rapidly growing Escherichia coli cells, the gene copy number can be 4-fold greater for genes located near the origin of DNA replication, rather than at the terminus of DNA replication. Elevating the gene copy number of a particular gene can increase the expression of the protein that it encodes.

[0085] In some embodiments, the methods of the present invention can be used to sequence a panel of loci. In one aspect, a panel of alleles may be sequenced. For example, the loci of interest include drug metabolizing enzyme and transporter (DMET). A common genomic technique is drug metabolizing enzyme and transporter (DMET) analysis, which is used to identify a subjects' likely response to a drug candidate based on the type and number of each of the two molecules that he or she produces. Many drug developers use DMET as part of wider absorption, distribution, metabolism, and elimination (ADME) analysis to identify patients who are most likely to derive benefits from a developmental treatment. The methods of the present invention may allow screening of about 2,000 drug metabolism markers across about 200 genes simultaneously. In other embodiments, the methods of the present invention can be used in advancing genome-wide association studies (GWAS) to identify common genetic factors that influence health and disease and sequence hundreds of high-value markers based on diseases.

[0086] In some embodiments, the loci-specific probes of the present invention are high fidelity oligonucleotide probes. The loci-specific probe set of the present invention may have very low proportion of incomplete and incorrect probes. In some embodiments, the loci-specific probe contains at least 10 nucleotides, for example, 50 nucleotides, i.e. 50-mer. The oligonucleotide probes of the present invention can range from 10 to 10,000 nucleotides (nt) in length, for example, from 50 to 2000 nt in length, including from 50-500 nt. The loci-specific probe can be 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 80, 90, 100, 110, 120, 130, 140, 150 nucleotides or longer. In one example, the loci-specific probe is a 50-mer. In some embodiments, the loci-specific probe is immobilized on a surface or a flow cell at its 5' end and the probe is extendable at the 3' hydroxy (OH) end. In one aspect, the loci-specific probes are allele-specific.

[0087] In one embodiment, assuming a locus or allele of interest is about 10 bp long. There can be 10 unique probe sequences per locus, in which 5 staggered probes are offset by one base. The probes can be in both forward and reverse orientations. An exemplary design of 10-bp locus of interest within genome 50-mer probes immobilized on surface at 5' end is shown in **Fig. 3**. In some embodiments, the probes are designed to avoid SNPs with respect to minor allele frequency cutoff values of > 0.05 at extreme 3' end of the probe sequences (**Fig. 4**). HapMap allele frequencies can be used. Exemplary sequence reads of about 15 bases or more and the expected distribution of read depth are shown in **Fig. 5**.

[0088] In some embodiments, "in situ" array features will each have at least hundreds of unique probe molecules with targets after hybridization and highly stringent washing. In some embodiments, at least two loci-specific or allele specific oligo probes are used to probe the same region on a target polynucleotide. The loci- or allele- specific or probes can be offset by 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30 bases or more to cover a certain region of the target polynucleotide, for example, genomic DNA.

[0089] In some embodiments, the present invention provides a method of performing a genomic sequencing reaction. Since the sequences of the probes are loci-specific and known, the number of duplicate probes for sequencing a certain

IDT Exhibit 1020.5219

region of the genome can be reduced. In one embodiment, there is no duplicate probe used in the subject methods. In another embodiment, 1, 2, 3, 4, or 5 duplicate probes are used.

[0090] **Figure 8** shows probe design sequences for five example loci from the PharmaADME panel (PharmaADME.org). Each locus is represented by sets of five overlapping probe sequences in both the forward and reverse orientations. The five loci include examples of a deletion, insertion, copy number, and two SNP variants. The probe sequences for the copy number example at CYP2A6 were modified from a primer sequence described in Fukami et al 2006 (*Pharmacogenomics Journal* 6: 401-412).

[0091] The loci-specific probes of the present invention may offer several advantages. Ordinary short oligonucleotide probes usually provide higher sequence-specificity but lower efficacy of hybridization than longer ordinary oligonucleotide probes where both are fully complementary to the target polynucleotide. The loci-specific, sequence-known oligonucleotide probes according to the present invention combine the hybridization efficacy of long probes with the sequence-specificity of short probes. In some embodiments, these loci-specific, sequence-known probes provide higher affinities toward their polynucleotide targets than short hybridization probes as well as increased sensitivity (signal-to-noise ratio) for target detection in comparison to ordinary long hybridization probes, which themselves are more sensitive than the shorter probes. In some embodiments, more than one probe that is specific for a region of the target polynucleotide is used. The probe is suitable for hybridization on complementary strands of a corresponding target nucleotide sequence to permit formation of a polymerase chain reaction product. In certain embodiments, an oligonucleotide probe according to the invention contains a single polynucleotide sequence, while in other embodiments, an oligonucleotide probe contains multiple polynucleotide sequences (so called bi-partite, tri- partite, tetra-partite, etc.). Oligonucleotide probes according to the present invention can include virtually any kind of nucleotide base, including (but not limited to) unmodified RNA bases, unmodified DNA bases or both (e.g., RNA-DNA chimeric polynucleotides) as well as one or more chemically modified RNA or DNA residue. Depending on their length and the presence of modified residues, the probes can be either chemically synthesized or prepared by enzymatic polymerization using techniques known in the art. In addition to ordinary RNA and DNA polymerases, mutated or engineered versions of polymerase enzymes can be used to incorporate into the probes variety of modified nucleotides.

[0092] The loci- and allele-specific oligonucleotide probes described herein can substitute for ordinary probes in commonly used hybridization methods such as dot/slot blots, Northern and Southern blots, in situ hybridization, sandwich hybridization, and gel-shift assays. These probes allow faster, more accurate and more sensitive detection and quantification of target polynucleotides with a higher level of multiplexing than ordinary hybridization probe.

[0093] In certain embodiments, the loci- and allele-specific oligonucleotide probes of the present invention can be modified to include additional functional moieties (also called modified oligonucleotide probes). Exemplary functional moieties include, without limitation, radioactive and fluorescent labels as well as anchor ligands such as biotin or digoxigenin. The functional moieties can be located internally or at either end of the probes. Probe modification can be carried out post- synthetically by chemical or enzymatic reactions such as ligation or polymerase-assisted extension. Alternatively, internal labels and anchor ligands can be incorporated into probes directly during enzymatic polymerization reactions using trace amounts of modified NTPs as substrates.

[0094] The capture oligonucleotides can be in the form of ribonucleotides, deoxyribonucleotides, modified ribonucleotides, modified deoxyribonucleotides, peptide nucleotide analogues, modified peptide nucleotide analogues, modified phosphate-sugar backbone oligonucleotides, nucleotide analogues, and mixtures thereof.

Hybridization

IDT Exhibit 1020.5220

[0095] In one aspect, the present invention provides a method for sequencing a target polynucleotide comprising the steps of: a) hybridizing a target polynucleotide to an allele-specific and sequence-defined capture probe; b) adding a first solution probe and ligase to the target-probe hybridization complexes to allow ligation of the solution probe to the substrate-attached probe and hybridization between the first solution probe and the target polynucleotide; c) capping the 3' end of the target-probe hybridization duplexes with a nucleotide reversible terminator; d) cleaving the 3' cap on the specific target-probe hybridization duplexes at the cleavage site present on the first solution probe; and (e) placing the specific target-probe hybridization duplexes under nucleic acid synthesis conditions in the presence of labeled nucleotides to allow extension of the probe that is complementary to the target polynucleotide.

[0096] In another aspect, the method further comprises the steps of: f) ligating a second solution probe to the first solution probe such that the second solution probe hybridizes to the target polynucleotide; g) washing the hybridization duplexes from step f) to remove nonspecific hybridization between the second solution probe and the target polynucleotide; h) ligating a set of third solution probes carrying a common primer at 3' end to the previously ligated solution probe such that the solution probe carrying the primer hybridizes to the target polynucleotide; i) washing the hybridization duplexes from step h) to remove the target polynucleotide from the hybridization duplexes; j) adding primers that are complementary to the common primer at 3' end of the third solution probes to allow primer annealing; and k) placing the single strand probe polynucleotide under nucleic acid synthesis conditions in the presence of free nucleotides to allow base extension from the annealed primer at the 3' end based on nucleic acid sequence of the probe strand that is complementary to the target polynucleotide. The methods of the present invention further comprise obtaining sequence information of the target polynucleotide via a computer, e.g. a computer software or algorithm.

[0097] "Hybridization" as used herein typically refers to the technique of allowing two single-stranded polynucleotide sequences with some degree of complementarity to bind to one another to form a stable double-stranded polynucleotide. "Complementary" and its equivalents as used herein generally refers to the hybridization or base pairing between nucleotides or nucleic acids, such as, for instance, between the two strands of a double stranded DNA molecule or between an oligonucleotide primer and a primer binding site on a single stranded nucleic acid to be sequenced or amplified. Complementary nucleotides are, generally, A and T (or A and U), or C and G. Two single stranded RNA or DNA molecules are said to be complementary when the nucleotides of one strand, optimally aligned and compared and with appropriate nucleotide insertions or deletions, pair with at least about 70% of the nucleotides of the other strand, usually at least about 80%, 85%, 90% to 95%, and more preferably from about 98 to 100%. See, M. Kanehisa Nucleic Acids Res. 12:203 (1984), incorporated herein by reference.

[0098] Hybridization usually involves the steps of 1) allowing binding between probe and target; and 2) washing away unbound or weakly bound probes under stringent conditions, wherein stringent hybridization conditions are those washing conditions that provide dissociation for imperfect complexes while preserving the intended complexes between target-specific probes and corresponding targets. Improvements in hybridization characteristics can be improvements in the selectivity of hybridization (sequence specificity and mismatch discrimination), the sensitivity of hybridization (ratio of absolute signal to background signal, signal-to-noise ratio), the affinity between probe and target (ratio of binding rate to dissociation rate between hybridization probes and targets); the stability of the duplex or complex (thermal stability, Tm; also kinetic inertness of dissociation or kinetic trap), or the efficiency or efficacy of hybridization (hybridization rate and/or yield of complex between probe and target for a fixed time of incubation under hybridization conditions). Methods for conducting polynucleotide hybridization assays have been well developed in the art. Hybridization assay procedures and conditions will vary depending on the application and are selected in accordance with the general binding methods known including those referred to in: Maniatis et al. Molecular Cloning: A Laboratory Manual (2nd Ed. Cold

IDT Exhibit 1020.5221

Spring Harbor, N.Y, 1989); Berger and Kimmel Methods in Enzymology, Vol. 152, Guide to Molecular Cloning Techniques (Academic Press, Inc., San Diego, Calif., 1987); Young and Davism, P.N.A.S, 80: 1194 (1983). Methods and apparatus for carrying out repeated and controlled hybridization reactions have been described in U.S. Pat. Nos. 5,871,928, 5,874,219, 6,045,996 and 6,386,749, 6,391,623 all of which are incorporated herein by reference.

[0099] In some embodiments, the primary polymerase chain reaction mixture is subjected to two or more polymerase chain reaction cycles involving a denaturation treatment, a hybridization treatment, and an extension treatment. During hybridization, target-specific portions of the probe hybridize to the target nucleotide sequences. The extension treatment causes hybridized primary oligonucleotide primers to be extended to form primary extension products complementary to the target nucleotide sequence to which the primary oligonucleotide primers are hybridized.

[00100] One notable feature of the present invention is that the hybridization probe is allele-specific and the sequence of the probe is known so that the capture oligonucleotide probe can hybridize with the target polynucleotide sequence in a stable fashion. Unless the oligonucleotides are designed in this fashion, false positive signals may result due to capture of adjacent unreacted oligonucleotides from the same oligonucleotide set which are hybridized to the target. In one aspect, the ligation step with the solution probes and the subsequent capping of the 3' end of the hybridization products and the cleaving of the 3' end cap of the specific probe-target hybridization complexes allow for higher specificity of hybridization to be achieved. In some embodiments, the hybridization specificity is greater than 95%, 96%, 97%, 98%, 99%, 99.5% or higher.

[00101] N-mer arrays (every possible sequence of a given length) can be used for sequencing by hybridization. N-mer arrays can also be used to sort a complex sample. This is particularly advantageous where they are linked to an anchor sequence, for example polyadenylation signal sequence or Poly A tail, or a sequence complementary to a clamp/adaptor sequence that has been ligated to target molecules. Each element of the spatially addressable array will contain a common anchor sequence and a unique member of the N-mer set. These probes can be used in hybridization, primer extension, ligation assays etc. In particular they can be used for priming sequencing by synthesis reactions, where for example the sequence has been fragmented and fragments have been ligated to a clamp. The advantage of the N-mer is that a certain amount of sequence information is already obtained from the target just by hybridization of the N-mer before a sequencing by synthesis reaction has been performed.

[00102] The stringency for target-probe hybridization can be adjusted and optimized. Hybridization stringency typically refers to the degree to which mismatches are tolerated in a hybridization assay. High stringency is achieved by using a high temperature and low salt concentration. Increasing the concentration of salt and reducing the temperature reduces the hybridization stringency, and enhances the stability of mismatched heteroduplexes. In some embodiments of the present invention, the highest possible stringency in hybridization and washing is used in the subject methods to increase hybridization specificity. In some embodiments of the ligation-captured sequencing, nonspecific hybridization products between a solution probe and the target polynucleotide can be washed off with high-stringency washing. The ionic strength of the buffers used for washing can be adjusted, for example, salt concentration can be lowered for extreme stringency. In some embodiments of the ligation-captured sequencing, the target polynucleotide, for example, genomic DNA can be washed off of the hybridization duplex to allow subsequent primer annealing and base extension (**Fig. 2B, C**).

Ligation

[00103] In one aspect, the present invention provides a ligation-assisted sequencing method in which a first solution is ligated to the capture probe that is attached to the substrate such that the solution probe hybridizes with the target polynucleotide (**Fig. 1**). In another aspect, the present invention provides a ligation-captured sequencing method

IDT Exhibit 1020.5222

comprising the steps depicted in **Fig. 2**. In some embodiments, a first solution probe hybridizes with the target polynucleotide before capping (**Fig. 2A**). A subsequent ligation with a second solution probe is performed after cleaving of the cap on the hybridization products, followed by washing of the ligation products to remove nonspecific hybridization products (**Fig. 2B**). In some embodiments, a third solution probe carrying a common 3' primer is added to allow ligation with the previous solution probe and hybridization between the 3' primer-carrying solution probe and the target polynucleotide. The target polynucleotide can be washed off from the complementary strand which contains the 3' primer. Subsequent primer annealing in the presence of free nucleotides may allow base extension of the complementary strand containing the 3' primer, thereby sequencing the target polynucleotide (**Fig. 2C**).

[00104] Ligation (chemical or enzymatic) provides for improving specificity and for trapping transient interactions. In some embodiments of the subject methods, the target strand is captured by the immobilized oligonucleotide capture probe. An oligonucleotide probe in solution, termed first solution probe, is ligated to the first capture probe, in a target dependent manner. There are several ways that this can be applied. In the first type of assay, the first solution probe is complementary in the region of the known polymorphisms under investigation, the first loci site. One oligo of either the array oligos or the first solution probe overlaps the SNP site and the other ends one base upstream of it. In the second type of assay, the first solution probe comprise the complete set with every sequence of a given length. This allows analysis of every position in the target.

[00105] In some embodiments, the first solution probe carries at least one base that is degenerate. Degenerate base typically refers to a position on a DNA sequence that can have multiple possible alternatives. For example, the 5' end of the first solution probe can be either A, T, G, or C. In some embodiments, the first base at 5' end of the first solution probe is degenerate. In some embodiments, the first solution probe is color labeled. The first solution probes carrying A, T, G, or C as the first base at the 5' end can be labeled with different colors such that the first base upon ligation to the probe in hybridization duplex with the target polynucleotide can be known by the color. In some embodiments, the first solution probe carries a cleavage site near its 3' end. Examples of a cleavage site include but are not limited to an amino-2-hydroxypropyl group. The cleavage site on the first solution probe ensures that the cap, i.e. reversible terminator nucleotides on the specific ligation products will be cleaved off while the cap on any non-specific hybridization complexes will remain. In some embodiments, the 3' end of a first solution probe is labeled and the 5' end has a phosphate group. The solution probe may contain 5-50 nucleotides. In some embodiments, the first solution probe contains 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, 35, 40, 45, or 50 nucleotides. In one example, the first solution probe contains 9 nucleotides, i.e. a 9-mer. In another example, the first solution probe is a 6-mer. In another example, the first solution probe is a 15-mer. In some embodiments, degenerate oligos are used as first solution probes. In one example, a degenerate 9-mer is used as the first solution probe. Since each base can be A, T, G, or C, 4^9 different first solution probes can be generated for specific ligation and hybridization to the target polynucleotide via the methods of the present invention.

[00106] One example of a typical ligation reaction is as follows: 5× ligation buffer, 100 mM Tris-HCL pH 8.3, 0.5% Triton X-100, 50 mM MgCl, 250 mM KCl, 5 mM AND+, 50 mM DTT, 5 mM EDTA, solution oligonucleotide 5-10 pmol. *Thermus thermophilus* DNA ligase (Tth DNA ligase) 1 U/ul, target sample, between 37° C. and 65° C. 1 hr. Alternatively, stacking hybridization can be performed first in high salt: 1M NaCl, 3-4.4M TMACl, 5-10 pmol solution oligonucleotide, target sample. After washing of excess reagents from the array under conditions that retain the solution oligonucleotide, the above reaction mix minus solution oligonucleotide and target sample is added to the reaction mix.

[00107] The solution probes of the present invention are suitable for ligation when hybridized to the target polynucleotide. During the initial hybridization, the more nucleotides in a given sequence that are complementary to

IDT Exhibit 1020.5223

another sequence, the more stringent the conditions can be for hybridization and the more specific will be the binding of the two sequences. Increased stringency is achieved by elevating the temperature, increasing the ratio of co-solvents, lowering the salt concentration, and the like. Two nucleic acid molecules may be hybridized although the base pairing is not completely complementary. Accordingly, mismatched bases do not prevent hybridization of two nucleic acid molecules provided that appropriate conditions, well known in the art, are used. In the ligation step of the subject method, there is a mismatch which interferes with such ligation when the oligonucleotide probes are hybridized non-specifically to any other nucleotide sequence other than the specific target polynucleotide present in the sample. Thus, the ligation step of the present invention provides higher specificity for the hybridization between the probe and the target polynucleotide. The ligation product of oligonucleotide probes may be distinguished from either individual probes or unspecific hybridization products. In some embodiments, the first solution probe, the probe-target polynucleotide complexes, the unspecific hybridization products, and a ligase are mixed to form a ligase detection reaction mixture to carry out the ligation reaction.

[00108] In some embodiments, the ligase detection reaction mixture is subjected to one or more ligase detection reaction cycles having a denaturation treatment and hybridization treatment substantially as described above. In the hybridization treatment, the probe hybridizes at adjacent positions in a base-specific manner to the respective target polynucleotide if present. As a result, adjacent probes ligate to the target polynucleotide to form a ligation product sequence. The ligation product may contain a detectable reporter label. The solid-phase bound oligonucleotide probe (capture probe) may have hybridized to nucleotide sequences other than the respective complementary target polynucleotides but do not ligate together with the solution probe due to the presence of one or more mismatches and individually separate during the denaturation treatment. In some embodiments, following the ligation detection reaction cycles, the reporter labels of the ligation product can be detected which indicates the presence of one or more target nucleotide sequences in the sample. Desirably, the first solution probe and capture probe, or adjacent additional solution probes (such as the first and second or second and third solution probes) are suitable for ligation together at a ligation junction when hybridized to a corresponding target nucleotide sequence due to perfect complementarity at the ligation junction. However, when the oligonucleotide probes in the set are hybridized to any other nucleotide sequence present in the sample, there is a mismatch at a base at the ligation junction which interferes with ligation. In some embodiments, the mismatch is at the base adjacent to the 3' base at the ligation junction. Alternatively, the mismatch can be at the bases adjacent to bases at the ligation junction.

[00109] One example of a ligation reaction of the subject method is illustrated as follows. After denaturation of the target nucleic acid, if present as a double stranded DNA molecule, at a temperature of 80-105°C, preferably 94°C, ligation detection reaction oligonucleotide probes for one strand of the target nucleotide sequence are added along with a ligase, for example, a thermostable ligase such as Thermus aquaticus ligase. The oligonucleotide probes are then allowed to hybridize to the target nucleic acid molecule and ligate together, typically, at a temperature of 45-85°C, preferably, 65°C. When there is perfect complementarity at the ligation junction, the oligonucleotides can be ligated together. Where the variable nucleotide is T or A, the presence of T in the target nucleotide sequence will cause the oligonucleotide probe with the F1 reporter label to ligate to the common oligonucleotide probe with the 5' poly A tail A_n, and the presence of A in the target nucleotide sequence will cause the oligonucleotide probe with the F2 reporter label to ligate to the common oligonucleotide probe with A_n. Similarly, where the variable nucleotide is A or G, the presence of T in the target nucleotide sequence will cause the oligonucleotide probe with F3AA reporter label (i.e. the F3 reporter label coupled to 2 additional bases forming a 5' poly A spacer) to ligate to the common oligonucleotide probe with the 5' poly A tail A_{n+4}, and the presence of C in the target nucleotide sequence will cause the oligonucleotide probe with the F3 reporter label to

IDT Exhibit 1020.5224

ligate to the common oligonucleotide probe with the 5' poly A tail A_{n+4}. Following ligation, the material is again subjected to denaturation to separate the hybridized strands. The hybridization/ligation and denaturation steps can be carried out through one or more cycles (e.g., 1 to 50 cycles) to amplify target signals. Equimolar ligation of both F3-labeled oligonucleotides indicates the individual is heterozygous for that locus, whereas ligation of only the F2 labeled oligonucleotides indicates the individual is homozygous for the other locus.

[00110] The ligase detection reaction is described generally in WO 90/17239 to Barany et al., F. Barany et al., "Cloning, Overexpression and Nucleotide Sequence of a Thermostable DNA Ligase-encoding Gene," Gene, 109:1-11 (1991), and F. Barany, "Genetic Disease Detection and DNA Amplification Using Cloned Thermostable Ligase," Proc. Natl. Acad. Sci. USA, 88:189-193 (1991), the disclosures of which are hereby incorporated by reference. In accordance with the present invention, the ligase detection reaction can use 2 sets of complementary oligonucleotides. This is known as the ligase chain reaction which is described in the 3 immediately preceding references, which are hereby incorporated by reference. Alternatively, the ligase detection reaction can involve a single cycle which is known as the oligonucleotide ligation assay. See Landegren, et al., "A Ligase-Mediated Gene Detection Technique," Science 241:1077-80 (1988); Landegren, et al., "DNA Diagnostics--Molecular Techniques and Automation," Science 242:229-37 (1988); and U.S. Pat. No. 4,988,617 to Landegren, et al., which are hereby incorporated by reference. During ligase detection reaction phases, the denaturation treatment can be carried out at a temperature of 80-105°C., while hybridization can take place at 50-85°C. Each cycle may comprise a denaturation treatment and a thermal hybridization treatment which in total is from about one to five minutes long. Typically, the ligation detection reaction involves repeatedly denaturing and hybridizing for 2 to 50 cycles. The total time for the ligase detection reaction phase is 1 to 250 minutes. The oligonucleotide probe sets or primers can be in the form of ribonucleotides, deoxynucleotides, modified ribonucleotides, modified deoxyribonucleotides, modified phosphate-sugar-backbone oligonucleotides, nucleotide analogs, and mixtures thereof. The oligonucleotide probe sets or primers may have a reporter label suitable for detection. Useful labels include but are not limited to chromophores, fluorescent moieties, enzymes, antigens, heavy metals, magnetic probes, dyes, phosphorescent groups, radioactive materials, chemiluminescent moieties, and electrochemical detecting moieties.

[00111] One example of a ligase that can be used in the present invention is thermostable ligase derived from Thermus aquaticus (M. Takahashi, et al., "Thermophillic DNA Ligase," J. Biol. Chem. 259:10041-47 (1984)), which is hereby incorporated by reference. Alternatively, it can be prepared recombinantly. Procedures for such isolation as well as the recombinant production of Thermus aquaticus ligase (as well as Thermus themophilus ligase) are disclosed in WO 90/17239 to Barany, et. al., and F. Barany, et al., "Cloning, Overexpression and Nucleotide Sequence of a Thermostable DNA-Ligase Encoding Gene," Gene 109:1-11 (1991), which are hereby incorporated by reference. These references contain complete sequence information for this ligase as well as the encoding DNA. Other suitable ligases include but are not limited to E. coli ligase, T4 ligase, and Pyococcus ligase.

[00112] The ligation detection reaction mixture may include a carrier DNA, such as salmon sperm DNA. The hybridization step in the ligase detection reaction, which is preferably a thermal hybridization treatment discriminates between nucleotide sequences based on a distinguishing nucleotide at the ligation junctions. The difference between the target nucleotide sequences can be, for example, a single nucleic acid base difference, a nucleic acid deletion, a nucleic acid insertion, or rearrangement. Such sequence differences involving more than one base can also be detected. Preferably, the oligonucleotide probe sets have substantially the same length so that they hybridize to target nucleotide sequences at substantially similar hybridization conditions. As a result, the process of the present invention is able to detect infectious diseases, genetic diseases, and cancer. It is also useful in environmental monitoring, forensics, and food science.

IDT Exhibit 1020.5225

Capping and Sequencing by Synthesis

[00113] In general, after the ligation step the non-specific hybridization is eliminated by a washing step. Thus, the only target polynucleotide captured are the allele and/or locus specific and are subject to sequencing using methods known in the art, preferably by sequencing-by-synthesis.

[00114] In some embodiments, the subject methods of the present invention comprise a capping step after the first ligation of the first solution probe to the target polynucleotide prevent base extension occurring in complexes formed by non-specific hybridization. For example, nonspecific hybridization may occur between an allele-specific oligo probe and the target polynucleotide. In some embodiments, the capping of the hybridization products uses nucleotide reversible terminators. Nucleotide reversible terminators prevent sequencing by synthesis. The array can be designed in such a way that each spot captures a consecutive fragment of DNA from the genome. For example probes can be 10-100 bases apart in the genome (or the part of the genome of interest). The intervening sequence can then be determined by sequencing by synthesis methods known in the art (for example, see WO9844152). The probes can also be part of an N-mer as disclosed herein. The underlying theory of procedures according to the invention is that, contrary to natural primer mediated template directed complementary DNA synthesis, only one base can be added at one time. Further additions are prevented by, for example, a blocking group such as a nucleotide reversible terminator, similar to having protecting groups in automated chemical DNA synthesis. The base added to each single molecule is detected and recorded after base addition. The blocking group such as a nucleotide reversible terminator is then removed allowing the next base to be added. As well as base by base, the procedure can also be performed dinucleotide by dinucleotide or oligomer by oligomer of any convenient length.

[00115] Nucleotide reversible terminators are nucleotide analogues, which are modified with a reversible chemical moiety capping the 3′-OH group to temporarily terminate the polymerase reaction. In this way, only one nucleotide is incorporated into the growing DNA strand even in homopolymeric regions. For example, the 3' end can be capped with an amino-2-hydroxypropyl group. An allyl or a 2-nitrobenzyl group can also be used as the reversible moiety to cap the 3′-OH of the four nucleotides. Examples of reversible terminators include but are not limited to 3′-*O*-modified nucleotides such as 3′-*O*-allyl-dNTPs and 3′-*O*-(2-nitrobenzyl)-dNTPs. After detection of the cleavage site present on the solution probe, the 3′-OH of the primer extension products is regenerated through different deprotection methods. The capping moiety on the 3′-OH of the DNA extension product can be efficiently removed after detection of a cleavage site by a chemical method, enzymatic reaction or photolysis, i.e. the cap will be cleaved from the cleavage site. To sequence DNA, in some embodiments, templates containing homopolymeric regions are immobilized on Sepharose beads, and then extension–signal detection–deprotection cycles are conducted by using the nucleotide reversible terminators on the DNA beads to unambiguously decipher the sequence of DNA templates. In some embodiments, this reversible-terminator-sequencing approach is used in the subject methods to accurately determine DNA sequences. (The cap may be referred to herein as a "protective group").

[00116] In some embodiments, the nucleotide reversible terminators are labeled with colors, for example, flurophores in four colors. In some embodiments, the nucleotide reversible terminator read length is at least 15 bases.

Base Extension

[00117] After the cap has been cleaved, base extension of the probe can be performed to continue the sequencing of the target polynucleotide. The base extension begins at the 3' end of the most distal solution probe from the capture probe **(Fig. 1)**.

[00118] Primer extension is another means for improving specificity of hybridization and genotyping at the free end of the immobilized probe and for trapping transient interactions. In general, the primers used according to the methods of

IDT Exhibit 1020.5226

the invention embrace oligonucleotides of sufficient length and appropriate sequence which provide specific initiation of polymerization of a significant number of nucleic acid molecules containing the target nucleic acid under the conditions of stringency for the reaction utilizing the primers. In this manner, it is possible to selectively amplify the specific target nucleic acid sequence containing the nucleic acid of interest. Specifically, the term "primer" as used herein refers to a sequence comprising two or more deoxyribonucleotides or ribonucleotides, preferably at least eight, which sequence is capable of initiating synthesis of a primer extension product that is substantially complementary to a target nucleic acid strand. The oligonucleotide primer typically contains 15-22 or more nucleotides, although it may contain fewer nucleotides as long as the primer is of sufficient specificity to allow essentially only the amplification of the specifically desired target nucleotide sequence (i.e., the primer is substantially complementary). The exact length of primer will depend on many factors, including temperature, buffer, and nucleotide composition. "Substantially complementary" refers to that the primers are sufficiently complementary to hybridize with their respective strands under conditions which allow the agent for polymerization to function. In other words, the primers should have sufficient complementarily with the flanking sequences to hybridize with and permit amplification of the nucleotide sequence. Preferably, the 3′ terminus of the primer that is extended has perfectly base paired complementarity with the complementary flanking strand. The oligonucleotide primers for use in the invention may be prepared using any suitable method, such as conventional phosphotriester and phosphodiester methods or automated embodiments thereof. In one such automated embodiment, diethylphosphoramidites are used as starting materials and may be synthesized as described by Beaucage, et al. (Tetrahedron Letters, 22:1859-1862, 1981). One method for synthesizing oligonucleotides on a modified solid support is described in U.S. Pat. No. 4,458,066. One method of amplification which can be used according to this invention is the polymerase chain reaction (PCR) described in U.S. Pat. Nos. 4,683,202 and 4,683,195.

[00119] More than one way of primer extension can be applied. The first is the multi-primer approach, where there are separate array elements containing single molecules for each allele. The second is the multi-base approach in which a single array contains a single species of primer whose last base is upstream of the polymorphic site. The different alleles are distinguished by incorporation of different bases each of which is differentially labeled. This approach is also known as mini-sequencing. For example, the following reaction mix and conditions can be used: 5 × polymerase buffer, 200 mM Tris-HCl pH 7.5, 100 mM $MgCl_2$, 250 mM NaCl, 2.5 mM DTT; ddNTPs or dNTPs (multibase); dNTPs (multiprimer), Sequenase V.2 (0.5 μ/μl) in polymerase dilution buffer, target sample, 37° C 1 hr. It can be advantageous to label the primer/capture probe to lend more confidence to an extension signal, if it co-localizes with labeled capture probe. Preferably no cold dNTP corresponding to the labeled dNTP is added. An exo-polymerase, for example, thermosequenase (Amersham) or Taquenase (Promega), can be used. The target can be capture immobilized and synthesis primed using an upstream primer. Multiple primers can prime synthesis at several points along the captured target. The target may or may not be horizontalized.

[00120] In some embodiments, in the polymerase chain reaction, one or a plurality of oligonucleotide primer sets can be provided. Each set can have an upstream primer containing the same sequence as the 5' upstream primer-specific portion of the ligation product sequence and a downstream primer complementary to the 3' downstream primer-specific portion of the ligation product sequence, where one primer has a detectable reporter label. In some embodiments, the ligase detection reaction mixture is blended with the one or a plurality of oligonucleotide primer sets and the polymerase to form a polymerase chain reaction mixture.

[00121] In some embodiments, the polymerase chain reaction mixture is subjected to one or more polymerase chain reaction cycles which include a denaturation treatment, a hybridization treatment, and an extension treatment. During the denaturation treatment, hybridized nucleic acid sequences are separated. The hybridization treatment causes primers to

IDT Exhibit 1020.5227

hybridize to their complementary primer-specific portions of the ligation product sequence. During the extension treatment, hybridized primers are extended to form extension products complementary to the sequences to which the primers are hybridized. In a first cycle of the polymerase chain reaction phase, the downstream primer hybridizes to the 3' downstream primer-specific portion of the ligation product sequence and is extended to form an extension product complementary to the ligation product sequence. In subsequent cycles, the upstream primer hybridizes to the 5' upstream primer-specific portion of the extension product complementary to the ligation product sequence and the 3' downstream primer hybridizes to the 3' downstream portion of the ligation product sequence. Following the polymerase chain reaction phase of this process, the reporter labels are detected and the extension products are distinguished to indicate the presence of one or more target nucleotide sequences in the sample.

[00122] The polymerase chain reaction process is fully described in H. Erlich, et. al., "Recent Advances in the Polymerase Chain Reaction," Science 252: 1643-50 (1991); M. Innis, et. al., PCR Protocols: A Guide to Methods and Applications, Academic Press: New York (1990); and R. Saiki, et. al., "Primer-directed Enzymatic Amplification of DNA with a Thermostable DNA Polymerase," Science 239: 487-91 (1988), which are hereby incorporated by reference.

Signal Detection and Imaging

[00123] In some embodiments, the subject methods provide sensitive signal detection that can detect at least hundreds of molecules with high resolution. In one embodiment, the subject methods can detect one molecule. In another embodiment, the subject methods can detect approximately 10 molecules. For example, the subject methods can detect 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more molecules. The ability of the subject methods to detect about 10-20 molecules can be critical to enable highly stringent washing as described herein. Without being bound by any theory, it is believed that highly stringent washing can wash off all or substantially all of the non-specific binding while leaving a few number of specific target molecules hybridized within the array feature of the present invention. In some embodiments, a highly sensitive detector, e.g. the Helicos system, can be used to detect the few molecules, e.g. 10-20 molecules that are specifically bound to the target polynucleotide.

[00124] Polynucleotide of the invention may be labeled. In some embodiments, a molecule or compound has at least one element, isotope or chemical compound attached to enable the detection of the compound. In general, labels of use in the invention include without limitation isotopic labels, which may be radioactive or heavy isotopes, magnetic labels, electrical labels, thermal labels, colored and luminescent dyes, enzymes and magnetic particles as well. Dyes of use in the invention may be chromophores, phosphors or fluorescent dyes, which due to their strong signals provide a good signal-to-noise ratio for decoding.

[00125] Many embodiments of the invention include the use of fluorescent labels. Suitable dyes for use in the invention include, but are not limited to, fluorescent lanthanide complexes, including those of Europium and Terbium, fluorescein, rhodamine, tetramethylrhodamine, eosin, erythrosin, coumarin, methyl-coumarins, pyrene, Malacite green, stilbene, Lucifer Yellow, Cascade Blue, Texas Red, and others described in the 6th Edition of the Molecular Probes Handbook by Richard P. Haugland, hereby expressly incorporated by reference in its entirety for all purposes and in particular for its teachings regarding labels of use in accordance with the present invention. Commercially available fluorescent nucleotide analogues readily incorporated into the labeling oligonucleotides include, for example, Cy3-dCTP, Cy3-dUTP, Cy5-dCTP, Cy5- dUTP (Amersham Biosciences, Piscataway, New Jersey, USA), fluorescein-12-dUTP, tetramethylrhodamine-6-dUTP, Texas Red®-5-dUTP, Cascade Blue®-7-dUTP, BODIPY® FL-14-dUTP, BODIPY®R-14-dUTP, BODIPY® TR-14-dUTP, Rhodamine GreenTM-5-dUTP, Oregon Green® 488-5-dUTP, Texas Red®-12-dUTP, BODIPY® 630/650-14-dUTP, BODIPY® 650/665-1 4-dUTP, Alexa Fluor® 488-5-dUTP, Alexa Fluor® 532-5-dUTP, Alexa Fluor® 568-5-dUTP, Alexa Fluor® 594-5-dUTP, Alexa Fluor® 546-1 4-dUTP, fluorescein-12-UTP,

IDT Exhibit 1020.5228

tetramethylrhodamine-6-UTP, Texas Red®-5-UTP, Cascade Blue®-7-UTP, BODIPY® FL-14-UTP, BODIPY® TMR-14-UTP, BODIPY® TR-14-UTP, Rhodamine GreenTM-5-UTP, Alexa Fluor® 488-5-UTP, Alexa Fluor® 546-1 4-UTP (Molecular Probes, Inc. Eugene, OR, USA). Other fluorophores available for post-synthetic attachment include, inter alia, Alexa Fluor® 350, Alexa Fluor® 532, Alexa Fluor® 546, Alexa Fluor® 568, Alexa Fluor® 594, Alexa Fluor® 647, BODIPY 493/503, BODIPY FL, BODIPY R6G, BODIPY 530/550, BODIPY TMR, BODIPY 558/568, BODIPY 558/568, BODIPY 564/570, BODIPY 576/589, BODIPY 581/591, BODIPY 630/650, BODIPY 650/665, Cascade Blue, Cascade Yellow, Dansyl, lissamine rhodamine B, Marina Blue, Oregon Green 488, Oregon Green 514, Pacific Blue, rhodamine 6G, rhodamine green, rhodamine red, tetramethylrhodamine, Texas Red (available from Molecular Probes, Inc., Eugene, OR, USA), and Cy2, Cy3.5, Cy5.5, and Cy7 (Amersham Biosciences, Piscataway, NJ USA, and others).

[00126] A number of multiplex detection formats can be used, including either labeled/tagged bead sets (e.g., those produced by Luminex), in which each label is assigned to the individual probe-specific primer, or oligonucleotide arrays on slides, in which in which specific oligonucleotide spot/position is assigned to the individual probe-specific primer. The limited sequence complexity of the recovered target- specific probes provides conditions for easier and higher level multiplexing, especially using with universal and Zip-code/ID sequence tags. After the hybridization of the primers to the target-probe complex, the primers are extended by a nucleotide polymerase. Polymerase chain reaction is a technique well known in the relevant art. In certain embodiments, the polymerase is selected from an RNA polymerase and a reverse transcriptase.

[00127] Where an array is utilized, the detection phase of the process may involve scanning and identifying target polynucleotide sequences in the test sample. Scanning can be carried out by scanning probe microscopy (SPM) including scanning tunneling microscopy (STM) and atomic force microscopy (AFM), scanning electron microscopy, confocal microscopy, charge-coupled device, infrared microscopy, electrical conductance, and fluorescent or phosphor imaging, for example fluorescence resonance energy transfer (FRET). Optical interrogation/detection techniques include but are not limited to near-field scanning optical microscopy (NSOM), confocal microscopy and evanescent wave excitation. More specific versions of these techniques include far-field confocal microscopy, two-photon microscopy, wide-field epi-illumination, and total internal reflection (TIR) microscopy. Many of the above techniques can also be used in a spectroscopic mode. The actual detection means include charge coupled device (CCD) cameras and intensified CCDs, photodiodes and photomultiplier tubes. These means and techniques are well-known in the art. Various detection methods are disclosed in U.S. Patent Application Publication No. US 2004/0248144, which is herein incorporated by reference.

[00128] For multicolor imaging, signals of different wavelength can be obtained by multiple acquisitions or by simultaneous acquisition by splitting the signal, using RGB detectors or analyzing the whole spectrum (Richard Levenson, Cambridge Healthtech Institutes, Fifth Annual meeting on Advances in Assays, Molecular Labels, Signaling and Detection, May 17-18th Washington D.C.). Several spectral lines can acquired by the use of a filter wheel or a monochromater. Electronic tunable filters such as acoustic-optic tunable filters or liquid crystal tunable filters can be used to obtain multispectral imaging (e.g. Oleg Hait, Sergey Smirnov and Chieu D. Tran, 2001, Analytical Chemistry 73: 732-739). An alternative method to obtain a spectrum is hyperspectral imaging (Schultz et al., 2001, Cytometry 43:239-247).

Processing of Raw Data and Analysis of Genetic Information with Computer Algorithm

[00129] Typically, identifying target polynucleotide sequence and integrating sequences to assemble genomic information is carried out with a computer (**Fig. 7**). In some embodiments, the present invention also encompasses

IDT Exhibit 1020.5229

computer software or algorithm designed to analyze and assemble sequence information obtained via the methods of the present invention.

[00130] In terms of sequence read interpretation for the "in situ" arrays, reads at array features correspond to X-Y coordinates that map to the loci of interest. A "read" typically refers to an observed sequence derived from raw data, such as the order of detected signals corresponding to the cyclical addition of individual nucleotides. In some embodiments, the reads are checked against the expected reference genome sequence at the 10-bp loci for quality control. A reference sequence enables the use of short read length. Reads that have passed the quality control check are then combined to generate a consensus sequence at each locus. In one example, there are 10 unique probes per locus of interest minus any reads that have failed the quality control checks.

[00131] In terms of sequence read interpretation for the "lawn" approach, the reads are at random locations on a surface, e.g. a flow cell. In some embodiments, the reads are checked against the expected subset of reference genome sequence at the loci of interest for quality control. Reads that have passed the quality control check are mapped to the individual locus of interest. Reads corresponding to each locus are then combined to generate a consensus sequence. In one embodiment, there are more than 3,000 reads per 10-bp locus.

[00132] In some embodiments, the present invention provides a method for sequencing a genome by generating reads that are greater than 5, 10, 15, 20, 30, 40, 50, 60, 74, 100, 150 bases. In some embodiments, the subject methods encompass sequencing a genome by sequencing a lawn of randomly placed probes, wherein each probe generates a read that is greater than 5 bases. The length of the reads is generally longer than the reads generated via other nucleotide sequencing techniques currently available. In other embodiments, the reads generated via the methods of the present invention can be 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 30, or more bases long.

[00133] In some embodiments, heterozygous insertions/deletions (indels) can be read and analyzed via the methods of the present invention (**Fig. 6**). Indels (insertions/deletions) are important DNA sequence variations because of the high frequency in the human genome, the deleterious effects on the reading frame and protein expression, and the association with disease and disease susceptibility of common diseases. In a recent study with a human individual with the whole genome sequenced, 292,102 heterozygous indels and 559,473 homozygous indels were identified. Decrypting such a large number of heterozygous indels is computationally intensive and requires efficient algorithms. In some embodiments of the "in situ" array of the present invention, probes have both maternal and paternal chromosome hybridized. The forward probe reads and the reverse probe reads of a paternal 3-base insertion are shown in **Fig. 6**. In some embodiments of a "lawn" flow cell, each probe has either maternal or paternal chromosome hybridized.

[00134] In some embodiments, copy number polymorphisms can be read and analyzed using the methods of the present invention. Copy number polymorphisms are present at CNVs of genomic DNA, for example, a 10-bp region within a larger common CNV. Count of reads at CNV can be proportional to an integer copy number. Copy number call can be based on comparison to estimated diploid read counts from all loci. The methods of the present invention do not generally map breakpoints.

[00135] Methods and apparatus for signal detection and processing of intensity data are disclosed in, for example, U.S. Pat. Nos. 5,143,854, 5,547,839, 5,578,832, 5,631,734, 5,800,992, 5,834,758; 5,856,092, 5,902,723, 5,936,324, 5,981,956, 6,025,601, 6,090,555, 6,141,096, 6,185,030, 6,201,639; 6,218,803; and 6,225,625, in U.S. Ser. No. 60/364,731 and in PCT Application PCT/US99/06097 (published as WO99/47964), each of which also is hereby incorporated by reference in its entirety for all purposes. Fluorescence imaging and software programs or algorithms for DNA sequence analysis and read interpretation are known to one of ordinary skill in the art and are disclosed in Harris TD, et al. "Single-Molecule DNA Sequencing of a Viral Genome" *Science* 4 April 2008: Vol. 320. no. 5872, pp. 106 – 109, which is herein

IDT Exhibit 1020.5230

incorporated by reference in its entirety. In some embodiments, Phred software is used for DNA sequence analysis. Phred reads DNA sequencer trace data, calls bases, assigns quality values to the bases, and writes the base calls and quality values to output files. Phred is a widely-used program for base calling DNA sequencing trace files. Phred can read trace data from SCF files and ABI model 373 and 377 DNA sequencer chromat files, automatically detecting the file format. After calling bases, phred writes the sequences to files in either FASTA format, the format suitable for XBAP, PHD format, or the SCF format. Quality values for the bases are written to FASTA format files or PHD files, which can be used by the phrap sequence assembly program in order to increase the accuracy of the assembled sequence. The quality value is a log-transformed error probability, specifically $Q = -10 \log_{10}(P_e)$ where Q and P_e are respectively the quality value and error probability of a particular base call. The phred quality values have been thoroughly tested for both accuracy and power to discriminate between correct and incorrect base-calls. Phred can use the quality values to perform sequence trimming.

[00136] Discrete groups of assay classification (e.g. nucleotide base calling) can be defined by various measures. A set of unique parameters are chosen to define each of several discrete groups. The result of interrogation of each individual molecule can be assigned to one of the discrete groups. One group can be assigned to represent signals that do not fall within known patterns. For example there may be groups for real base additions, a, c, g, and t in extension assays.

[00137] The single molecule approach enables direct counting and classification of individual events. A general algorithm for single molecule counting, once the single molecules have been labeled by, for example, thresholding, is:

Loop through all pixels, p(x,y) left to right, top to bottom

If p(xy)=0, do nothing

If p(xy)=1, add to counter

[00138] In some embodiments, the methods of this invention require basic image processing operations and counting, measuring and assignment operations to be performed on the raw images that are obtained. The invention includes the adaptation and application of general methods including software and algorithms, known in the art for digital signal processing, counting, measuring and making assignments from the raw data. This includes Bayesian, heuristic, machine learning and knowledge based methods. Moreover, digital data processing facilitates error correction and temporal resolution of reactions at the array surface. Thus, time-resolved microscopy techniques can be used to differentiate between bona-fide reactions between probe and sample and "noise" due to aberrant interactions which take place over extended incubation times. The use of time-gated detection or time-correlated single-photon counting is particularly preferred in such an embodiment.

[00139] In some embodiments, the present invention also provides a method for sorting signals obtained from single molecule analysis according to the confidence with which the signal may be treated. A high confidence in the signal leads to the signal being added to a PASS group and counted; signals in which confidence is low are added to a FAIL group and discarded, or used in error assessment and as a resource for assay design (for example the propensity of a particular primer sequence to give rise to errors in primer extension, can be used to inform primer design in future experiments.

[00140] The reaction is controlled by adjusting reaction components, for example salt concentration, ddNTP concentration, temperature or pH such that the incorporations occur within the time window analyzed. A subroutine can be included to check that the fluorescence shows single-step photobleaching characteristic, but ignoring short-scale fluctuations which are likely to be due to blinking. If a single dye molecule, which photobleaches after a time, is associated with each ddNTP, then an additional sub-process/routine can be added which eliminates signals that after an initial burst re-occur in the same pixel after such a number oftime points that the absence cannot be attributed to blinking.

IDT Exhibit 1020.5231

This is likely to be non-specific absorption at the same foci as a legitimate extension. A sub-routine can be included to eliminate any fluorescence that occurs in multiple filters, above the level expected for the dye being analyzed. Fluorescence due to a single dye molecule can be distinguished from particulate contamination by analyzing the concentration dependence of the signal. This can be done if each sequence is arrayed at two or more concentrations. Signals that remain at equal concentration across the array dilution are artifacts, and real signals are those whose frequency changes in line with changes in array probe concentration. If the array is composed of elements an additional process can be used to organize the data into groupings representing the array elements. In some embodiments, detection events are generated by labeling the sample nucleic acids and/or the probe molecules, and imaging the labels on the array using a suitable detector. Preferred labeling and detection techniques are described herein.

[00141] In some embodiments, the present invention also encompasses computer algorithm or software designed to analyze and assemble genomic information obtained via the methods of the present invention. In some embodiments, information on SNPs, CNVs, or DMET markers of a sample is obtained via the methods of the present invention.

[00142] During the initial hybridization, non-specific strand can be captured by the array probes. Using the ligation assisted or ligation captured methods of the present invention, the non-specific hybridization complexes can be easily eliminated from further sequencing. Regions of the genome that are known to be duplicated or repetitive may be avoided. In some embodiments, the use of markers other than the labels may aid in the integration of genomic information from pieces of target polynucleotide strands, for example, marking the ends of the molecule or other sites, including SNP sites with markers that can be distinguishable from the SNP color tags.

In some embodiments, specificity of read mapping can be increased by ranking probes based on number of hits to a reference genome. For example, in silico, tally hit counts of 15 to 20 base reads at X% identity, where X can vary from 95 to 99%. Empirically, defined subsets of probes can be run in experiments to determine the number of read hits to regions other than the locus of interest. Based on the results from a number of subsets, the problematic probes or loci can be assessed. In other embodiments, high hit count probes or loci can be mitigated. For example, these high hit count probes can be replaced with nearby probe sequences. It may require longer read lengths if new probes are further away from the locus of interest. In another embodiment, problematic probes or loci can be grouped and run in separate flow cell to maintain high read mapping accuracy for non-problematic probes. This may increase costs due to the need to run more than one flow cell per sample and the need to manufacture flow cells with different sets of probes. In another embodiment, the problematic loci may be omitted from the panel.

II. Ligation Captured Sequencing

[00143] Another aspect of the present invention relates to a method for determining or identifying nucleotide sequence via ligation captured sequencing. This method generally comprises the steps of a) immobilizing a set of loci-specific probes (also referred to as "capture probes") on a substrate, wherein the probes have defined sequences; b) hybridizing a sample containing a target polynucleotide to the set of allele-specific probes; c) ligating a first solution probe to the loci-specific probe, wherein the solution probe carries a cleavage site and hybridizes to the target polynucleotide; d) providing a nucleotide reversible terminator to the hybridization duplexes resulted from steps b) and c), wherein the nucleotide reversible terminator prevents base extension of nonspecific hybridization complexes; e) cleaving the nucleotide reversible terminator at the cleavage site present on the solution probe to allow further hybridization; f) ligating a second solution probe to the first solution probe such that the second solution probe hybridizes to the target polynucleotide; g) washing the hybridization duplexes from step f) to remove nonspecific hybridization between the second solution probe and the target polynucleotide; h) ligating a set of solution probes carrying a common primer at 3' end to the previously

IDT Exhibit 1020.5232

ligated solution probe such that the solution probe carrying the primer hybridizes to the target polynucleotide; i) washing the hybridization duplexes from step h) to remove the target polynucleotide from the hybridization duplexes; j) adding primers that are complementary to the common primer sequence at 3' end of the solution probes to allow primer annealing; and k) placing the single strand probe polynucleotide under nucleic acid synthesis conditions in the presence of free nucleotides to allow base extension from the annealed primer at the 3' end based on nucleic acid sequence of the probe strand that is complementary to the target polynucleotide. In some embodiments, the method further comprises obtaining sequence information of the target polynucleotide via a computer system. In some embodiments, the target polynucleotide is genomic DNA. In some embodiments, the solution probes are labeled and have at least one base that is degenerate. For example, the first base at 5' end of the solution probes is degenerate, i.e. it can be A, T, G, or C. In one embodiment, the solution probe is a 9-mer, i.e. contains 9 nucleotides. In some embodiments, the washing condition is sufficiently stringent such that the solution probes that have nonspecifically hybridized to the target polynucleotide can be washed off. In other embodiments, the washing condition is sufficiently stringent such that the target polynucleotide can be washed off from the hybridization duplex. The cycle of ligation between the solution probes and the hybridization between solution probes and the target polynucleotide followed by stringent washing to remove nonspecific ligation can be repeated more than one time. The subject ligation-captured sequencing method provides higher hybridization specificity. The specificity can be at least 95%, 96%, 97%, 98%, 99%, 99.5% or more.

[00144] In some embodiments, ligation-captured sequencing comprises ligating a second set of solution probes to the probe-target duplexes after cleaving of the 3' end reversible terminator to form double ligated probe-target duplexes, as depicted in **Fig. 2B**. Nonspecific hybridization between the second solution probe and the target polynucleotide may form. In some embodiments, the double ligation step is followed by washing. The hybridization products resulted from the second ligation reaction can be washed by non-stringent wash and stringent wash. For example, non-stringent wash is normally 2X SSC, 65°C, Eff T_m = 81.5 + 16.6[log(0.33)] + 0.41(45%)= 92.0°C, % Homology = 100 - [(92-65)/1.4] = 80.7%. An example of stringent wash can be 0.1X SSC, 65°C, Eff T_m = 81.5 + 16.6[log(0.0165)] + 0.41(45%) = 70.4°C, %Homology = 100 -[(70.4-65)/1.4] = 96.1%. Changes in temperature and salt concentrations affect the stringency of washes and therefore the hybridization specificity. Determining the stringency of washes after polynucleotide hybridization is well within the knowledge of one skilled in the art. In some embodiments, the washing condition is sufficiently stringent such that the second solution probes that have nonspecifically hybridized to the target polynucleotide can be washed off.

[00145] In some embodiments, the ligation step f) and the subsequent washing step g) can be repeated more than one time. For example, the ligation reaction with a solution probe followed by stringent washing to remove nonspecific hybridization can be performed 2, 3, 4, 5 times or more. The hybridization specificity can be enhanced via multiple rounds of ligation as disclosed herein.

[00146] In some embodiments, the ligation duplexes resulted from step g) are then provided with another set of solution probes which carry a common primer at 3' end. These primer-carrying solution probes can be labeled with different colors and have at least one degenerate base. In some embodiments, the first base at 5' end of the primer-carrying solution probes is degenerate, i.e. it can be A, T, G, or C. In some embodiments, the primer-carrying solution probes ligate to the previously ligated solution probe and hybridize to the target polynucleotide as depicted in **Fig. 2B**. This ligation between the primer-carrying solution probe and the target polynucleotide can be followed by stringent washing. In some embodiments, the washing condition is sufficiently stringent such that all the target polynucleotides can be washed off from the hybridization duplexes, leaving only the probe strands attached to the substrate, as shown in **Fig. 2C**. The probe strand is complementary to the target polynucleotide.

IDT Exhibit 1020.5233

[00147] In some embodiments, in the subsequent step (i.e. step j), primers that are complementary to the common 3' end primer present on the solution probes are added to the single-stranded probe strand, which is complementary to the target polynucleotide, to allow primer annealing. After the primers have annealed, free labeled nucleotides can be added under nucleic acid synthesis conditions to allow base extension from the annealed primer at the 3' end based on the sequence of the probe strand that is complementary to the target polynucleotide. In this way, the nucleic acid sequence of the target polynucleotide can be accurately determined.

[00148] In some embodiments, the sequence of the target polynucleotide is determined from 5' end. In other embodiments, the sequence of the target polynucleotide is determined from 3' end.

III. Utility

[00149] The methods of the present invention provide several advantages. First, the loci- and allele-specific, sequence-known probes provide a superior combination of efficiency, sensitivity and specificity of hybridization: they provide for hybridization that exceeds the sequence-specificity of ordinary short oligonucleotide probes, that exceed the kinetics of ordinary long oligonucleotide probes, that exceeds the affinity to polynucleotide and oligonucleotide target, and that exceeds the sensitivity (signal-to-noise ratio) of ordinary short and long oligonucleotide probes. Furthermore, they are SNP capable, and can carry multiple signal- reporting labels or ligands, provide for a higher level of multiplexing of diverse target sequences.

[00150] In addition, the ligation-assisted sequencing or ligation-captured sequencing comprising ligation of target polynucleotide with solution probes, capping of the 3' end of hybridization duplexes, specific cleaving of the 3' cap on nonspecific hybridization products, and/or subsequent ligation with the solution probes, stringent washing and primer extension of the complementary strand provide higher hybridization specificity and ensure more accurate sequencing of the target polynucleotide. The hybridization specificity achieved via the methods of the present invention is at least 95%, 96%, 97%, 98%, 98.5%, 99%, 99.5% or more. The present invention provides low-cost, high-throughput and accurate methods for sequencing target polynucleotide.

[00151] Moreover, the methods of the invention can be multiplexed to a very high degree. Samples can comprise pooled genomes of target and control subject populations respectively, since accurate analysis of allele frequencies can be accurately determined by single molecule counting. Since more than a single site on each molecule can be probed, haplotype information is easily determined. There is also the possibility of obtaining haplotype frequencies. Such methods are particularly applicable in association studies, where SNP frequencies are correlated with diseases in a population. The expense of single SNP typing reactions can be prohibitive when each study requires the performance of millions of individual reactions; the present invention permits millions of individual reactions to be performed and analyzed on a single array surface.

[00152] The methods of the present invention are useful in identifying high value polymorphisms located in regulatory elements and coding regions for a number of drug metabolizing enzyme and transporter (DMET) genes. Expression of these DMET genes will give information on the absorption, distribution, metabolism, and excretion profiles of a drug. Interpretation of complex transcriptional responses to various drugs and subsequent prediction of physiological effects is important for the development of effective therapeutics. The methods of the present invention can help draw links between gene expression profiles and physiological effects including a subjects' likely response to a drug candidate.

[00153] A wide variety of diseases can be detected by the process of the present invention, for example, infectious diseases caused by bacterial, viral, parasite, and fungal infectious agents. The resistance of various infectious agents to drugs can also be determined using the present invention.

IDT Exhibit 1020.5234

[00154] Genetic diseases can also be detected by the process of the present invention. This can be carried out by prenatal or post-natal screening for chromosomal and genetic aberrations or for genetic diseases. Examples of detectable genetic diseases include: 21 hydroxylase deficiency, cystic fibrosis, Fragile X Syndrome, Turner Syndrome, Duchenne Muscular Dystrophy, Down Syndrome or other trisomies, heart disease, single gene diseases, HLA typing, phenylketonuria, sickle cell anemia, Tay-Sachs Disease, thalassemia, Klinefelter Syndrome, Huntington Disease, autoimmune diseases, lipidosis, obesity defects, hemophilia, inborn errors of metabolism, and diabetes.

[00155] Cancers which can be detected by the process of the present invention generally involve oncogenes, tumor suppressor genes, or genes involved in DNA amplification, replication, recombination, or repair. Examples of these include: BRCA1 gene, p53 gene, APC gene, Her2/Neu amplification, Bcr/Ab1, K-ras gene, and human papillomavirus Types 16 and 18. Various aspects of the present invention can be used to identify amplifications, large deletions as well as point mutations and small deletions/insertions of the above genes in the following common human cancers: leukemia, colon cancer, breast cancer, lung cancer, prostate cancer, brain tumors, central nervous system tumors, bladder tumors, melanomas, liver cancer, osteosarcoma and other bone cancers, testicular and ovarian carcinomas, head and neck tumors, and cervical neoplasms.

[00156] In the area of environmental monitoring, the present invention can be used for detection, identification, and monitoring of pathogenic and indigenous microorganisms in natural and engineered ecosystems and microcosms such as in municipal waste water purification systems and water reservoirs or in polluted areas undergoing bioremediation. It is also possible to detect plasmids containing genes that can metabolize xenobiotics, to monitor specific target microorganisms in population dynamic studies, or either to detect, identify, or monitor genetically modified microorganisms in the environment and in industrial plants.

[00157] The present invention can also be used in a variety of forensic areas, including for human identification for military personnel and criminal investigation, paternity testing and family relation analysis, HLA compatibility typing, and screening blood, sperm, or transplantation organs for contamination.

[00158] In the food and feed industry, the present invention has a wide variety of applications. For example, it can be used for identification and characterization of production organisms such as yeast for production of beer, wine, cheese, yogurt, bread, etc. Another area of use is with regard to quality control and certification of products and processes (e.g., livestock, pasteurization, and meat processing) for contaminants. Other uses include the characterization of plants, bulbs, and seeds for breeding purposes, identification of the presence of plant-specific pathogens, and detection and identification of veterinary infections.

[00159] The present invention is further illustrated by the following non-limiting examples.

EXAMPLES

Example 1 – Genomic DNA Preparation

[00160] Genomic DNA is prepared from the blood of two normal human volunteers, one male and one female, according to standard techniques. Briefly, approximately 12 ml of blood is obtained in EDTA-containing blood collection tubes. Red blood cells are lysed by mixing the blood samples with 4 volumes of lysis buffer (10 mM Tris pH 8.0, 10 mM EDTA). After 10 min on ice with occasional agitation, the suspensions are centrifuged and the supernatants are decanted. The white blood cell pellets are resuspended in 20 ml of lysis buffer, and the above process is repeated. Each cell pellet is then suspended in 15 ml of digestion buffer (50 mM Tris pH 8.0, 5 mM EDTA, 100 mM NaCl, 1% SDS) and 3 mg (0.2 mg/ml) of proteinase K is added. The cells are digested at 37°C for 5 hours. The digests are extracted twice with equal volumes of phenol, then once with equal volumes of a 1:1 phenol:chloroform mixture and finally once with equal

IDT Exhibit 1020.5235

volumes of chloroform, each time centrifuging the mixture and removing the aqueous phase for the next extraction. After the final extraction and removing the aqueous phases, one tenth volume of 3 M sodium acetate, pH 6.5, is added. Two volumes of ice cold 100% EtOH are then added to each solution to precipitate the genomic DNAs, which are spooled out of solution on glass pipettes. The DNA precipitates are washed twice in 0.75 ml volumes of 70% EtOH, briefly centrifuging each time to allow removal of the supernatants. After removing the supernatants for the second time, the remaining EtOH is allowed to evaporate and the DNA is suspended in 0.5 ml of TE (10 mM Tri-HCl pH 8.0 containing 1 mM EDTA) solution. A fifth dilution of each DNA solution is also prepared in TE.

[00161] To determine the concentrations of the one fifth DNA solutions, 1, 2, and 4 .mu.l aliquots of each are loaded on a 1% agarose gel with a known amount of HindIII digested lambda DNA as a control. The gel is run at 150 Volts for 2 hours with ethidium bromide in the electrophoresis buffer. After photographing the gel and comparing the intensities of the DNA bands, the one fifth dilutions are judged to have concentrations of approximately 100 ng/ml. DNA solutions extracted from various tumor cell lines are the generous gifts of other laboratories. The concentrations of these solutions are checked in a similar fashion and solutions of 100 ng/ml in TE are prepared.

[00162] To digest the genomic DNAs with Taq I, 25 µl of the 100 ng/µl solutions is mixed with 5 µl of 10x medium salt buffer (0.5 M NaCl, 0.1 M $MgCl_2$, 0.1 M Tris, pH 8.0), 20 µl of water-ME (i.e. water containing 6 mM ME (i.e., mercaptoethanol)), and 400 U of Taq I restriction endonuclease. The digests are covered with mineral oil and incubated at 65°C for 1 hour. The reactions are stopped by adding 1.2 µl of 500 mM EDTA and heating the specimens to 85°C for 10 min. Complete digestion of the DNA is checked by electrophoresing aliquots on a 1% agarose gel.

Example 2 - Locus Specific Sequencing

[00163] A DNA chip containing over 10 thousand SNPs, serving as capture probes, was mixed with target polypeptide. A first solution probe was added for hybridizing with the target polypeptide. Different first solution probes were used with a difference in the base to be ligated to the SNP. At a given SNP site, the sample DNA should be a perfect match with a given probe. After a stringent wash, a first run of ligation was conducted. The chip was then washed, stained and scanned. After cleavage, through the addition of RNase, followed by phosphorylation, a second round ligation was conducted, followed by additional washing and scanning. A score, referred to as the call rate, was recorded. The call rate refers to the signal recorded of the ligation event occurring from the specific probe ligation events minus the signal from the reference probe ligation events, divided by total signal of both specific and reference probes combined. The call rate for the first round of ligation was 75% and the call rate for the second round of ligation was 45%.

[00164] While preferred embodiments of the present invention have been shown and described herein, it will be obvious to those skilled in the art that such embodiments are provided by way of example only. Numerous variations, changes, and substitutions will now occur to those skilled in the art without departing from the invention. It should be understood that various alternatives to the embodiments of the invention described herein may be employed in practicing the invention. It is intended that the following claims define the scope of the invention and that methods and structures within the scope of these claims and their equivalents be covered thereby.

IDT Exhibit 1020.5236

WHAT IS CLAIMED IS:

1. A method of sequencing comprising:

(a) forming a hybridization complex comprising:

(i) a first set of capture probes;

(ii) a target polynucleotide; and

(iii) a first solution probe,

wherein at least one of the first set of capture probes binds a first region of the target polynucleotide, the first solution probe binds a second region of the target polynucleotide, the first solution probe has a base at the 3' end complementary to a first loci site of the target polynucleotide, therefore when hybridization occurs said at least one of the first set of capture probes and the first solution probe can be ligated, wherein said first solution probe comprises a cleavable first cap;

(b) ligating said at least one of the first set of capture probes and said first solution probe;

(c) adding a second cap to capture probes of said first set of capture probes not ligated to said first solution probe;

(d) removing said first cap; and

(e) sequencing said target polynucleotide.

2. The method of claim 1, further comprising:

(f) forming a hybridization complex comprising a second solution probe that hybridizes to the target polynucleotide, wherein the second solution probe binds to a third region of the target polynucleotide, wherein the second solution probe has a base complementary to a second loci site of the target polynucleotide;

(g) ligating the second solution probe to the first solution probe;

(h) forming a hybridization complex comprising a third solution probe that hybridizes to the target polynucleotide, wherein the third solution probe binds to a fourth region of the target polynucleotide, wherein the third solution probe has a base complementary to a third loci site of the target polynucleotide, wherein the third solution probe comprises a universal primer sequence;

(i) ligating the third solution probe to the second solution probe;

(j) removing the target polynucleotide; and

(k) sequencing with a primer that hybridizes to said universal primer sequence of said third solution probe.

IDT Exhibit 1020.5237

3. The method of any of claims 1-2, wherein sequencing comprises placing the target-probe hybridization complex under nucleic acid synthesis conditions in the presence of free nucleotides to allow extension of the probe that is complementary to the target polynucleotide.

4. The method of any of claims 1-3, further comprising removing non-specific hybridization.

5. The method of any of claims 1-4, wherein said first cap comprises a capping dye and said first solution probe comprises a cleavage site.

6. The method of any of claims 1-5, wherein said second cap comprises a protection group.

7. The method of any of claims 1-6, wherein said first set of capture probes is attached to a flat surface or a bead.

8. The method of claim 7, wherein said first set of capture probes is synthesized or spotted on said flat surface.

9. The method of claim 8, wherein said first set of capture probes is spotted randomly at a controlled density.

10. The method of claim 8, wherein said first set of capture probes is spotted at known locations on the flat surface.

11. The method of any of claims 1-10, wherein said first set of capture probes comprise polynucleotides.

12. The method of claim 11, wherein each capture probe of said first set of capture probes contains 50-150 polynucleotides.

13. The method of any of claims 1-12, wherein said first, second, or third solution probe comprises a polynucleotide.

14. The method of claim 13, wherein said first, second, or third solution probe contains 5 to 50 nucleotides.

15. The method of any of claims 1-14, wherein said target polynucleotide comprises a single nucleotide polymorphism (SNP), an insertion or deletion, or a copy number variation.

16. The method of claim 2, further comprising removal of the target polynucleotide from the hybridization duplex.

17. The method of any of claims 1-16, wherein said first set of capture probes comprises allele-specific capture probes.

18. The method of claim 17, wherein said allele-specific capture probes are known alleles of said first loci.

IDT Exhibit 1020.5238

19. The method of any of claims 1-18, wherein said first cap comprises a nucleotide reversible terminator.

20. The method of claim 19, wherein the nucleotide reversible terminator comprises an amino-2-hydroxypropyl group.

21. The method of any of claims 1-20, wherein at least 50% of a genome is sequenced using no more than 500 million probes.

22. The method of claim 21, wherein 99.5% of the genome is sequenced.

23. The method of any of claims 1-22, wherein the nucleic acid sequence of said first, second or third solution probe is known.

24. The method of any of claims 1-23, wherein the number of total probes used is about 5 million to 500 million.

25. The method of any of claims 1-24, wherein the number of duplicate probes is less than 5.

26. The method of any of claims 1-25, wherein multiple loci in the genome are sequenced.

27. The method of any of claims 1-26, wherein the sequencing avoids SNP with minor allele frequency of greater than 0.05 at the extreme 3' end of said first, second, third or fourth solution probe or first or second capture probe sequences.

28. The method of any of claims 1-27, wherein sequencing said target polynucleotide comprises generating reads that are greater than 5-50 bases.

29. The method of claim 7, wherein the flat surface is a flow cell.

30. The method of any of claims 1-29, wherein the first, second, third or fourth solution probe is labeled.

31. The method of any of claims 1-30, wherein the first, second, third or fourth solution probe has at least one base that is degenerate.

32. The method of any of claims 1-31, further comprising obtaining sequence information of the target polynucleotide or genome via a computer system.

33. The method of claim 32, wherein the computer system comprises an algorithm for detecting, processing, analyzing and/or assembling nucleotide sequence information.

IDT Exhibit 1020.5239

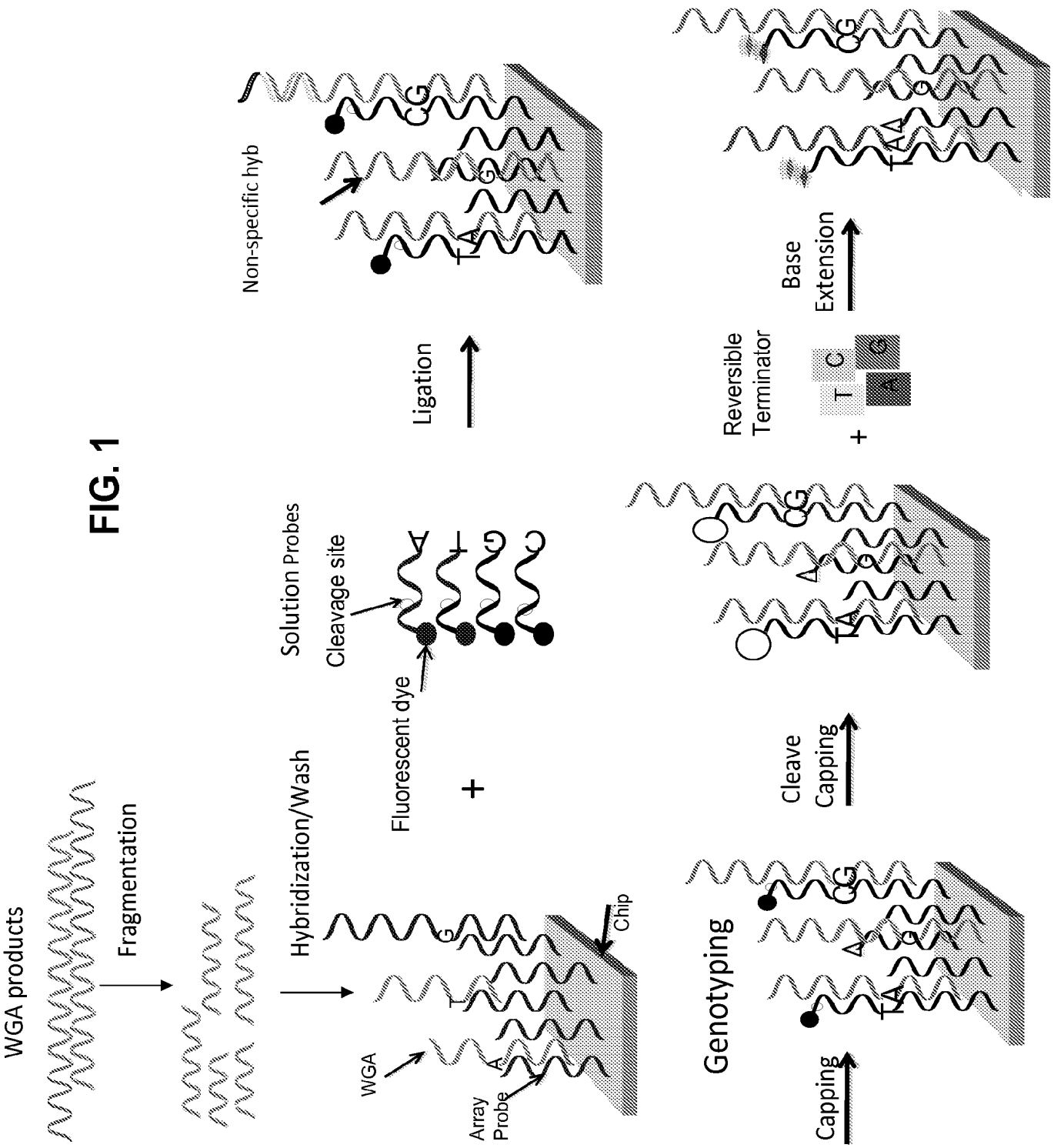

WGA products
Fragmentation
Hybridization/Wash
WGA
Array Probe
Chip
Solution Probes
Cleavage site
Fluorescent dye
+
Ligation
Non-specific hyb
Cleave
Capping
Capping
Genotyping
Base Extension
Reversible Terminator
T C G
A
+

FIG. 1

FIG. 2A

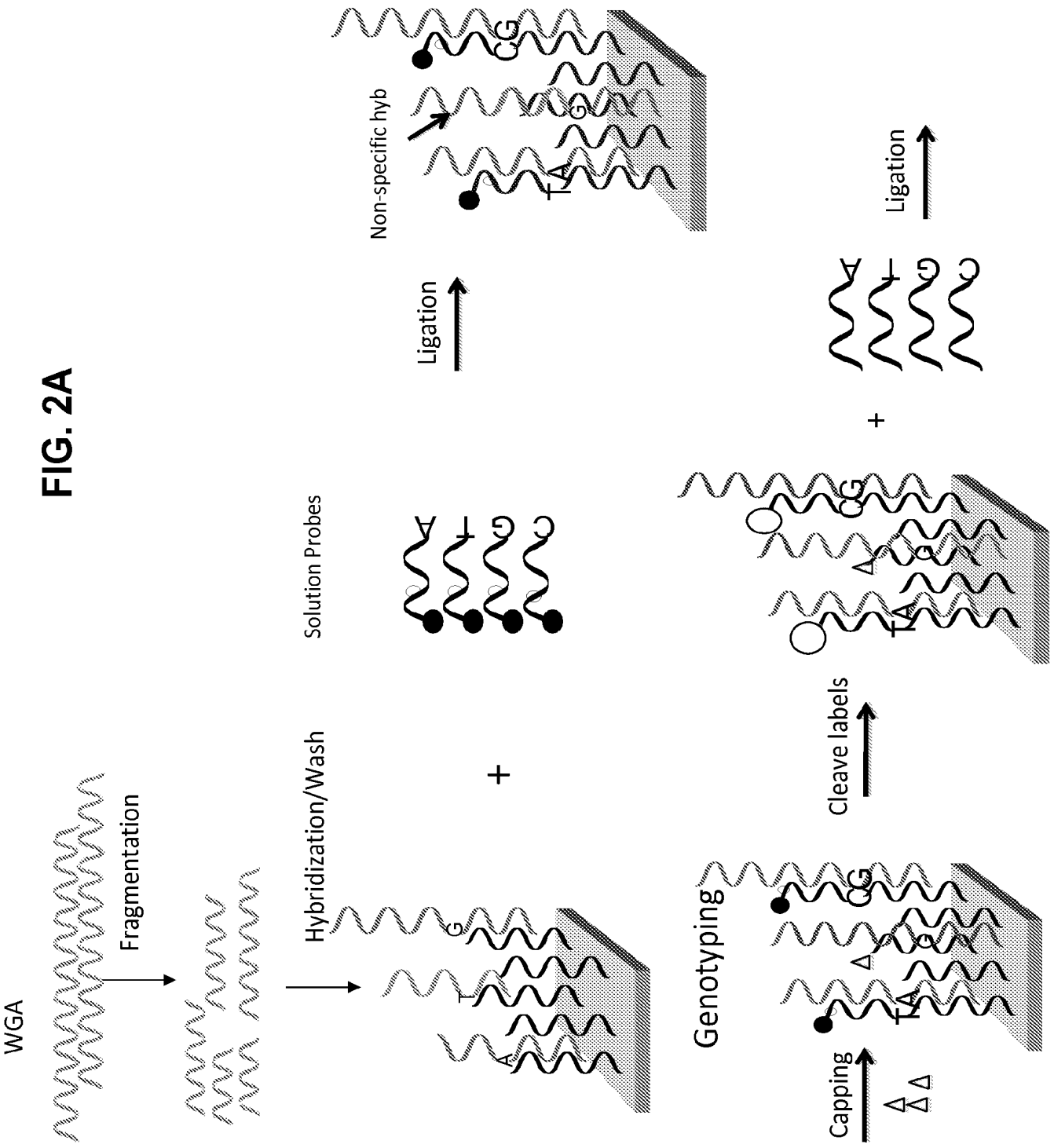

WGA
Fragmentation
Hybridization/Wash
Solution Probes
+
Non-specific hyb
Ligation
Ligation
+
Cleave labels
Genotyping
Capping

FIG. 2B

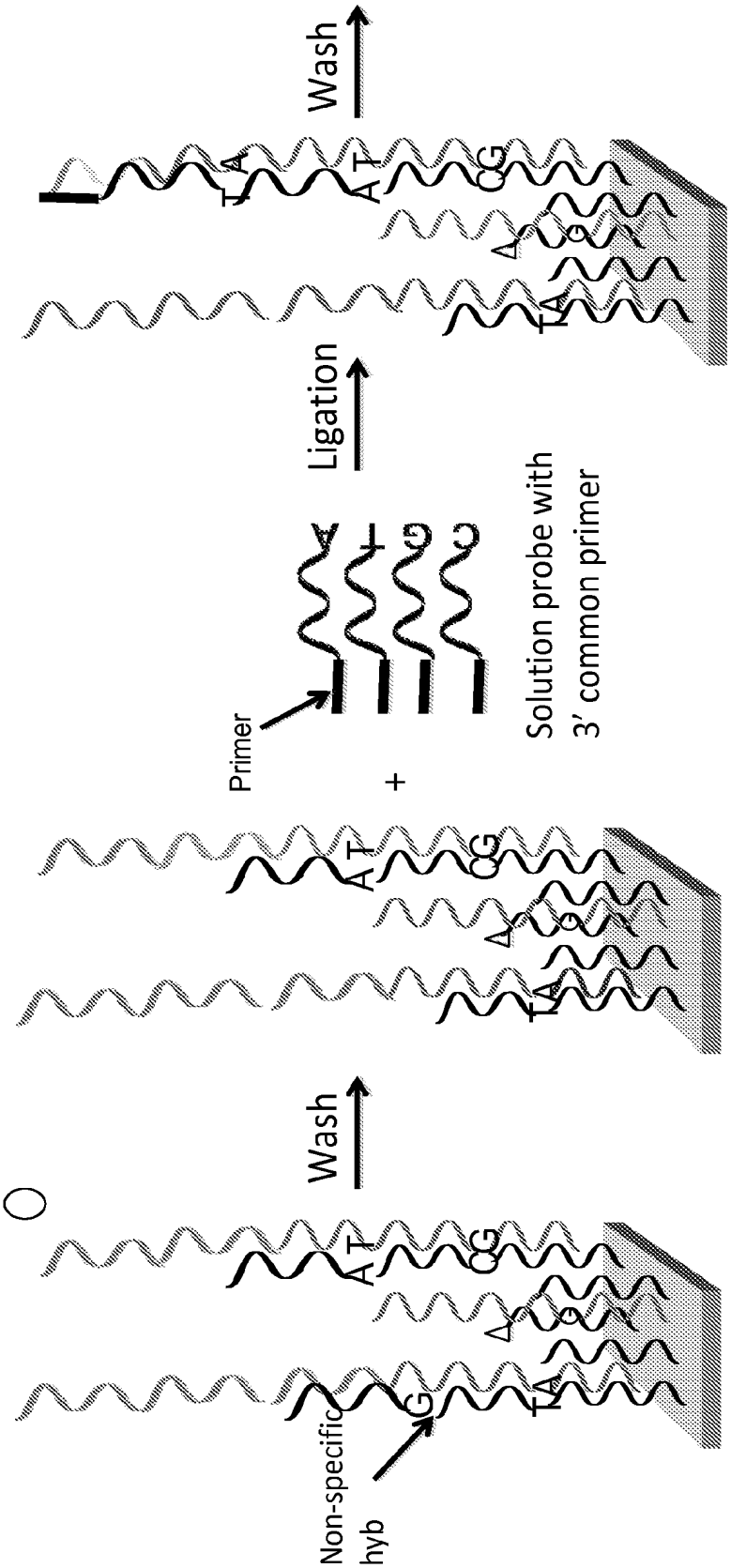

Wash
Ligation
Solution probe with
3' common primer
Primer
+
Wash
Non-specific
hyb

FIG. 2C

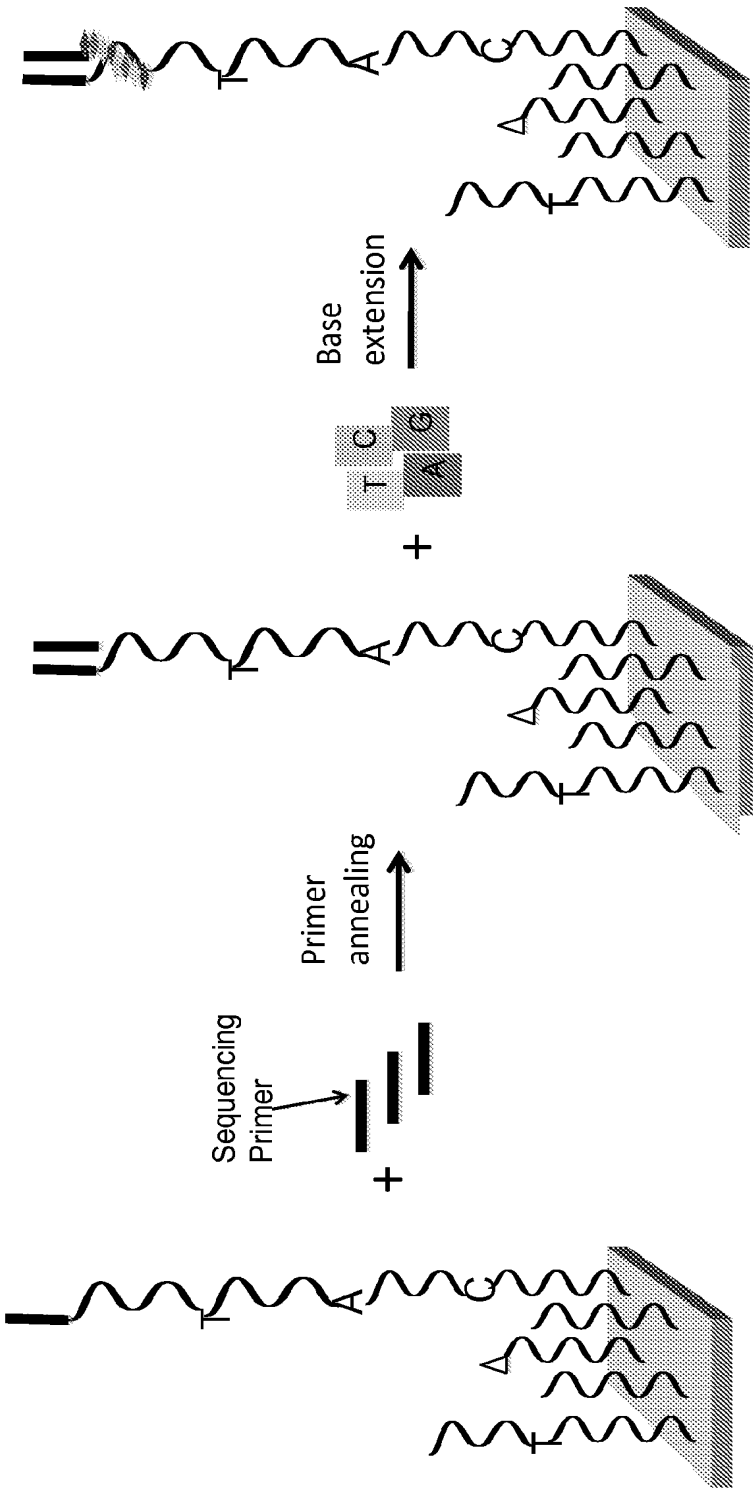


FIG. 3

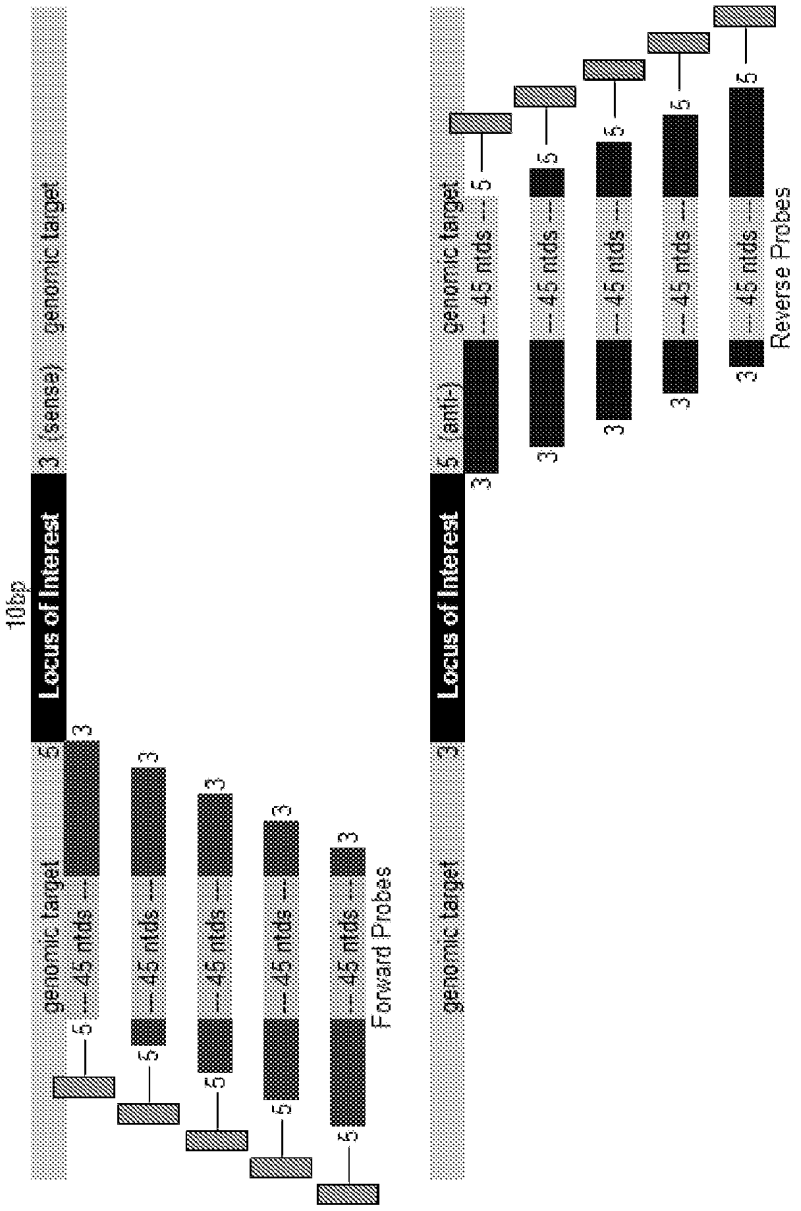

10bp
Locus of Interest
genomic target
(sense)
45 ntds
Forward Probes
Locus of Interest
genomic target
(anti-)
45 ntds
Reverse Probes

FIG. 4

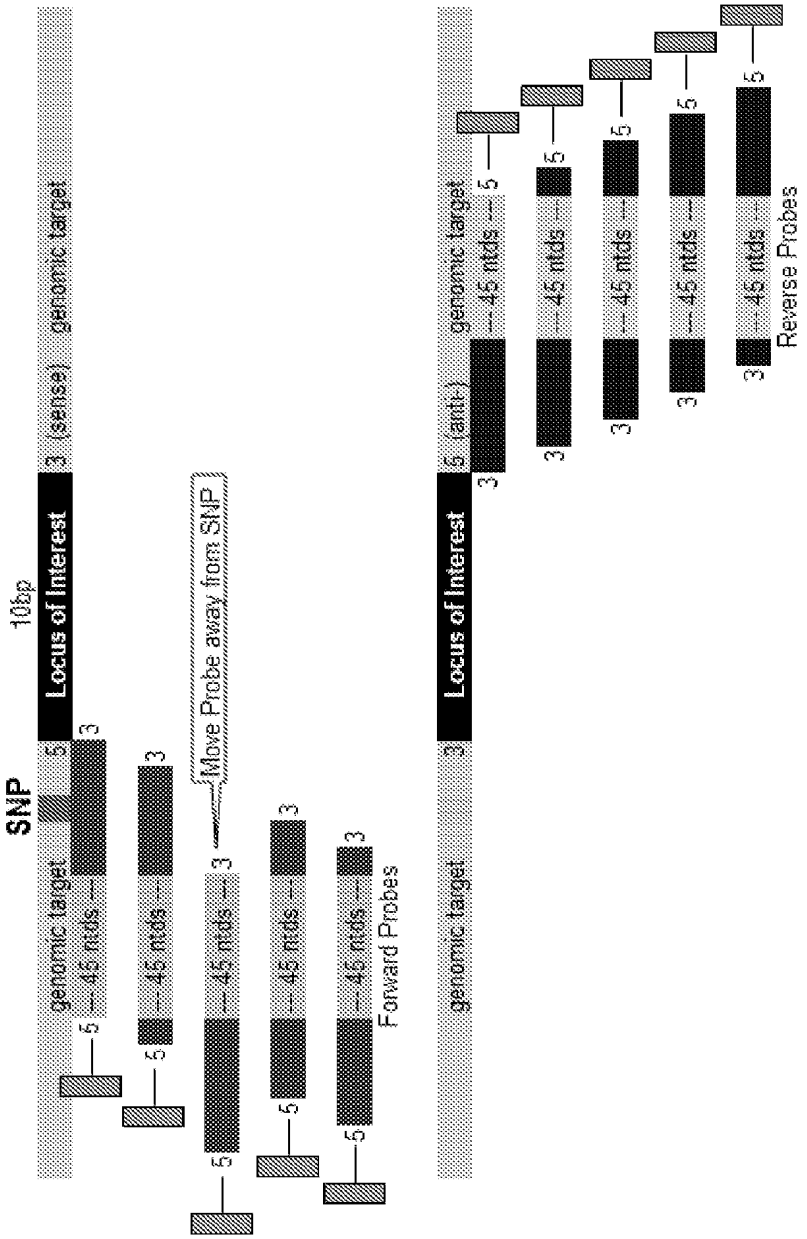

SNP
10bp
Locus of Interest
genomic target
5
3 (sense)
genomic target
45 ntds
Move Probe away from SNP
Forward Probes
Locus of Interest
genomic target
5 (anti-)
genomic target
Reverse Probes

FIG. 5

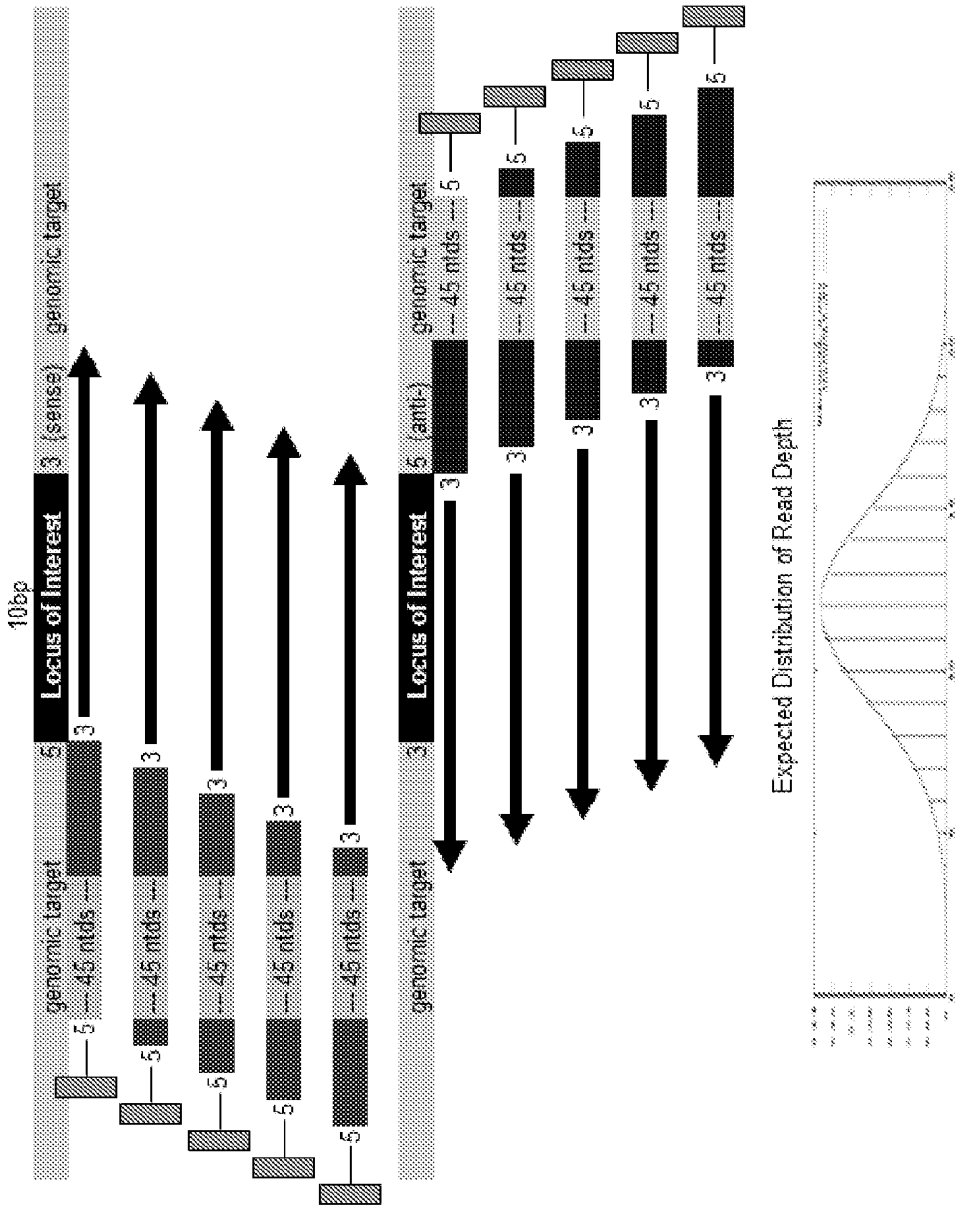

10bp
Locus of Interest
genomic target
(sense)
(anti-)
45 ntds
Expected Distribution of Read Depth

FIG. 6

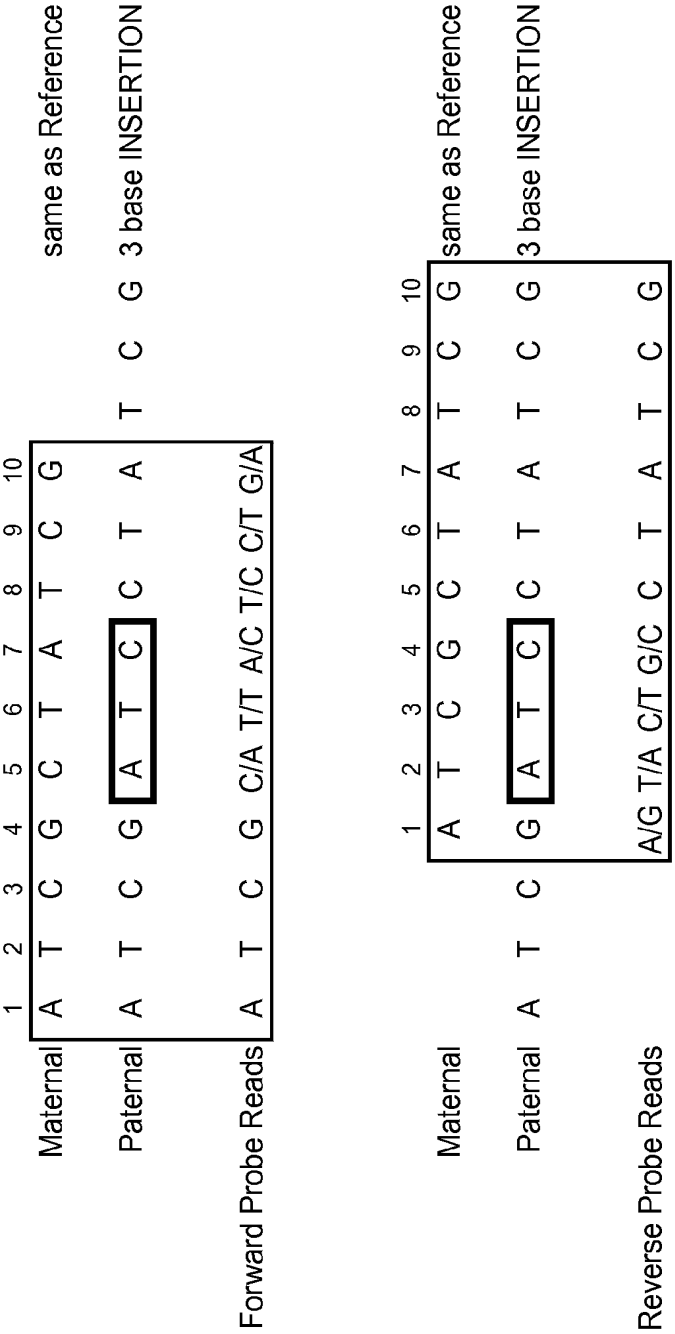

1 2 3 4 5 6 7 8 9 10
Maternal A T C G C T A T C G same as Reference
Paternal A T C G A T C C T A T C G 3 base INSERTION
Forward Probe Reads A T C G C/A T/T A/C T/C C/T G/A
1 2 3 4 5 6 7 8 9 10
Maternal A T C G C T A T C G same as Reference
Paternal A T C G A T C C T A T C G 3 base INSERTION
Reverse Probe Reads A/G T/A C/T G/C C T A T C G

FIG. 7

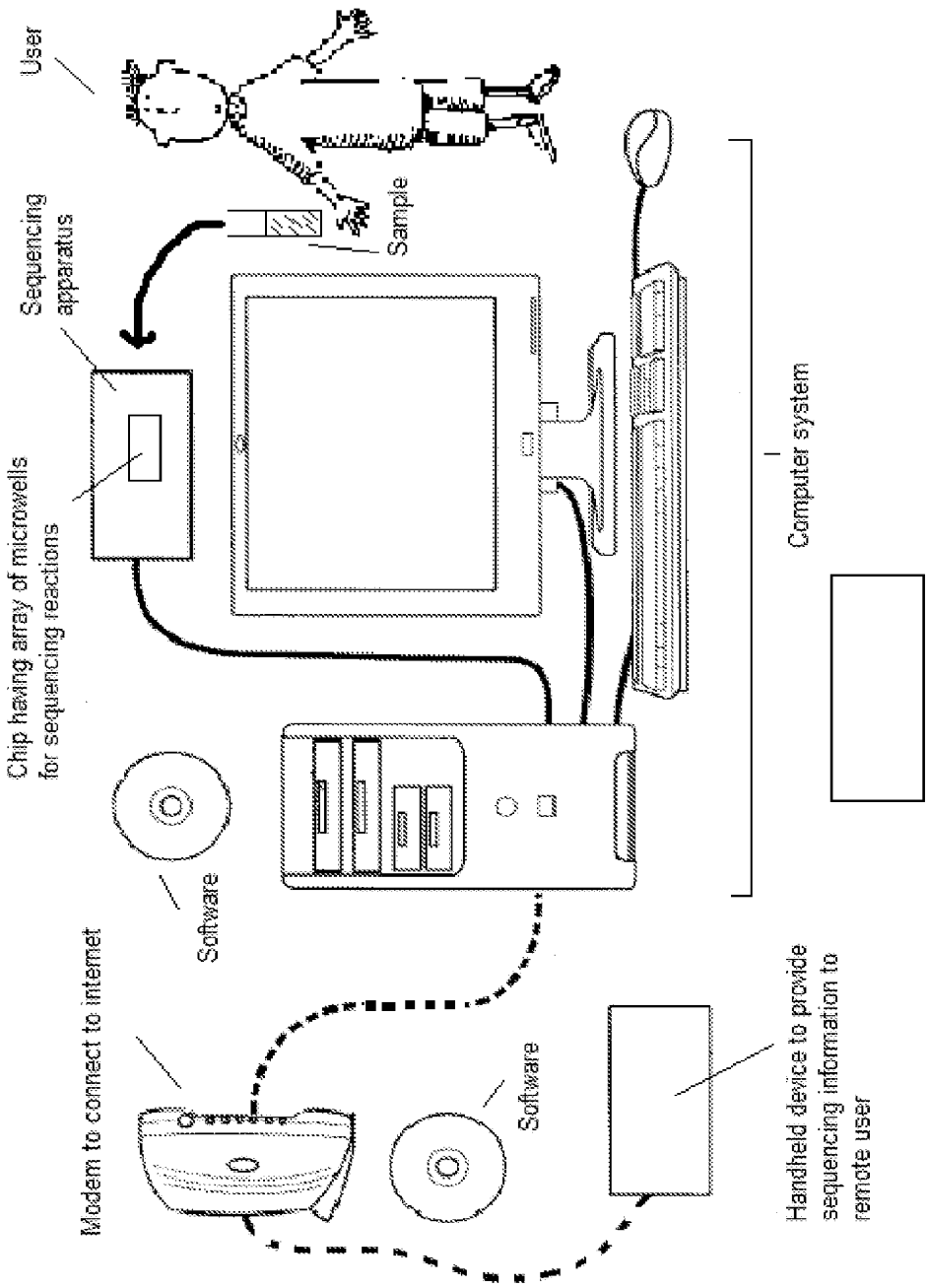
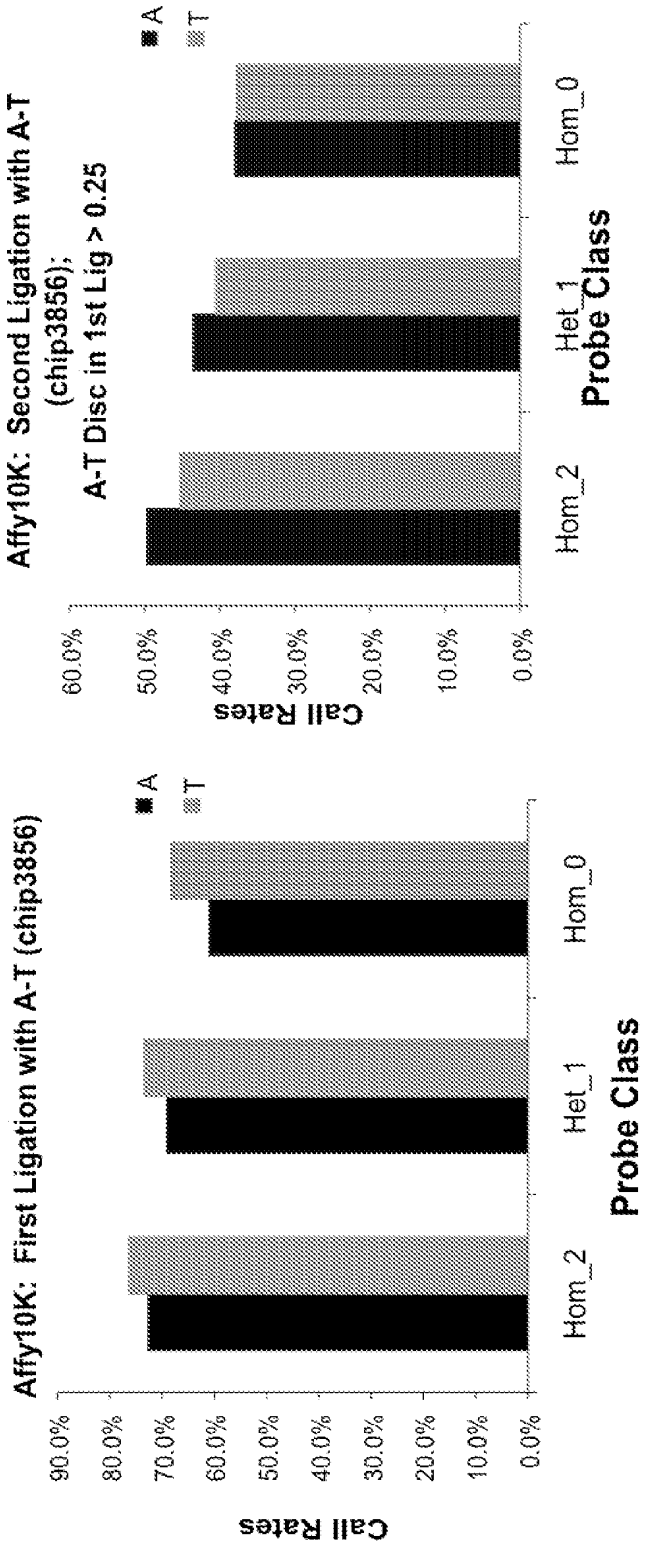

User
Sample
Sequencing apparatus
Chip having array of microwells for sequencing reactions
Computer system
Software
Modem to connect to internet
Software
Handheld device to provide sequencing information to remote user

FIG. 8

Gene	Variation	Alleles	Ref	rs_id	hg18_position	Forward_Probes	Reverse_Probes
CYP2C8	*6	-/A	A	rs3333131	chr10:96839029	AATCTAATGGACATGAACAACCTCAGGACTTTATGATTGTTCCTGATG	TCAACAAATCACAAATTCACAAGCACTCACATAACTAAGCTTTGTTTAC
						ATCAATGGACATGAACAACCTCAGGACTTTATGATTGTTCCTGATGA	CAACAAATCACAAATTCACAAGCACTCACATAACTAAGCTTTGTTTACA
						TCAATGGACATGAACAACCTCAGGACTTTATGATTGTTCCTGATGAA	AACAAATCACAAATTCACAAGCACTCACATAACTAAGCTTTGTTTACAT
						CAATGGACATGAACAACCTCAGGACTTTATGATTGTTCCTGATGAAA	ACAAATCACAAATTCACAAGCACTCACATAACTAAGCTTTGTTTACATT
						AATGGACATGAACAACCTCAGGACTTTATGATTGTTCCTGATGAAAA	CAAATCACAAATTCACAAGCACTCACATAACTAAGCTTTGTTTACATTT
CYP3A4	*6	-/A	-	rs4646436	chr7:99220970-99221971	GCCCTCAGAACACTCTGGTTACTCTTTGTGGACTAGTTTCTTTTAAT	TCACTGAATTTCTCTTTTGCTTCTTTTGCTCTACCGGTGAATTCCTTCAGC
						CCCTCAAAACACTCTGTTACTCTTTGTGGACTCAGTTTCTTTTGATT	CACTCAAATTTCTCTTTTGCTTCTTTTGCTCACCGGTGAATTCCTTCAGCT
						CCTCAGAACACTCTGTTACTCTTTGTGGACTCAGTTTCTTTTGAATTC	AGTCAAATTTCTCTTTTGCTTCTTTTGCTCACCGGTGAATTCCTTCAGCTG
						CTCAGAACACTCTGTTACTCTTTGTGGACTCAGTTTCTTTTGAATTCT	GTGAAATTTCTCTTTTGCTTCTTTTGCTCACCGGTGAATTCCTTCAGCTGA
						TCAGAACACTCTGTTACTCTTTGTGGACTCAGTTTCTTTTGAATTCTG	TGAAATTTCTCTTTTGCTTCTTTTGCTCACCGGTGAATTCCTTCAGCTGAT
CYP2A6	*9	G/T	T	rs28399433	chr19:46043219	TGAGGCCAGCATGGTGATAGTGATGAGATGAATGGTGACCCTGGGGTGG	GAGTGAATGAGGTAATTATGTAATCAGCCAAAGTCCATCCCTCTTTTT
						GAGGCCAGCATGGTGGTAGTGGTGAGATGAATGGTGACCCTGGGGTGGT	AGTCAAATGAGGTAATTATGTAATCAGCCAAAGTCCATCCCTCTTTTTC
						AGGCCAGCATGGTGTAGTGGTGAGATGAATGGTGACCCTGGGGTGGTT	GGTCAAATGAGGTAATTATGTAATCAGCCAAAGTCCATCCCTCTTTTTCA
						GGCCAGCATGGTGTAGTGGTGAGATGAATGGTGACCCTGGGGTGGTTT	GTCAAATGAGGTAATTATGTAATCAGCCAAAGTCCATCCCTCTTTTTCAG
						GCCAGCATGGTGGTAGTGGTGATGAGATGAATGGTGACCCTGGGGTGGTTG	TCAAATGAGGTAATTATGTAATCAGCCAAAGTCCATCCCTCTTTTTCAGG
4SCB1	T1236delC	C/T	T	rs1135503	chr7:87017537	CCACAGCCACTGTTTCCAACAGGGCCACCGTCTGCCCATCTGCACCTT	CTGTGTCTGTGAATTGCCTTGAAGTTTTTTTCTCACTCGTCCTGGTAGAT
						CACAGCCACTGTTTCCAACAGGGCCACCGTCTGCCCACTCTGCACCTTC	TGTGTCTGTGAATTGCCTTGAAGTTTTTTTCTCACTCGTCCTGGTAGATC
						ACAGCCACTGTTTCCAACAGGGCCACCGTCTGCCCACTCTGCACCTTCA	GTGTCTGTGAATTGCCTTGAAGTTTTTTTCTCACTCGTCCTGGTAGATCT
						CAGCCACTGTTTCCAACAGGGCCACCGTCTGCCCACTCTGCACCTTCAG	TGTCTGTGAATTGCCTTGAAGTTTTTTTCTCACTCGTCCTGGTAGATCTT
						AGCCACTGTTTCCAACAGGGCCACCGTCTGCCCACTCTGCACCTTCAGG	GTCTGTGAATTGCCTTGAAGTTTTTTTCTCACTCGTCCTGGTAGATCTTG
CYP2A6	*4	Del/Het	Del	chr19:46042070-46042089 (Fukami et al. 2008; 246n667)	AGGTGTAGCAATGGGAAGGTTTAGATGTCTTTCCTGCCCAGGTACACTTGC	AGGAGATGACGGGACACAGCAGTCATATTTGCAAGTGTACCTGGCAGGAAAGC	
						GGTGTAGCAATGGGAGGTTTAGATGTCTTTCCTGCCAGGTACACTTGCA	GGAGATGACGGGACACAGCAGTCATATTTGCAAGTGTACCTGGCAGGAAAGCA
						GTGTAGCAATGGGAGGTTTAGATGTCTTTCCTGCCAGGTACACTTGCAA	GAGATGACGGGACACAGCAGTCATATTTGCAAGTGTACCTGGCAGGAAAGCAC
						TGTAGCAATGGGAGGTTTAGATGTCTTTCCTGCCAGGTACACTTGCAAA	AGATGACGGGACACAGCAGTCATATTTGCAAGTGTACCTGGCAGGAAAGCACA
						GTAGCAATGGGAGGTTTAGATGTCTTTCCTGCCAGGTACACTTGCAAAT	GATGACGGGACACAGCAGTCATATTTGCAAGTGTACCTGGCAGGAAAGCACA

FIG. 9


Affy10K: First Ligation with A-T (chip3856)
Call Rates
90.0%
80.0%
70.0%
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
A
T
Hom_2
Het_1
Hom_0
Probe Class
Affy10K: Second Ligation with A-T (chip3856); A-T Disc in 1st Lig > 0.25
Call Rates
60.0%
50.0%
40.0%
30.0%
20.0%
10.0%
0.0%
A
T
Hom_2
Het_1
Hom_0
Probe Class

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/48526

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12Q 1/68 (2010.01)
USPC - 435/6
According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
USPC: 435/6

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
USPC: 435/91.2,283.1,287.1,288.4,287.2, 6 (text search)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
Electronic data bases searched: PubWEST (PGPB, UPST, EPAB, JPAB); Google Scholar
Search terms: Targeted sequencing, oligonucleotide probe, oligonucleotide ligation assay, ligation assisted sequencing, ligation captured sequencing, microarray captured template, capture probe, genotyping, SNP, reversible nucleotide terminator

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	US 4,988,617 A (LANDEGREN et al.) 29 January 1991 (29.01.1991), Especially col 3 ln 1-17, co. 3 ln 51-52.	1-3,16
Y	TURCATTI et al. A new class of cleavable fluorescent nucleotides: synthesis and optimization as reversible terminators for DNA sequencing by synthesis. Nucl Acid Res 7 February 2008 Vol. 36, No. 4 e25 Pages 1-13, Especially pg 7 fig 3.	1-3,16
Y	US 6,025,139 A (YAGER et al.) 15 February 2000 (15.02.2000). Especially col 2 ln 54-63, col 4 ln 33-37	2,16
A	HACIA et al. Resequencing and mutational analysis using oligonucleotide microarrays. Nature Genetics January 1999 Vol 21 Suppl 1 Pages 42-47, Especially pg 42 fig 1, pg 44 fig 3.	1-3,16

☐ Further documents are listed in the continuation of Box C. ☐

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of another citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents, such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
3 November 2010 (03.11.2010)	22 NOV 2010
Name and mailing address of the ISA/US Mail Stop PCT, Attn: ISA/US, Commissioner for Patents P.O. Box 1450, Alexandria, Virginia 22313-1450 Facsimile No. 571-273-3201	Authorized officer: Lee W. Young PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US 10/48526

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☐ Claims Nos.:
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:

3. ☒ Claims Nos.: 4-15 and 17-33
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required additional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying additional fees, this Authority did not invite payment of additional fees.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.

☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.

☐ No protest accompanied the payment of additional search fees.

IDT Exhibit 1020.5252

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
19 September 2013 (19.09.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/138510 A1

(51) International Patent Classification:
C12Q 1/68 (2006.01) *C12N 15/11* (2006.01)

(21) International Application Number:
PCT/US2013/031014

(22) International Filing Date:
13 March 2013 (13.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/609,985 13 March 2012 (13.03.2012) US

(72) Inventor; and
(71) Applicant : **PATEL, Abhijit Ajit** [CA/US]; 27 Allison Drive, Madison, Connecticut 06443 (US).

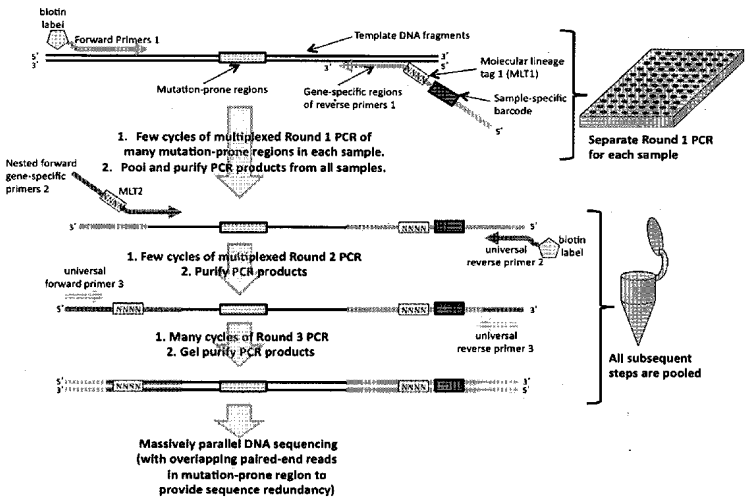
(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:
— *with international search report (Art. 21(3))*
— *with sequence listing part of description (Rule 5.2(a))*

(54) **Title**: MEASUREMENT OF NUCLEIC ACID VARIANTS USING HIGHLY-MULTIPLEXED ERROR-SUPPRESSED DEEP SEQUENCING

Figure 1



(57) **Abstract**: Methods and compositions are disclosed for measuring low-abundance DNA variants from a complex mixture of DNA molecules. Embodiments of the methods allow for extremely sensitive detection and can distinguish true variants from sequencer misreads and PCR misincorporations.

WO 2013/138510 A1

MEASUREMENT OF NUCLEIC ACID VARIANTS USING HIGHLY-MULTIPLEXED ERROR-SUPPRESSED DEEP SEQUENCING

Governmental Interests

The research leading to this application was funded by the National Institutes of Health from grant RR014139. The government may have certain rights in this invention.

Background

Tumor-derived DNA is released into the bloodstream from dying cancer cells in patients with various types of malignancies. Such circulating tumor DNA (ctDNA) is showing excellent promise as a non-invasive cancer biomarker. However, an assay that is capable of exploiting ctDNA for early cancer detection presents several challenges. In the bloodstream, ctDNA can be distinguished from normal background DNA based on the presence of tumor-specific mutations. However, mutant ctDNA is usually only present in small amounts, having been previously reported to comprise an average of 0.2% of total plasma DNA (Diehl et al., *Nat Med*. 2008; 14: 985-990). If variant DNA sequences are low in abundance, detecting and quantifying these variants can be more challenging. Small amounts of mutation-harboring ctDNA can be obscured by a relative excess of background wild-type plasma DNA. Thus, an assay with extremely high detection sensitivity is required.

There is a need for a method that is able to detect and quantify rare variant sequences to detect cancers in situations where the amount of DNA in a given sample is limited. Unlike existing approaches, a test should be able to evaluate an entire panel of mutation-prone regions without needing to divide DNA samples into separate reactions (which could reduce detection sensitivity by providing fewer template DNA copies per reaction). Methods and compositions are described herein that provide a multiplex assay to detect minute amounts of ctDNA and address the current deficiencies to assay ctDNA.

Summary

Described herein are compositions and methods relating to next-generation sequencing and medical diagnostics. Methods include identifying and quantifying nucleic acid variants, particularly those available in low abundance or those obscured by an abundance of wild-type sequences. Also described herein are methods related to identifying and quantifying specific sequences from a plurality of sequences amid a plurality of samples. Methods as described herein also include detecting and distinguishing true nucleic acid variants from misincorporation errors,

sequencer errors, and sample misclassification errors. Methods include early attachment of barcodes and molecular lineage tags (MLTs) to targeted nucleic acids within a sample. Methods also include clonal overlapping paired-end sequencing to achieve sequence redundancy.

In an embodiment, a method includes measuring nucleic acid variants by tagging and amplifying low abundance template nucleic acids in a multiplexed primer extension or PCR. Low abundance template nucleic acids may be fetal DNA, circulating tumor DNA (ctDNA), viral RNA, viral DNA, DNA from a rejected transplanted organ, or bacterial DNA. A multiplex PCR may include gene specific primers, wherein primers are specific for a mutation prone region (e.g., within *KRAS*, *EGFR*, etc.). In an embodiment, a mutation prone region may be associated with cancer. As disclosed herein, a multiplex PCR can include more than one round of PCR and/or primer-extension. In an embodiment, a multiplex PCR can include two or three rounds of PCR.

In an embodiment, primers comprise a barcode and/or a molecular lineage tag (MLT). In an embodiment, a MLT can be 2-10 nucleotides. In another embodiment, a MLT can be 6, 7, or 8 nucleotides. In an embodiment, a barcode can identify the sample of template nucleic acid. In an embodiment, a PCR reaction mixture includes template nucleic acids from multiple samples (e.g., patients), wherein the barcode identifies the sample origin of the template nucleic acid. In an embodiment, a primer extension reaction employs targeted early barcoding. In targeted early barcoding, a plurality of different primers specific for different nucleic acid regions all have an identical barcode. An identical barcode identifies the nucleic acids from a particular sample. In an embodiment, primers used for targeted early barcoding are produced by combining a unique barcode-containing oligonucleotide segment with a uniform mixture of gene-specific primer segments in a modular fashion.

In an embodiment, multiplex assays described herein can be used for clinical purposes. In an embodiment, nucleic acid variants within blood can be identified and measured before and after treatment. In an example of cancer, a nucleic acid variant (e.g., cancer-related mutation) can be identified and/or measured prior to treatment (e.g., chemotherapy, radiation therapy, surgery, biologic therapy, combinations thereof). Then after treatment, the same nucleic acid variant can be identified or measured. After treatment, a decrease or absence of the nucleic acid variant can indicate that the therapy was successful.

Brief Description of the Drawings:

Fig. 1 is a schematic of a copying, tagging, amplification, and sequencing process. Two rounds of limited-cycle PCR were performed to attach barcode sequences and molecular lineage tag

sequences to copies of targeted template DNA fragments. Stringent purification of the PCR products was carried out between rounds in order to remove unextended primers, spurious extension products, and template DNA. A final round of PCR was performed using universal primers to further amplify the purified products from Round 2. Final amplification products were gel-purified and subjected to clonal overlapping paired-end massively parallel sequencing. Use of primers synthesized from modular segments allowed early barcoding of targeted template DNA during the first round of PCR, enabling subsequent steps to be performed in a combined reaction volume.

Fig. 2 illustrates a general approach to combining modular oligonucleotide segments to produce mixtures of gene-specific barcoded primers. Primers produced by combining modular segments allow primer-extension and early barcoding of multiple targeted gene regions in a given sample. A primer mix used for a particular sample may have a unique barcode and multiple gene-specific primer sequences. For a different sample, the mix can have the same set of gene-specific sequences in identical ratios, but a different barcode.

Fig. 3 illustrates a method for producing combinations of modular oligonucleotide segments using an automated oligonucleotide synthesizer. First, gene-specific 3'-segments of oligonucleotides were synthesized on solid supports on separate synthesis columns. The oligonucleotides were synthesized in a 3' to 5' direction. The synthesis was then paused, and the partially-synthesized oligonucleotides were left in a protected state on solid support particles. The contents of all columns were evenly mixed, and the mixture of solid support particles was then dispensed into separate fresh columns. Synthesis of the barcode-containing 5'-segment of the oligonucleotides was then continued in the new columns. A uniquely barcoded 5'-segment was added in each column. After cleavage, deprotection and purification, the resulting barcoded oligonucleotide mixtures all had identical ratios of 3'-segments.

Figs. 4A and Fig. 4B are schematics of error-suppressed multiplexed deep sequencing. Fig. 4A shows cell-free DNA purified from plasma undergoing two rounds of amplification by PCR. The first round amplifies mutation hotspot regions of several genes from a given sample in a single tube. The second round separately amplifies each hotspot region using nested primers incorporating unique combinations of barcodes to label distinct samples. The barcoded PCR products are then pooled and subjected to deep sequencing. Millions of sequences are sorted and counted to determine the ratio of mutant to wild-type molecules derived from each sample. The total number of plasma DNA fragments is measured by real-time PCR and can be used to calculate the absolute concentration of mutant ctDNA. Fig. 4B shows sequence redundancy in mutation hotspot regions is produced by partial overlap of paired-end reads from the forward and reverse strands of each clone.

This yields highly accurate base-calls, permitting detection and quantitation of rare mutations with greater sensitivity.

Figs. 5A-C are graphs depicting suppression of spurious mutation counts to reveal low-abundance variants. Each bar indicates the frequency of a particular deviation from the wild-type sequence occurring within the codon 12/13 hotspot region of KRAS. The tested sample contains 0.2% DNA derived from a lung cancer cell line that is known to be homozygous for a KRAS Gly12Ser mutation. Fig. 5A shows filtered reads from one end of the amplicon have relatively frequent mismatches when directly compared to the wild-type sequence. Data from 3 replicate amplifications are shown. Fig. 5B indicates sequencer errors are greatly reduced by requiring both partially overlapping paired-end reads from each clone to exactly match a specific mutation. The Gly12Ser mutation is now readily distinguished from the remaining low-level errors that were likely introduced during DNA amplification and processing. Insertions and deletions are no longer seen in this region after requiring agreement of overlapped reads. Fig. 5C shows a further reduction in the relative error level can be achieved by calculating the mean values of 3 replicate measurements, since mutations found in the original DNA sample should produce more consistent counts than randomly occurring errors.

Figs. 6A-C indicates the performance of error-suppressed deep sequencing . Measurements of DNA extracted from mutant and wild-type cancer cell lines mixed in various ratios ranging from 1:10,000 to 10,000:1 show a high degree of accuracy and reproducibility. Fig. 6A is a linear plot of DNA from the KRAS-mutant cell line over the range of concentrations tested. Fig. 6B and Fig. 6C show that BRAF- and EGFR-mutant lines, respectively, contained a small amount of wild-type DNA, thereby yielding a plateau at higher mutant to wild-type ratios. Non-linear least-squares fits were performed using the equation y = 10^(slope*log((1-C)*x/(C*x+1))+intercept) where C was the fraction of wild-type molecules found in DNA extracted from mutant cell lines. Error bars indicate the standard deviation of 3 measurements.

Figs. 7A-C show the changes in ctDNA levels with treatment or disease progression. Measurements of mutant ctDNA from patients with NSCLC are shown at various times in relation to therapeutic interventions and disease status. ctDNA was considered undetectable if sequence counts yielded a quantity of less than one mutant molecule per sample. Median genome equivalents per sample, as determined by real-time PCR were 9602 (IQR = 5412-11513) Fig. 7A shows the timeline of treatment for Patient 3 who had stage IV lung adenocarcinoma with a 4.3 cm right upper lobe tumor and large metastases in the abdomen and supraclavicular region. She was treated concurrently with an experimental histone deacetylase (HDAC) inhibitor and palliative radiation therapy directed at her painful 6.9 cm supraclavicular lesion. She began chemotherapy

treatment shortly afterwards. Fig. 7B shows the timeline of treatment for Patient 5 who had a 7.5 cm lung adenocarcinoma with eight small brain metastases ranging from 3 mm to 15 mm in size at presentation. He was treated with palliative whole-brain radiation therapy, followed by long-term weekly chemotherapy. Follow-up imaging revealed an excellent, durable response with shrinkage of the lung tumor to ~15% of its original volume at 7 months after diagnosis. No evidence of disease progression was seen during this time period. Fig. 7C shows the timeline of treatment for Patient 9 who underwent definitive radiation treatment for locally advanced, stage IIIB undifferentiated NSCLC. Other health conditions prevented him from undergoing surgery or concurrent chemotherapy. Blood sample collection commenced upon completion of his treatment. Although his disease was confined to the thorax prior to initiating radiation therapy, a PET scan performed 8 weeks after treatment showed marked progression of disease with multiple osseous, hepatic, and subcutaneous metastases. He expired 10 weeks after completing treatment.

Figs. 8A-C show the ctDNA levels in patients with fewer than 3 time-points. Fig. 8A shows the timeline of treatment for Patient 11 who had stage IV lung adenocarcinoma with widespread metastatic disease in the bones of her spine, ribs, sternum, clavicle, humerus, and pelvis. She was treated with a short course of palliative radiation therapy for a pathologic fracture in her lumbar spine. A single blood sample was obtained on her last day of treatment. She passed away approximately 1 week after completion of therapy. Fig. 8B shows the timeline of treatment for Patient 14 who had stage IV lung adenocarcinoma. He received palliative radiation therapy to a painful 8.9 x 5.9 x 4.9 cm lesion in his left posterior chest wall, given concurrently with an experimental histone deacetylase (HDAC) inhibitor. He had additional metastatic lesions in his liver, kidneys, and peri-splenic region. He was hospitalized for profound weakness 10 days post-treatment, and expired shortly afterwards. Fig. 8C shows the timeline of treatment for Patient 15 who had stage IV undifferentiated NSCLC with metastasis in the supraclavicular and inguinal regions, as well as several small tumors in his brain. His brain lesions were treated with single-fraction stereotactic radiosurgery. He then began palliative radiation therapy for a painful 7.1 cm left upper lobe lung mass, which was threatening to obstruct his left mainstem bronchus. He received concurrent treatment with a HDAC inhibitor as part of a clinical trial. He passed away unexpectedly after receiving 8 of 10 planned radiation treatments.

Fig. 9 shows a scheme for appending modular barcodes to gene-specific primers. An ability to combine barcodes and gene-specific primers in a modular fashion provides flexibility to modify or expand a panel of genes or number of samples being tested. A gene-specific primer was added to the 3'-end of a barcoded oligonucleotide by polymerization on a biotinylated template. A biotin tag was used to capture a double-stranded product onto streptavidin resin. A barcoded

primer was then released into solution by heat-denaturation. In a similar manner, a mixture of biotinylated templates can be used to produce a mixture of gene specific primers, all having the same barcode. Separate reactions use different barcoded oligonucleotides to produce uniquely barcoded primer mixes that can be used for targeted early barcoding.

Fig. 10 is a schematic of a process described in Example 2. First, a primer extension step was carried out using primers that assign sample-specific barcodes and MLT sequences to copied template DNA. For a given sample, multiple targeted sequences were copied using multiple gene-specific primers, all bearing the same sample-specific barcode. After stringent purification of specifically extended products, a pre-amplification step was performed in order to produce many copies of the tagged molecules. This allowed splitting of the products into different tubes for separate amplification of each target site, while ensuring that copies of the original templates are adequately sampled. The products of the final PCRs were combined and subjected to clonal overlapping paired-end deep sequencing. Nested primers were used to enhance target specificity at each step.

Fig. 11 shows a workflow of a process described in Example 2. Separate primer-extension reactions were initially carried out for each sample. Barcoded products were then be mixed into a single volume for purification and pre-amplification steps. Purified products were then split into separate tubes and underwent final single-target PCR in separate reaction volumes.

Fig. 12 shows an example of a Round 1 reverse primer sequence, highlighting various elements of the sequence. Note that the gene-specific sequence at the 3'-end can act as a primer for either PCR or primer-extension by a DNA polymerase. The 5'-segment contains a sample-specific barcode sequence, a Molecular Lineage Tag (MLT), as well as adapter sequences required by the next-generation sequencing platform. In this example, a gene-specific segment is specific for a mutation prone-region of TP53. For Round 1 PCR or primer-extension of a given sample, a mixture of several reverse primers would be used, all having the same sample-specific barcode sequence, and multiple different gene-specific sequences. A similar mixture, but with another barcode, would be used for Round 1 PCR or primer-extension of a different sample.

Fig. 13 is a schematic of a process of splint-mediated ligation of modular oligonucleotide segments. A 5'-segment containing a particular barcode sequence can be ligated to a mixture of 3'-segments having a variety of gene-specific primer sequences using a biotinylated splint oligonucleotide. Hybridization to the splint oligonucleotide is mediated by common annealing sequences. A 5'-phosphate is necessary on the 3'-segments to permit enzymatic ligation. The biotinylated splint can be used to capture and wash and elute the ligated products.

Fig. 14 shows elements of a sequence output using the Illumina® platform. Read 1 and read 3 are from opposite strands, and provide sequence redundancy via overlap in the mutation-prone region. This clonal redundancy allowed sequences resulting from sequencer errors to be identified and discarded, permitting greater sensitivity for detection of rare sequence variants.

Fig. 15 shows hypothetical processing of data from sequences assigned to a single gene and a single barcode. The example illustrates how analysis of variant sequences and associated molecular lineage tags can be performed.

Fig. 16 shows processed data from sequences generated using methods described in Example 2. Data are shown for a single gene and a single barcode. Symbols used in the mismatch table are as defined in Fig. 15. MLT counts associated with variant sequences are displayed in the format "N x Z" where N is the number of copies of a particular MLT sequence, and Z is the number of different MLT sequences having N copies.

Fig. 17 shows an ethidium bromide-stained 2% agarose gel containing products of Round 3 PCR. A marker lane contained a 100 base-pair ladder for size comparison. The gel shows a diffuse band containing amplified products within the expected size range, and very little spurious product migrating at a different size.

Fig. 18 shows processed data from sequences generated from the methods described in Example 3. Data are shown for a single gene and a single barcode. Results are displayed in a format similar to Fig. 15, but in this case analysis of two separate MLT sequence regions (MLT-1 and MLT-2) was performed for variant and wild-type sequences. To report these counts in a succinct format, MLT counts are binned by powers of two. For example, an MLT-1 count of 13 would be placed into bin 4 (because 2^4 is the smallest power of 2 that is greater than or equal to 13). Thus, a report of 4x5 means that there were five instances of counts in the range of 9 to 16. Similarly, a report of 3x6 means that there were six instances of counts in the range of 5 to 8. For a given collection of MLT-1 counts, the associated MLT-2 counts were reported in a similar format, to the right of the MLT-1 counts and separated by colons. For example, 4x5:2x3:1x7 meant that among 5 sets of MLT-1 sequences occurring between 9 and 16 times, there were 3 instances of MLT-2 sequences that occurred between 3 and 4 times, and 7 instances of MLT-2 sequences that occurred twice. Different MLT-1 bins were separated by a space.

Detailed Description

Definitions

The terms "nucleic acid," "nucleotide," "polynucleotide," and "oligonucleotide" are used interchangeably. They refer to a polymeric form of nucleotides of any length, either

deoxyribonucleotides or ribonucleotides, or analogs thereof. Polynucleotides may have any three-dimensional structure, and may perform any function, known or unknown. The following are non-limiting examples of polynucleotides: coding or non-coding regions of a gene or gene fragment, loci (locus) defined from linkage analysis, exons, introns, messenger RNA (mRNA), transfer RNA, ribosomal RNA, ribozymes, cDNA, recombinant polynucleotides, branched polynucleotides, plasmids, vectors, isolated DNA of any sequence, isolated RNA of any sequence, nucleic acid probes, and primers. A polynucleotide may comprise modified nucleotides, such as methylated nucleotides and nucleotide analogs. If present, modifications to the nucleotide structure may be imparted before or after assembly of the polymer. The sequence of nucleotides may be interrupted by non-nucleotide components. A polynucleotide may be further modified after polymerization, such as by conjugation with a labeling component.

The term "base", in its singular form, refers to a single residue within a nucleic acid molecule or to a single position within a nucleic acid sequence read.

The term "biological sample" refers to a body sample from any animal, but preferably is from a mammal, more preferably from a human. Such samples include biological fluids such as serum, plasma, vitreous fluid, lymph fluid, synovial fluid, follicular fluid, seminal fluid, amniotic fluid, milk, whole blood, urine, cerebro-spinal fluid, saliva, sputum, tears, perspiration, mucus, and tissue culture medium, as well as tissue extracts such as homogenized tissue, and cellular extracts.

As used herein, "buffer" refers to a buffered solution that resists changes in pH by the action of its acid-base conjugate components. Buffers may optionally comprise a salt such as $MgCl_2$, $MnCl_2$, or the like. Buffers may also optionally comprise other constituents to improve the efficiency of reverse transcription or amplification, including, but not limited to, betaine, dimethyl sulfoxide, surfactant, bovine serum albumin, etc.

The term "cDNA" refers to a complementary DNA molecule synthesized using a ribonucleic acid strand (RNA) as a template. RNA may be mRNA, tRNA, rRNA, microRNA, or another form of RNA, such as viral RNA. The cDNA may be single-stranded, double-stranded or may be hydrogen-bonded to a complementary RNA molecule as in an RNA/cDNA hybrid.

The term "polymerase chain reaction" or "PCR" refers to a procedure or technique in which minute amounts of nucleic acid, RNA and/or DNA, are amplified as described in U.S. Pat. No. 4,683,195 issued Jul. 28, 1987. Generally, sequence information from the ends of the region of interest or beyond needs to be available, such that oligonucleotide primers can be designed; these primers will be identical or similar in sequence to opposite strands of the template to be amplified. The 5' terminal nucleotides of the two primers may coincide with the ends of the amplified material. PCR can be used to amplify specific RNA sequences, specific DNA sequences from total genomic

DNA, and cDNA transcribed from total cellular RNA, bacteriophage or plasmid sequences, etc. See generally Mullis *et al.*, Cold Spring Harbor Symp. Quant. Biol., 51:263 (1987); Erlich, ed., PCR Technology, (Stockton Press, NY, 1989).

The term "reverse transcription polymerase chain reaction" or "RT-PCR" refers to the transcription of cDNA from a RNA template by the enzyme reverse transcriptase. The cDNA is then amplified by known PCR methods.

The term "primer-extension" refers to an enzymatic process whereby a primer is hybridized to a template nucleic acid strand and is polymerized using said strand as a template. Polymerization can be mediated by enzyme classes including but not limited to DNA polymerases or reverse transcriptases. Primer-extension can take place as an isolated reaction (single extension of a primer on a template), or as part of a repetitive process such as PCR.

The term "primer" refers to an oligonucleotide capable of acting as a point of initiation of synthesis along a complementary strand when conditions are suitable for synthesis of a primer extension product. The synthesizing conditions include the presence of four different deoxyribonucleotide triphosphates (dNTPs) and at least one polymerization-inducing agent such as reverse transcriptase or DNA polymerase. These are present in a suitable buffer, which may include constituents which are co-factors or which affect conditions such as pH and the like at various suitable temperatures. A primer is preferably a single strand sequence, such that amplification efficiency is optimized, but double stranded sequences can be utilized. A primer can have some sequences that are not designed to hybridize to the targeted template DNA, including sequences at the 5'-end of the primer that becomes incorporated into the amplified products. Such sequences can include universal primer binding sites to be used in subsequent amplifications, sample-specific barcodes, or molecular lineage tags. In addition to serving the purpose of copying a nucleic acid template, a primer can also be used to append labels or other sequences to the copied products. Primers and other synthetic oligonucleotides disclosed herein have undergone either polyacrylamide gel purification or reverse-phase cartridge purification unless otherwise specified. A primer can also be modified by attachment of one or more chemical moieties including but not limited to biotin, a fluorescent tag, a phosphate, or a chemically reactive group.

The term "gene-specific primer" refers to a primer that is designed to hybridize to and be extended on a particular nucleic acid target. The 3'-segment of a gene-specific primer is complementary to its targeted RNA or DNA sequence, but other portions of the primer need not be complementary to any target. The target need not be a "gene" in the strict sense of the word. Possible targets include but are not limited to genomic DNA, mitochondrial DNA, viral DNA, mRNA, microRNA, viral RNA, tRNA, rRNA, and cDNA.

The term "nested primer" refers to a primer that is designed to hybridize to a primer-extended or PCR amplified product at a position that is either entirely or partially within the target region that was flanked by the original primers. The 3'-end of a nested primer is complementary to target sequences that would not have been contained within the original primers, but rather would have been copied by extension of the original primers on the desired template. Nested primers thus provide additional specificity for copying or amplifying a desired target after an initial round of primer-extension or PCR.

The terms "reaction mixture" or "PCR reaction mixture" or "PCR master mix" refer to an aqueous solution of constituents in a PCR or RT-PCR reaction that can be constant across different reactions. An exemplary PCR reaction mixture includes buffer, a mixture of deoxyribonucleoside triphosphates, reverse transcriptase, primers, probes, and DNA polymerase. Generally, template DNA is the variable in a PCR reaction.

The terms "sequence variant" or "mutation" are used interchangeably and refer to any variation in a nucleic acid sequence including but not limited to single point-mutations, multiple point-mutations, insertions/deletions (indels), and single-nucleotide polymorphisms (SNPs). These terms are used interchangeably in this document, and it is understood that when reference is made to a method for evaluating one type of variant, it could be equally applied to evaluation of any other type of variant. The term "variant" can also be used to refer to a single molecule whose sequence deviates from a reference sequence, or a collection of molecules whose sequences all deviate from the reference sequence in the same way. Similarly, "variant" can refer to a single sequence (or read) that deviates from a reference sequence or a set of sequences that deviate from a reference sequence.

The terms "mutation-prone region" and "mutation hotspot" are used interchangeably, and refer to a sequence region of a nucleic acid obtained from a biological source that has a higher probability of being mutated than surrounding sequence regions within the same nucleic acid. In the case of tumor-derived DNA, mutation-prone regions can be found in certain cancer-related genes. The mutation-prone region can be of any length, but mutation-prone regions that are analyzed using the methods disclosed herein are less than 100 nucleotides long. A mutation can be found anywhere within a mutation-prone region.

The term "target region" refers to a region of a nucleic acid that is targeted for primer extension or PCR amplification by specific hybridization of complementary primers.

The term "clonal overlapping paired-end sequencing" refers to a massively parallel sequencing method in which paired-end reads are obtained for each clonal sequence such that portions of the two reads from opposite strands are able to cover the same region of DNA. This

approach is used to reduce or suppress or distinguish sequencer-derived errors, thereby allowing base-calls to be made with greater confidence. The region of DNA that is covered by the overlapping reads is effectively read twice in opposite directions, once from each strand of the duplex. Thus, by including the mutation-prone region within the area of sequence overlap, the mutation prone region is read in one direction and then proofread in the opposite direction. Read-pairs that do not have perfect sequence consistency in the overlapping region (after obtaining a reverse-complement of one of the reads) can be attributed to sequencer error and can be discarded from the analysis. This approach greatly reduces the background of sequencer-generated errors and allows rare mutant molecules to be detected with greater sensitivity.

The terms "barcode", "tag", and "index" are used interchangeably and refer to a sequence of bases at certain positions within an oligonucleotide that is used to identify a nucleic acid molecule as belonging to a particular group. A barcode is often used to identify molecules belonging to a certain sample when molecules from several samples are combined for processing or sequencing in a multiplexed fashion. A barcode can be any length, but is usually between 6 and 12 bases long (need not be consecutive bases). Barcodes are usually artificial sequences that are chosen to produce a barcode set, such that each member of the set can be reliably distinguished from every other member of the set. Various strategies have been used to produce barcode sets. One strategy is to design each barcode so that it differs from every other barcode in the set at a minimum of 2 distinct positions.

The term "sample-specific barcode" refers to a barcode sequence that is assigned to molecules that are derived from a particular sample.

The term "template nucleic acid" refers to any nucleic acids that can serve as targets for primer-extension, reverse-transcription, or PCR. A template nucleic acid can be DNA or RNA. Methods described herein for analysis of DNA can also be applied to the analysis of RNA after reverse-transcribing the RNA to produce cDNA. Methods for evaluating DNA can be equally applied to the evaluation of RNA.

The terms "deep sequencing" and "ultra-deep sequencing" are used interchangeably herein and refer to approaches that use massively parallel sequencing technologies to obtain large numbers of sequences corresponding to relatively short, targeted regions of the genome. A targeted region can include, for example, an entire gene or small segment of a gene (such as a mutation hotspot). In some cases, many thousands of clonal sequences are obtained from a short targeted segment allowing identification and quantitation of sequence variants.

The term "clonal sequence" refers to a sequence that is derived from a single molecule within a sample that is subjected to massively parallel sequencing. Specifically, each clonal sequence

that is generated by massively parallel sequencing is derived from a distinct DNA molecule within a sample that serves as the "input" for the sequencing workflow.

The terms "targeted early barcoding", "early barcoding", "attachment of early barcodes", and "assignment of early barcodes" are used interchangeably and refer to assignment of barcodes to selected nucleic acid targets within a sample by specific hybridization and polymerization of barcode-containing primers at an early processing step. Preferably, barcode assignment occurs during the first enzymatic step that is performed after template nucleic acid molecules are purified from a biological sample. This first enzymatic step can be primer-extension, reverse-transcription, or PCR. When multiple different target sequences are to be tagged and copied from a given sample, a mixture of several different target-specific primers are used in a single reaction volume, with every primer in the mixture having the same sample-specific barcode. Separate early barcoding reactions are carried out for each sample, using similar mixtures of primers bearing distinct barcodes for each sample. Targeted early barcoding allows molecules from different samples to be combined into a single volume for all subsequent processing steps.

The term "degenerate sequence" refers to a stretch of sequence in which, within a population of nucleic acid molecules, two or more different bases can be found at each position. Most often, degenerate sequences are produced such that there is an approximately equal probability of each position having A, C, G, or T (in the case of DNA), or having A, C, G, or U (in the case of RNA). However, in certain situations, different bases can be incorporated in varying ratios at different positions, and some bases can be omitted at certain positions if desired. A degenerate sequence can be of any length.

The terms "molecular lineage tag", "MLT", and "MLT sequence" are used interchangeably and refer to a stretch of degenerate sequence that is contained within a synthetic oligonucleotide (e.g. a primer) and is used to assign a set of diverse sequence tags to copies of template nucleic acid molecules. A molecular lineage tag is designed to have between 2 and 10 degenerate base positions, but preferably has between 6 and 8 base positions. The bases need not be consecutive, and can be separated by constant sequences. The number of possible MLT sequences that can be generated in a population of oligonucleotide molecules is generally determined by the length of the MLT sequence and the number of possible bases at each degenerate position. For example, if an MLT is 8 bases long, and has an approximately equal probability of having A, C, G, or T at each position, then the number of possible sequences is 4^8 = 65,536. A molecular lineage tag is not designed to assign a completely unique sequence tag to each molecule, but rather is designed to have a low probability of assigning any given sequence tag to a particular molecule. The greater the number of possible MLT sequences, the lower the probability of any particular sequence being

assigned to a molecule. When many template molecules are copied and tagged, the same MLT sequence can be assigned to more than one template molecule. MLT sequences are used to track the lineage of molecules from initial copying through amplification, processing and sequencing. They can be used to distinguish sequences that arise from polymerase misincorporations or sequencer errors from sequences that are derived from true mutant template molecules. MLTs can also be used to distinguish sequences that have the wrong barcode assignment as a result of cross-over of barcodes during pooled amplification. Because the same MLT sequence can be assigned to more than one template molecule, meaningful analysis of MLT sequences requires first identifying variant target sequences and then analyzing the distribution of MLT sequences associated with those variants.

The term "molecular lineage tagging" refers to the process of assigning molecular lineage tags to nucleic acid templates molecules. MLTs can be incorporated within primers, and are attached to copies made from targeted nucleic acids by specific extension of primers on the templates.

The term "include" and its derivations should be understood to mean "including, but not limited to". The words "a", "an", and "the" include both singular and plural referents unless the context indicates otherwise.

Embodiments of the Methods

Methods and compositions are disclosed herein for identifying and quantifying nucleic acid sequence variants. Methods disclosed herein can identify and quantify low-abundance sequence variants from complex mixtures of DNA or RNA. Embodiments of the methods can measure small amounts of tumor-derived DNA that can be found in the circulation of patients with various types of cancer.

Assessment of rare variant DNA sequences is important in many areas of biology and medicine. Small amounts of fetal DNA can be found in the circulation of pregnant women. An embodiment includes analyzing rare fetal DNA that can be used to assess disease-associated genetic features or the sex of the fetus. An organ that is undergoing rejection by the recipient can release small amounts of DNA into the blood, and this donor-derived DNA can be distinguished based on genetic differences between the donor and the recipient. An embodiment includes measuring donor-derived DNA to provide information about organ rejection and efficacy of treatment. In another embodiment, nucleic acids can be detected from an infectious agent (e.g., bacteria, virus, fungus, parasite, etc.) in a patient sample. Genetic information about variations in pathogen nucleic acids can help to better characterize the infection and to guide treatment decisions. For instance,

detection of antibiotic resistance genes in the bacterial genome infecting a patient can direct antibiotic treatments.

Detection and measurement of low-abundance mutations has many important applications in the field of oncology. Since tumors are known to acquire somatic mutations, some of which promote the unregulated proliferation of cancer cells, identifying and quantifying these mutations has become a key diagnostic goal. Companion diagnostics have become an important tool in identifying the mutational cause of cancer and then administering effective therapy for that particular mutation. Furthermore, some tumors acquire new mutations that confer resistance to targeted therapies. Thus, accurate determination of a tumor's mutation status can be a critical factor in determining the appropriateness of particular therapies for a given patient. However, detecting tumor-specific somatic mutations can be difficult, especially if tumor tissue obtained from a biopsy or a resection has few tumor cells in a large background of stromal cells. Tumor-derived mutant DNA can be even more challenging to measure when it is found in very small amounts in blood, sputum, urine, stool, pleural fluid, or other biological samples.

Tumor-derived DNA is released into the bloodstream from dying cancer cells in patients with various types of malignancies. Detection of circulating tumor DNA (ctDNA) has several applications including, but not limited to, detecting presence of a malignancy, informing a prognosis, assessing treatment efficacy, tracking changes in tumor mutation status, and monitoring for disease recurrence or progression. Since unique somatic mutations can be used to distinguish tumor-derived DNA from normal background DNA in plasma, a new class of highly specific DNA-based cancer biomarkers are described with clinical applications that may complement those of conventional serum protein markers. In an embodiment, methods include screening ctDNA for presence of tumor-specific, somatic mutations. In such embodiments, false-positive results are very rare since it would be very unlikely to find cancer-related mutations in the plasma DNA of a healthy individual. Described herein are methods that specifically and sensitively measure rare mutant DNA molecules that are shed into blood from cancer cells. Achieving extremely high detection sensitivity is especially important for detection of a small tumor at an early (and more curable) stage.

Since somatic mutations can occur at many possible locations within various cancer-related genes, a clinically useful test for analyzing ctDNA would need to be able to evaluate mutations in many genes simultaneously, and preferably from many samples simultaneously. In embodiments, analysis of a plurality of mutation-prone regions from a plurality of samples allows more efficient use of large volumes of sequence data that can be obtained using massively parallel sequencing technologies. In an embodiment, labeling molecules arising from a given sample with a sample-specific DNA sequence tag, also known as a barcode or index, facilitates simultaneous analysis of

more than one sample. By using distinct barcode sequences to label molecules derived from different samples, it is possible to combine molecules and to carry out massively parallel sequencing on a mixture. Resultant sequences can then be sorted based on barcode identity to determine which sequences were derived from which samples. To minimize chances of misclassification, barcodes are designed so that any given barcode can be reliably distinguished from all other barcodes in the set by having distinct bases at a minimum of two positions.

In most protocols that are currently used to prepare samples for massively parallel sequencing, barcodes are attached after several steps of sample processing (e.g. purification, amplification, end repair, etc). Barcodes can be attached either by ligation of barcoded sequencing adapters or by incorporation of barcodes within primers that are used to make copies of nucleic acids of interest. Both approaches typically require several processing steps to be performed separately on nucleic acids derived from each sample before barcodes can be attached. Only after barcodes are attached can samples be mixed.

In an embodiment, barcodes are assigned to targeted molecules at a very early step of sample processing. Targeted early barcode attachment not only permits sequencing of multiple samples to be performed in batch, it also enables most antecedent processing steps to be performed in a combined reaction volume. Once barcodes are attached to nucleic acid molecules in a sample-specific manner, molecules can be mixed, and all subsequent steps can be carried out in a single tube. If a large number of samples are analyzed, targeted early barcoding can greatly simplify the workflow. Since all molecules can be processed under identical conditions in a single tube, the molecules would experience uniform experimental conditions, and inter-sample variations would be minimized. In an embodiment, tagging of nucleic acids from different samples can be achieved in consistent proportions and then used to enable quantitative comparisons of nucleic acid concentrations across samples. In addition to quantifying DNA, targeted early barcoding can enable quantifying RNA (e.g., RNA expression levels across different samples). Once barcodes are attached, targeted nucleic acids bearing different sample-specific barcodes can be amplified in a combined reaction volume by competitive end-point PCR, and relative counts of different barcodes in amplified products could be used to quantify associated nucleic acids in various samples. Thus, early barcoding can be used to quantify a total amount of various targeted nucleic acids, and not just variants, across many samples.

In an embodiment, well-defined mixtures of primers are produced containing combinations of sample-specific barcodes and consistent ratios of gene-specific segments. Such primers can be used for targeted early barcoding and subsequent batched sample processing. These primers can also be used for quantitation of DNA or RNA in different samples. In an embodiment, such primers

allow parallel processing and analysis of multiple mutation-prone genomic target regions from multiple samples in a simplified and uniform manner.

Embodiments include methods that accurately quantify mutant DNA rather than simply determining its presence or absence. In an embodiment, an amount of mutant DNA provides information about tumor burden and prognosis. Embodiments are capable of analyzing DNA that is highly fragmented due to degradation by blood-borne nucleases as well as due to degradation upon release from cells undergoing apoptotic death. Since somatic mutations can occur at many possible locations within various cancer-related genes, an embodiment can evaluate mutations in many genes simultaneously from a given sample. Embodiments are capable of finding mutations in ctDNA without knowing beforehand which mutations are present in a patient's tumor. An embodiment is able to screen for many different types of cancer by evaluating multiple regions of genomic DNA that are prone to developing tumor-specific somatic mutations. An embodiment includes multiple samples combined together in the same reaction tube to minimize inter-sample variations.

Although the methods described herein have been optimized for measurement of small amounts of mutant circulating tumor DNA (ctDNA) in a background of normal (wild-type) cell-free DNA in the plasma or serum of a patient having cancer, it is understood that they could be applied more broadly to the analysis of nucleic acid variants from a variety of sources. Examples of such sources include, but are not limited to lymph nodes, tumor margins, pleural fluid, urine, stool, serum, bone marrow, peripheral white blood cells, cheek swabs, circulating tumor cells, cerebrospinal fluid, peritoneal fluid, amniotic fluid, cystic fluid, frozen tumor specimens, and tumor specimens that have been formalin-fixed and paraffin-embedded.

Features:

Methods include identifying and measuring low-abundance variants occurring in multiple mutation-prone regions of genomes from multiple samples in parallel. One aspect includes early attachment of sample-specific DNA barcodes to a plurality of nucleic acid targets that are derived from a plurality of samples. Specifically, a mixture of gene-specific primers, all bearing the same barcode, are used to make tagged copies of several different genomic target regions from nucleic acids in a given sample in a single reaction volume. For each additional sample, this process is repeated in a separate reaction volume using a similar mixture of gene-specific primers bearing a different barcode. All members of a given primer mix have the same sample-specific barcode, but different primer mixes have different barcodes. Once barcodes have been attached, the DNA from multiple samples can be combined into a single volume for further processing.

If many DNA targets from many samples are to be analyzed, large numbers of primers would need to be produced, each having different combinations of barcoded 5' segments and gene-specific 3' segments. Targeted early barcoding allows combining nucleic acids from different samples and processing of the nucleic acids together in a combined reaction volume. Batched processing has an advantage of simplified workflow and greater experimental consistency and uniformity across different samples. Batched processing decreases potential quantitative variability arising from very small inter-sample concentration or temperature differences. Although the variability may be small at time of initial input, the end result may have substantial variability due to the exponential nature of PCR. Amplification of differently barcoded nucleic acid copies in a combined reaction volume by competitive end-point PCR followed by high throughput sequencing of the products would allow direct enumeration of the various barcodes associated with a given genomic target region. The relative quantity of each targeted nucleic acid in the different samples could be deduced from the relative abundance of the various barcodes within the sequence data.

Another aspect includes producing primers by combining modular oligonucleotide segments. Implementing targeted early barcoding requires generating well-defined mixtures of large numbers of primers. Primer mixtures are produced in such a way that each mixture contains identical proportions of 3'-gene-specific segments, ensuring that target nucleic acids from different samples are copied in consistent ratios. This makes it possible to quantitatively compare nucleic acid concentrations across different samples. In an embodiment, combining modular oligonucleotide segments is used. More specifically, to generate each mixture, a portion of a uniform pool of various gene-specific 3' oligonucleotide segments is joined to a single, uniquely-barcoded 5' segment. Since the 3' segments used to produce each final mixture are derived from a common pool (or master-mix), each uniquely barcoded primer mix has similar proportions of the different 3' gene-specific segments. Several approaches are described herein for joining the modular 5' and 3' segments. This modular approach to producing primer mixes allows the production of thousands of primer and barcode combinations that would have otherwise been very costly and laborious to produce. Furthermore, the consistency of gene-specific primer ratios that can be achieved across different mixes would not be possible by mixing individually synthesized primers. Methods described herein utilize next-generation, high-throughput DNA sequencing technologies to identify and quantify nucleic acid variants. These technologies are able to quickly and inexpensively produce sequences from millions of DNA molecules in a massively parallel fashion. By oversampling sequences of a large number of DNA molecules from a particular genomic region using ultra-deep sequencing, it would be possible to identify and enumerate rare sequence variants. The sensitivity of the sequencing is limited by the inherent error rate of the sequencer since incorrectly read bases might be mistaken

for true mutant DNA copies. Mutant ctDNA has been reported to comprise on average 0.2% of total plasma DNA (Diehl et al., *Nat Med*. 2008; 14: 985-990) – a range in which sequencer misreads can be problematic. This is a limitation of massively parallel sequencing to measure very low-abundance mutations.

Herein methods are described that use clonal overlapping paired-end sequencing to achieve sequence redundancy in mutation-prone regions, thereby allowing base calls to be made with much greater confidence. Embodiments include methods of reducing, suppressing, and distinguishing sequencer-derived errors. Using an Illumina® next-generation sequencing platform, an embodiment includes obtaining a read in one direction from a clonal cluster of DNA molecules, and then subsequently obtaining a read in the opposite direction (from the opposite strand of the duplex). The length of each "paired-end" read can be 36, 50, 75, 100, or 150 bp or longer. An embodiment includes sequencing short PCR amplicons in a paired-end fashion to obtain overlapping reads from both strands of a clone. By designing the mutation-prone region to be in the area of sequence overlap, clonal sequence redundancy can be achieved in this region. Thus, each clonal sequence from a mutation-prone region is read in one direction, and then is proofread in the other direction. Read-pairs that do not have perfect agreement in the overlapping region (after obtaining a reverse-complement of one of the reads) can be attributed to sequencer error, and can be ignored in the final analysis. In this way, sequencer-generated errors in a region of interest can be reduced since a probability of finding the same sequencer error in reads from both strands of a clone is exceedingly low. By reducing the background of sequencer errors, it becomes possible to achieve better detection sensitivity for rare mutant molecules. Detection sensitivity is especially important in patients with early-stage cancers who are likely to have a very low concentration of mutant ctDNA molecules in their blood.

Another aspect includes distinguishing nucleotide misincorporation errors that can be introduced during DNA copying, amplification, or processing. After suppression of sequencer-derived errors, variant sequences are still found that do not correspond to authentic mutations arising from mutant template DNA molecules. A majority of these variant sequences arise from incorporation of incorrect nucleotides when DNA template molecules are copied or amplified. Possible causes of such misincorporation errors include but are not limited to DNA damage (for example, cytosine deamination during heating) or polymerase-induced errors.

To distinguish variant sequences arising from true mutant template molecules versus those arising from misincorporation errors, an embodiment includes molecular lineage tagging. In molecular lineage tagging, a degenerate sequence called a molecular lineage tag (MLT) is incorporated into primers that make a small number of copies (between 2 to 20) of an original

template DNA molecule. An MLT is a stretch of degenerate sequence having an approximately equal probability of having A, T, C, or G at each position and can be about 2 to about 10 bases in length, but preferably would be 6, 7, or 8 bases long. An MLT sequence can also be split into segments that are separated by non-degenerate positions within an oligonucleotide.

It is not necessary that each template molecule be tagged with a unique MLT, but only that each template molecule should have a low probability of being tagged with any given MLT-sequence. For example, if the MLT region consisted of 8 degenerate positions, then 4^8 = 65,536 possible MLT sequences could be generated. MLT-containing primers are used to make a limited number of copies of the template DNA molecules, via either a few cycles (2 to 4) of PCR or primer-extension. Thus, each template copy would be tagged with one of 65,536 possible MLT sequences. When these tagged copies are amplified by PCR, the "progeny" molecules derived from amplification of a given "parent" copy should retain the same identifying MLT sequence as the parent molecule. If a variant sequence arose from a true mutant template molecule, then many copies of a given MLT sequence should be associated with that variant sequence (since that MLT was associated with the mutant copy at the beginning of the amplification process). On the other hand, if an error was introduced during amplification or processing, one would expect a smaller number of copies of a given MLT to be associated with the erroneous variant sequence (unless the error occurred at a very early cycle of amplification). It is important to note that if several thousand template molecules are tagged with MLTs, there is a high probability that some MLT sequences may be assigned to more than one template molecule.

With non-unique MLT's, it is less informative to evaluate the percentage of mutant and wild-type sequences associated with a particular MLT sequence. Rather, it is preferable to identify mutant sequences, and then to evaluate distribution of MLT sequences associated with those variants. If the number of sampled clonal sequences (post-amplification) is several-fold greater than the number of tagged template copies, then variant sequences arising from true mutant template molecules would be associated with multiple copies of a given MLT sequence, whereas variants arising from misincorporation errors would be likely to be associated with fewer copies of any given MLT. Analysis of MLT distributions (number of different MLT sequences and number of copies of each sequence) associated with a particular variant made it possible to identify the majority of variants arising from misincorporation errors, thereby further improving the sensitivity for detecting true template-derived mutations.

Another aspect includes distinguishing sequences that are misclassified as belonging to a wrong sample. Such incorrect classification of a sequence can occur if it is associated with an inappropriate barcode. Since barcodes are designed to differ from all other barcodes in a set at a

minimum of two distinct positions, misclassification due to barcode sequence errors would be rare. However, cross-over of barcodes has been observed from differently barcoded molecules that undergo combined polymerization or amplification in the same reaction volume. This can happen, for example, if primer-extension stalls before a polymerase has completed extending on a template during a given cycle of PCR. That partially-extended strand (possibly containing a mutant or wild-type sequence) could then anneal to a different template during the next cycle of PCR, and could incorporate an inappropriate barcode. Alternatively, if two strands of DNA containing different barcodes are annealed to each other via a common complementary sequence, the 3'-5' exonuclease activity of a proofreading polymerase can digest the barcode on one strand and then extend that strand using the opposite strand's barcode as a template. MLT sequences can be used to distinguish sequences derived from such barcode "cross-over" events. If an MLT region is positioned in proximity to or adjacent to a barcode sequence, then it can be used to track the lineage of the barcode. If a variant is tagged with an inappropriate barcode as a result of cross-over during the process of amplification, then one would expect fewer than average copies of a particular MLT sequence to be associated with that barcode/variant combination. To further aid in distinguishing cross-over sequences, a second MLT can be positioned on the opposite side of the mutation-prone region (so that the sequence order, for example, could be MLT-1/Barcode/mutation-prone region/MLT-2). In this case, DNA molecules that undergo cross-over between a barcode and a mutation prone region would also undergo cross-over of MLT-1 between MLT-2. Thus, such crossed-over sequences could be identified because the number of copies of a particular MLT-1/MLT-2 combination would be lower than for sequences that did not undergo cross-over. Thus, MLT sequences can allow differently barcoded molecules to be amplified in a combined reaction volume while maintaining accurate assignment of mutations to specific samples.

Another aspect includes highly-specific tagging, copying, and amplification of several genomic target regions from several samples simultaneously in a single reaction volume while minimizing accumulation of unwanted, spurious amplification products. Such highly multiplexed processing and amplification is prone to accumulation of spurious products because of the presence of large numbers of different primers. Having a complex mixture of primers with different combinations of barcodes, degenerate sequence regions, and gene-specific regions in a single PCR amplification can lead to formation of many primer dimers and non-specific amplification products. An embodiment includes multi-step tagging and amplifying without having to compromise primer concentrations. An embodiment of a process includes highly stringent purification of desired amplification products between each amplification step to remove unextended primers, spurious extension products, and genomic template DNA as well as enzyme, buffer, and nucleotides. An

embodiment utilizes biotin-tagged oligonucleotides to mediate specific isolation of desired products. Another embodiment utilizes high-temperature washes when using biotin-tagged oligonucleotides. Another embodiment includes digesting unwanted single-stranded products and primers with an exonuclease to further improve amplification specificity. An embodiment also uses nested primers to provide further selectivity for desired products. An embodiment includes universal PCR primers for the final amplification. Under the stringent conditions described herein, universal PCR primers can be used for the final amplification without significant accumulation of spurious products.

Methods:

Producing combinations of modular oligonucleotide segments for tagging of nucleic acids

In an embodiment, tagged copies of multiple nucleic acid targets are made from template DNA or RNA derived from a given sample. To produce such tagged copies, a mixture of primers is used in which the 3'-segments of the primers are able to hybridize to RNA or DNA targets by sequence complementarity (as illustrated, for example, by the reverse primers 1 in Figure 1). A polymerase (such as a reverse transcriptase or a DNA polymerase) can then be used to extend the primers in the 5' to 3' direction using the targeted nucleic acids as templates. In an embodiment, a sample-specific DNA barcode sequence can be incorporated into the 5'-segment of each primer such that the barcode becomes attached to the copy of the target template after undergoing primer-extension. Since multiple target templates are to be copied from a given sample in a single reaction tube, a mixture of primers is required having various target-specific sequences in their 3'-segments and all having the same sample-specific barcode sequence in their 5'-segments. If several different samples are to be analyzed, then similar mixtures of primers must be made for each sample, with each mixture containing a unique, sample-specific barcode sequence in the 5'-segment. In some embodiments, the 5'-segment of each primer can also contain other elements such as sequencing adapters (to facilitate sequencing of the copied DNA), binding sites for PCR primers, or stretches of degenerate sequence (having equal probability of A, C, T, or G bases at each position) that can serve as tags to follow the lineage of molecules during copying, amplification, and sequencing.

In an embodiment, a barcode comprises a unique sequence (typically 6 to 12 nucleotides long) that is used to identify molecules derived from a particular sample after molecules from multiple samples are pooled and sequenced in batch. In an embodiment, a computer program can be used to sort clonal sequences derived from each molecule based on barcode identity. In order to minimize the chance that a sequence derived from one sample might be misclassified as being derived from another sample, each barcode sequence is designed to differ from all other barcodes in

the set by at least 2 nucleotides (so that a single sequencing error would not lead to misclassification).

In an embodiment, multiple gene-specific primer regions (at the 3'-ends of primers) are attached in separate batches, to unique sample-specific barcodes (near the 5'-regions of primers). If many genomic targets are to be analyzed from many samples, the number of combinations of primer 3'-ends and 5'-ends can become very large. For example, if 40 target gene regions are to be evaluated from 96 different samples, 40 x 96 = 3,840 different oligonucleotides would need to be made, each with a unique combination of 3' gene-specific sequence and 5' barcode. If conventional oligonucleotides were individually synthesized, a mixture of 40 different gene-specific primers having a particular barcode would be used to primer-extend nucleic acid targets from a given sample within a single tube. Thus, all 40 target regions would be tagged with the same sample-specific barcode. However, synthesis and purification of 3,840 oligonucleotides individually would be impractical. Because termination sequences would be abundant when making long primers, full-length oligonucleotides would have to be purified by methods including but not limited to polyacrylamide gel electrophoresis, high performance liquid chromatography, or reverse-phase cartridge purification.

To address the need for producing uniform mixtures of multiple gene-specific primer, with each mixture having a unique barcode sequence, embodiments are described in which combinations of modular gene-specific 3' oligonucleotide segments can be attached to a modular barcoded 5' oligonucleotide segment. In various embodiments, production of modular oligonucleotides allows multiple gene-specific 3' segments to be synthesized and uniformly mixed, and then attached in separate batches to each barcoded 5' segment (Figure 2). Resulting uniquely barcoded primer mixes would each have consistent ratios of gene-specific primer sequences. In some embodiments, such uniform primer mixes could be used to copy and label DNA or RNA molecules from different samples with consistent efficiency (so that the resulting tagged copies would be proportionate to the amount of target nucleic acids in each sample). Subsequent pooled amplification of differently barcoded molecules by competitive PCR, followed sequencing to count barcoded sequences, would enable accurate quantitation of DNA or RNA targets in the various samples. Such uniform primer mixes would be very difficult to achieve by simply mixing individually synthesized primers.

In some embodiments, multiple 3' oligonucleotide segments are produced and mixed, and then the mixture is joined in separate batches to unique 5' oligonucleotide segments. In an embodiment (Figure 3), different 3'-segments can be synthesized in separate columns on an automated oligonucleotide synthesizer (on solid supports). Synthesis can then be paused and the solid supports from the different columns can be uniformly mixed. Then the mixture can be

dispensed into several fresh columns. Synthesis can then be continued, adding a uniquely barcoded 5' segment to each new column. After cleavage, deprotection, and purification, the desired uniquely barcoded uniform primer mixtures are obtained. In another embodiment, mixtures of 5'-phosphorylated 3'-segments can be ligated to different barcoded oligonucleotides using splint-mediated enzymatic ligation. In another embodiment, primer extension can be used to produce combinations of modular segments. In this approach, a single barcoded 5' oligonucleotide can be hybridized via a common sequence to a mixture of complementary templates. These templates can be designed to produce various gene-specific 3'-ends when the barcoded oligonucleotide is primer extended on these templates. In an embodiment, biotin tags on the templates can be used to separate the templates from the desired uniquely barcoded primer mix. Similar primer extension reactions can be performed in separate reaction volumes using different barcoded oligonucleotides to produce several similar uniform primer mixes. In yet another embodiment, mixtures of 3' oligonucleotide segments having reactive chemical conjugation moieties on their 5'-end can be combined in separate batches with uniquely barcoded 5' oligonucleotides segments having reactive conjugation moieties on at their 3' ends. Such chemical conjugation would allow post-synthesis combination of oligonucleotide segments. Special conjugation chemistries have been previously described that can conjugate two oligonucleotide segments together leaving a phosphodiester bond at the junction (or similar bond that would be compatible with subsequent enzymatic processes).

Isolation of Template DNA

Embodiments provide methods for purification or isolation of DNA or RNA from various clinical or experimental specimens. Many kits and reagents are commercially available to facilitate nucleic acid purification. Depending on the type of sample to be analyzed, appropriate nucleic acid isolation techniques can be selected. Substances that might inhibit subsequent enzymatic reaction steps (such as polymerization) should be removed or reduced to non-inhibitory concentrations in purified DNA or RNA samples. A yield of nucleic should be maximized whenever possible. It would be disadvantageous to lose DNA during purification, wherein the lost DNA might include rare variant DNA. When isolating DNA from plasma, about 10 ng to 100 ng of cell-free DNA can be purified from 1 mL of plasma, which corresponds to 3,500 to 35,000 genome copies. To note, DNA yields can vary dramatically, especially in patients with an ongoing disease process such as cancer.

In an embodiment, DNA can also be analyzed from other sample types, including but not limited to the following: pleural fluid, urine, stool, serum, bone marrow, peripheral white blood cells, circulating tumor cells, cerebrospinal fluid, peritoneal fluid, amniotic fluid, cystic fluid, lymph

nodes, frozen tumor specimens, and tumor specimens that have been formalin-fixed and paraffin-embedded.

Producing a limited number of tagged copies of targeted nucleic acids

In an embodiment, a limited number of tagged copies (e.g., fewer than 20) of targeted nucleic acid molecules are made at an early step in the process. After DNA or RNA is purified from the original sample, targeted nucleic acid template molecules can be copied by specifically hybridizing and polymerizing tagged primers. When a plurality of target regions are to be copied and tagged from a given sample, a mixture of modular barcoded primers can be used (as described above). In an embodiment, targeted nucleic acid regions are mutation-prone regions (also called mutation hotspots). A mixture of primers for a given sample can contain sequences at their 3'-ends that specifically hybridize to an area of DNA near or adjacent to a target region. All primers used for a given sample would have the same sample-specific barcode sequence, and different samples would have different barcodes. In some embodiments, the primers can also contain stretches of degenerate bases known as molecular lineage tags (MLTs) that can be helpful in distinguishing sequences arising from true mutant template molecules versus those arising from misincorporation errors occurring during amplification or processing. The MLTs can also help to identify sequences that are assigned to the wrong barcode due to cross-over of barcodes during pooled amplification of differently barcoded molecules. In an embodiment, primers can also contain adapter sequences that are necessary for sequencing, and universal primer binding sites that can be used in subsequent amplifications.

In an embodiment, a DNA polymerase can be used to extend the primers on hybridized templates, thus producing copies of the target nucleic acids with sample-specific barcodes attached. A DNA polymerase can be a thermostable or non-thermostable enzyme, and may or may not have proofreading activity. Examples of polymerases include, but are not limited to, *Taq*, Phusion®, Vent$_R$®, *Pfu*, *Pfx*, DNA Polymerase I (Klenow fragment), or reverse transcriptase. When specific primer annealing and extension is to be carried out at temperatures above 50°C, thermostable polymerases with hot-start capability are preferred in order to minimize spurious polymerization at room temperature during reaction set-up. Copies of template nucleic acids can be made by a single primer extension step, by a few cycles of primer extension (1 to 10 cycles, with heat-denaturation of the extended products between cycles), or by a few cycles of PCR in which opposite primers are also added (2 to 5 cycles). A few tagged copies of each template molecule can be produced so that a complete sampling of sequences can be obtained even if there is some loss of copies during the various purification, processing, and amplification steps. However, the number of tagged copies

must be limited to avoid assigning too many different MLTs to each template molecule, which would require greater sequencing depth for analysis. In an embodiment, after a limited number of tagged copies are made, the polymerase is inactivated, and barcoded copies from different samples can then be pooled into a combined volume for further processing.

Purification of tagged copies

In an embodiment, tagged, primer-extended copies of target sequences are purified away from un-extended and non-specifically extended primers and from excess template nucleic acids. Purification also removes other reaction components such as buffer, dNTPs, and polymerase. Removal of un-extended primers and non-specifically extended primers is preferred so that they are not carried over to the next polymerization step. Also, removal of excess primers and template molecules allows greater specificity of polymerization in subsequent steps.

In an embodiment, purification of specifically tagged and extended products is mediated by capture using biotin-labeled complementary oligonucleotides that hybridize to the specifically extended products. Oligonucleotides can be designed to anneal to sequences produced when tagged primers are extended beyond the mutation-prone region (or target region). Such hybridization of the biotin-labeled capture oligonucleotides to the extended tagged copies can be achieved either by using the biotinylated primers in PCR (Figure 1), or by subsequently annealing them to primer-extended copies (Figure 10). In an embodiment, immobilized streptavidin (or an analogue with affinity for biotin) is used to isolate and purify the tagged, extended copies that hybridize to the capture oligonucleotides. Immobilized streptavidin is available in many forms, including but not limited to surface-bound, agarose bead-bound, magnetic bead-bound, or filter-bound. In some embodiments, removal of non-specifically annealed nucleic acids can be achieved by washing the bound molecules at room temperature or at elevated temperatures that would selectively disrupt short, non-specific stretches of hybridization, but would not disrupt specifically-annealed products. In some embodiments, nuclease treatment of the bound molecules can also be used to digest non-specifically annealed products. Nucleases that could be used include but are not limited to Exonuclease I, Exonuclease VII, and Rec Jf. In some embodiments, elution of specifically-annealed copies can be achieved by heat-denaturation or by alkaline-denaturation to separate biotin-labeled strands from the desired single-stranded tagged copies. Biotin labeled strands should remain attached to the immobilized streptavidin since the biotin-streptavidin interaction is not substantially disrupted by heat or moderate alkaline conditions.

In another embodiment, specifically primer-extended copies can be purified by carrying out limited cycles of PCR and then digesting single-stranded nucleic acids to remove un-extended

primers. In yet another embodiment, oligonucleotides can be specifically hybridized to primer-extended products to protect their 3'-ends from digestion by a 3' to 5' single-stranded exonuclease such as Exonuclease I.

Double-stranded products that survive digestion can be purified by a variety of approaches, including but not limited to ethanol precipitation, silica membrane partitioning, or binding to magnetic Solid Phase Reversible Immobilization (SPRI) beads.

Second round of tagging, copying, and purification

In an embodiment, the tagged, pooled, and purified DNA copies from multiple samples can be subjected to another round of limited-cycle primer-extension or limited-cycle PCR (similar number of cycles as described for the first round). Primers used in this second round would be designed to incorporate MLTs on the opposite side of the mutation-prone region relative to the MLTs incorporated in the first round (Figure 1). This second MLT region could be used to distinguish sequences arising from barcode cross-over events that occurred during pooled amplification or processing. Use of nested primers in the second round of PCR or primer-extension would provide additional selectivity for the targeted genomic sequences. In an embodiment, primers used in the second round could contain universal primer binding sites that would be used for subsequent amplification with universal primers. In an embodiment, primers could also contain adapter sequences that facilitate sequencing using a next-generation sequencer.

In an embodiment, a limited number of specifically primer-extended copies produced in the second round could be purified away from un-extended or non-specifically extended primers and other reaction components using similar approaches as described for the first round. In an embodiment, purification can be achieved using biotinylated capture oligonucleotides designed to specifically hybridize to sites on the opposite primer-extended strands (relative to the hybridization sites of the biotinylated oligonucleotides used in the first round). In an embodiment, nuclease treatment may be used to digest un-extended or non-specifically extended primers.

Amplification and purification of specifically copied and tagged products

In an embodiment, products from the first two rounds of copying, tagging, and purification are used as templates for further PCR amplification. In an embodiment, universal primers are used for PCR that are designed to bind to sequences introduced by primers in the first two rounds. Since universal primers are used, it is very important that only desired targeted products remain as templates for the final PCR after the second-round purification. Presence of even small amounts of primer dimers or other spurious products could lead to competitive amplification of undesired

templates by the universal primers. In an embodiment, this round of PCR can be carried out for a larger number of cycles than were used in the first 2 rounds. A total of 5 to 40 PCR cycles may be used, depending on the amount of template nucleic acid present and the number of samples being multiplexed. A final PCR is designed to produce sufficient DNA as required for massively parallel sequencing (which can differ depending on the sequencing platform being used). In some embodiments, a final PCR may not be necessary if the required input of the sequencer is satisfied by the amount of DNA product generated after the first 2 rounds. In some embodiments, the DNA products are gel-purified to select products of the desired size and to eliminate unused primers before subjecting to massively parallel sequencing. In some embodiments, other approaches to purification could be used, including but not limited to high-performance liquid chromatography, capillary electrophoresis, silica membrane partitioning, or binding to magnetic Solid Phase Reversible Immobilization (SPRI) beads.

Massively parallel sequencing and data analysis

In an embodiment, a next-generation sequencer is used to obtain large numbers of sequences from the tagged, amplified, and purified products. Clonal sequences (each sequence arising from a single nucleic acid molecule) produced by such a sequencer can be used to identify and quantify variant molecules using an approach known as ultra-deep sequencing. In principle, because large numbers of sequences can be obtained for each target site and for each sample, rare variants can be detected and measured. However, the error rate of the sequencer can limit the sensitivity of detection because such errors might be mistaken as true variants. To minimize the contribution of sequencer errors, an embodiment uses clonal overlapping paired-end sequences. By separately sequencing opposite strands of DNA from each clonal population, and comparing the overlapping regions of the sequences, the vast majority of variants arising from sequencer errors can be eliminated. In an embodiment, the region of sequence overlap is designed to be in the mutation-prone area. In an embodiment, only read-pairs that perfectly match in the overlapping region are retained for further analysis. For such analysis, instruments that produce clonal paired-end reads (such as the Illumina platform) are preferred. In some embodiments, other massively parallel sequencing platforms that provide sequence redundancy can also be utilized.

In an embodiment, errors introduced into the DNA during amplification or processing can be distinguished from true template-derived mutant sequences by analyzing the distribution of molecular lineage tags (MLTs) associated with variant sequences. In an embodiment, MLTs can also be used to distinguish sequences bearing incorrect barcodes due to cross-over events during pooled amplification.

The present technology may be better understood by reference to the following examples. These examples are intended to be representative of specific embodiments of the invention and are not intended to limit the scope of the invention.

EXAMPLES

Example 1

This example demonstrates application of a deep sequencing approach in which 3 mutation hotspot regions were analyzed from multiple plasma samples. The method in this example includes redundancy within each clonal sequence to produce extremely high quality base-calls in short, mutation-prone regions of plasma DNA. Amplification of both mutated and wild-type sequences was carried out by unbiased PCR in the same tube, ensuring highly accurate and reproducible quantitation. The scheme was designed to have flexibility to simultaneously analyze mutations in several genes from multiple patient samples, making it practically feasible to screen plasma samples for mutant ctDNA without prior knowledge of the tumor's mutation profile.

MATERIALS AND METHODS

Patient plasma and tumor samples

Under the approval of the Human Investigation Committees at the Yale School of Medicine and at Lawrence & Memorial Hospital, plasma samples were obtained from 30 patients with stage I-IV non-small cell lung cancer (NSCLC) between July 2009 and July 2010. Informed consent was obtained from all patients. Most patients were recruited in the radiation oncology clinic, and underwent treatment with radiation therapy, chemotherapy, targeted systemic therapy, and/or surgery. Whenever possible, blood samples were collected from patients before starting the current course of treatment and then at subsequent times during and after treatment. A total of 117 samples were obtained. Formalin-fixed, paraffin-embedded tumor specimens were obtained for all patients with non-squamous histology whose tumors had not already been tested for mutations by a clinical laboratory, and for whom sufficient tissue was available in the block after standard pathology evaluation.

Extraction and amplification of plasma DNA

Blood was collected in EDTA-containing tubes (Becton Dickinson) and was centrifuged at 1000 *g* for 10 minutes within 3 hours of collection. Plasma was transferred to cryovials, being careful to avoid the buffy coat, and was stored at -80°C until further processing. Frozen plasma aliquots stored at -80°C were thawed to room temperature, and DNA was purified using the

QIAamp® DNA Blood Mini kit (Qiagen Sciences, Valencia, CA) as per the manufacturer's instructions. 5 µg of carrier RNA was added to each 200 µL plasma sample as recommended to improve adsorption of low-concentration nucleic acids to the silica membrane.

Purified plasma DNA was then subjected to 2 rounds of amplification by PCR (in triplicate) using primers designed to amplify short DNA segments that included codons 12 and 13 of *KRAS*, codon 858 of *EGFR*, and codon 600 of *BRAF*. The sequences of the primers used in both rounds of PCR are listed in Table 1.

Table 1: Primers used in first and second rounds of PCR.

PCR Primer	Sequence (5' – 3')	SEQ ID NO:
Round 1 Forward *KRAS**	GGCCTGCTGAAAATGACTGAATATAAAC	1
Round 1 Reverse *KRAS**	TTCGTCCACAAAATGATTCTGAATTAGC	2
Round 1 Forward *BRAF**	TCATGAAGACCTCACAGTAAAAATAGGTG	3
Round 1 Reverse *BRAF**	CACAAAATGGATCCAGACAACTGTTC	4
Round 1 Forward *EGFR**	GTACTGGTGAAAACACCGCAGCAT	5
Round 1 Reverse *EGFR**	CTTACTTTGCCTCCTTCTGCATGGTATT	6
Round 2 Forward *KRAS*	GG[**FWD BC**]CGAACAGTCTCCGAATATAAACTTGTGG TAGTTGG	7
Round 2 Reverse *KRAS*	GC[**REV BC**]GGATGAGTGCAGTGAATTAGCTGTATCG TCAAG	8
Round 2 Forward *BRAF*	GG[**FWD BC**]CGAACAGTCTCCAAATAGGTGATTTT GGTCTAGC	9
Round 2 Reverse *BRAF*	GC[**REV BC**]GGATGAGTGCAGCCAGACAACTGTTCAA ACTGA	10
Round 2 Forward *EGFR*	GG[**FWD BC**]CGAACAGTCTCCCAGCATGTCAAGAT CACAGATT	11
Round 2 Reverse *EGFR*	GC[**REV BC**]GGATGAGTGCAGGCATGGTATTCTTT CTCTTCC	12

* Primers were gel-purified prior to use.
FWD BC = Forward barcode
REV BC = Reverse barcode

In a first round of PCR, all hotspot regions from a given sample were amplified in a multiplexed fashion. Three aliquots of purified plasma DNA from each sample were used as templates in three identical multiplexed PCRs containing 1x Kapa Fidelity buffer (Kapa Biosystems, Inc., Woburn, MA), 300 µM each dNTP, 50 nM each primer (Round 1 Forward and Reverse *KRAS*, *BRAF*, and *EGFR* primers), and 1 unit/50 µL HiFi Hotstart DNA polymerase (Kapa Biosystems).

Mineral oil was added to all PCR tubes to minimize evaporation during heating. Temperature cycling parameters were 95°C for 2 minutes, followed by 35 cycles of 98°C for 20 sec, 64°C for 20 sec, and 72°C for 30 sec. A final extension was performed at 72°C for 1 minute, prior to cooling the reaction at 4°C. EDTA was then added at a final concentration of 5 mM to stop polymerase activity.

The amplicons from each first round PCR were diluted 5000-fold and used as templates for 3 separate second round PCRs to individually amplify the hotspot regions of *KRAS*, *BRAF*, or *EGFR*. To promote specific amplification, the second-round primers were nested relative to the primers used in the first round of PCR. The nested primers were labeled with sample-specific barcode sequences to allow multiplexed sequencing of DNA from many samples. The barcode sequences were 6 nucleotides in length, and were designed to differ from all other barcodes in the set at a minimum of 2 positions so that a single sequencing error would not lead to misclassification of samples. Different combinations of 16 forward and 16 reverse barcoded primers could be used to uniquely identify up to 256 different samples. PCR was carried out using the same reaction conditions as were used in the first round, with the following modifications: the annealing temperature was increased to 65°C, and the 3 pairs of multiplexed primers were replaced with a single pair of barcoded primers (Round 2 Forward and Reverse *KRAS*, *BRAF*, or *EGFR* primers listed in Table 1) at a final concentration of 200 nM each. After addition of 5 mM EDTA, the PCR products were mixed together to produce 3 pools, one for each of the 3 replicate reactions. All PCR steps were carried out using a high-fidelity polymerase (HiFi HotStart, Kapa Biosystems).

Production of barcoded PCR primers

In order to build flexibility and scalability into the design of the deep sequencing scheme, barcoded oligonucleotides and gene-specific PCR primers were combined in a modular fashion, as illustrated in Fig. 9. A set of 16 unique barcodes was produced for the forward primers, and a different set of 16 barcodes was produced for the reverse primers. These barcodes were 6 nucleotides in length, and were designed to differ from all other barcodes in the set by at least 2 nucleotides (to minimize the probability that miscalled bases would cause misclassification of sequences). Each barcode was incorporated into an oligonucleotide, which was primer-extended using a partially complementary single-stranded template containing the reverse-complement of the gene-specific primer sequence. The sequences of the template and barcode-containing oligonucleotides are listed in Table 2.

Table 2: Oligonucleotides used to produce barcoded primers.

Oligonucleotide	Sequence (5' – 3')	SEQ ID NO:
Forward *KRAS* template	Biotin-CCAACTACCACAAGTTTATATTCGGAGACTGTTCG	13
Forward *EGFR* template	Biotin-AATCTGTGATCTTGACATGCTGGGAGACT GTTCG	14
Forward *BRAF* template	Biotin-GCTAGACCAAAATCACCTATTTGGAGAC TGTTCG	15
Forward barcode 1 oligo	GG**AACCTT**CGAACAGTCTCC	16
Forward barcode 2 oligo	GG**AACGTA**CGAACAGTCTCC	17
Forward barcode 3 oligo	GG**AAGCAT**CGAACAGTCTCC	18
Forward barcode 4 oligo	GG**AAGGAA**CGAACAGTCTCC	19
Forward barcode 5 oligo	GG**ATCCAT**CGAACAGTCTCC	20
Forward barcode 6 oligo	GG**ATCGAA**CGAACAGTCTCC	21
Forward barcode 7 oligo	GG**ATGCAA**CGAACAGTCTCC	22
Forward barcode 8 oligo	GG**ATGGTA**CGAACAGTCTCC	23
Forward barcode 9 oligo	GG**TACCTA**CGAACAGTCTCC	24
Forward barcode 10 oligo	GG**TACGAA**CGAACAGTCTCC	25
Forward barcode 11 oligo	GG**TACGTT**CGAACAGTCTCC	26
Forward barcode 12 oligo	GG**TAGCTT**CGAACAGTCTCC	27
Forward barcode 13 oligo	GG**TAGGAT**CGAACAGTCTCC	28
Forward barcode 14 oligo	GG**TTCGAT**CGAACAGTCTCC	29
Forward barcode 15 oligo	GG**TTGCAT**CGAACAGTCTCC	30
Forward barcode 16 oligo	GG**TTGCTA**CGAACAGTCTCC	31
Reverse *KRAS* template	Biotin-C**TT**GACGATACAGCTAATTCACTGCACTCATCC	32
Reverse *EGFR* template	Biotin-GGAAGAGAAAGAATACCATGCCTGCACTCATCC	33
Reverse *BRAF* template	Biotin-TCAGTTTGAACAGTTGTCTGGCTGCACTCATCC	34
Reverse barcode 1 oligo	GC**AATCAA**GGATGAGTGCAG	35
Reverse barcode 2 oligo	GC**AATGAT**GGATGAGTGCAG	36

Reverse barcode 3 oligo	GC**AAGATA**GGATGAGTGCAG	37
Reverse barcode 4 oligo	GC**AACATT**GGATGAGTGCAG	38
Reverse barcode 5 oligo	GC**ATCATA**GGATGAGTGCAG	39
Reverse barcode 6 oligo	GC**ATAGTT**GGATGAGTGCAG	40
Reverse barcode 7 oligo	GC**ATCAA**TGGATGAGTGCAG	41
Reverse barcode 8 oligo	GC**ATGTAT**GGATGAGTGCAG	42
Reverse barcode 9 oligo	GCT**AACA**TGGATGAGTGCAG	43
Reverse barcode 10 oligo	GC**TAGTAA**GGATGAGTGCAG	44
Reverse barcode 11 oligo	GC**TATGTA**GGATGAGTGCAG	45
Reverse barcode 12 oligo	GC**TTACAA**GGATGAGTGCAG	46
Reverse barcode 13 oligo	GC**TTCATT**GGATGAGTGCAG	47
Reverse barcode 14 oligo	GC**TTAGTA**GGATGAGTGCAG	48
Reverse barcode 15 oligo	GC**TTCTAA**GGATGAGTGCAG	49
Reverse barcode 16 oligo	GC**TACAA**TGGATGAGTGCAG	50

Barcode sequences are boldfaced and underlined.

Each forward barcode oligo (8 μM) was annealed to each forward template oligo (8 μM) in separate reaction tubes containing 1x NEBuffer 2 (New England Biolabs, Ipswich, MA), 200 μM each dNTP, and 1 mM dithiothreitol. Annealing was carried out by heating the solution to 95°C for 2 minutes, 60°C for 1 minute, and then slowly cooling to 25°C over approximately 15 minutes. All possible combinations of forward barcode and template oligos were produced. The set of reverse oligos were annealed in a similar manner. 1 unit/10 μL of DNA polymerase I, Large (Klenow) Fragment (New England Biolabs) was added to each tube, and the reaction was incubated at 25°C for 30 minutes. The reaction was stopped by adding 25 mM ethylenediaminetetraacetic acid (EDTA) and heating to 75°C for 20 minutes. A biotin tag attached to the 5'-end of the template oligonucleotide was used to purify the primer-extended products from the reaction mix by binding to high capacity streptavidin-coated agarose resin (ThermoFisher Scientific, Wilmington, MA) (5 μL resin slurry added per 50 μL reaction). The resin particles were agitated constantly in the solution at room temperature for 8 hours. The resin was washed three times in buffer containing 10 mM Tris pH 7.6 and 50 mM NaCl. The barcoded PCR primers were then released from the resin-bound

template oligos into a fresh 40 µL volume of the same buffer by heat denaturation at 95°C for 1 minute. After concentration adjustment, the primers were ready for use in PCR.

Analysis of cell line DNA

Genomic DNA was purified from human cancer cell lines using the same method used for purifying plasma DNA, after suspending cells in 0.2 mL of phosphate-buffered saline. The following cell lines were used: A549 (having a KRAS Gly12Ser mutation), H1957 (having an EGFR Leu858Arg mutation), and YUSAC (having a BRAF Val600Glu mutation). Cells were passed in culture for no more than 6 months after being thawed from original stocks. Because cell lines were used only for analysis of short regions of genomic DNA, authentication of lines by our laboratory was limited to sequencing of those regions. To test the performance of the deep sequencing method for a particular gene, DNA derived from cells known to be either mutant or wild-type with respect to that gene was mixed in various ratios between 10,000:1 and 1:10,000. Cell line DNA samples were then amplified and sequenced according to the same methods that were used for plasma DNA.

Ultra-deep sequencing

Barcoded PCR products from all samples were mixed to produce 3 separate pools, each corresponding to one set of replicate reactions. Uniquely indexed TruSeq® adapters (Illumina, Inc., San Diego, CA) were ligated to each of the 3 pools of PCR amplicons using a modified version of the manufacturer's protocol. Amplicon pools were purified by phenol-chloroform-isoamyl alcohol (PCA, Sigma-Aldrich Co., St. Louis, MO) extraction followed by ethanol precipitation. Addition of deoxyadenosine to the 3'-ends of the blunt-ended amplicons was performed according to Illumina's recommendations. PCA extraction and ethanol precipitation were again used for purification. TruSeq adapters were ligated and the products were purified on a 2% agarose gel according to the standard protocol. DNA concentration was estimated using a Bioanalyzer 2100 (Agilent Technologies, Santa Clara, CA). Without further amplification, the 3 pools were combined and loaded onto a single lane of an Illumina HiSeq® 2000 instrument. Prior to loading, the samples were diluted by adding between 2- and 8-fold excess Phi-X DNA to improve cluster discrimination. Sequencing was carried out in multiplexed, 75 base pair, paired-end mode at the Yale Center for Genomic Analysis.

Data analysis

A computer script was written to filter, assort, align, and count millions of paired-end sequences. First, a read-pair was assigned to a data bin based on the barcode of each read in the

pair. Then, based on PCR primer sequences, the pair was assigned to one of the reference genes. Next, the longest stretch of perfect sequence agreement between each pair of reads was determined, and this was used to align the reads to the reference sequence for the gene. A read pair was discarded if either member did not pass Illumina filtering or a nucleotide was reported to be "."; if there was an inconsistency in barcodes, strands, or PCR tags; or if their region of perfect sequence agreement was less than 36 nucleotides in length. Finally, variant sequences confirmed by reads from both strands were identified and counted within each data bin based on comparison to the reference sequence. A module used to perform sequence alignments using a Smith-Waterman algorithm was taken, with permission, from Dr. Conrad Huang, Resource for Biocomputing, Visualization & Informatics, University of California, San Francisco. A module used to determine the longest common substring was taken from a web resource.

Confirmation of mutations in tumor tissue

Genomic DNA was isolated from paraffin-embedded tumor tissue samples using the QuickExtract™ FFPE DNA Extraction Kit (Epicentre Biotechnologies, Madison, WI). Mutation hotspot regions of *KRAS*, *BRAF*, and *EGFR* were amplified using the same PCR primers that were used in the first round of PCR described above. Sanger sequencing was performed on gel-purified amplicons, and mutations were identified from chromatograms using Mutation Surveyor software (SoftGenetics LLC, State College, PA).

Determining the absolute concentration of mutant DNA in plasma.

Real-time quantitative PCR was used to measure the concentration of *KRAS* DNA fragments in each patient's plasma sample. This value was multiplied by the fraction of mutant molecules as determined by deep sequencing in order to calculate the absolute mutant *KRAS* DNA concentration. PCR conditions were the same as those used in the first round of amplification described above except for the use of a single pair of primers (Round 1 *KRAS* Fwd and Rev) at 200 nM final concentration, and the addition of SYBR® Green dye (Stratagene, La Jolla, CA) at 1:60,000 final dilution. Amplification was carried out using an IQ5 Real-time PCR Detection System with version 2.1 software (Bio-Rad Laboratories, Hercules, CA). To enable determination of absolute copy numbers, a standard curve was generated using known concentrations of a cartridge-purified oligonucleotide that was designed to mimic the fragment of *KRAS* DNA being amplified from plasma. The sequence of the oligonucleotide was: 5'-AAGGCCTGCTGAAAATGACTGAATATAAACTTGTGGTAGATGGAGCTGGTGGCGTAAGCAAGAGTG CCTTGACGATACAGCTAATTCAGAATCATTTTGTGGACGAATA-3' (SEQ ID No: 51). Real-time PCRs were

performed in triplicate, and the *KRAS* DNA concentration was determined using the mean of the 3 measurements.

RESULTS

Error suppression reveals low-abundance variants

To determine the relative abundance of tumor-specific mutations, massively parallel sequencing was performed on PCR amplicons derived from plasma DNA fragments containing known mutation hotspots. Thousands of clonal sequence reads from each plasma sample were compared to reference sequences in order to identify and quantify variants. For proof of concept, analysis was restricted to frequently mutated codons within 3 oncogenes that commonly develop somatic mutations in various malignancies: codons 12 and 13 of *KRAS*, codon 600 of *BRAF*, and codon 858 of *EGFR*. By designing PCR primers that flank very short regions (<50 bp) surrounding these mutation hotspots, adequate amplification of highly fragmented plasma DNA could be ensured and greater sequence depth could be achieved. Modular attachment of DNA barcode tags to the 5'-ends of the PCR primers allowed sequencing of up to 256 DNA samples in batch (Fig. 4A and Fig. 9). A median depth of 108,467 read pairs was obtained per mutation site per sample after filtering and de-multiplexing a total of 86,359,980 raw sequences generated on a single lane of an Illumina HiSeq® 2000 flow cell.

Importantly, the design of short PCR amplicons enabled us to devise a sequencing strategy that could distinguish mutant from wild-type DNA molecules with very high confidence. Illumina's paired-end sequencing mode was modified to achieve partial overlap of 75 base-pair bidirectional reads obtained sequentially from the forward and reverse strands of each clonal DNA cluster on the flow cell (Fig. 4B). Mutation hotspots were included in the overlapping sequence region so that the hotspot within each clone would be read from one strand and then proofread from the opposite strand. By discarding clones that did not have perfect sequence agreement between the two paired-end reads, the vast majority of sequencer-generated errors were eliminated. Imperfect sequence agreement was found in 22% of read pairs that had already passed Illumina's chastity filter. A median error frequency of 0.31% per base was observed when directly comparing single reads derived from either strand of wild-type control samples to known reference sequences. The frequency of such errors was reduced to 0.07% per base in the region of overlap after removing read pairs that lacked sequence consistency.

Any remaining errors were highly unlikely to be caused by coincidentally consistent misreads from opposite ends of a clone. Rather, most of these errors were probably present within the DNA molecules being sequenced, introduced by polymerase misincorporations or DNA damage. To

further discriminate true mutations from such errors, all amplification and processing steps were performed in triplicate, and the mean of the three mutation counts was determined. This was done based on the premise that true mutations would be reproducibly counted in all three instances, whereas counts from randomly occurring errors would be more variable (recognizing that the distribution of errors is not entirely random). Using this approach, the frequency of miscalls of specific mutations from known wild-type samples was reduced to a median value of 0.014% (interquartile range [IQR]: 0.0052% to 0.023%; Table 3). Suppression of errors in this manner permitted rare mutations to be identified with a high degree of certainty (Fig. 5).

Table 3: Background level of spurious mutation counts obtained from known wild-type samples.

Mutation Type	Fraction of mutant : wild-type counts (from 3 replicate PCRs)
BRAF Val600Glu	0.00013
EGFR Leu858Arg	0.000047
KRAS Gly12Ser	0.00029
KRAS Gly12Val	0.00018
KRAS Gly12Arg	0.000057
KRAS Gly12Asp	0.00015
KRAS Gly12Ala	0.000014
KRAS Gly12Cys	0.00025
KRAS Gly13Ser	0.00015
KRAS Gly13Val	0.00029
KRAS Gly13Arg	0.000050
KRAS Gly13Asp	0.000044
KRAS Gly13Ala	0.000058
KRAS Gly13Cys	0.00049
Median Value	**0.00014**
Interquartile range	**0.000052 to 0.00023**

Sensitive and accurate quantitation of mutant DNA

Next, mutant and wild-type DNA levels were measured over a broad range of relative concentrations. Genomic DNA from *KRAS*-, *BRAF*-, or *EGFR*-mutant cancer cell lines was mixed in different ratios, and then subjected to amplification and deep sequencing. Mutant DNA could be accurately and reproducibly measured in a linear manner over approximately 8 orders of magnitude and down to levels as low as 1 in 10,000 molecules (Fig. 6). Also, by testing combinations of DNA from multiple mutant cell lines, the assay was able to simultaneously quantify more than one mutation from a given sample.

Monitoring ctDNA levels in cancer patients

To compare with clinical samples, plasma collected from patients with non-small cell lung cancer (NSCLC) at various times before, during, or after treatment was analyzed. Patients were enrolled in the study (and their plasma DNA was tested) without prior knowledge of the mutation status of their tumors. A total of 117 samples were obtained from 30 patients (17 patients with adenocarcinoma, 9 with undifferentiated NSCLC, and 4 with squamous cell carcinoma). *KRAS* Gly12Asp, Gly12Val, Gly12Cys, or Gly13Asp point-mutations were detectable in the plasma DNA of 6 patients out of 26 with adenocarcinoma or undifferentiated NSCLC. As expected, no *KRAS* mutations were found in specimens from patients with squamous cell carcinoma. *BRAF* and *EGFR* mutations were not detectable in any plasma samples. This was somewhat surprising for *EGFR,* which has a reported prevalence of activating mutations in NSCLC of approximately 10% (Lynch et al., *N Engl J Med*. 2004; 350: 2129-2139; Paez et al., *Science*. 2004; 304: 1497-1500; Pao et al., *Proc. Natl. Acad. Sci. USA*. 2004; 101: 13306-13311). However, evaluation of 21 available tumor tissue specimens confirmed the absence of *EGFR* mutations in this population (mutations occurring outside of the sequenced hotspot region may have been missed). The presence or absence of *KRAS* mutations in all tested tumor samples was tested to be concordant with the findings in plasma: 5 patients had identical *KRAS* mutations in both tumor and plasma, and 16 patients had no *KRAS* mutations detected from either source. Tumor tissue was unavailable or insufficient for 1 patient with mutant *KRAS* in the plasma, and for 4 patients with no plasma mutations. Table 4 lists the clinical characteristics and mutation findings for all enrolled patients.

Table 4. Clinical characteristics and mutation findings.

				Plasma Samples		Tumor Tissue			
Patient No.	**Sex**	**Age**	**Stage**	**Mutation Type***	**No. of Samples**	**NSCLC Histology**	**Tissue Source**	**Tissue Mutation**	**Method of Mutation Testing**
1	M	82	IV	*KRAS* WT *EGFR* WT *BRAF* WT	2	Adeno	Cells in pleural fluid	*KRAS* WT *EGFR* WT	Clinical lab
2	M	68	IV	*KRAS* WT *EGFR* WT *BRAF* WT	2	Adeno	Lung core Bx	Tissue not available from outside hospital	
3	F	51	IV	***KRAS* Gly12Asp** *EGFR* WT *BRAF* WT	3	Adeno	Tracheal Bx	***KRAS* Gly12Asp** *EGFR* WT *BRAF* WT	Sanger seq.
4	M	71	IIIB	*KRAS* WT *EGFR* WT *BRAF* WT	12	Squam	Para-tracheal lymph node Bx	Not tested (squamous histology)	

5	M	44	IV	***KRAS* Gly12Val** *EGFR* WT *BRAF* WT	5	Adeno	Lung core Bx	***KRAS* Gly12Val** *EGFR* WT *BRAF* WT	Sanger seq.
6	M	68	Lung: IA Prost.: II Esoph. : III	*KRAS* WT *EGFR* WT *BRAF* WT	13	Adeno	Lung core Bx	Excess tissue not available	
7	M	59	IIIA	*KRAS* WT *EGFR* WT *BRAF* WT	8	Squam	Lung core Bx	Not tested (squamous histology)	
8	F	70	IIIA	*KRAS* WT *EGFR* WT *BRAF* WT	3	Adeno	Lung core Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
9	M	72	IIIB	***KRAS* Gly12Val** *EGFR* WT *BRAF* WT	4	Undiff	Bronchial brushing	***KRAS* Gly12Val** *EGFR* WT *BRAF* WT	Sanger seq.
10	F	62	Lung: IV Breast : I	*KRAS* WT *EGFR* WT *BRAF* WT	1	Lung Adeno and Breast Adeno	Iliac wing core Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
11	F	79	IV	***KRAS* Gly12Val** *EGFR* WT *BRAF* WT	1	Adeno	Lung fine needle aspirate	***KRAS* Gly12Val** *EGFR* WT *BRAF* WT	Sanger seq.
12	M	69	IV	*KRAS* WT *EGFR* WT *BRAF* WT	3	Adeno	Scapula mass Bx	*KRAS* WT *EGFR* WT	Clinical lab
13	F	61	Lung: IV Breast : I	*KRAS* WT *EGFR* WT *BRAF* WT	3	Undiff	Pre-tracheal lymph node needle aspirate	*KRAS* WT *EGFR* WT	Clinical lab
14	M	77	IV	***KRAS* Gly12Cys** *EGFR* WT *BRAF* WT	2	Adeno	Calf mass excision	***KRAS* Gly12Cys** *EGFR* WT *BRAF* WT	Sanger seq.
15	M	65	IV	***KRAS* Gly13Asp** *EGFR* WT *BRAF* WT	2	Undiff	Bronchial brushing	Excess tissue not available	
16	F	73	IV	*KRAS* WT *EGFR* WT *BRAF* WT	3	Undiff	Bronchial Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
17	F	65	IA	*KRAS* WT *EGFR* WT *BRAF* WT	5	Adeno	Lung core Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
18	F	77	IV	*KRAS* WT *EGFR* WT *BRAF* WT	1	Adeno	Lung core Bx	Excess tissue not available	
19	F	75	IV	*KRAS* WT *EGFR* WT	2	Adeno	Bronchial Bx	*KRAS* WT *EGFR* WT	Clinical lab

				BRAF WT					
20	M	73	IB	*KRAS* WT *EGFR* WT *BRAF* WT	5	Squam	Lung lobect-omy	Not tested (squamous histology)	
21	M	73	IIB	*KRAS* WT *EGFR* WT *BRAF* WT	4	Adeno	Lung core Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
22	F	68	IV	*KRAS* WT *EGFR* WT *BRAF* WT	3	Undiff	Lung tumor excision	*KRAS* WT *EGFR* WT	Clinical lab
23	F	79	IA	*KRAS* WT *EGFR* WT *BRAF* WT	3	Undiff	Lung core Bx	Tissue not available from outside hospital	
24	M	64	IIIB	*KRAS* WT *EGFR* WT *BRAF* WT	8	Squam	Lung core Bx	Not tested (squamous histology)	
25	F	73	Locally recur. IB	*KRAS* WT *EGFR* WT *BRAF* WT	8	Undiff	Lung lobect-omy	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
26	F	63	IIIB	*KRAS* WT *EGFR* WT *BRAF* WT	1	Adeno	Lung core Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
27	F	74	IIIA	*KRAS* WT *EGFR* WT *BRAF* WT	4	Undiff	Para-tracheal lymph node Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.
28	F	61	IV	*KRAS* WT *EGFR* WT *BRAF* WT	3	Adeno	Spine Met Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq. and Clinical lab
29	F	82	IV	*KRAS* WT *EGFR* WT *BRAF* WT	2	Undiff	Lung core Bx	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq. and Clinical lab
30	F	69	Lung: IV Breast : IV	*KRAS* WT *EGFR* WT *BRAF* WT	1	Adeno	Lung fine needle aspirate	*KRAS* WT *EGFR* WT *BRAF* WT	Sanger seq.

The list is ordered by date of first specimen collection.

* Plasma DNA was only tested for mutations at codons 12 and 13 of *KRAS*, 858 of *EGFR*, and 600 of *BRAF*.

Squam = Squamous cell carcinoma

Adeno = Adenocarcinoma

Undiff = Undifferentiated NSCLC (not otherwise specified)

WT = Wild-type

Bx = Biopsy

Sanger Seq. = Direct Sanger sequencing of tissue-derived PCR amplicons by our laboratory.

Clinical lab = Mutations tested for clinical purposes in a laboratory certified under the Clinical Laboratory Improvement Amendments of 1988 (CLIA). Tissue was not tested for *BRAF* mutations by clinical laboratories because of low prevalence in NSCLC.

For patients with detectable plasma DNA mutations, changes in measured ctDNA levels were followed in the context of therapeutic interventions or disease progression. To determine the absolute concentration of mutant *KRAS* DNA fragments in a plasma sample, the total concentration of *KRAS* fragments was measured by real-time PCR and then multiplied by the fraction of mutant molecules determined by deep sequencing. The median concentration among samples with detectable mutations was 5,694 mutant KRAS molecules per mL (IQR: 2,655 to 25,123). Time-courses of mutant ctDNA measurements for patients who had 3 or more samples collected are shown in Fig. 7 (data for patients with fewer measurements are shown in Fig. 8). In two cases, the ctDNA level decreased upon treatment with radiation and/or systemic therapy. Aggressive progression of metastatic disease in a different patient was accompanied by a substantial rise in ctDNA. In another two cases, ctDNA levels increased shortly after initiating treatment, perhaps because more tumor DNA was released into the bloodstream as cancer cells were being killed.

Example 2

This example includes methods that incorporate elements of Example 1, but also includes several modifications. (Figs. 10 and 11). In this example, 40 different genomic target regions were analyzed. Of the 40 genomic target regions, 38 were prone to developing somatic mutations, and 2 were included as controls that were not expected to be mutated.

Preparation of mixtures of primers having combinations of modular oligonucleotide segments

As described previously, early tagging of targeted DNA template molecules required the production of mixtures of primers having a common barcode in their 5' region, and having several different gene-specific primer segments at their 3' end. Herein modular oligonucleotide segments were combined during oligonucleotide synthesis on an automated synthesizer, "modular automated synthesis and purification", and the approach is illustrated in (Fig. 3).

Each different gene-specific 3'-portion was synthesized on separate oligonucleotide synthesis columns. Standard phosphoramidite chemistry was used, and the oligonucleotides were grown on a solid support. Both polystyrene and controlled-pore-glass were used as solid supports, but polystyrene was preferable. Both types of supports performed similarly. The solid support consisted of small particles that appeared as a powder. The powder was contained within an oligonucleotide synthesis column, sandwiched loosely between two frits. Multiple different 3'-segments were grown (oligomerized by chemical coupling of phosphoramidite monomers) in separate synthesis columns on an automated synthesizer in the 3' to 5' direction. The synthesis was

paused, and partially synthesized oligonucleotides were left on the column in the protected state with the trityl group left on.

"Pipette tip"-style oligonucleotide synthesis columns were utilized with sufficient controlled-pore glass (1000 angstrom pore size) or polystyrene to synthesize oligos at the 40 nanomole or 200 nanomole scale (3-Prime, Aston, PA). Forty different partial 3' oligonucleotide segments were synthesized on 40 separate columns using a Dr. Oligo 192 automated synthesizer. The oligonucleotides were not cleaved from the solid supports, were not deprotected, and the trityl group was left on so that further synthesis could be continued. The sequences of these 40 different 3' segments are listed in Table 5.

Table 5. List of forty 3' oligonucleotide segments synthesized in separate columns for first phase of modular automated synthesis.

Name	DNA Sequence	SEQ ID NO:
3'segment1	AGACGTGTGCTCTTCCGATCTNNNNNNCTGTGCTGTGACTGCTTG	52
3'segment2	AGACGTGTGCTCTTCCGATCTNNNNNNTAGCACATGACGGAGGTT	53
3'segment3	AGACGTGTGCTCTTCCGATCTNNNNNNACAAATACTCCACACGCAAATT	54
3'segment4	AGACGTGTGCTCTTCCGATCTNNNNNNATATTTGGATGACAGAAACACTT	55
3'segment5	AGACGTGTGCTCTTCCGATCTNNNNNNCTGTGATGATGGTGAGGATGG	56
3'segment6	AGACGTGTGCTCTTCCGATCTNNNNNNCTGGGACGGAACAGCTTTGAG	57
3'segment7	AGACGTGTGCTCTTCCGATCTNNNNNNTGCAATTTCTACACGAGATCCTCT	58
3'segment8	AGACGTGTGCTCTTCCGATCTNNNNNNTCTTTGGAGTATTTCATGAAACAAAT GA	59
3'segment9	AGACGTGTGCTCTTCCGATCTNNNNNNAACAGTAAAAATAGGTGATTTTGGTC TA	60
3'segment10	AGACGTGTGCTCTTCCGATCTNNNNNNTGCAACTACTGGACGCTGGAC	61
3'segment11	AGACGTGTGCTCTTCCGATCTNNNNNNCTCAATTTTGTTTCAGGACCTGCT	62
3'segment12	AGACGTGTGCTCTTCCGATCTNNNNNNCTGGCAGCAACAGTCTTACCT	63
3'segment13	AGACGTGTGCTCTTCCGATCTNNNNNNACCC AGC TTG GAG GCT GC	64
3'segment14	AGACGTGTGCTCTTCCGATCTNNNNNNAGCCAGGCCGCTGAAGACA	65
3'segment15	AGACGTGTGCTCTTCCGATCTNNNNNNGGCAATTCACTGTAAAGCTGGAAAG	66
3'segment16	AGACGTGTGCTCTTCCGATCTNNNNNNATGAAGATATATTCCTCCAATTCAGG AC	67
3'segment17	AGACGTGTGCTCTTCCGATCTNNNNNNGCGTTTCCTTTAACCACATAATTAGA ATC	68
3'segment18	AGACGTGTGCTCTTCCGATCTNNNNNNGTTTTCCCTTTCTCCCCACAG	69
3'segment19	AGACGTGTGCTCTTCCGATCTNNNNNNGTTCCTGTAGCAAAACCAGAAATC	70
3'segment20	AGACGTGTGCTCTTCCGATCTNNNNNNCGGTGAGAAAGTTAAAATTCCCGTC	71
3'segment21	AGACGTGTGCTCTTCCGATCTNNNNNNAAGCATGTCAAGATCACAGATTTTG	72
3'segment22	AGACGTGTGCTCTTCCGATCTNNNNNNCTCACCTCCACCGTGCAGCT	73
3'segment23	AGACGTGTGCTCTTCCGATCTNNNNNNGACCACCCGCACGTCTGT	74
3'segment24	AGACGTGTGCTCTTCCGATCTNNNNNNTCTTCCATACTTGATTCATGATATTTT ACT	75
3'segment25	AGACGTGTGCTCTTCCGATCTNNNNNNGACCTCCTCAAACAGCTCAAAC	76
3'segment26	AGACGTGTGCTCTTCCGATCTNNNNNNATGGGAGATCTTCACGCTGG	77
3'segment27	AGACGTGTGCTCTTCCGATCTNNNNNNTCCCTGAGCGTCATCTGCC	78
3'segment28	AGACGTGTGCTCTTCCGATCTNNNNNNCGCTGGTGGAGGCTGACGA	79
3'segment29	AGACGTGTGCTCTTCCGATCTNNNNNNGTTCCCTATCAAATATGTCAACGACT	80
3'segment30	AGACGTGTGCTCTTCCGATCTNNNNNNAATTTTGGTCTTGCCAGAGACA	81

3'segment31	AGACGTGTGCTCTTCCGATCTNNNNNNTATCGACTCCACCGAGGTCA	82
3'segment32	AGACGTGTGCTCTTCCGATCTNNNNNNATACTTGGAGGACCTGCACG	83
3'segment33	AGACGTGTGCTCTTCCGATCTNNNNNNGTCGTCAAGGCACTCTTGCCT	84
3'segment34	AGACGTGTGCTCTTCCGATCTNNNNNNCGATATTCTCGACACAGCAGGT	85
3'segment35	AGACGTGTGCTCTTCCGATCTNNNNNNATCAGTGCGCTTTTCCCA	86
3'segment36	AGACGTGTGCTCTTCCGATCTNNNNNNTGACATACTGGATACAGCTGGA	87
3'segment37	AGACGTGTGCTCTTCCGATCTNNNNNNGTGGTCAGCGCACTCTTGCCC	88
3'segment38	AGACGTGTGCTCTTCCGATCTNNNNNNTCATCCTGGATACCGCCGGC	89
3'segment39	AGACGTGTGCTCTTCCGATCTNNNNNNATCCTGTTTATAATATTGACAAAACA CCT	90
3'segment40	AGACGTGTGCTCTTCCGATCTNNNNNNATCAGGACAAAGTCCGGATTGA	91

These oligonucleotides were synthesized at the 200 nanomole scale, with the oligo left on the column in the protected state with the trityl group left on. Positions marked "N" have equal probability of being A, C, G, or T.

The solid supports of all 40 partially synthesized oligonucleotides were dried by blowing argon gas through the columns, and then the controlled-pore glass or polystyrene powder from all 40 columns was mixed by pouring the contents of each column (after cutting the tops off of the columns) into a common container (such as a glass vial). The solid support particles were then suspended in a solvent of similar density so that the particles could be thoroughly mixed and then the mixture could be dispensed into fresh oligonucleotide synthesis columns. When using polystyrene supports, a 3:1 mixture of dichloromethane : acetonitrile was used as the suspension liquid, and when using controlled-pore glass supports, a 5:1 mixture of 1,2-dibromoethane : acetonitrile was used as the suspension liquid. The particles were maintained as a uniform slurry in the liquid by constantly swirling or agitating the vial while using a pipette to dispense equal volumes of the slurry into fresh columns (with the bottom frit already in place). The slurry was dispensed into 96 fresh columns. The particles settled onto the frits, while the liquid drained out from the bottom of the columns by gravity. To ensure that the particles had all settled onto the frit, the columns were filled with acetonitrile and this was again allowed to drain out from the bottom by gravity. After the acetonitrile had fully drained out, the top frits were put in place to secure the powder into the columns.

The new columns were then placed back on the automated synthesizer, and the oligonucleotide synthesis was continued. Each column was assigned a different barcode sequence that was incorporated into the 5' oligonucleotide segment. A "dummy base" was added to the 3' end of the 5' segment sequence when programming the synthesizer in order to account for the partially synthesized oligonucleotides that were already present on the solid supports. The sequences of the 96 different 5' segments consisted of the following common sequence with each of 96 different barcodes inserted in the position marked [BC1-96]. One unique barcode was used per oligonucleotide synthesis column.

5'-CGAGACGGATCAAGCAGAAGACGGCATACGAGATNN[BC1-96]GTGACTGGAGTTC(T)-3' (SEQ ID NO:92)

The "T" in parentheses at the 3'-end of the sequence is the "dummy base". The 96 barcodes that were used are listed in Table 6. The automated synthesizer was programmed to carry out synthesis at a 40 nmole scale (which determines the volume of reagents passed through the columns), although the actual amount of solid support in each column was likely to produce less than 40 nmoles of oligonucleotides.

Table 6. List of 96 sample-specific barcodes

Barcode #	Sequence (BBBBBBBB)
1	CCGATATT
2	GCCATATT
3	TCTGGATT
4	ACTCGATT
5	TCACGATT
6	ACTGCATT
7	TCAGCATT
8	TCTCCATT
9	ATTGGTGT
10	TTAGGTGT
11	ATACGTGT
12	ATAGCTGT
13	ATTCCTGT
14	TTACCTGT
15	CTCTATGT
16	CTCTTAGT
17	CTGATAGT
18	GTCATAGT
19	ATAGGAGT
20	ATTCGAGT
21	TTACGAGT
22	ATTGCAGT
23	TTAGCAGT
24	ATACCAGT
25	AGTGGTCT
26	TGAGGTCT
27	TGTCGTCT
28	TGTGCTCT
29	AGTCCTCT
30	TGACCTCT
31	CGCTATCT
32	CGCTTACT
33	CGGATACT
34	GGCATACT

35	TGTGGACT
36	AGTCGACT
37	TGACGACT
38	AGTGCACT
39	TGAGCACT
40	TGTCCACT
41	TCATTGTG
42	TCTATGTG
43	ACTATCTG
44	ACTTACTG
45	TCATACTG
46	ATTATCGG
47	TTATACGG
48	TGATTGCG
49	TGTATGCG
50	AGTATCCG
51	AGTTACCG
52	TGATACCG
53	ACTATGTC
54	ACTTAGTC
55	TCATAGTC
56	TCATTCTC
57	TCTATCTC
58	ATTATGGC
59	TTATAGGC
60	TTATTCGC
61	AGTATGCC
62	AGTTAGCC
63	TGATAGCC
64	TCTGGTTA
65	TCACGTTA
66	TCAGCTTA
67	TCTCCTTA
68	CCGTATTA
69	GCCTATTA
70	CCGTTATA
71	GCCTTATA
72	TCAGGATA
73	TCTCGATA
74	TCTGCATA
75	TCACCATA
76	CTGATTGA
77	GTCATTGA
78	TTACGTGA
79	TTAGCTGA
80	CTGTATGA
81	GTCTATGA
82	CGGATTCA

83	GGCATTCA
84	TGTGGTCA
85	TGACGTCA
86	TGAGCTCA
87	TGTCCTCA
88	CGGTATCA
89	GGCTATCA
90	CGGTTACA
91	GGCTTACA
92	CGCATACA
93	TGAGGACA
94	TGTCGACA
95	TGTGCACA
96	TGACCACA

After completion of the second phase of the modular synthesis, the oligos were cleaved off the solid supports with the trityl group still left on. They underwent rapid deprotection followed by purification on a separate Glen-Pak DNA reverse-phase cartridge for each of the 96 oligonucleotide mixtures (Glen Research, Sterling, VA). The trityl group at the 5'-end of completed oligonucleotides was selectively retained by the cartridge, enriching for full-length products and removed failure sequences that did not contain the trityl group. The trityl group was removed upon completion of purification. The purified oligonucleotides were then dried and re-suspended in 10 mM Tris pH 7.6 to produce a 33 micromolar working stock solution. Polyacrylamide gel purification was used in some cases to further purify the full-length oligonucleotides.

Collection and processing of patient plasma samples

Blood was collected by venipuncture into a vacuum tube containing potassium-EDTA. Various tube sizes were used, typically between 3mL and 10mL. Blood was inverted in the tube several times at the time of collection to ensure even mixing of the K_2-EDTA. Samples were stored temporarily and transported at room temperature (20-25°C) prior to separation of plasma. Plasma was separated and frozen as soon as possible after blood collection, preferably within 3 or 4 hours. The collection tubes were centrifuged at 1000 x g for 10 minutes in a clinical centrifuge with a swinging bucket rotor with slow acceleration and deceleration (brake off). Plasma was removed from the red blood cells and buffy coat using a 1 mL pipette, being careful not to disturb the cells at the bottom of the tube (to avoid aspirating white blood cells which would lead to increased background wild-type DNA levels). The plasma was dispensed into 1.5 mL cryovials in 0.5 to 1 mL aliquots. The plasma was then frozen at -80°C until needed for further processing.

Extraction and purification of DNA from plasma

Plasma was removed from the -80°C freezer and was thawed at room temperature for 15 to 30 minutes before proceeding with DNA extraction. Thawed plasma was then centrifuged at 6800 x g for 3 minutes to remove any cryoprecipitate. The supernatant was transferred to a fresh tube for further processing.

The QiaAmp® DNA Blood Mini Kit (Qiagen) was used for purification from plasma volumes up to 200 µL (elution volume of 50 µL), and the QiaAmp® MinElute® Virus Vacuum Kit (Qiagen) has also been used for plasma volumes up to 1 mL (elution volume as low as 20 µL). For larger volumes of a particular sample of plasma, more than one column of the QiaAmp® MinElute® Virus Vacuum Kit was used for purification. All kits were used according to the manufacturer's instructions, generally eluting the DNA into the lowest recommended volume (preferably 20 µL). To process 1 mL of plasma using the QiaAmp® MinElute® Virus Vacuum Kit, 5 micrograms of carrier RNA (cRNA; Qiagen) were added per mL, and the user-developed protocol found on the Qiagen website was followed.

Primer-extension reaction

Specific mutation-prone regions of purified, plasma-derived template DNA molecules were copied using targeted gene-specific primers. The number of different gene-specific primer sequences used in each tube depended on the number of targeted DNA regions within the genome. A combination of 40 different gene-specific primers were used in each sample to target 40 different gene regions. As described previously, each set of gene-specific primers had a unique, sample-specific DNA sequence (a barcode) near the 5'-end of the primers that were incorporated in a modular fashion. Each sample underwent primer-extension using an approximately equimolar concentration of 40 different gene-specific primers, all of which had the same sample-specific barcode. These primers also included degenerate sequence regions known as molecular lineage tags (MLTs) as well as common sequences at the 5'-end that allowed for hybridization of "universal" PCR primers in subsequent steps.

Control DNA molecules containing known mutations were spiked into each primer extension reaction to serve as internal quantitative standards. These DNA molecules were cartridge-purified oligonucleotides that were synthesized to contain variations from the wild-type sequence at two distinct positions (which would be extremely unlikely to occur in plasma-derived DNA). These variations allowed the control sequences to be readily distinguished from other variants within DNA purified from a clinical sample. The sequences of the top strands of these control DNA oligonucleotides are listed in Table 7. Reverse complements of these 40 sequences were also separately synthesized to produce bottom strands. In order to make the control DNA as similar as

possible to the clinically-derived DNA, both strands were annealed to make them double-stranded before adding them to the primer-extension reaction. The double-stranded DNA was quantified by UV spectrometry and then diluted to the desired concentration. To each primer-extension reaction, approximately 200 copies of the double-stranded control DNA fragments corresponding to each of the 40 gene target sites were added.

Table 7. List of spiked-in quantitative standard oligonucleotides containing mutations at 2 distinct sites relative to wild-type.

Name	DNA Sequence	SEQ ID NO:
CFTP-1	TCAACAAGATGTTTTGCCAACTGGCCAAGACCTGCCCTGTGCAGCTGT GGGTTGATTCCACACCCCCGCCCGGCACCCGCGTCCGCGTCATGACCA TCTACAAGCAGTCACAGCACATGA	93
CFTP-2	TCACAGCACATGACGGAGGTTGTGAGGCGCTGCCACCACCATGTGCG CTGCTCAGATAGCGATGGTGAGCAGCTGGGGCTGGAGAGACGACAG GGCTG	94
CFTP-3	ACAGGGCTGGTTGCCCAGGGTCCCCAGGCCTCTGATTCCTCACTGATT GCTCTTAGGTCTGGCCCCTCCTCAGCATCATATCCGAGTCGAAGGAAA TTTGCGTGTGGAGTATTTGGATG	95
CFTP-4	GAGTATTTGGATGACAGAAACACTTTTCGACACAGTGTGGTGATGCC CTATGAGCCGCCTGAGGTCTGGTTTGCAACTGGGGTCTCTGGGAGGA GGGGTTAAGGGTGGTTGTCAGTGGCCCTC	96
CFTP-5	CTGGCCTCATCTTGGGCCTGTGTTATCTCCTAGGTTGGCTCTGACTGT ACCACCATCCACTACAACGACATGTGTAACTGTTCCTGCATGGGCGGC ATGAACCGGAGGCCCATCCTCACCATCATCACACTGG	97
CFTP-6	TACTGGGACGGAACAGCTTTGAGGTGCGTGTTTGTGCATGTCCTGGG ACAGACCGGCGCACAGAGGAAGAGAATCTCCGCAAGAAAGGGGAGC CTCACCACGAGCTGCCCCCAG	98
CFPIK-1	CAAAGCAATTTCTACACGAGATCCTCTCTCTGAAGTCAGTGAGCAGGA GAAAGATTTTCTATGGAGTCACAGGTAAGTGCTAAAATGGAGATTCT CTGTTTCTTTTTC	99
CFPIK-2	GAGGCTTTGGAGTATTTCATGAAACAAATGAATCATACACATCATGGT GGCTGGACAACAAAAATGGATTGGATCTTCCACACAATTAAACAGCA TGCATTGAACTGAAAAG	100
CFBRAF	CCTCACAGTAAAAATAGGTGATTTTGGTCTAGCGACAGTGAAAGCTC GATGGAGTGGGTCCCATCAGTTTGAACAGTTGTCTGGATCCATTTTGT GGATGGTAAGAATT	101
CFFox	AAGGGCAACTACTGGACGCTGGACCCGACCTGCGCAGACATGTTCGA GAAGGGCAACTACCGGCGCCGCCGCCGCATGAAGAGGCCCTTCCGG CCG	102
CFGNAS	ACCTCAATTTTGTTTCAGGACCTGCTTCACTGCCGTATCCTGACTTCTG GAATCTTTGAGACCAAGTTCCAGGTGGACAAAGTCAACTTCCAGTAA GCCAACT	103
CFCTNN	CACTGGCAGCAACAGTCTTACCTGGACTCTGGAATCCATTCTGATGCC ACTACCACAGATCCTTCTCTGAGTGGTAAAGGCAATCCTGAGGAAGA GGATGTGGATACCTCCCAAGTCCTGTAT	104
CFPPP-1	CTGCCTGCTGCCTCAGGATCCCCGTCCCCGACTCCCAGGTACTTCCGG	105

	AACCTGTGCTCAGATGACACCCCCACGGTGCGGCGGACCGCAGCCTC CAAGCTGGGGGAG	
CFPPP-2	CTGCGCCAGGCCGCTGAAGACAAGACCTGGCGCATCCGCTACATGGT GGCTGACAAGTTCACAGAGGTAGATGAGCGACCGTTGACATTGTCCC ACTGGT	106
CFPTEN-1	TGCAGCAATTCACTGTAAAGCTGGAAAGGGACGAACAGGTGTAATGA CATGTGCATATTTATTACATCGGGGCAAATTTTTAAAGGCACAAGAGG CCCTAGATTTCTATGGGGAAG	107
CFPTEN-2	AGGTGAAGATATATTCCTCCAATTCAGGACCCTCACGACGGGTAGAC AAGTTCATGTACTTTGAGTTCCCTCAGCCGTTACCTGTGTGTGGTGAT ATCAAAGTAGAGTTCT	108
CFKIT-1	GAGACTTGGCAGCCAGAAATATCCTCCTTACTCATGGTCGGATCACAA AGATTTGTGATTTTGGTCTAGCCATAGACATCACGAATGATTCTAATT ATGTGGTTAAAGGAAACGTGAG	109
CFKIT-2	TATTTTTCCCTTTCTCCCCACAGAAACCTATGTATGAAGTACAGTGGAA GGATGTTGAGGAGATAAATGGAAACAATTATGTTTACATAGACCCAA CACAACTTCCTTATGATCACAAATGGGAGTTTC	110
CFKIT-3	GTTTTCCTGTAGCAAAACCAGAAATCCTGACTTACGACAGGCTAGTGA ATGGCATGCTCCAATGTGTGGCAGCAGGATTCCCAGAGCCCACAATA GATTGGTATTTTT	111
CFEG-1	AGAAGGTGAGAAAGTTAAAATTCCCGTCGCTATGAAGGAATTAAGAG AAGCAACATCTCCGTAAGCCAACAAGGAAATCCTCGATGTGAGTTTCT GCTTTGCTGTGTGGGGGTC	112
CFEG-2	CCGCAGCATGTCAAGATCACAGATTTTGGGCTGGACAAACAGCTGGG TGCGGAAGAGAAAGAATACCATGCAGAAGGAGGCAAAGTAAGGAG GTGGCTTTAG	113
CFEG-3	GCCTCACCTCCACCGTGCAGCTCATGACGTAGCTCATGCCCTTCGGCT GCCTCCTGGACTATGTCCGGGAACACAAAGACAATA	114
CFAKT1	TCTCACCACCCGCACGTCTGTAGAGGACTACATCAAGACCTGGCGGC CACGCTACTTCCTCCTCAAGAATGATGGCACCTTCATTGG	115
CFATM	TGTACTTCCATACTTGATTCATGATATTTTACTCCTAGATACGAATGAA TCATGGAGAAATCTGCTTTCTACACATGTTCAGGGATTTTTCACCAGCT GTCTTCGACACTTCTCGC	116
CFAPC	CACCACCTCCTCAAACAGCTCAAACCATGCGATAAGTACCTAAAAATA AAGCACCTACTGCTGAAAAGAGAGAGAGTGGACCTAAGCAAGCTGC AGT	117
CFFGFR-1	GCTCTGGGAGATCTTCACGCTGGGGGACTCCCCGTATCCCGGCATCCC TGTGGAGGAGCTCTTCAAGCTGCTGAAGGAGGGCCACCGCATGGAC AAGCCCGCCA	118
CFFGFR-2	TGGCCCCTGAGCGTCATCTGCCCCCACTGAGCGCTCCACGCACCGGCC CATCCTGCAGGCGGGGCTGCCGGCCAACCAGACGGCGGTGCTGGGC AGCGACGTGGAGTTCC	119
CFFGFR-3	AGGAGCTGGTGGAGGCTGACGAGGCGGGCAGTATGTATACAGGCAT CCTCAGCTACGGGGTGGGCTTCTTCCTGTTCATCCTGGTGGTGGCGGC TGTGAC	120
CFMET-1	GCATTCCCTATCAAATATGTCAACGACTTCATCAACAAGATAGTCAAC AAAAACAATGTGAGATGTCTCCAGCATTTTTACGGACCCAATCATGAG CACTGCTTTAATAGGGTAA	121
CFMET-2	GCTGATTTTGGTCTTGCCAGAGACATGTATCATAAACAATACTATAGT GTACACAACAAAACAGGTGCAAAGCTGCCAGTGAAGTGGATGGCTTT GGAAAGTCTG	122

CFSTK-1	CCGCATCGACTCCACCGAGGTCATCTACCAGCCGAGCCGCATGCGGG CCAAGCTCATCGGCAAGTACCTGATGGGGGACCTGCTGGGGGAAGG CTCTTACGGCAAGGTGAAGGAGGTGCTGGACTCGGAG	123
CFSTK-2	CCGTACTTGGAGGACCTGCACGGCGCGGATGAGGACGAGGACCACT TCGACATCGAGGATGACATCATCTACACTCAGGACTTCACGGTGCCCG GTGAGTCTGGCGGGGG	124
CFKRAS-1	TATAGTCACATTTTCATTATTTTTATTATAAGGCCTGCTGAAAATGACT GAATATAAACTTGTGGTAGTTGCAGATGGTGGCGTAGGCAAGAGTGC CTTGACGATAC	125
CFKRAS-2	CTTGGATATTCTCGACACAGCAGGTCAAGACGAGTACTGTGCAATGA GGGACCAGTACATGAGGACTGGGGAGGGCTTTCTTTGTGTATTTGCC ATAAATAATACTAAA	126
CFNRAS-1	TGTAGATGTGGCTCGCCAATTAACCCTGATTACTGGTTTCCAACAGGT TCTTGCTGGTGTGAAATGACTGAGTACAAACTGGTCGTGGATGGAGC AGGTGGTGTTGGGAAAAGCGCACTGACAAT	124
CFNRAS-2	GTTGGACATACTGGATACAGCTGGACAAGAAGAGCACAGTGACATG AGAGACCAATACATGAGGACAGGCGAAGGCTTCCTCTGTGTATTTGC CATCAATAATAGCAAGTCAT	128
CFHRAS-1	GGTGGGGCAGGAGACCCTGTAGGAGGACCCCGGGCCGCAGGCCCCT GAGGAGCGATGACGGAATATAAGCTGGTGGTCGTGGACGCCGGCGG TGTGGGCAAGAGTGCGCTGACCATCC	129
CFHRAS-2	TGGACATCCTGGATACCGCCGGCCAGGAGTACTACAGCGCCATGCGG GACCAGTACATGCGCACCGGGGAGGGCTTCCTGTGTGTGTTTGCCAT CAACAACACCAAGTCTTTTGAGGA	130
CFK-Ctrl	TGTTCCTGTTTATAATATTGACAAAACACCTTAGCGGATGACATTTAA GAATTCTAAAAGTCCTAATATATGTAATATATATTCAGTTGCCTGAAG AGAAACATAAAGAATCCTTTCTTAAT	131
CFB-Ctrl	ATGTCAGGACAAAGTCCGGATTGAATATAACTCTGCTTTATATTATAG GCCTATGAAGAATACACCAGCAAGCTAGATGCACTCCAACAAAGAGA ACAACAGTTATTGGAATCTCTGGG	132

*All oligos synthesized at 40 nmole scale with cartridge purification.

Conditions were optimized so that on average, more than one copy of each original DNA template molecule would be present at the beginning of the next amplification step. Typically between 2 and 10 cycles of primer-extension were carried out. Primer extension was performed using Accuprime *Taq* polymerase (Invitrogen) as described below.

Primer-extension reaction setup (30 µL reaction):

Purified template DNA (with co-eluted carrier RNA [cRNA])	20 µL (or less)
100 copies of control mutant DNA in 10 mM Tris (with 300 ng per mL cRNA)	(as needed)
10 mM Tris with cRNA (300 ng per mL)	(as needed for final 30 uL volume)
10 x concentrated Accuprime Buffer #2	3 µL
Mix of 40 modular barcoded primers (50 µM total stock) (final ~200 nM each)	4.8 µL
Accuprime *Taq* polymerase	0.6 µL

Total 30 µL

Temperature cycling conditions (carried out on a BioRad iCycler®):

a. 94°C for 120 sec
b. 94°C for 20 sec
c. 60°C for 20 sec (this step provides more time for annealing)
d. 55°C for 1 min (may decrease this temperature to improve primer annealing)
e. 72°C for 20 sec
f. repeat steps b-e for 3 more cycles (total 4 cycles)
g. 4°C for up to 20 minutes

As quickly as possible once the reactions had reached 4°C, 1 µL of 300 mM EDTA was added (to make a final concentration of 10 mM) to terminate the activity of the polymerase. Each tube was agitated gently to ensure even mixing of the EDTA. Because the primer-extended molecules had sample-specific barcodes attached, the products of all reactions that were derived from different samples could be pooled together into a single tube.

Purification of primer-extended products

The purification of primer-extended products was achieved via pull-down and elution steps using complementary biotinylated "capture" oligonucleotides and streptavidin-agarose beads (Thermo-Fisher). First, a mixture of complementary biotinylated oligonucleotides was added to the pooled primer-extension products. These oligonucelotides were designed to anneal to the specific sequences that should be produced if the primers were extended using the intended genomic DNA target region as their templates. A list of the 40 biotinylated oligos that were used in the present example is included in Table 8. By capturing with these biotinylated oligos, it was possible to ensure that only the specifically extended primers were isolated, and that any un-extended primers and any primers that were extended on non-specific DNA templates were not pulled down. For every 30 microliter reaction volume (plus 1 microliter of EDTA added), a final concentration of 200 nM of each biotinylated oligo was added (by addition of 3.5 µL of an 80 micromolar oligonucleotide mix for a final total concentration of 8 micromolar biotinylated oligos [all 40 oligos]). Annealing of the biotinlyated capture oligos with the primer-extended products was achieved by heating the mixture to 95°C for 30 seconds, then to 70°C for 20 seconds, then cooling by 2.5°C every 20 seconds until the mixtures reached 25°C.

Table 8. List of biotinylated target-specific capture DNA oligonucleotides

NAME	DNA OLIGO SEQUENCE	SEQ ID NO:
BNTP53-1	5'-BIO-CAAGATGTTTTGCCAACTGGCC	133
BNTP53-2	5'-BIO-CCTGTCGTCTCTCCAGCCCCAG	134
BNTP53-3	5'-BIO-GGCTGGTTGCCCAGGGTCCC	135
BNTP53-4	5'-BIO-GCCACTGACAACCACCCTTAACC	136
BNTP53-5	5'-BIO-CCTCATCTTGGGCCTGTGTTATCT	137
BNTP53-6	5'-BIO-GGGCAGCTCGTGGTGAGGC	138
BNPIK-1	5'-BIO-AAGAAACAGAGAATCTCCATTTTAGCAC	139
BNPIK-2	5'-BIO-TCAGTTCAATGCATGCTGTTTAATTGTG	140
BNBRAF	5'-BIO-CTTACCATCCACAAAATGGATCCAGAC	141
BNFoxL2	5'-BIO-CGGAAGGGCCTCTTCATGCGGC	142
BNGNAS	5'-BIO-GGCTTACTGGAAGTTGACTTTGTCCAC	143
BNCTNN	5'-BIO-AGGACTTGGGAGGTATCCACATCC	144
BNPPP-1	5'-BIO-CTGCTGCCTCAGGATCCCCGTCC	145
BNPPP-2	5'-BIO-GTGGGACAATGTCAACGGTCGCT	146
BNPTEN-1	5'-BIO-CCCATAGAAATCTAGGGCCTCT	147
BNPTEN-2	5'-BIO-CTCTACTTTGATATCACCACACACAGG	148
BNKIT-1	5'-BIO-CTTGGCAGCCAGAAATATCCTCCTTACTC	149
BNKIT-2	5'-BIO-CTCCCATTTGTGATCATAAGGAAGTTG	150
BNKIT-3	5'-BIO-ATACCAATCTATTGTGGGCTCTGG	151
BNEG-1	5'-BIO-CCCACACAGCAAAGCAGAAAC	152
BNEG-2	5'-BIO-AGCCACCTCCTTACTTTGCCTCC	153
BNEG-3	5'-BIO-GTCTTTGTGTTCCCGGACATAGTCC	154
BNAKT1	5'-BIO-TGAAGGTGCCATCATTCTTGAGGAG	155
BNATM	5'-BIO-GAAGTGTCGAAGACAGCTGGTGAA	156
BNAPC	5'-BIO-CAGCTTGCTTAGGTCCACTCTCTC	157
BNFGFR-1	5'-BIO-GGGCTTGTCCATGCGGTGGCC	158
BNFGFR-2	5'-BIO-CTCCACGTCGCTGCCCAGCACC	159
BNFGFR-3	5'-BIO-CAGCCGCCACCACCAGGATGAAC	160
BNMET-1	5'-BIO-CCTATTAAAGCAGTGCTCATGATTGG	161
BNMET-2	5'-BIO-CTTTCCAAAGCCATCCACTTCAC	162
BNSTK-1	5'-BIO-GAGTCCAGCACCTCCTTCACCTTG	163
BNSTK-2	5'-BIO-CGCCAGACTCACCGGGCACC	164
BNKRAS-1	5'-BIO-GTCACATTTTCATTATTTTTATTATAAGGCCTGC	165
BNKRAS-2	5'-BIO-GTATTATTTATGGCAAATACACAAAGAAAGC	166
BNNRAS-1	5'-BIO-GATGTGGCTCGCCAATTAACCCTGA	167
BNNRAS-2	5'-BIO-CTTGCTATTATTGATGGCAAATACACAG	168
BNHRAS-1	5'-BIO-GGGCAGGAGACCCTGTAGGAG	169
BNHRAS-2	5'-BIO-CAAAAGACTTGGTGTTGTTGATGGCA	170
BNK-Ctrl	5'-BIO-AGAAAGGATTCTTTATGTTTCTCTTCAGG	171
BNB-Ctrl	5'-BIO-GAGATTCCAATAACTGTTGTTCTCTTTGT	172

5'-Bio = 5'-Biotin

Then, 7 µL of high capacity streptavidin-agarose bead slurry (Thermo-Fisher) was added (per 30 µL primer-extension reaction). Tubes were turned end-over-end constantly for at least 2 hours to promote binding of biotinylated oligos to the streptavidin beads. Beads were then centrifuged briefly, and any unbound supernatant was carefully removed, avoiding aspiration of any beads. The beads were then washed in about 200 µL of 10 mM Tris pH 7.6 and 50 mM NaCl (referred to hereafter as wash buffer). Beads were suspended in wash buffer by gentle agitation, then were briefly centrifuged, and the supernatant wash buffer was removed and discarded. A second wash was performed in the same way, except that once the beads were suspended, they were incubated at 45°C for 30 minutes while the tube was turned end-over-end (this was to promote dissociation of any DNA molecules that may have annealed non-specifically to the biotinylated capture oligos). The beads were again centrifuged briefly, and the supernatant wash buffer was removed. The captured primer-extended products were eluted from the surface of the washed beads by heat-denaturation. Since the biotin-streptavidin interaction was not substantially disrupted by heating at 95°C, only the captured primer-extended products were eluted from the beads, whereas the biotinylated capture oligos remained bound to the beads. Elution was carried out directly into the pre-amplification PCR cocktail as described below.

Multiplexed pre-amplification PCR

The purified primer-extension products were eluted directly into a cocktail of buffer, nucleotides, and primers that was used to carry out the multiplexed pre-amplification reaction. The primer-extended DNA was eluted into the following cocktail:

Molecular grade water	(enough to make 100 µL total reaction volume)
10x Accuprime *Taq* PCR buffer #1 or pfx buffer (with dNTPs already added)	10 µL
Forward primer mix for 40 amplicons (20 uM Fwd mix stock, 200 nM final each)	40 µL
Universal reverse primer - ExtV2Rev (2 uM stock, 200 nM final)	10 µL
Total	100 µL

The beads in the pre-amplification cocktail were heated at 95°C for 30 seconds, were quickly and gently centrifuged, and the supernatant was transferred to a clean PCR tube. When the cocktail reached room temperature, 2 µL of Accuprime hotstart *Taq* polymerase (or 1 µL Accuprime *Pfx*) was added to the tube, and mixed by pipetting up and down. Then 30 µL of mineral oil was added to prevent evaporation during thermal cycling which was carried out as follows:

a. 94°C for 2 minutes (95°C if using Accuprime *Pfx*)

b. 94°C for 20 seconds (95°C if using Accuprime *Pfx*)

c. 63°C for 30 seconds

d. 72°C for 20 seconds

e. repeat (b) to (d) for a total of 15 cycles

f. 72°C for 2 minutes

Then, 11 µL of 100 mM EDTA was added (10 mM final concentration) to the completed reaction to chelate divalent cations and thus terminate polymerase activity.

The forward primers used in this pre-amplification reaction were designed to hybridize to regions on the target sequences that were nested relative to the binding sites of the biotinylated capture oligonucleotides that were used in the first primer extension reaction. This nested design provided an additional level of specificity so that the desired target DNAs would be preferentially amplified. The sequences of the universal pre-amplification reverse primer (ExtV2Rev), and the 40 different nested forward primers are listed in Table 9.

Table 9. List of 40 forward primers and the single universal reverse primer (ExtV2Rev) used for the pre-amplification reaction

Name	DNA Sequence	SEQ ID NO:
ExtV2REV	CGAGACGGATCAAGCAGAAGACG	173
ExF-TP53-1	GCCAACTGGCCAAGACCTGC	174
ExF-TP53-2	CTCCAGCCCCAGCTGCTCAC	175
ExF-TP53-3	GTCCCCAGGCCTCTGATTCCTC	176
ExF-TP53-4	CCTCCCAGAGACCCCAGTTGC	177
ExF-TP53-5	TGGGCCTGTGTTATCTCCTAGGTTG	178
ExF-TP53-6	GCAGCTCGTGGTGAGGCTCC	179
ExF-PIK3CA-1	AGAAACAGAGAATCTCCATTTTAGCACTTACC	180
ExF-PIK3CA-2	TTCAATGCATGCTGTTTAATTGTGTGGAAG	181
ExF-BRAF	TCCACAAAATGGATCCAGACAACTGTTC	182
ExF-FoxL2	GGCGCCGGTAGTTGCCCTTC	183
ExF-GNAS	GGAAGTTGACTTTGTCCACCTGGAAC	184
ExF-CTNNB1	GGAGGTATCCACATCCTCTTCCTCAG	185
ExF-PPP2R1A-1	CGACTCCCAGGTACTTCCGGAAC	186
ExF-PPP2R1A-2	TGTCAACGGTCGCTCATCTACCTC	187
ExF-PTEN-1	CCATAGAAATCTAGGGCCTCTTGTGC	188
ExF-PTEN-2	CACCACACACAGGTAACGGCTG	189
ExF-KIT-1	GAAATATCCTCCTTACTCATGGTCGGATCA	190
ExF-KIT-2	CCCATTTGTGATCATAAGGAAGTTGTGTTG	191
ExF-KIT-3	GTGGGCTCTGGGAATCCTGCTG	192
ExF-EGFR-1	CCACACAGCAAAGCAGAAACTCAC	193

ExF-EGFR-2	ACCTCCTTACTTTGCCTCCTTCTGC	194
ExF-EGFR-3	GTGTTCCCGGACATAGTCCAGGAG	195
ExF-AKT1	GCCATCATTCTTGAGGAGGAAGTAGC	196
ExF-ATM	AGACAGCTGGTGAAAAATCCCTGAAC	197
ExF-APC	TGCTTAGGTCCACTCTCTCTCTTTTCAG	198
ExF-FGFR3-1	GCGGTGGCCCTCCTTCAGCAG	199
ExF-FGFR3-2	CCAGCACCGCCGTCTGGTTG	200
ExF-FGFR3-3	CCACCAGGATGAACAGGAAGAAGC	201
ExF-MET-1	CAGTGCTCATGATTGGGTCCGT	202
ExF-MET-2	GCCATCCACTTCACTGGCAGC	203
ExF-STK11-1	CTTCACCTTGCCGTAAGAGCCTTC	204
ExF-STK11-2	CTCACCGGGCACCGTGAAGTC	205
ExF-KRAS-1	CATTATTTTTATTATAAGGCCTGCTGAAAATGACTGA	206
ExF-KRAS-2	TGGCAAATACACAAAGAAAGCCCTCC	207
ExF-NRAS-1	CAATTAACCCTGATTACTGGTTTCCAACAG	208
ExF-NRAS-2	GGCAAATACACAGAGGAAGCCTTCG	209
ExF-HRAS-1	CAGGAGACCCTGTAGGAGGACC	210
ExF-HRAS-2	TGATGGCAAACACACACAGGAAGC	211
ExF-KRAS-Cntrl	AGGATTCTTTATGTTTCTCTTCAGGCAACTG	212
ExF-BRAF-Cntrl	ACTGTTGTTCTCTTTGTTGGAGTGCATC	213

Purification of the products of the pre-amplification reaction

The products of the pre-amplification reaction were purified using a QIAquick® PCR purification kit (Qiagen) according to the manufacturer's instructions. This removed the enzyme, dNTPs, and unincorporated primers from the double-stranded reaction products. Elution of the DNA from the column was carried out in 60 µL of EB buffer (composed of 10 mM Tris). This elution volume allowed 1 µL to be used in each of the 40 individual PCRs (see next section), with approximately 20 µL left over in case any failed reactions need to be repeated. The purified DNA can be stored at 4°C for several days if necessary. Extra care was taken when handling any of the amplified products to avoid contamination of these products into the reagents used for reaction set-up (separate work-spaces were maintained for reagents and for amplification products).

Separate PCR amplification of individual gene targets (mutation hotspots)

After purification, products of the pre-amplification reaction were subjected to further amplification by PCR in separate tubes (one tube for each of the 40 target gene regions). These individual PCRs were performed in order to provide an additional layer of amplification specificity, since the multiplexed pre-amplification reaction was likely to have produced many spurious products in addition to the amplicons of interest. Using PCR primers that were nested relative to the primers used in the previous pre-amplification step allowed the desired target DNAs to be preferentially amplified. Also, by carrying out each individual PCR to saturation and using the same concentration

of primers in each reaction, similar numbers of copies of each target region could be produced. Normalization of molecular counts in this way allowed a similar sequencing depth to be achieved for each target.

A different gene-specific forward primer was paired with a universal reverse primer in each of the 40 PCR tubes. Both primers were nested relative to the primers used in the pre-amplification reaction so that further amplification specificity could be achieved (a nested primer is designed so that its 3'-end hybridizes to a region within the desired target sequence that was flanked by the primers used in the earlier round of amplification). The forward primers contained extra sequences on their 5'-ends that were necessary for subsequent sequencing on an Illumina flow cell. The reverse primer was also designed to produce a product that was compatible with the Illumina sequencer without the need for attachment of additional adapter sequences. The sequences of the universal reverse PCR primer (called IntV2Rev) and the 40 different, target-specific forward PCR primers are listed in Table 10. A 4 nucleotide stretch of degenerate sequence was included in the forward primer to provide greater sequence diversity at the first few read positions, thereby improving cluster discrimination on the Illumina sequencer. Although these primers were designed to be compatible with the Illumina next-generation sequencing system, the method can relatively easily be adapted to other sequencing platforms. The PCR setup of each individual tube was as follows:

Molecular grade water	4.8 µL
10x Accuprime Taq Buffer #1	1 µL
Forward gene-specific primer (1 **u**M stock, 200 nM final)	2 µL
Universal reverse primer IntV2Rev (2uM stock, 200 nM final)	1 µL
Template DNA purified after pre-amplification reaction	1 µL
Accuprime Taq DNA polymerase	0.2 µL
Total 10 µL	

Table 10. List of 40 nested forward primers and the single nested universal reverse primer (IntV2Rev)

IntV2REV	CAAGCAGAAGACGGCATACGAGA	217
InF-TP53-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTGCAGCTGTGGGTTGATTCCAC	218
InF-TP53-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCAGCTGCTCACCATCGCTATCT	219
InF-TP53-3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTCCTCACTGATTGCTCTTAGGTCTGG	220
InF-TP53-4	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCAAACCAGACCTCAGGCGGCTC	221
InF-TP53-5	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGCTCTGACTGTACCACCATCCAC	222
InF-TP53-6	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTCCCCTTTCTTGCGGAGATTCTCT	223

InF-PIK3CA-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCACTTACCTGTGACTCCATAGAAAATCTTTC	224
InF-PIK3CA-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGAAGATCCAATCCATTTTTGTTGTCCAGC	225
InF-BRAF	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCAAACTGATGGGACCCACTCCATC	226
InF-FoxL2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCGGTAGTTGCCCTTCTCGAACATG	227
InF-GNAS	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTACTTGGTCTCAAAGATTCCAGAAGTCAG	228
InF-CTNNB1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCTCAGGATTGCCTTTACCACTCAG	229
InF-PPP2R1A-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCGGAACCTGTGCTCAGATGACAC	230
InF-PPP2R1A-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCTGTGAACTTGTCAGCCACCATGTAG	231
InF-PTEN-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCTTTAAAAATTTGCCCCGATGTAATAAATATGC	232
InF-PTEN-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGGCTGAGGGAACTCAAAGTACATGAAC	233
InF-KIT-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGGATCACAAAGATTTGTGATTTTGGTCTAGC	234
InF-KIT-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGGTCTATGTAAACATAATTGTTTCCATTTATCT	235
InF-KIT-3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGCCACACATTGGAGCATGCCA	236
InF-EGFR-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGAAACTCACATCGAGGATTTCCTTGTTG	237
InF-EGFR-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTGCATGGTATTCTTTCTCTTCCGCAC	238
InF-EGFR-3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGAGGCAGCCGAAGGGCATGAG	238
InF-AKT1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGTGGCCGCCAGGTCTTGATG	240
InF-ATM	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCATGTGTAGAAAGCAGATTTCTCCATGATTC	241
InF-APC	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTTCAGCAGTAGGTGCTTTATTTTTAGGTAC	242
InF-FGFR3-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCAGCTTGAAGAGCTCCTCCACAG	243
InF-FGFR3-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCTGCAGGATGGGCCGGTG	244
InF-FGFR3-3	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCACCCCGTAGCTGAGGATGC	245
InF-MET-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGCTGGAGACATCTCACATTGTTTTTGTTG	246
InF-MET-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGCTTTGCACCTGTTTTGTTGTGTACAC	247
InF-STK11-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCCATCAGGTACTTGCCGATGAG	248
InF-STK11-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCTGAGTGTAGATGATGTCATCCTCGATG	249
InF-KRAS-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTGCTGAAAATGACTGAATATAAACTTGTGGTA	250
InF-KRAS-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCAGTCCTCATGTACTGGTCCCTCATT	251
InF-NRAS-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTCTTGCTGGTGTGAAATGACTGAGTAC	252
InF-NRAS-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTCGCCTGTCCTCATGTATTGGTCT	253
InF-HRAS-1	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTCCTGAGGAGCGATGACGGAATATAAG	254
InF-HRAS-2	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTATGTACTGGTCCCGCATGGCG	255
InF-KRAS-Cntrl	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTAGGACTTTTAGAATTCTTAAATGTCATCCGC	256
InF-BRAF-Cntrl	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCTNNNNACTTCTAGCTTGCTGGTGTATTCTTCATAGG	257

Mineral oil (20 µL) was added to each tube to prevent evaporation during PCR. Again, both Accuprime *Taq* polymerase as well as Accuprime *Pfx* were tested for PCR amplification, and both worked. The temperature cycling conditions used for PCR were as follows:

a. 94°C for 2 minutes (95°C if using Accuprime *Pfx*)

b. 94°C for 20 seconds (95°C if using Accuprime *Pfx*)

c. 64°C for 30 seconds

d. 72°C for 20 seconds

e. repeat b to d for a total of 36 to 45 cycles (36 cycles for *Taq* and 45 cycles for *Pfx*).

f. 72°C for 2 minutes

g. 4°C until removed from thermal cycler

While the PCR tubes were still at 4°C, 5 µL of 30mM EDTA was added to inactivate the polymerase (10 mM EDTA final). This was added under the mineral oil layer, and was pipetted up and down to mix. Products from all 40 reactions were pooled into a single tube (equal volumes from each of the 40 reactions were added to the final mix).

Preparation of DNA for next-generation sequencing

The pooled PCR reaction products were purified on a 2% agarose gel with ethidium bromide and 1x TBE buffer. Since all PCR products were of a similar final length, the pooled products appeared on the gel as a somewhat diffuse band. This diffuse band was excised from the gel using a fresh scalpel blade, ensuring that the gel was cut a few millimeters above and below the visible band to include any low-intensity bands that may have run faster or slower and were not well-visualized. Using a QIAquick® Gel Extraction kit (Qiagen) according to the manufacturer's instructions, the DNA was isolated from the gel slice. The DNA was eluted into 50 µL of elution buffer, EB.

Next-generation sequencing

To prepare the sample for loading onto an Illumina HiSeq flow cell, the concentration of the DNA was measured using an Agilent Bioanalyzer®, and the DNA was diluted to the concentration recommended by Illumina. In order to increase sequence diversity on the flow cell, Phi-X control DNA (Illumina) was added so that the total molar amount of Phi-X DNA was approximately 30% of the final sample that was loaded onto the flow cell.

Cluster formation was carried out on the flow cell according to Illumina's protocol. The sample was loaded onto a single lane of a flow cell. The sequencing was performed on a HiSeq® 2000 instrument in multiplexed paired-end mode, with a read length of 75 base pairs in each direction. An index read was also performed, and the length of the index read was increased from the standard 7 cycles up to 13 cycles so that our longer custom barcodes and MLT sequences could be appropriately read. A control lane was designated that contained either phi-X DNA or genomic DNA so that matrix generation for phasing/prephasing would be based on a sample having greater

sequence diversity than was present in our sample. Demultiplexing of the sequences was performed using custom computer code.

Outline of algorithm for sequence analysis

The sequences that survived the filtering process were comprised of the PCR amplicons of interest as well as sequences derived from control Phi-X DNA. Our algorithm effectively ignored Phi-X sequences because those sequences did not conform to the filtering requirements described below.

A computer algorithm was designed to sort, align, and count the millions of sequences that were generated by the high-throughput sequencer. The sequence elements used in the algorithm are identified in Figure 14. The following steps provide an outline of the process and rationale used to analyze the sequence data:

1. Only clonal sequences that had passed Illumina's chastity filter were included in the analysis. Any sequences that had a "." at any position were eliminated (these were counted as dot rejects). If an unusually large number of dots were found at a particular sequence position (indicating sequencer failure at that cycle), the filter was modified in order to avoid filtering out an unreasonably large fraction of sequences.

2. The 8 nucleotide barcode from the index read (read #2) was used to assign each filtered clonal sequence to a sample-specific bin. The sequence in the region of the barcode was expected to be in the format, BBBBBBBBNNAT, where "B" was a barcode nucleotide and "N" was a nucleotide belonging to the molecular lineage tag (a position designated as N had an approximately equal probability of being A, C, G, or T in any given molecule. In order for the clonal sequence to be assigned to a sample-specific bin, the following conditions had to be satisfied:

   a. the sequence BBBBBBBB at positions 1-8 had to exactly match the reverse complement of one of the 96 barcodes listed in Table 6;

   AND

   b. the nucleotides at position 11 and 12 of the index read had to be AT. If a clonal sequence failed to satisfy both above conditions, it was classified as a barcode reject. In case the lack of sequence diversity at these positions 11 and 12 caused the read quality to be greatly diminished, leading to a high rate of miscalls or "." calls, requirement (b) was optionally modified or eliminated.

3. Each clonal sequence that was assigned to a sample-specific bin was further sub-classified according to the targeted gene segment from which it arose. The primer sequences from both the

forward and reverse reads were used to assign each clone to a particular gene segment. In the present example, 40 distinct gene segments were analyzed. In order to assign a clonal sequence to a gene target bin, the conditions (a), (b), and (c) had to be satisfied.

a. In the forward read, the first 8 nucleotides of the primer sequence (designated by a "F" in Figure 14) had to exactly match the first 8 nucleotides of one of the 40 forward gene-specific primer sequences.

b. In the reverse read, the first 8 nucleotides of the primer sequence (designated by a "R" in Figure 14) had to exactly match the first 8 nucleotides of one of the 40 reverse gene-specific primer sequences.

c. The forward primer and reverse primer reads had to lead to assignment of each clone to the same gene segment. Assignment of a single clonal sequence to more than one gene segment bin was not permissible.

If a clonal sequence failed to satisfy these conditions, it was classified as a gene segment reject.

4. Each clonal sequence that was successfully assigned to a sample-specific barcode bin and to a gene segment bin then had its forward and reverse reads aligned to each other using a Smith-Waterman algorithm, as described in Example 1 (the reverse-complement of read #3 was derived to facilitate alignment). This enabled identification of the region of overlap between the forward and reverse reads. Different lengths of overlap were expected for different gene segments since the forward and reverse read-lengths were constant but the PCR amplicon length differed for different gene segments. The length of overlap could also vary because of the presence of insertion or deletion mutations. The forward and reverse reads were also aligned to the wild-type reference sequence for its assigned gene segment (the full-length wild-type reference sequences are listed in Table 11).

Table 11. List of wild-type reference sequences for all 40 targeted gene segments

Gene Segment	Reference Sequence (FFFF....FFFF[xxxx.....xxxx]RRRR.....RRRR)	
TP53-1	TGCAGCTGTGGGTTGATTCCAC[acccccgcccggcacccgcgtccgcgccatggccatcta]CAAGCAGTCACAG CACAG	258
TP53-2	CCAGCTGCTCACCATCGCTATCT[gagcagcgctcatggtgggggcagcgcctcac]AACCTCCGTCATGTGCTA	259
TP53-3	TCCTCACTGATTGCTCTTAGGTCTGG[cccctcctcagcatcttatccgagtggaagga]AATTTGCGTGTGGAGT ATTTGT	260
TP53-4	CAAACCAGACCTCAGGCGGCTC[atagggcaccaccacactatgtcgaa]AAGTGTTTCTGTCATCCAAATAT	261
TP53-5	GCTCTGACTGTACCACCATCCAC[tacaactacatgtgtaacagttcctgcatgggcggcatgaaccggaggc]CC ATCCTCACCATCATCACAG	262
TP53-6	TCCCCTTTCTTGCGGAGATTCTCT[tcctctgtgcgccggtctctcccaggacaggcacaaacacgcac]CTCAAA GCTGTTCCGTCCCAG	263
PIK3CA-1	CACTTACCTGTGACTCCATAGAAAATCTTTC[tcctgctcagtgatttcagag]AGAGGATCTCGTGTAGAAATTG CA	264
PIK3CA-2	GAAGATCCAATCCATTTTTGTTGTCCAGC[caccatgatgtgcatcat]TCATTTGTTTCATGAAATACTCCAAAG A	265
BRAF	CAAACTGATGGGACCCACTCCATC[gagatttcactgtagc]TAGACCAAAATCACCTATTTTTACTGTT	266

FoxL2	CGGTAGTTGCCCTTCTCGAACATG [tcttcgcaggccgg] GTCCAGCGTCCAGTAGTTGCA	267
GNAS	ACTTGGTCTCAAAGATTCCAGAAGTCAG [gacacggcagcga] AGCAGGTCCTGAAACAAAATTGAG	268
CTNNB1	CCTCAGGATTGCCTTTACCACTCAG [agaaggagctgtggtagtggcaccagaatggattccagagtcc] AGGTAA GACTGTTGCTGCCAG	269
PPP2R1A -1	CCGGAACCTGTGCTCAGATGACAC [ccccatggtgcggcgggcc] GCAGCCTCCAAGCTGGGT	270
PPP2R1A -2	CTGTGAACTTGTCAGCCACCATGTAG [cggacgcgccaggact] TGTCTTCAGCGGCCTGGCT	271
PTEN-1	CCTTTAAAAATTTGCCCCGATGTAATAAATATGC [acatatcattacaccagttcgtcc] CTTTCCAGCTTTACAG TGAATTGCC	272
PTEN-2	GGCTGAGGGAACTCAAAGTACATGAAC [ttgtcttcccgtcgtgtgg] GTCCTGAATTGGAGGAATATATCTTCAT	273
KIT-1	GGATCACAAAGATTTGTGATTTTGGTCTAGC [cagagacatcaagaat] GATTCTAATTATGTGGTTAAAGGAAAC GC	274
KIT-2	GGTCTATGTAAACATAATTGTTTCCATTTATCT [cctcaacaaccttccactgtacttcatacatgggttt] CTGT GGGGAGAAAGGGAAAAC	275
KIT-3	GCCACACATTGGAGCATGCCA [ttcacgagcctgtcgtaagtcag] GATTTCTGGTTTTGCTACAGGAAC	276
EGFR-1	GAAACTCACATCGAGGATTTCCTTGTTG [gctttcggagatgttgcttctcttaattccttgatagc] GACGGGAA TTTTAACTTTCTCACCG	277
EGFR-2	TGCATGGTATTCTTTCTCTTCCGCAC [ccagcagtttggccagcc] CAAAATCTGTGATCTTGACATGCTT	278
EGFR-3	GAGGCAGCCGAAGGGCATGAG [ctgcgtgatg] AGCTGCACGGTGGAGGTGAG	279
AKT1	GTGGCCGCCAGGTCTTGATG [tactcccct] ACAGACGTGCGGGTGGTC	280
ATM	CATGTGTAGAAAGCAGATTTCTCCATGATTC [atttgtatcttgg] AGTAAAATATCATGAATCAAGTATGGAAGA	281
APC	TTCAGCAGTAGGTGCTTTATTTTTAGGTAC [ttctcgcttg] GTTTGAGCTGTTTGAGGAGGTC	282
FGFR3-1	CAGCTTGAAGAGCTCCTCCACAG [ggatgccggggtacggggagcccc] CCAGCGTGAAGATCTCCCAT	283
FGFR3-2	CCTGCAGGATGGGCCGGTG [cggggagcgctctgtggg] GGCAGATGACGCTCAGGGA	284
FGFR3-3	CCACCCCGTAGCTGAGGATGC [ctgcatacacactgcccgcc] TCGTCAGCCTCCACCAGCG	285
MET-1	GCTGGAGACATCTCACATTGTTTTTGTTG [acgatcttgttgaaga] AGTCGTTGACATATTTGATAGGGAAC	286
MET-2	GCTTTGCACCTGTTTTGTTGTGTACAC [tatagtattctttatcataca] TGTCTCTGGCAAGACCAAAATT	287
STK11-1	CCCATCAGGTACTTGCCGATGAG [cttggcccgcttgcggcgcggctggtaga] TGACCTCGGTGGAGTCGATA	288
STK11-2	CTGAGTGTAGATGATGTCATCCTCGATG [tcgaagaggtcctcgtcctcgtccgcgc] CGTGCAGGTCCTCCAAGT AT	289
KRAS-1	GCTGAAAATGACTGAATATAAACTTGTGGTA [gttggagctggtggcgt] AGGCAAGAGTGCCTTGACGAC	290
KRAS-2	CAGTCCTCATGTACTGGTCCCTCATT [gcactgtactcctcttg] ACCTGCTGTGTCGAGAATATCG	291
NRAS-1	TCTTGCTGGTGTGAAATGACTGAGTAC [aaactggtggtggttggagcaggtggtgt] TGGGAAAAGCGCACTGAT	292
NRAS-2	TCGCCTGTCCTCATGTATTGGTCT [ctcatggcactgtactcttcttg] TCCAGCTGTATCCAGTATGTCA	293
HRAS-1	CCTGAGGAGCGATGACGGAATATAAG [ctggtggtggtgggcgccggcggtgt] GGGCAAGAGTGCGCTGACCAC	294
HRAS-2	ATGTACTGGTCCCGCATGGCG [ctgtactcctcctg] GCCGGCGGTATCCAGGATGA	295
KRAS-Cntrl	AGGACTTTTAGAATTCTTAAATGTCATCCGC [at] AGGTGTTTTGTCAATATTATAAACAGGAT	296
BRAF-Cntrl	TCTAGCTTGCTGGTGTATTCTTCATAGG [cctataaaataaagcagacttatat] TCAATCCGGACTTTGTCCTGA T	297

Note: Forward and reverse primer sequences are in capital letters, to the left and right of the square brackets, respectively. The actual reverse primer sequence would be the reverse-complement of that shown above. The genomic wild-type amplicon target sequence is in lower case letters within the square brackets.

5. Next, any variants or mutations that existed within the amplicon target region for each gene segment were identified and quantified (nucleotides in this region were designated by a "X" in Figure 14). Wild-type sequences of the amplicon target regions (region between flanking primers) for all 40 gene segments are listed in lower case letters within square brackets in Table 11. All clones belonging to a particular sample-specific bin and a particular gene segment bin were compared to the wild-type sequence in the amplicon target region. If a clonal sequence had perfect agreement

between its two overlapping reads in the amplicon target region, but deviated from the wild-type sequence, then that clone was identified and counted as a "consistent variant". If a clonal sequence had perfect agreement between its two overlapping reads in the amplicon target region, and was perfectly consistent with the wild-type sequence, then it was identified and counted as an "exact match to wild type". If a clonal sequence had deviations from the amplicon target reference sequence seen in either or both of the forward and reverse reads, but the two reads were not perfectly consistent with each other, then that clone was identified and counted as an "inconsistent variant". Any mismatches, insertions, or deletions relative to the reference sequence (whether found in both reads or in a single read) were counted and tabulated for each position within the amplicon target region for all sequences in a given bin (For purposes of illustration, results of a hypothetical experiment are shown in Figure 15).

6. In order to distinguish mutant sequences that were present in the original template DNA molecules from those arising due to sequencing errors or errors introduced during PCR amplification or sample processing, sequences called "molecular lineage tags" (MLTs) were used. As shown in Figure 14, the sequence for MLT-1 was comprised of a total of 8 degenerate nucleotide positions (derived by concatenating 6 positions of MLT-1a and 2 positions of MLT-1b). Each of the eight N positions had an approximately equal likelihood of having an A, C, G, or T nucleotide, so that 4^8 = 65,536 possible MLT sequences could be generated. Thus, a particular primer molecule would be expected to have any one of the 65,536 possible MLT sequences. Prior to amplification by PCR, the DNA template molecules were copied by primer-extension, and a MLT-1 sequence became attached to each primer-extended copy. Thus, each template copy was tagged with one of 65,536 possible MLT-1 sequences.

To identify variants arising from mutant template DNA molecules, first a list of all "consistent variants" was generated. If a "consistent variant" sequence was seen in more than one clone within a bin of sequences, then the number of copies of such variants was counted. These variants were listed along with the number of clonal copies (in descending order of frequency) as shown in Figure 15. Then, for all clones belonging to a particular "consistent variant" sequence, a list of MLT-1 sequences associated with the clones was generated (the actual list of MLT-1 sequences was not displayed). Within each list, any MLT sequence that was found to be associated with more than one clone was classified as a "multiply occurring MLT". A histogram of such multiply occurring MLTs was generated for each variant (as shown in Figure 15). The count of different MLT-1 sequences occurring "N" times for a given variant was listed in a numerical table (where N was the number of copies of the same MLT). An alternate way to present the MLT-1 counts was to list the "N" value and the number of different MLTs having that number of copies (e.g. N x Z, where Z is the number of

different MLT sequences having N copies). For variant sequences arising from one or more mutant DNA template molecules, there was a high probability of finding multiply occurring MLTs with a high "N" value (because the number of clonal sequences sampled post-amplification was several-fold greater than the number of template DNA molecules that were copied and tagged in the primer-extension reaction). In contrast, for variant sequences arising from errors introduced during amplification of wild-type template molecules, it was unlikely to find MLT-1 sequences with "N" values as high as those associated with true mutant templates. Since known mutant template oligonucleotides had been spiked into each sample prior to amplification, these internal standards were used to determine the range of "N" values that should be expected for variant sequences derived from unknown mutant templates. Values falling below that range were presumed to be associated with variants arising from errors of amplification or sequencing.

7. A "mutation authenticity score" (MAS) can be used to facilitate the identification of variant sequences arising from mutant template DNA molecules. The MLT copy numbers, "N", that are associated with the spiked-in mutant internal control oligos (having mutations at two distinct positions) can be evaluated (Table 7). The variant sequences that arise from these authentic mutant templates are associated with MLT-1 sequences having relatively high "N" values. The value "N_{auth}" is in one embodiment the mean "N" value for these known authentic mutant templates. The "N_{auth}" value can be weighted or unweighted. If a mutation were introduced during the first cycle of PCR, the "N" value of such a variant sequence would be approximately (1/1.7) x N_{auth} (if each cycle of PCR yields approximately 1.7-fold amplification). Similarly, a variant sequence would have "N" values of approximately $(1/1.7^{Y})$ x N_{auth} (if a mutation were introduced during the yth cycle of PCR). Thus, a mutation authenticity score is calculated for each "consistent variant" sequence based on how close the "N" values of its MLTs are to the "N_{auth}" values of the authentic mutants that are spiked into the reaction. A variant would be likely to be authentic if its "N" values were distributed within a defined range of the N_{auth} values.

Results

A set of control plasma-derived DNA samples was tested. These samples contained various ratios of normal plasma DNA spiked with known amounts of mutant oligonucleotides (listed in Table 7). It was consistently observed that the PCR products were formed in a highly specific manner for all 40 gene segments included in the panel. The methods were extensively tested using a real-time quantitative thermal cycler, and comparisons to negative controls having no plasma DNA or having mouse DNA confirmed that the intended targets were being amplified. The products of all 40 PCRs were run on an agarose gel, and the production of appropriate-sized amplicons was confirmed.

Sequencing of the 40 pooled amplicons from multiple barcoded samples on the Illumina HiSeq® 2000 platform further confirmed that all intended gene segments were amplified. The total number of raw clonal sequences yielded was 282,965,036. After filtering, the rejected sequences were as follows:

Failed Illumina's chastity filter:	79,320,290
Positions 5-7 in forward read were not "ACT":	23,751,168
Rejected because of the presence of an N in position 8 or beyond:	27,477,576
Failed to recognize forward primer:	6,304,833
Barcode did not exactly match one in our set of 96:	15,921,744
Positions 11-12 of index read were not "AT":	2,903,342
Failed to recognize reverse primer:	17,064,651
Remaining filtered reads:	110,221,432

The total number of filtered counts assigned to each of the 40 gene segments is listed in Table 12. These data revealed a relatively even distribution of counts across the various amplicons.

Table 12. Number of filtered sequence counts associated with each targeted gene segment.

AKT1	1927236
APC	3263261
ATM	2988621
BRAF	2644671
BRAF-Cntrl	2827920
CTNNB1	3387874
EGFR-1	2582553
EGFR-2	2670482
EGFR-3	1908549
FGFR3-1	2441848
FGFR3-2	1907661
FGFR3-3	2154173
FoxL2	2782971
GNAS	2481154
HRAS-1	2173717
HRAS-2	2456244
KIT-1	1960170
KIT-2	4202032
KIT-3	2739896
KRAS-1	5647782
KRAS-2	3421539
KRAS-Cntrl	3076757

MET-1	2923088
MET-2	2956664
NRAS-1	4037334
NRAS-2	2462906
PIK3CA-1	2124016
PIK3CA-2	2807504
PPP2R1A-1	2087213
PPP2R1A-2	1966662
PTEN-1	3125952
PTEN-2	2562118
STK11-1	2969119
STK11-2	2613351
TP53-1	2494793
TP53-2	2248788
TP53-3	3074251
TP53-4	2964973
TP53-5	2632869
TP53-6	2522720
Total	110221432

The sequence data were processed using a modified version of the computer code that was used in Example 1. The results demonstrated that control double-mutant oligonucelotides that were spiked into plasma DNA could be reliably detected and quantified. Requiring consistency of overlapping paired-end reads appeared to eliminate the vast majority sequencer errors. Also, analysis of the MLT sequences associated with "consistent variants" made it possible to distinguish sequences arising from authentic mutant templates from those introduced during amplification or sequencing. An example of processed data for the *BRAF* gene target region for a sample in which approximately equal numbers of copies of normal plasma DNA and double-mutant control oligos were mixed is shown in Figure 16. This output represented the analysis of data from a single bin (single gene segment, single barcode). In this example, a total count of 103,742 clonal sequences were assigned to the bin, and 65,143 of these counts arose from amplification of the double-mutant oligonucleotide that was spiked in. The double-mutant sequences comprised approximately half of the total sequences in this bin. The MLT counts associated with each consistent variant were listed in the format N x y where y was the number of unique MLT sequences that had N copies associated with that particular consistent variant. It was observed that the variant sequence arising from spiked-in control mutant templates was associated with several MLT-1 sequences having high copy numbers (N). The highest value of N for this variant was 3742, and there were many distinct MLTs that had copy numbers in the thousands. In contrast, the next most abundant variant had only 965 total counts and was associated with only a few distinct MLTs. The highest value of N for this variant

was 576, and only one other MLT had a count in the hundreds range. All other consistent variants in the list were associated with very low MLT copy numbers. Based on these observations, it could be confirmed that only the spiked-in control oligonucleotides produced variant sequences that had high MLT counts. MLT counts associated with variant sequences that likely arose from errors of PCR or sequencing were much lower, as predicted.

Example 3

This example demonstrates the application of methods that incorporated methods of Example 2, and included modifications thereof. A modification included elimination of separate PCRs for each target DNA in the final step. Instead, the final amplification was performed in a single tube using universal PCR primers. This also eliminated the requirement for a pre-amplification step. Pooled amplification was made possible by copying, tagging, and purifying the targeted DNA regions in a highly selective manner; spurious templates that could be amplified by universal primers in the final PCR would be minimized (Figure 1). In this example, the same 40 genomic target regions were analyzed as in Example 2.

Methods

Preparation of mixtures of primers having combinations of modular oligonucleotide segments

Mixtures of primers having combinations of modular barcode segments and gene-specific segments were prepared as described in Example 2. The preferred approach, called "modular automated synthesis and purification", is schematized in Figure 3, and is described in detail in Example 2.

Collection and processing of patient plasma samples

Blood was collected and processed as described in Example 2.

Extraction and purification of DNA from plasma

DNA was extracted from plasma as described in Example 2.

Round 1 PCR

In order to make a limited number of tagged copies (fewer than 20) of the plasma-derived template DNA molecules, a few cycles of PCR were performed (in contrast to primer extension that was performed in Example 2). The reverse primers used in the first round of PCR were the modular barcoded mixtures of gene-specific primers as described above (same as the primers used in the

primer extension reaction in Example 2). For forward primers, the same oligonucleotides were used as the biotinylated capture oligonucleotides that had been used in Example 2 to purify the primer-extension products. The sequences of the forward primers are listed in Table 8.

Forty different gene regions were targeted, and therefore a combination of 40 different biotinylated forward primers and 40 different modular barcoded gene-specific reverse primers were used in the Round 1 PCR for each sample. For a given sample, the mixture of gene-specific reverse primers all had the same, sample-specific barcode in the 5' segment. The primer mixes were produced so that an approximately equimolar concentration of 40 different forward and 40 different reverse primers would be present in the reaction (final concentration of approximately 100 nM each primer). In addition to sample-specific barcodes, the reverse primers also contained degenerate sequence regions known as molecular lineage tags (MLTs) as well as common sequences at the 5'-end that allowed for hybridization of "universal" PCR primers in subsequent steps. The MLT assigned to each copy in Round 1 PCR was referred to as MLT-1.

Control DNA molecules containing known mutations were spiked into each Round 1 PCR to serve as internal quantitative standards. As described in Example 2, these DNA molecules were cartridge-purified oligonucleotides that were synthesized to contain variations from the wild-type sequence at two distinct positions. These variations allowed the control sequences to be readily distinguished from other variants within DNA purified from a clinical sample. The sequences of the top strands of these control DNA oligonucleotides are listed in Table 7. Bottom strands were also synthesized corresponding to the reverse complements of these 40 sequences. In order to make the control DNA as similar as possible to the clinically-derived DNA, both strands were annealed to make them double-stranded before adding them to the primer-extension reaction. The double-stranded DNA was quantified by UV spectrometry and then diluted to the desired concentration. To each PCR, approximately 200 copies of the double-stranded control DNA fragments corresponding to each of the 40 gene target sites were added.

The Round 1 PCR amplification consisted of the following components: (1) template DNA purified from plasma and eluted in 20 microliters of Qiagen elution buffer AVE, (2) 1 x Phusion® buffer HF, (3) 200 mM of each dNTP (dATP, dCTP, dGTP, and dTTP), (4) mixture of 40 reverse barcoded primers, 100 nM each, (5) mixture of 40 forward biotinylated primers, 100 nM each, (6) 200 copies of double-stranded control DNA, (7) molecular grade water as needed to make the desired total volume, and (8) Phusion® Hot Start Flex DNA polymerase, (0.04 U/μL). The total volume of each reaction was 40 microliters (for each 20 μ eluted plasma DNA sample). A separate reaction was set up for each sample.

Thermal cycling was carried out on a BioRad iCycler® using the following protocol: (1) 98°C for 45 seconds, (2) 98°C for 10 seconds, (3) 70°C for 30 seconds, (4) slowly cooling by 1°C every 30 seconds down to 56°C, (5) 55°C for 2 minutes, (6) 72°C for 1 minute, (7) repeat steps 2 to 6 for 3 cycles total, and (8) hold temperature at 72°C indefinitely.

As quickly as possible, while the reaction was still at 72°C, EDTA (10 mM final concentration) was added to terminate the polymerase activity. Each tube was agitated gently to ensure even mixing of the EDTA. Since the PCR products now had sample-specific barcodes attached, the products of all reactions could be pooled together into a single tube.

Purification of Round 1 PCR products

Since the forward primers used in the Round 1 PCR contained biotin tags at their 5'-ends, these tags were incorporated into the PCR products and were used to purify the products. To capture the biotin-tagged PCR products, 10 µL of high capacity streptavidin-agarose bead slurry (Thermo-Fisher) was added (per 40 µL PCR). Thus, for example, if fifty Round 1 PCRs were performed in a volume of 40 µL each, then the volume after combining all samples would be 2 mL, and 500 µL of bead slurry would be used. Tubes were turned end-over-end constantly for at least 2 hours at room temperature to promote binding of biotinylated DNA to the streptavidin beads. Beads were then gently and briefly centrifuged at low speed, and any unbound supernatant was carefully removed, avoiding aspiration of any beads. The beads were then washed in 200 µL of buffer containing 10 mM Tris pH 7.6, 50 mM NaCl, and 1 mg/mL salmon sperm DNA ("wash buffer"). Beads were suspended in wash buffer by gentle agitation, were gently centrifuged, and then the supernatant wash buffer was discarded. A second wash was performed in the same way, except that the suspended beads were incubated at 50°C for 25 minutes followed by 60°C for 5 minutes while the tube was turned end-over-end to promote dissociation of any DNA molecules that may have annealed non-specifically to the biotinylated oligonucleotides. The beads were again centrifuged gently, and the supernatant wash buffer was removed.

Optionally, between the first and second washes, the beads were treated with Exonuclease I (New England Biolabs) in order to digest any single stranded DNA (including un-extended biotinylated primer) that was bound to the beads. For the tested samples, it was found that this nuclease treatment was not necessary following the first Round of PCR. For digestion, the beads were suspended in 1x Exonuclease I buffer (2 µL for every 1 µL of beads), and then Exonuclease I enzyme was added to a final concentration of 0.5 µL. The reaction was incubated at 37°C for 30 minutes. The beads were then centrifuged, the supernatant was discarded, and the beads then were subjected to the second wash.

The captured PCR products were then eluted from the surface of the washed beads by heat-denaturation. Elution was carried out by heating the beads to 95°C for 30 seconds directly in Round 2 PCR cocktail (as described below), gently centrifuging the beads, and harvesting the eluted DNA within the supernatant cocktail. Note that only one strand of the PCR product was eluted because the biotin-streptavidin interaction was not substantially disrupted by heating at 95°C, and thus the biotinylated strand would remain bound to the beads. Likewise, any un-extended biotinylated oligonucleotides would also remain bound to the beads.

Round 2 PCR

The second round of PCR was also performed for only a few cycles (between 2 and 4). This PCR provided additional selectivity by using a mixture of 40 nested forward primers that would specifically hybridize to the desired genomic target sequences. This step also provided a second molecular lineage tag on the other side of the mutation-prone target sequence (opposite to the barcode and MLT-1). The forward primers contained a stretch of degenerate positions, called "molecular lineage tag – 2" (MLT-2), which was useful in determining which sequences had become labeled with the wrong barcode due to sequence crossover during pooled amplification. The forward primers also contained a common sequence at their 5'-ends which served as a universal primer binding site in the third and final round of PCR. This common sequence also provided some of the adapter sequences required for sequencing on the Illumina platform. The reverse primer used in Round 2 PCR had a biotin tag at its 5'-end which was used for purification of the Round 2 PCR products.

The purified Round 1 PCR products were eluted directly into a cocktail that was used for Round 2 PCR. For every 10 µL of bead slurry that was used, 40 µL of PCR cocktail was used for elution. The Round 2 PCR cocktail consisted of the following components: (1) 1 x Phusion® buffer HF, (2) 200 mM of each dNTP (dATP, dCTP, dGTP, and dTTP), (3) a mixture of 40 nested forward primers, 100 nM each, (4) 10 ng/ µL salmon sperm DNA, and (5) molecular grade water as needed to make the desired total volume.

After elution of the single-stranded PCR product from the beads into the above cocktail (and removal of the beads), a biotinylated universal reverse primer was added to achieve a final concentration of 200 nM. This biotinylated primer had to be added to the cocktail after removal of the streptavidin-agarose beads to prevent the biotin from binding to the beads. Finally, Phusion® Hot Start Flex DNA polymerase was added to the cocktail to a final concentration of 0.04 units per microliter, and was mixed by gently pipetting the cocktail up and down. If the total volume was

greater than recommended for a single PCR tube, then the cocktail was split into the appropriate number of identical reaction volumes.

Thermal cycling was carried out on a BioRad iCycler® using the following protocol: (1) 98°C for 45 seconds, (2) 98°C for 10 seconds, (3) 70°C for 30 seconds, (4) slowly cooling by 1°C every 30 seconds down to 61°C, (5) 60°C for 2 minutes, (6) 72°C for 1 minute, (7) repeat steps 2 to 6 for 3 cycles total and, (8) hold temperature at 72°C indefinitely.

As quickly as possible, while the reaction was still at 72°C, EDTA (10 mM final concentration) was added to terminate the polymerase activity. The tube was agitated gently to ensure even mixing of the EDTA.

The sequences of the 40 nested forward primers were the same as those provided in Table 10, except that the sequence "NNNNACT" in each primer was replaced by "NNNNNN". The common "ACT" sequence was removed because it led to poor sequence diversity which produced low-quality base-calls on the Illumina sequencer. Instead, the stretch of degenerate positions was increased from 4 to 6 bases to provide a greater number of sequence combinations at MLT-2. The sequence of the biotinylated reverse primer used in Round 2 PCR (called BioV2rev) was as follows: 5'-Biotin-CGAGACGGATCAAGCA GAAGACG-3' (SEQ ID NO:214).

Purification of Round 2 PCR products

The biotin tag at the 5'-end of the reverse primer used in Round 2 PCR was used to capture and purify the products of Round 2 PCR. This step removed any un-extended forward primers, as well as many spurious products that might have been produced during the amplification, which prevented inappropriate incorporation of new MLTs during the next round of amplification.

The capture, washing, digestion, and elution of the Round 2 PCR products was performed in a manner that was essentially identical to the process described above for the purification of Round 1 PCR products. In Round 2 PCR purification, the Exonuclease I step was not optional. Thus, the beads were washed once in wash buffer at room temperature, then were treated with Exonuclease I, and then were washed a second time at elevated temperature (50°C for 25 minutes followed by 60°C for 5 minutes) to remove non-specific DNA. Fewer beads were used for a given volume of Round 2 PCR reaction. Five microliters of bead slurry was used for every 40 µL of PCR reaction volume.

Elution of the captured PCR products was also performed in a manner that was essentially the same as that used for purification of the Round 1 PCR products. The streptavidin-agarose beads were heated to 95°C for 30 seconds to elute the product directly into a cocktail that was used for

Round 3 PCR (described below). The biotinylated strand of the PCR product remained bound to the beads, while the opposite strand was eluted into the Round 3 PCR cocktail.

Round 3 PCR

The third and final round of PCR amplified the DNA molecules that were specifically tagged, copied, and purified in the first 2 rounds of PCR. To provide sufficient DNA for visualization by ethidium bromide staining on an agarose gel, the amount of PCR product from Round 3 had to be substantial (at least 0.5 microgram). Thus, the final PCR amplification was carried to saturation or beyond (typically 15 to 35 cycles, depending on the amount of template DNA in each sample and the total number of samples that were pooled).

In contrast to the final PCRs in Example 2 which were performed separately for each genomic target site, the final PCR in the present Example was performed in a combined reaction volume for all genomic targets and for all samples. This extremely high level of multiplexing was only possible because of the highly selective methods used for amplification and purification in the prior two rounds of PCR.

As described above, the round 2 PCR products were eluted directly into Round 3 PCR cocktail. The volume of this cocktail depended on the volume of beads used. For every 5 µL of bead slurry, 20 µL of PCR cocktail was used. The Round 3 PCR cocktail consisted of the following components: (1) 1 x Phusion® buffer HF, (2) 200 mM of each dNTP (dATP, dCTP, dGTP, and dTTP), (3) Universal forward and reverse primers, 200 nM each, (4) 10 ng/ µL salmon sperm DNA, and (5) molecular grade water as needed to make the desired total volume.

After elution of the single-stranded PCR product from the beads into the above cocktail, Phusion® Hot Start Flex DNA polymerase was added to the cocktail to a final concentration of 0.04 U/µL, and was mixed by gently pipetting the cocktail up and down. If the total volume was greater than recommended for a single PCR tube, then the cocktail was split into the appropriate number of identical reaction volumes. Mineral oil (20 µL) was added to the tube(s) to prevent evaporation during PCR.

Thermal cycling was carried out on a BioRad® iCycler using the following protocol: (1) 98°C for 45 seconds, (2) 98°C for 10 seconds, (3) 62°C for 30 seconds, (4) 72°C for 20 seconds, (5) repeat steps 2 to 4 for 35 cycles total, and (8) hold temperature at 4°C indefinitely.

Soon after the reaction had reached 4°C, EDTA (10 mM final concentration) was added to terminate the polymerase activity. Since the PCR product was under mineral oil, a pipette with a filtered tip was used to evenly mix the EDTA. Special care was taken to avoid contamination of other reagents and workspaces with PCR products.

Preparation of DNA for Next-generation sequencing

The product of the Round 3 PCR was purified on a 2% agarose gel, as described in Example 2. Since the products were not of a homogeneous length, a somewhat diffuse band was seen on the gel. The band was cut with a few mm margin above and below to ensure inclusion of any low-intensity bands that may have been difficult to visualize. A QIAquick® Gel Extraction kit (Qiagen) was used to isolate the DNA from the gel slice. The DNA was eluted into 50 µL of EB buffer (supplied in the kit).

Next-generation sequencing

Next generation sequencing was performed as described in Example 2, using the Illumina HiSeq® 2000 platform. In the present example, the Illumina MiSeq® instrument was also used with similar success for samples requiring less sequence depth. In contrast to Example 2, addition of Phi-X DNA to improve sequence diversity was not necessary in the present Example because modification of the Round 2 PCR forward primers to remove the common "ACT" sequence and to lengthen MLT-2 resulted in adequate sequence diversity.

Outline of algorithm for sequence analysis

Essentially the same algorithm that was described in Example 2 was applied to the data generated in Example 3. Although many of the processing steps used in Example 3 differ from those used in Example 2, the structure of the final double stranded DNA products are virtually identical. Thus, a very similar algorithm can be applied for sorting, aligning, and counting the resulting sequences. As noted above, the region of MLT-2 which was "NNNNACT" in Example 2 was replaced with "NNNNNN" in Example 3, and this change was accounted for in the modified algorithm.

To minimize the probability of mis-classifying a variant sequence as belonging to the wrong sample, MLT-1 and MLT-2 sequences were used to distinguish sequences in which barcode "cross-over" may have occurred during pooled amplification. Since a portion of MLT-1 is adjacent to the barcode sequence, and MLT-2 is on the other side of the target region (Figure 14), molecules that undergo such cross-over between the barcode and the mutation-prone region would also undergo cross-over between MLT-1 and MLT-2 (or between the two separate regions of MLT-1). Such "crossed-over" sequences would be expected to have a low number of copies having a given combination of MLT-1 and MLT-2 sequences. In contrast, sequences arising from an authentic mutant template that remained attached to its originally assigned barcode would be expected to have greater copies of a given MLT-1 and MLT-2 combination.

The algorithm in Example 3 was modified to facilitate evaluation of the relationship between MLT-1 and MLT-2 sequence counts for each "consistent variant" and also for the wild-type sequences. In order to report these counts in a reasonably succinct format, it was necessary to bin MLT counts by powers of two. For example, an MLT-1 count of 13 would be placed into bin 4 (because 2^4 is the smallest power of 2 that is greater than or equal to 13). Thus, a report of 4x5 meant that there were five instances of counts in the range of 9 to 16. Similarly, a report of 3x6 meant that there were six instances of counts in the range of 5 to 8. For a given collection of MLT-1 counts, the associated MLT-2 counts were reported in a similar format , to the right of the MLT-1 counts and separated by colons. For example, 4x5:2x3:1x7 meant that among 5 sets of MLT-1 sequences occurring between 9 and 16 times, there were 3 instances of MLT-2 sequences that occurred between 3 and 4 times, and 7 instances of MLT-2 sequences that occurred twice. Different MLT-1 bins were separated by a space.

Results

Purified DNA that was obtained from 0.5 mL of plasma of healthy volunteers was mixed with various amounts of the control mutant oligonucleotides listed in Table 7. Between 200 and 5,000 copies of each of the 40 control oligonucleotides were added to each purified plasma DNA sample. These mixtures were subjected to 3 rounds of PCR and purification as described in the methods. The highly multiplexed Round 3 PCR in which multiple gene targets from multiple samples were amplified in a single tube, resulted in the specific production of amplicons of the expected size. As shown in Figure 17, a relatively broad band was seen migrating at a size corresponding to between 200 and 300 base pairs on a 2% non-denaturing agarose gel (approximately centered at 250 base pairs). The primers and target regions were of variable length, and the amplification products spanned a range of sizes. Extensive testing of negative controls confirmed that these products were specific and were absent when mouse DNA or no template DNA was substituted for human plasma DNA in Round 1 PCR.

The gel-extracted PCR products were subjected to next-generation sequencing using an Illumina MiSeq® instrument. The total number of raw clonal paired-end sequences was 20,511,389. After application of the various filters described above, the remaining sequences numbered 11,184,975. The 40 different gene target regions were fairly evenly represented among the filtered sequences. The median sequence count for the 40 gene-specific regions (all barcodes) was 166,867.

After processing of the sequence data using the computer algorithm described above, control mutant oligonucleotides that were spiked into the plasma DNA were identified and quantified. Importantly, they were readily distinguished from the vast majority of errors introduced

during amplification, processing, or sequencing. As observed previously in Example 1 and Example 2, the sequence redundancy provided by the clonal overlapped paired-end reads was able to virtually eliminate sequencer-generated errors in the mutation-prone sequence regions. The "consistent variants" were then analyzed for the distribution of their associated MLT sequences. As an example, the summary output for analysis of sequences belonging to a single barcode and target gene region KRAS-2 (region surrounding codon 61 of the KRAS gene) is shown in Figure 18. In this sample, approximately 200 copies of the mutant oligonucleotides were spiked in (each having two distinct mutations relative to the wild-type sequence). The mismatches of the "consistent variants" relative to the reference wild-type sequence are displayed in the lower portion of Figure 18. This single data bin contained 13,315 total sequences, of which 10,815 were exact matches to the wild-type sequence and 1,767 were exact matches to the spiked-in mutant sequence (an exact match requires that the overlapping portions of the paired-reads agree with each other). The spiked-in mutant sequences comprised approximately 10% to 15% of the total DNA in the sample, which is in the expected range (there should be approximately 1,000 to 2,000 genome copies of fragmented DNA in 0.5 mL of plasma). The counts of MLT-1 and MLT-2 are reported for each "consistent variant" according to the scheme described above. The MLT-1 counts associated with sequences arising from the spiked-in control mutant oligonucleotides were generally higher than those associated with other variant sequences, as expected. This made it possible to distinguish many of the "consistent variants" arising from polymerase misincorporations that might have otherwise been mistaken for sequences arising from true mutant template molecules. Analysis of MLT-2 counts associated each group of MLT-1 counts provided insight into the efficiency of molecular tagging and copying at PCR Rounds 1 and 2. It also helped to distinguishing variants assigned to the wrong sample due to barcode cross-over during pooled amplification.

Example 4: Splint-mediated enzymatic ligation of modular oligonucleotide segments

In this example, an alternative approach is described for the production of mixtures of primers in which each mixture had a common 5' barcode segment and a variety of gene-specific 3' segments. Enzymatic ligation was used to concentrate modular oligonucleotide segments. More specifically, in each ligation, a uniquely barcoded 5' oligonucleotide segment was ligated to a uniform mixture of different gene-specific 3' segments. A DNA splint was used to faciliate the ligation.

Gene-specific oligonucleotides with a common sequence at the 5'-end (and a 5'-phosphate group added during oligonucleotide synthesis) were mixed in equimolar ratios. The uniform mixture

was divided into separate tubes and was ligated to a uniquely barcoded oligonucleotide in each tube using a biotin-tagged DNA splint as illustrated in Figure 13. The sequences of the 5'-phosphorylated gene-specific oligonucleotides are listed in Table 13.

Table 13. List of chemically 5'-phosphorylated gene-specific oligonucleotides used for splint-mediated modular ligation.

Name	DNA Sequence	
Ph-TP53-1	X-AGACGTGTGCTCTTCCGATCTGTGCTGTGACTGCTTG	298
Ph-TP53-2	X-AGACGTGTGCTCTTCCGATCTAGCACATGACGGAGGTT	299
Ph-TP53-3	X-AGACGTGTGCTCTTCCGATCTCAAATACTCCACACGCAAATT	300
Ph-TP53-4	X-AGACGTGTGCTCTTCCGATCTATTTGGATGACAGAAACACTT	301
Ph-TP53-5	X-AGACGTGTGCTCTTCCGATCTTGTGATGATGGTGAGGATGG	302
Ph-TP53-6	X-AGACGTGTGCTCTTCCGATCTGGGACGGAACAGCTTTGAG	303
Ph-PIK3CA-1	X-AGACGTGTGCTCTTCCGATCTGCAATTTCTACACGAGATCCTCT	304
Ph-PIK3CA-2	X-AGACGTGTGCTCTTCCGATCTCTTTGGAGTATTTCATGAAACAAATGA	305
Ph-BRAF	X-AGACGTGTGCTCTTCCGATCTACAGTAAAAATAGGTGATTTTGGTCTA	306
Ph-FoxL2	X-AGACGTGTGCTCTTCCGATCTGCAACTACTGGACGCTGGAC	307
Ph-GNAS	X-AGACGTGTGCTCTTCCGATCTCAATTTTGTTTCAGGACCTGCT	308
Ph-CTNNB1	X-AGACGTGTGCTCTTCCGATCTGGCAGCAACAGTCTTACCT	309
Ph-PPP2R1A-1	X-AGACGTGTGCTCTTCCGATCTCCC AGC TTG GAG GCT GC	310
Ph-PPP2R1A-2	X-AGACGTGTGCTCTTCCGATCTGCCAGGCCGCTGAAGACA	311
Ph-PTEN-1	X-AGACGTGTGCTCTTCCGATCTGCAATTCACTGTAAAGCTGGAAAG	312
Ph-PTEN-2	X-AGACGTGTGCTCTTCCGATCTGAAGATATATTCCTCCAATTCAGGAC	313
Ph-KIT-1	X-AGACGTGTGCTCTTCCGATCTCGTTTCCTTTAACCACATAATTAGAATC	314
Ph-KIT-2	X-AGACGTGTGCTCTTCCGATCTTTTCCCTTTCTCCCCACAG	315
Ph-KIT-3	X-AGACGTGTGCTCTTCCGATCTTCCTGTAGCAAAACCAGAAATC	316
Ph-EGFR-1	X-AGACGTGTGCTCTTCCGATCTGGTGAGAAAGTTAAAATTCCCGTC	317
Ph-EGFR-2	X-AGACGTGTGCTCTTCCGATCTAGCATGTCAAGATCACAGATTTTG	318
Ph-EGFR-3	X-AGACGTGTGCTCTTCCGATCTCACCTCCACCGTGCAGCT	319
Ph-AKT1	X-AGACGTGTGCTCTTCCGATCTACCACCCGCACGTCTGT	320
Ph-ATM	X-AGACGTGTGCTCTTCCGATCTCTTCCATACTTGATTCATGATATTTTACT	321
Ph-APC	X-AGACGTGTGCTCTTCCGATCTACCTCCTCAAACAGCTCAAAC	322
Ph-FGFR3-1	X-AGACGTGTGCTCTTCCGATCTTGGGAGATCTTCACGCTGG	323
Ph-FGFR3-2	X-AGACGTGTGCTCTTCCGATCTCCCTGAGCGTCATCTGCC	324
Ph-FGFR3-3	X-AGACGTGTGCTCTTCCGATCTGCTGGTGGAGGCTGACGA	325
Ph-MET-1	X-AGACGTGTGCTCTTCCGATCTTCCCTATCAAATATGTCAACGACT	326
Ph-MET-2	X-AGACGTGTGCTCTTCCGATCTATTTTGGTCTTGCCAGAGACA	327
Ph-STK11-1	X-AGACGTGTGCTCTTCCGATCTATCGACTCCACCGAGGTCA	328
Ph-STK11-2	X-AGACGTGTGCTCTTCCGATCTACTTGGAGGACCTGCACG	329
Ph-KRAS-1	X-AGACGTGTGCTCTTCCGATCTCGTCAAGGCACTCTTGCCT	330
Ph-KRAS-2	X-AGACGTGTGCTCTTCCGATCTGATATTCTCGACACAGCAGGT	331
Ph-NRAS-1	X-AGACGTGTGCTCTTCCGATCTTCAGTGCGCTTTTCCCA	332
Ph-NRAS-2	X-AGACGTGTGCTCTTCCGATCTGACATACTGGATACAGCTGGA	333
Ph-HRAS-1	X-AGACGTGTGCTCTTCCGATCTGGTCAGCGCACTCTTGCCC	334
Ph-HRAS-2	X-AGACGTGTGCTCTTCCGATCTCATCCTGGATACCGCCGGC	335
Ph-KRAS-C	X-AGACGTGTGCTCTTCCGATCTCCTGTTTATAATATTGACAAAACACCT	336
Ph-BRAF-C	X-AGACGTGTGCTCTTCCGATCTCAGGACAAAGTCCGGATTGA	337

X = 5'-posphate added chemically during oligonucleotide synthesis.

The sequences of the 96 different barcoded oligonucleotides contained the following common sequence, with each oligonucleotide containing a different 8-nucleotide barcode from the list in Table 6 inserted into the position marked [BC1-96]:

5'-CGAGACGGATCAAGCAGAAGACGGCATACGAGAT[BC1-96]NNNNNNGTGACTGGAGTTC-3' (SEQ ID NO:215)

The sequence of the 3'-biotin tagged splint oligonucleotide was:

5'-

ATCGGAAGAGCACACGTCTGAACTCCAGTCACAAAAAAAAAAAAATCTCGTATGCCGTCTTCTGCTTGATCCG TCTCG-3'-Biotin-TEG (SEQ ID NO:216)

The barcoded oligonucleotides were cartridge purified to ensure that they were mostly full-length. They were synthesized at the 40 nmole scale, with an expected full-length yield of approximately 50 to 60%. The phosphorylated gene-specific oligonucleotides and splint oligonucleotide were purified on a polyacrylamide gel (as described in Sambrook JJ, et al., *Molecular Cloning: A Laboratory Manual*, Cold Spring Harbor Press, 2001) .

The ligation reactions were carried out using the following conditions:

Molecular grade water		38.15 µL
Dithiothreitol (100 mM stock)		2.5 µL
NEB Ligase buffer (10x)		10 µL
5'-Phosphorylated oligo mix (263 uM stock) (7uM final)		2.7 µL
splint oligo (248 uM stock) (14 uM final)		5.65 µL
Barcoded oligo (50 uM stock) (20uM final)		40 µL
	Total	100 µL

The 5'-phosphorylated oligonucleotide mix consisted of an equimolar mixture of 40 different gene-specific oligos. 96 different reactions were set up in separate tubes, each one with a different barcoded oligonucleotide. To anneal the oligonucleotides to the splint, the reaction mixes were heated on a thermal cycler as follows: 95°C for 30 sec, 70°C for 20 sec, then the temperature was decreased by 2.5°C every 20 sec until the samples reached 25°C.

Then 2 µL of T4 DNA Ligase (400,000 U/mL, New England Biolabs) was added to each reaction, and after mixing, the reactions were incubated at 25°C for at least 2 hours.

Then 20 µL of streptavidin-agarose high-capacity bead slurry (Thermo Scientific, Pierce) was added, and the samples were incubated at room temperature while being turned end-over-end on a rotisserie for at least 2 hrs.

The streptavidin-agarose beads were then washed three times with 200 microliters of Tris 10 mM pH 7.6, NaCl 50 mM. The ligated (and unligated) DNA molecules were then eluted from the beads by heat-denaturation of the DNA duplex. The majority of the biotinylated splint oligo remained attached to the beads because the biotin-steptavidin interaction was not significantly disrupted by heating at 95°C. The elution was carried out in 2 steps. In the first step, the beads were heated to 95°C in 40 microliters of the Tris/NaCl buffer for 30 seconds. The beads were quickly spun down by brief centrifugation, and then the supernatant containing was removed and stored. In the second step, the same elution process was carried out, but with heating to 95°C for 45 seconds in order to remove any remaining ligated DNA from the beads. The supernatants containing the ligated (and unligated) DNA from the first and second elution steps were combined into a total volume of 80 microliters. This process yielded approximately 600 to 700 picomoles of ligated oligonucleotides in 80 microliters of buffer, for a final concentration of approximately 7-8 micromolar.

CLAIMS

1. A method of measuring nucleic acid variants comprising

amplifying a low abundance template nucleic acid in a multiplexed polymerase chain reaction (PCR),

wherein the PCR comprises gene specific primers comprising a) a barcode; and b) a molecular lineage tag (MLT),

wherein the gene specific primers are specific for a mutation prone region.

2. The method of claim 1, wherein the primers are specific for more than one mutation prone region.

3. The method of claim 1 or claim 2, wherein the primers are specific for at least 40 mutation prone regions.

4. The method of any one of claims 1 to 3, further comprising targeted early barcoding.

5. The method of any one of claims 1 to 4, wherein the MLT is 2 to 10 nucleotides

6. The method of claim 5, wherein the MLT is 6 to 8 nucleotides.

7. The method of any one of claims 1 to 6 further comprising a second multiplexed PCR comprising nested gene specific primers comprising a second MLT, wherein the first and second MLTs at least partially flank the barcode and the mutation-prone region.

8. The method of any one of claims 1 to 7, wherein the low abundance template nucleic acid is circulating tumor DNA (ctDNA).

9. The method of any one of claims 1 to 8 further comprising clonal overlapping paired-end sequencing.

10. The method of claim 1, wherein the nucleic acid variant is measured before and after treatment.

11. A method of primer extension comprising targeted early barcoding.

12. The method of claim 11, wherein DNA copies were tagged by early barcoding in a first primer-extension comprising a plurality of gene specific primers, wherein the plurality of gene specific primers in each sample comprise an identical barcode.

13. The method of any one of claims 1 to 12, wherein modular oligonucleotide segments are combined to produce the gene specific primers.

14. The method of any one of claims 1 to 13, wherein a single barcoded oligonucleotide segment is combined with a uniform mixture of gene-specific segments to produce a uniquely barcoded mixture of the gene specific primers.

15. A method of distinguishing true nucleic acid variants from nucleotide misincorportations during PCR comprising

a) amplifying template nucleic acids with primers comprising MLTs, wherein there is a diversity of MLT sequences and a MLT sequence can be assigned to more than one template molecule; and

b) identifying misincorporation errors wherein sampled clonal sequences are several-fold greater than tagged template copies.

Figure 1

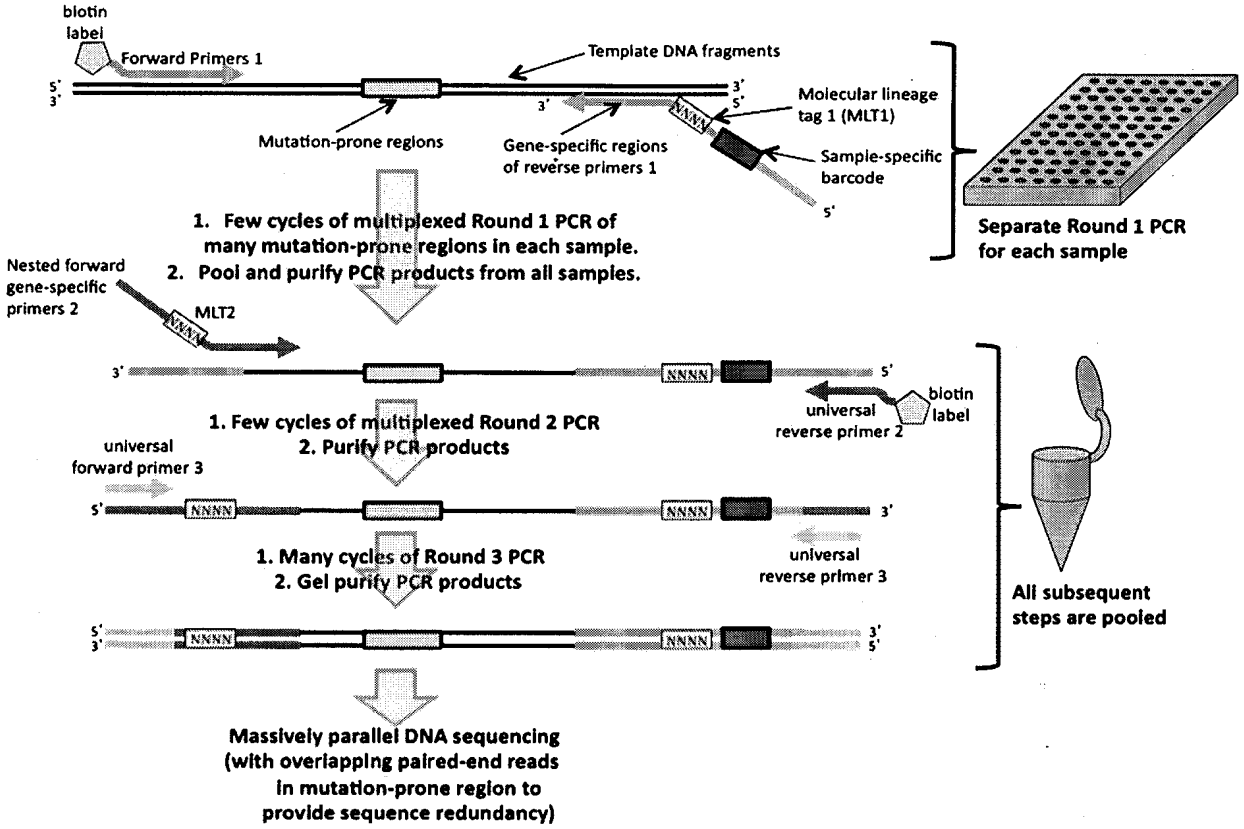
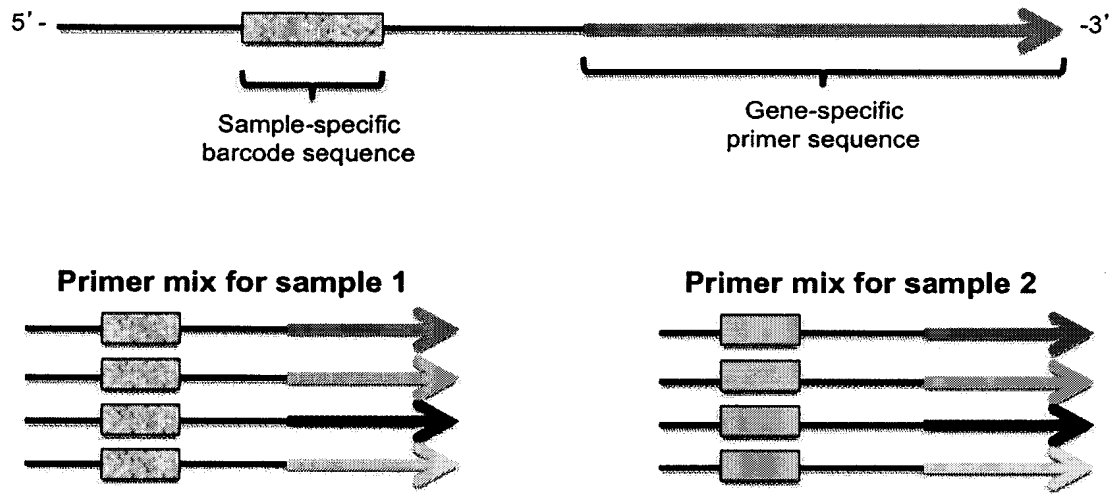
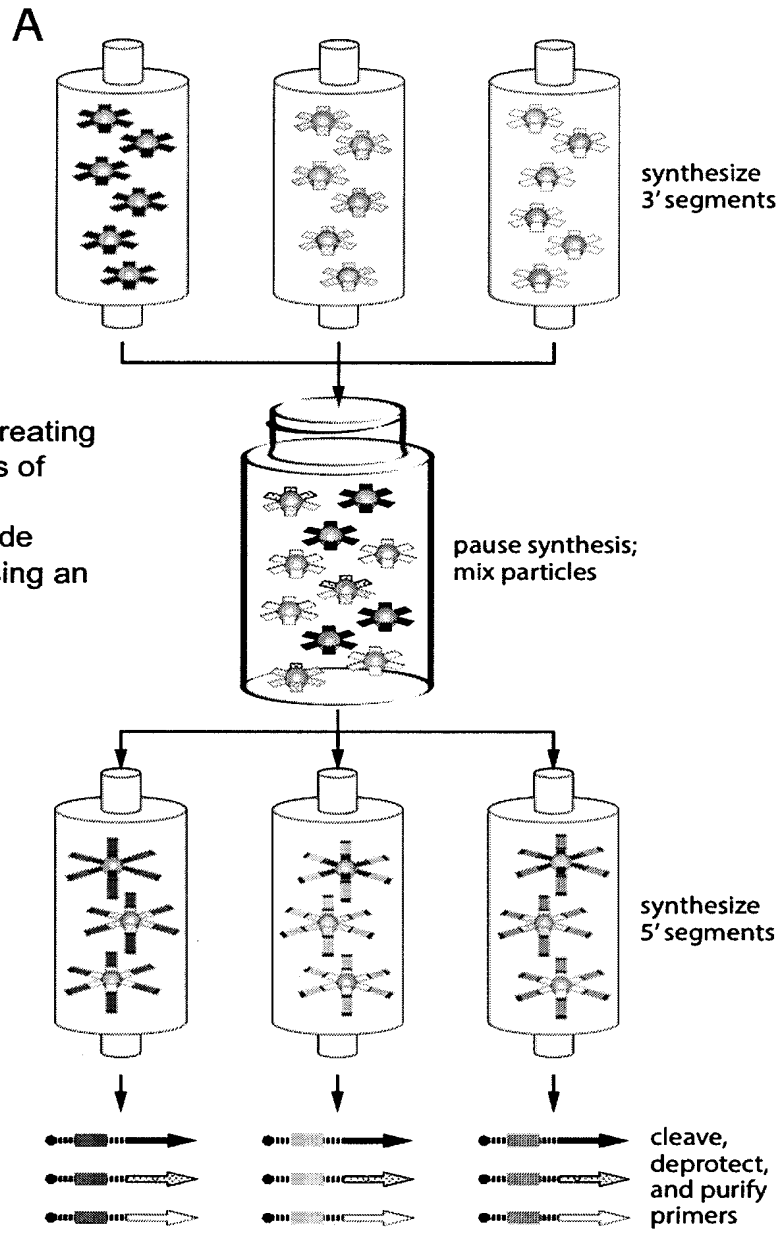

biotin label
Forward Primers 1
Template DNA fragments
Molecular lineage tag 1 (MLT1)
Mutation-prone regions
Gene-specific regions of reverse primers 1
Sample-specific barcode
Separate Round 1 PCR for each sample
1. Few cycles of multiplexed Round 1 PCR of many mutation-prone regions in each sample.
2. Pool and purify PCR products from all samples.
Nested forward gene-specific primers 2
MLT2
universal reverse primer 2
biotin label
1. Few cycles of multiplexed Round 2 PCR
2. Purify PCR products
universal forward primer 3
universal reverse primer 3
1. Many cycles of Round 3 PCR
2. Gel purify PCR products
All subsequent steps are pooled
Massively parallel DNA sequencing (with overlapping paired-end reads in mutation-prone region to provide sequence redundancy)

Figure 2 Combining modular oligonucleotide segments



Primer mixtures for samples 1 and 2 contain identical ratios of multiple gene-specific sequences. Each mixture has a distinct sample-specific barcode.

Figure 3



Method for creating combinations of modular oligonucleotide segments using an automated synthesizer.

Figure 4

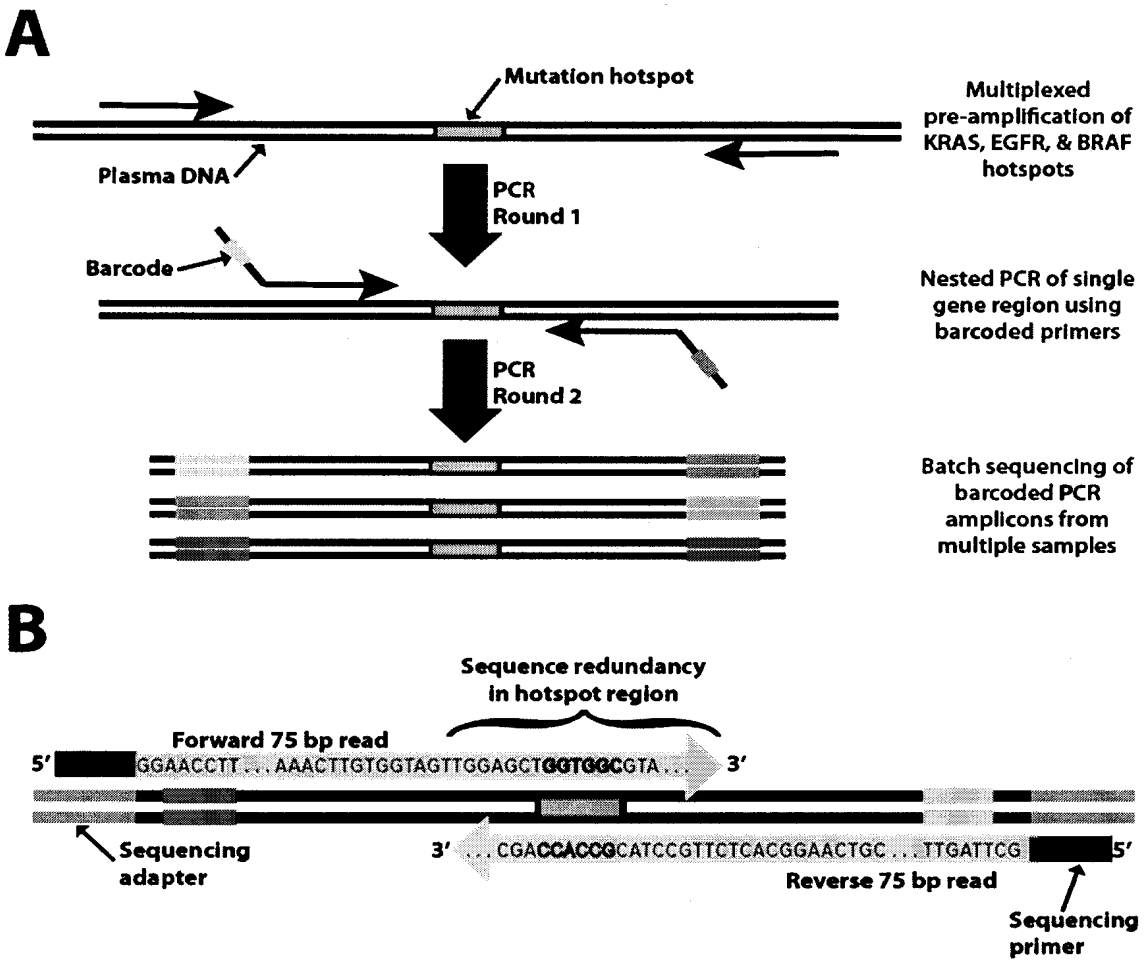

A
Mutation hotspot
Plasma DNA
PCR
Round 1
Barcode
PCR
Round 2
Multiplexed pre-amplification of KRAS, EGFR, & BRAF hotspots
Nested PCR of single gene region using barcoded primers
Batch sequencing of barcoded PCR amplicons from multiple samples
B
Sequence redundancy in hotspot region
Forward 75 bp read
5′ GGAACCTT . . . AAACTTGTGGTAGTTGGAGCTGGTGGCGTA . . . 3′
Sequencing adapter
3′ . . . CGACCACCGCATCCGTTCTCACGGAACTGC . . . TTGATTCG 5′
Reverse 75 bp read
Sequencing primer

Figure 5

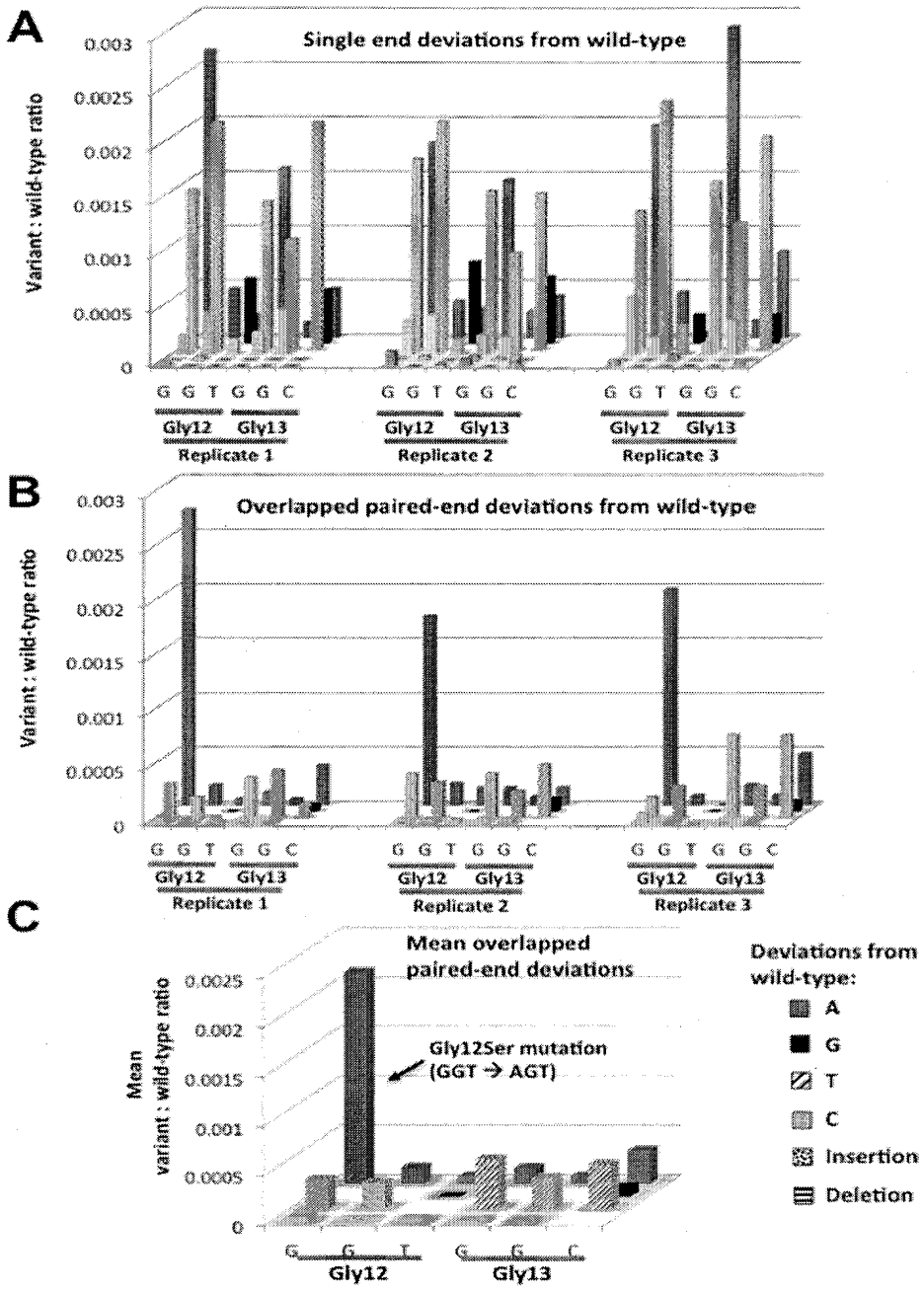

A
Single end deviations from wild-type
Variant : wild-type ratio
0.003
0.0025
0.002
0.0015
0.001
0.0005
0
G G T G G C
Gly12 Gly13
Replicate 1
Replicate 2
Replicate 3
B
Overlapped paired-end deviations from wild-type
C
Mean overlapped paired-end deviations
Mean variant : wild-type ratio
Gly12Ser mutation (GGT → AGT)
Deviations from wild-type:
A
G
T
C
Insertion
Deletion

Figure 6

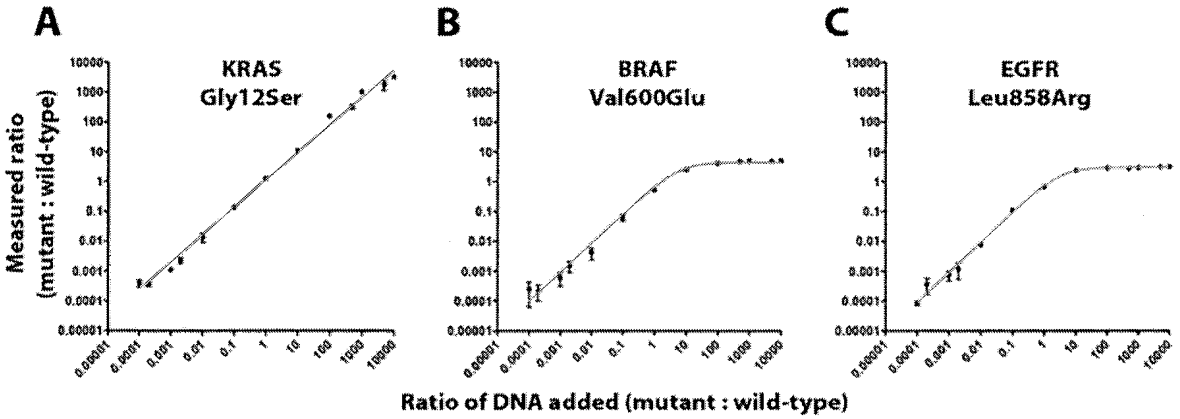

A
KRAS
Gly12Ser
B
BRAF
Val600Glu
C
EGFR
Leu858Arg
Measured ratio
(mutant : wild-type)
Ratio of DNA added (mutant : wild-type)

Figure 7

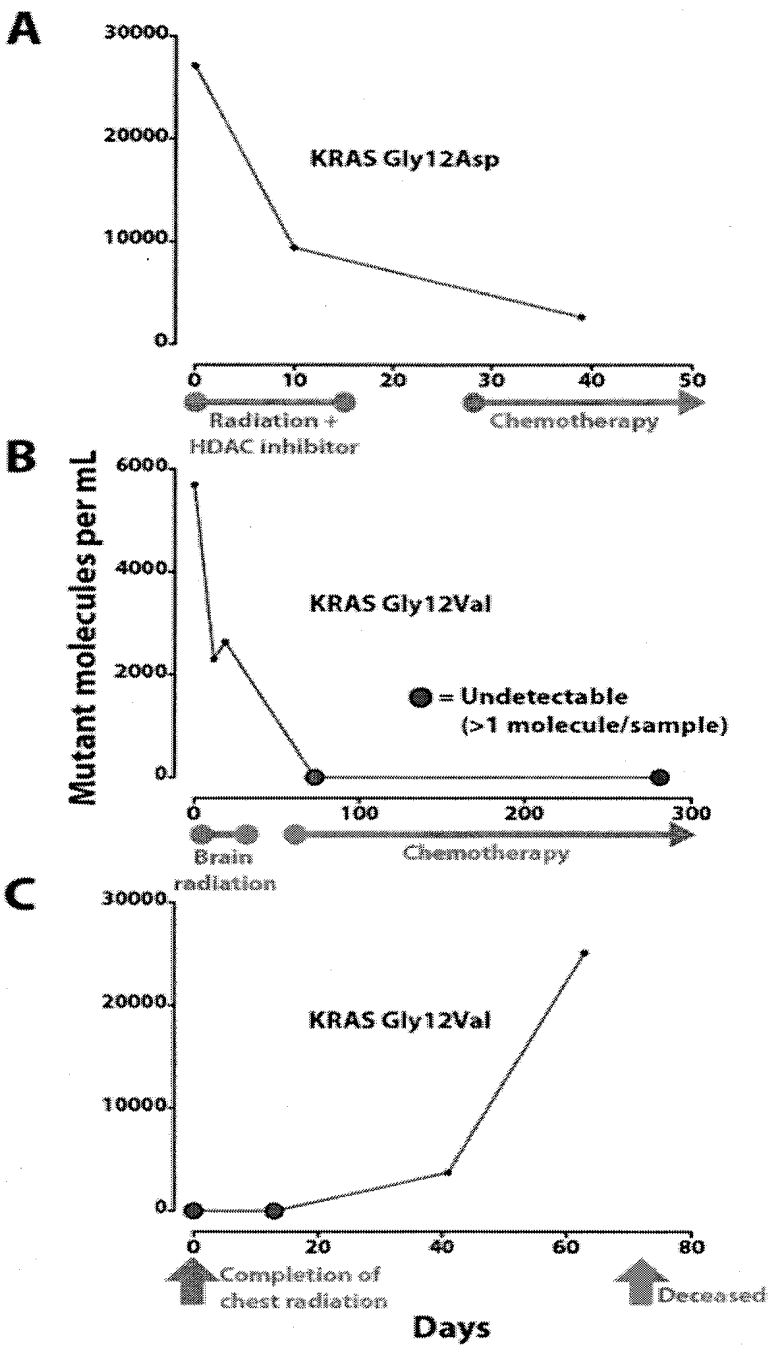

A
30000
20000
10000
0
KRAS Gly12Asp
0
10
20
30
40
50
Radiation +
HDAC inhibitor
Chemotherapy
B
Mutant molecules per mL
6000
4000
2000
0
KRAS Gly12Val
= Undetectable
(>1 molecule/sample)
0
100
200
300
Brain
radiation
Chemotherapy
C
30000
20000
10000
0
KRAS Gly12Val
0
20
40
60
80
Completion of
chest radiation
Deceased
Days

Figure 8

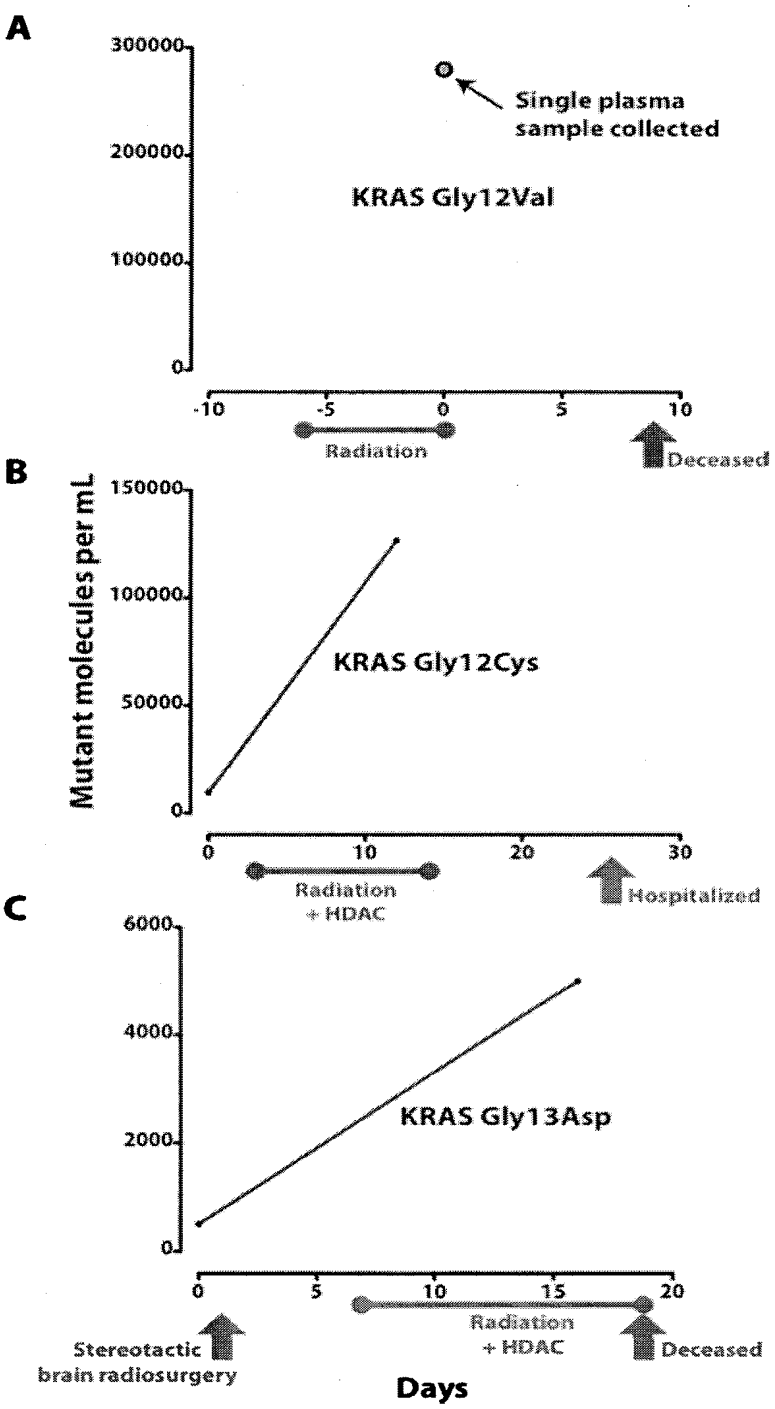

A
300000
200000
100000
0
Single plasma sample collected
KRAS Gly12Val
-10
-5
0
5
10
Radiation
Deceased
B
Mutant molecules per mL
150000
100000
50000
0
KRAS Gly12Cys
0
10
20
30
Radiation + HDAC
Hospitalized
C
6000
4000
2000
0
KRAS Gly13Asp
0
5
10
15
20
Radiation + HDAC
Stereotactic brain radiosurgery
Deceased
Days

Figure 9

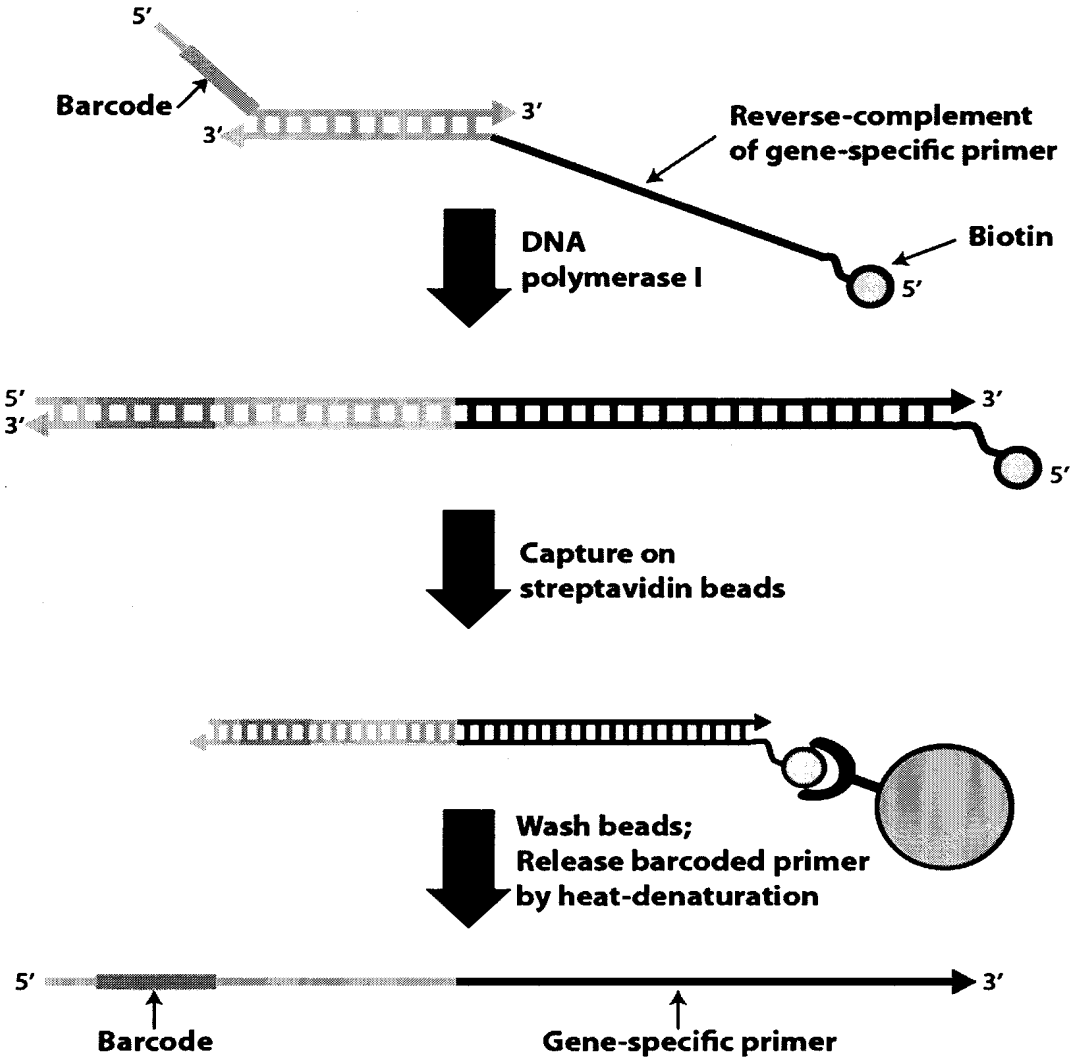

5′
Barcode
3′
3′
Reverse-complement
of gene-specific primer
DNA
polymerase I
Biotin
5′
5′
3′
3′
5′
Capture on
streptavidin beads
Wash beads;
Release barcoded primer
by heat-denaturation
5′
3′
Barcode
Gene-specific primer

Figure 10 Schematic of process described in Example 2

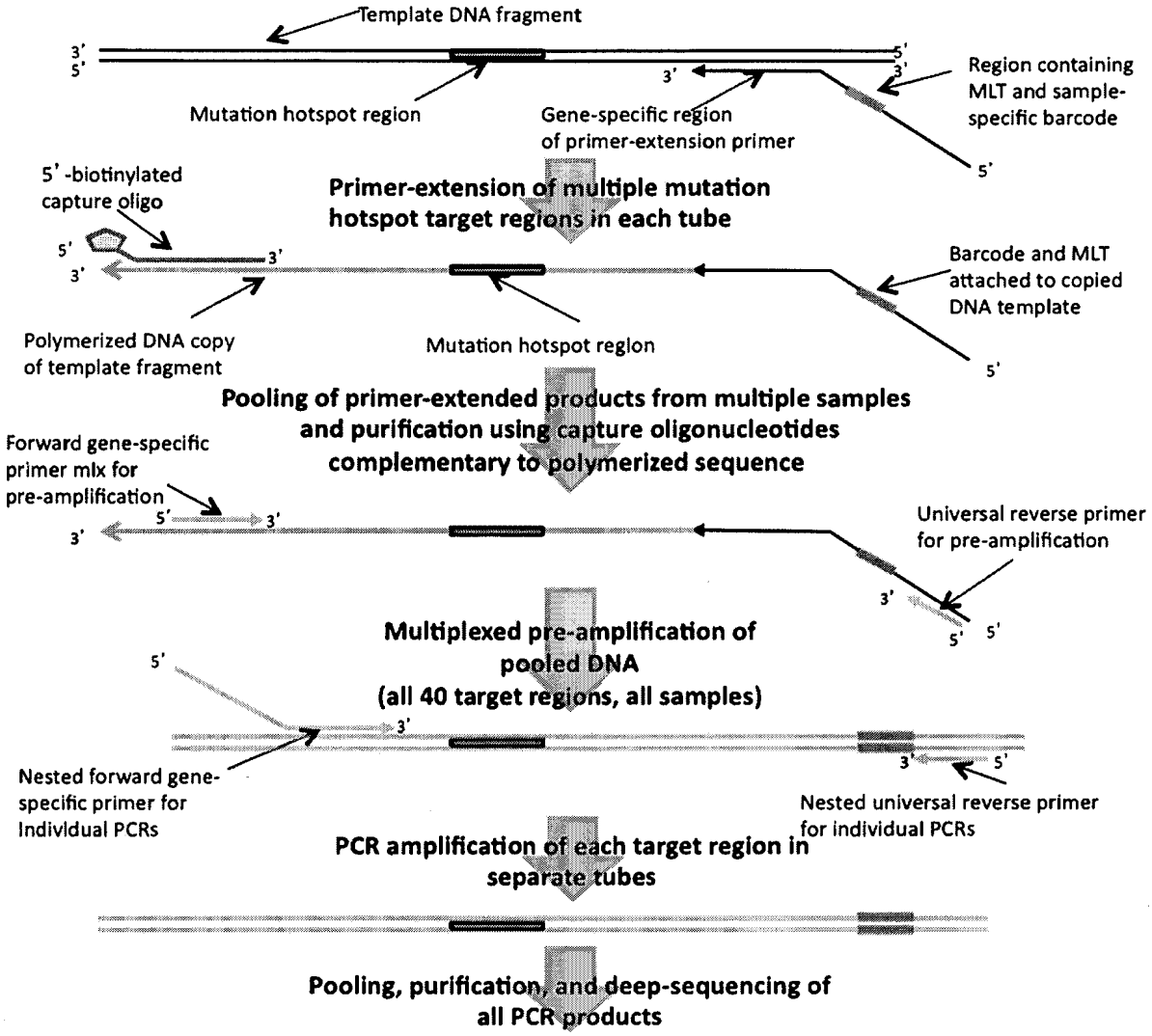


Figure 11 Workflow of process described in Example 2

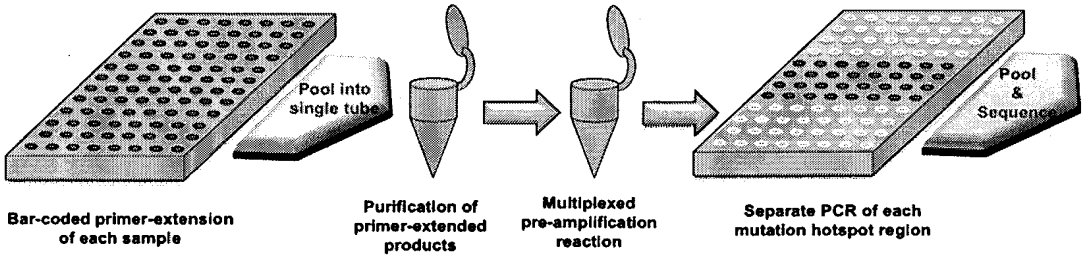
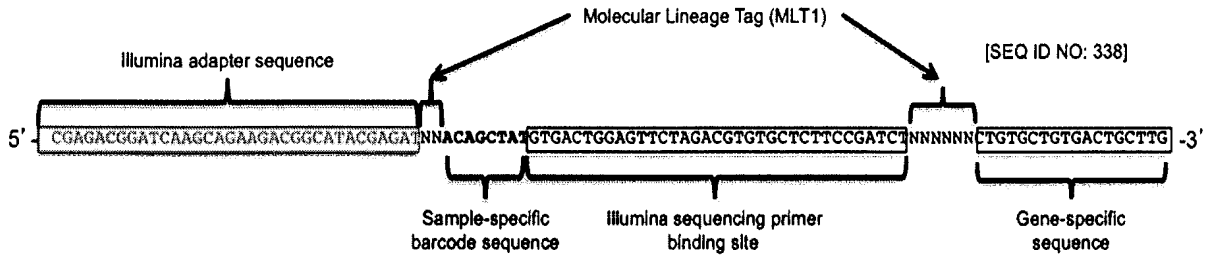


Figure 12



N = degenerate sequence position at which there is an approximately equal probability of an A, C, G, or T being present.

Figure 13 Schematic of splint-mediated ligation of modular oligonucleotides

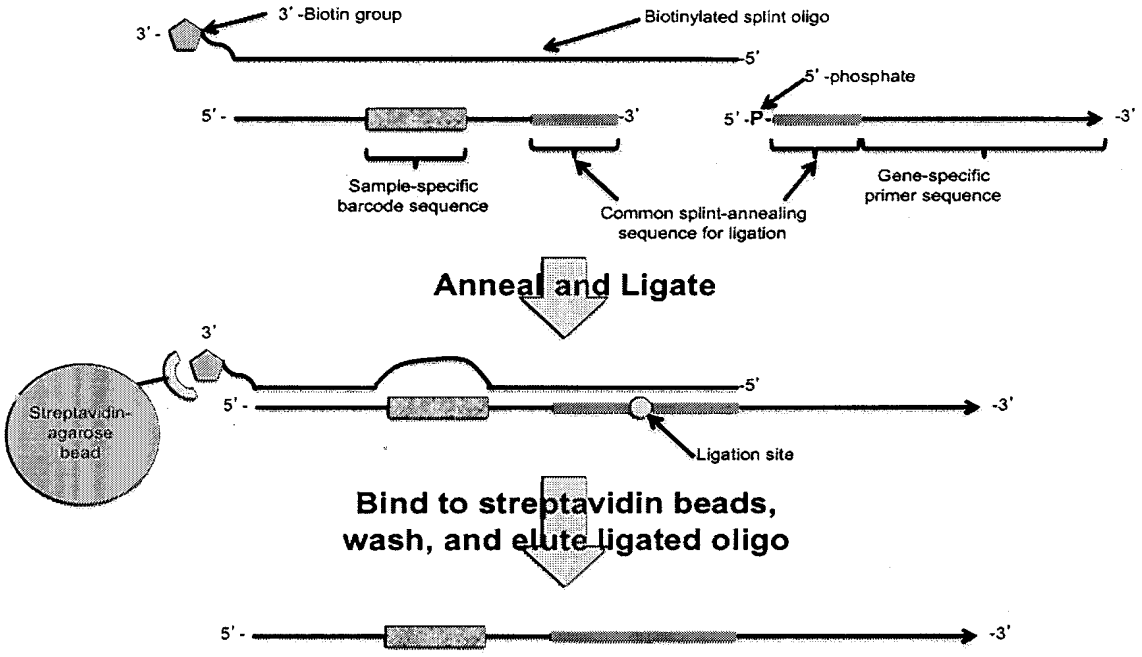
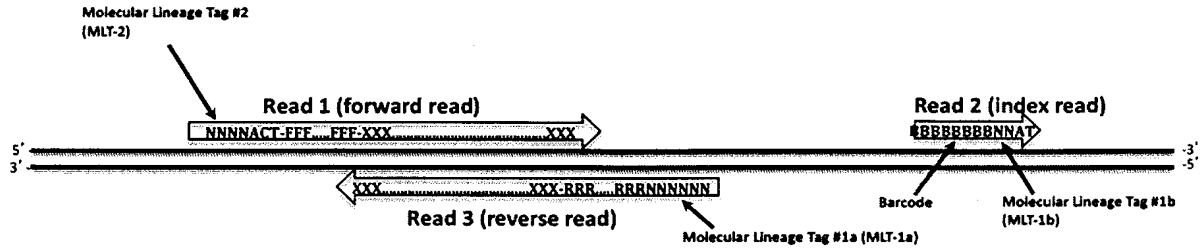


Figure 14



X= amplicon target region (area to be analyzed for presence of mutations)
F= forward primer sequence
R= reverse primer sequence
B= barcode sequence (specifies the sample of origin)
N= Degenerate sequence (equal probability of A,C, G, or T at each position; allows tracking of molecular lineage; also promotes more accurate cluster discrimination during Illumina sequencing).
Note: MLT-1 is created by concatenating the sequences of MLT-1a and MLT-1b
Note: In Example 3, the sequence "NNNNACT" in the region of MLT-2 is replaced by "NNNNNN"

Molecular Lineage Tags permit variants derived from true mutant template molecules to be distinguished from errors of sequencing or amplification; they also serve to minimize misclassification of molecules caused by replacement of barcodes during multiplexed amplification.

Figure 15

Gene Segment: TP53-1
Sample Barcode: 16
Total clones in bin: 58378
Exact matches to Wild-type: 51246
Consistent variants: 2395
Inconsistent variants: 4737

Mismatch Table:

Pos:		*	-	+	#A|#C|#G|#T	A	C	G	T
1:A	58372	3	1	0	00|00|00|00	0	2	0	0
2:C	58294	53	4	1	00|00|01|00	2	0	23	1
3:C	58376	2	0	0	00|00|00|00	0	0	0	0
4:C	58320	43	5	0	00|00|00|00	5	0	3	2
5:C	58370	6	0	0	00|00|00|00	1	0	1	0

etc...

Consistent variants:

			Multiply occurring MLTs: (count of MLTs occurring N times)													
#	ACCCCCGCCCGGCACCCGCGTCCGCGCCATGGCCATCTA	[SEQ ID NO: 339]	N: 15	14	13	12	11	10	09	08	07	06	05	04	03	02	Total
167	=A==========	[SEQ ID NO: 340]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
145	----------	[SEQ ID NO: 341]	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1
120	----------	[SEQ ID NO: 342]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
113	T==========	[SEQ ID NO: 343]	0	0	0	0	0	0	0	0	0	0	0	0	0	2	2
96	----------	[SEQ ID NO: 344]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
84	**====C==========A===**	[SEQ ID NO: 345]	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**3**	**6**	**2**	**0**	**0**	**0**	**0**	**11**
81	----------	[SEQ ID NO: 346]	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

etc...

- Mismatches to the wild-type reference sequence are indicated by:
  - A "*" if the mismatch is found in a read from one strand, but not confirmed by the read from the opposite strand.
  - A "–" if the mismatch is a deletion that is confirmed by overlapping paired-end reads.
  - A "+" if the mismatch is an insertion that is confirmed by overlapping paired-end reads (the inserted nucleotides counts are listed under #A, #C, #G, or #D).
  - An "A" if the mismatch is a point mutation in which the nucleotide is replaced by an "A" and has been confirmed by overlapping paired-end reads.
  - A "C" if the mismatch is a point mutation in which the nucleotide is replaced by a "C" and has been confirmed by overlapping paired-end reads.
  - A "G" if the mismatch is a point mutation in which the nucleotide is replaced by a "G" and has been confirmed by overlapping paired-end reads.
  - A "T" if the mismatch is a point mutation in which the nucleotide is replaced by a "T" and has been confirmed by overlapping paired-end reads.
- Variant sequences that are likely to be arising from mutant template DNA molecules rather than from errors of amplification or sequencing are boldfaced.

Figure 16. Example of processed data for BRAF target region

```
0 exact matches to wild out of 103742. 34272 exact overlap matches.

Mismatch Table:
Pos:                *          -          +          #A|#C|#G|#T          A          C          G          T
  1:        69470    0          0          0          00|00|00|00          0          0          0          0
  2:        69436    34         0          0          00|00|00|00          0          0          0          0
  3:        69374    96         0          0          00|00|00|00          0          0          0          0
  4:        68686    784        0          0          00|00|00|00          0          0          0          0
  5:        69371    99         0          0          00|00|00|00          0          0          0          0
  6:        69171    302        0          0          00|00|00|00          0          0          0          0
  7:        69393    78         0          0          00|00|00|00          0          0          0          0
  8:        69404    69         0          0          00|00|00|00          0          0          0          0
  9:        69268    202        0          0          00|00|00|00          0          0          0          0
 10:        69383    88         0          0          00|00|00|00          0          0          0          0
 11:        69389    184        0          0          00|00|00|00          0          0          0          0
 12:        69112    359        0          0          00|00|00|00          0          0          0          0
 13:        69083    390        0          0          00|00|00|00          0          0          0          0
 14:        69439    31         0          0          00|00|00|00          0          0          0          0
 15:        69409    64         0          0          00|00|00|00          0          0          0          0
 16:        69419    51         0          0          00|00|00|00          0          0          0          0
 17:        67849    1621       0          0          00|00|00|00          0          0          0          0
 18:        69418    53         0          0          00|00|00|00          0          0          0          0
 19:        69353    123        0          0          00|00|00|00          0          0          0          0
 20:        69316    154        0          0          00|00|00|00          0          0          0          0
 21:        69344    126        0          0          00|00|00|00          0          0          0          0
 22:        69401    69         0          0          00|00|00|00          0          0          0          0
 23:        69395    76         0          0          00|00|00|00          0          0          0          0
 24:        69406    64         0          0          00|00|00|00          0          0          0          0
 25:        68619    851        0          0          00|00|00|00          0          0          0          0
 26:        68282    989        0          0          00|00|00|00          0          7          191        1
 27:        69423    1          0          0          00|00|00|00          44         1          0          1
 28:        1148     1          1          0          00|00|00|00          0          68267      38         15
 29:        69243    0          0          0          00|00|00|00          9          204        14         0
 30:        69246    0          0          0          00|00|00|00          9          215        0          0
 31:        69190    0          0          0          00|00|00|00          25         251        4          0
 32:        69436    0          0          0          00|00|00|00          1          0          3          30
 33:        69245    0          1          0          00|00|00|00          0          7          204        13
 34:        69446    0          0          0          00|00|00|00          0          0          2          22
 35:        69256    0          0          0          00|00|00|00          23         185        6          0
 36:        69394    0          0          0          00|00|00|00          75         0          0          1
 37:        69328    0          0          0          00|00|00|00          11         134        5          0
 38:        988      7          0          0          00|00|00|00          0          68342      46         82
 39:        69404    7          0          0          00|00|00|00          55         0          0          4
 40:        69399    71         0          0          00|00|00|00          0          0          0          0
 41:        68774    696        0          0          00|00|00|00          0          0          0          0
 42:        69405    65         0          0          00|00|00|00          0          0          0          0
 43:        69404    66         0          0          00|00|00|00          0          0          0          0
 44:        69441    29         0          0          00|00|00|00          0          0          0          0
 45:        69406    65         0          0          00|00|00|00          0          0          0          0
 (list truncated)
Variants detected: 69470
Known:
Unknown:
                CAAACTGATGGGACCCACTCCATCGAGATTTCACTGTAGCTAGACCAAAATCACCTATTTTTTACTGTT  (SEQ ID NO: 347)
 63797  65143                                  ==-C-==========C--  (SEQ ID NO: 348)      MLTs: 3742X1 2954X1 2801X1 2765X1
2403X1 2109X1 2055X1 1878X1 1839X1 1753X1 1702X1 1634X1 1626X1 1599X1 1580X1 1552X1 1357X1 1336X1 1313X1 1287X1 1159X1
1114X1 1089X1 1058X1 1054X1 1042X1 1040X1 1031X1 962X1 917X1 852X1 821X1 645X1 641X1 605X1 582X1 488X1 477X1 451X1
442X1 432X1 430X2 407X1 385X1 376X1 361X1 322X1 308X1 287X1 269X1 235X1 226X1 188X1 182X1 181X1 166X1 161X1 154X1 137X1
136X1 135X1 125X1 121X1 111X1 107X1 105X1 104X1 94X1 81X1 71X1 66X1 57X2 45X1 44X1 42X1 41X1 28X1 25X2 24X1 22X3 21X2
20X2 19X1 18X1 17X1 15X2 14X2 13X1 12X6 11X1 10X3 9X8 8X5 7X13 6X21 5X25 4X45 3X82 2X104 1X532
   906    965                                  G-=C-==========C--  (SEQ ID NO: 349)      MLTs: 576X1 324X1 6X1 4X1 3X2 2X5 1X39
   395    478                                  A-=C-==========C--  (SEQ ID NO: 350)      MLTs: 30X1 25X1 22X1 18X1 17X2 15X1
14X2 13X1 12X4 11X1 10X3 9X1 8X6 7X3 6X3 5X5 4X5 3X5 2X11 1X26
   182    278                                  ==-==========C--  (SEQ ID NO: 351)      MLTs: 13X2 11X1 10X3 9X1 8X5 7X3 6X5
5X3 4X5 3X7 2X11 1X33
   162    211                                  A-==========----  (SEQ ID NO: 352)      MLTs: 22X1 21X1 16X1 12X1 11X1 10X1 9X1
7X3 6X5 5X2 4X3 3X5 2X2 1X18
   122    194                                   --C-==========C--  (SEQ ID NO: 353)      MLTs: 14X1 13X1 11X1 9X1 8X1 7X1 6X5
5X6 4X2 3X5 2X12 1X25
    57    183                                  ---C-C-========C--  (SEQ ID NO: 354)      MLTs: 20X1 7X1 5X6 4X6 3X5 2X11 1X25
    87    142                                  ---C-----G-====C--  (SEQ ID NO: 355)      MLTs: 41X1 8X1 7X1 6X1 5X5 4X3 3X2 2X10
1X17
    67    140                                  -G-C-==========C--  (SEQ ID NO: 356)      MLTs: 46X1 16X1 5X1 4X6 3X3 2X9 1X22
    62    137                                  ---C-==========--  (SEQ ID NO: 357)      MLTs: 8X1 7X3 6X3 5X3 4X4 3X6 2X8 1X25
(list truncated)
```

Figure 17

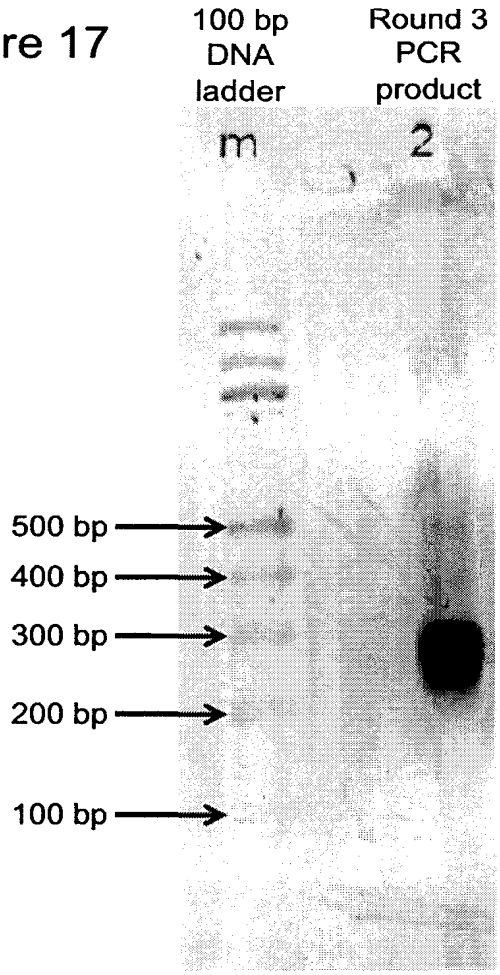
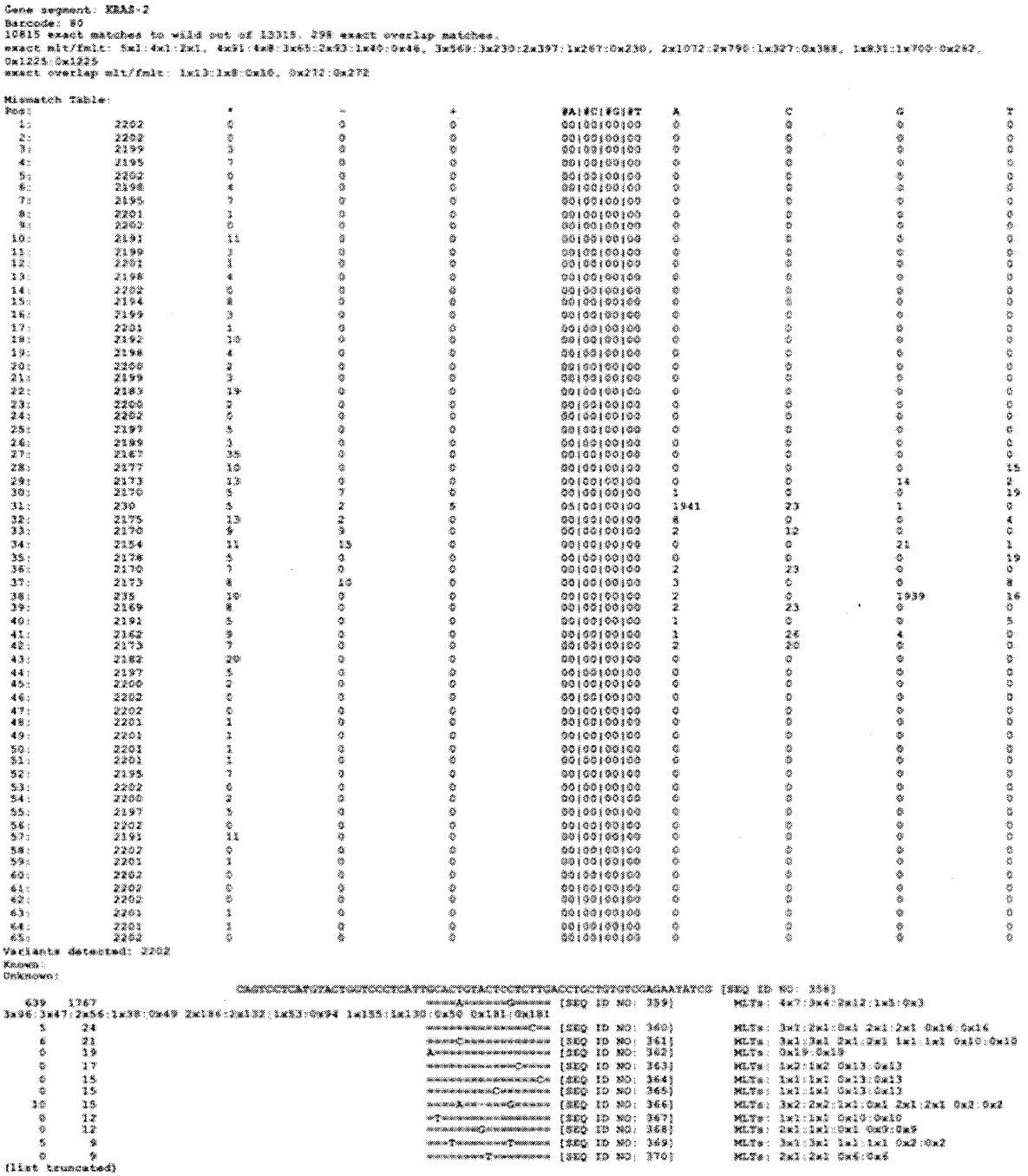

100 bp DNA ladder
Round 3 PCR product
m
2
500 bp
400 bp
300 bp
200 bp
100 bp

Figure 18



```
Gene segment: KRAS-2
Barcode: 80
10815 exact matches to wild out of 13315. 29% exact overlap matches.
exact mlt/fmlt: 5x1:4x1:2x1, 4x91:4x8:3x65:2x93:1x40:0x46, 3x569:3x230:2x397:1x267:0x230, 2x1072:2x790:1x327:0x388, 1x831:1x700:0x262,
0x1225:0x1225
exact overlap mlt/fmlt: 1x13:1x8:0x16, 0x272:0x272

Mismatch Table:
Pos:                    *          -          +          #A|#C|#G|#T     A          C          G          T
  1:          2202      0          0          0          00|00|00|00      0          0          0          0
  2:          2202      0          0          0          00|00|00|00      0          0          0          0
  3:          2199      3          0          0          00|00|00|00      0          0          0          0
  4:          2195      7          0          0          00|00|00|00      0          0          0          0
  5:          2202      0          0          0          00|00|00|00      0          0          0          0
  6:          2198      4          0          0          00|00|00|00      0          0          0          0
  7:          2195      7          0          0          00|00|00|00      0          0          0          0
  8:          2201      1          0          0          00|00|00|00      0          0          0          0
  9:          2202      0          0          0          00|00|00|00      0          0          0          0
 10:          2191     11          0          0          00|00|00|00      0          0          0          0
 11:          2199      3          0          0          00|00|00|00      0          0          0          0
 12:          2201      1          0          0          00|00|00|00      0          0          0          0
 13:          2198      4          0          0          00|00|00|00      0          0          0          0
 14:          2202      0          0          0          00|00|00|00      0          0          0          0
 15:          2194      8          0          0          00|00|00|00      0          0          0          0
 16:          2199      3          0          0          00|00|00|00      0          0          0          0
 17:          2201      1          0          0          00|00|00|00      0          0          0          0
 18:          2192     10          0          0          00|00|00|00      0          0          0          0
 19:          2198      4          0          0          00|00|00|00      0          0          0          0
 20:          2200      2          0          0          00|00|00|00      0          0          0          0
 21:          2199      3          0          0          00|00|00|00      0          0          0          0
 22:          2183     19          0          0          00|00|00|00      0          0          0          0
 23:          2200      2          0          0          00|00|00|00      0          0          0          0
 24:          2202      0          0          0          00|00|00|00      0          0          0          0
 25:          2197      5          0          0          00|00|00|00      0          0          0          0
 26:          2199      3          0          0          00|00|00|00      0          0          0          0
 27:          2167     35          0          0          00|00|00|00      0          0          0          0
 28:          2177     10          0          0          00|00|00|00      0          0          0         15
 29:          2173     13          0          0          00|00|00|00      0          0         14          2
 30:          2170      5          7          0          00|00|00|00      1          0          0         19
 31:           230      5          2          5          05|00|00|00   1941         23          1          0
 32:          2175     13          2          0          00|00|00|00      8          0          0          4
 33:          2170      9          3          0          00|00|00|00      2         12          0          0
 34:          2154     11         15          0          00|00|00|00      0          0         21          1
 35:          2178      5          0          0          00|00|00|00      0          0          0         19
 36:          2170      7          0          0          00|00|00|00      2         23          0          0
 37:          2173      8         10          0          00|00|00|00      3          0          0          8
 38:           235     10          0          0          00|00|00|00      2          0       1939         16
 39:          2169      8          0          0          00|00|00|00      2         23          0          0
 40:          2191      5          0          0          00|00|00|00      1          0          0          5
 41:          2162      9          0          0          00|00|00|00      1         26          4          0
 42:          2173      7          0          0          00|00|00|00      2         20          0          0
 43:          2182     20          0          0          00|00|00|00      0          0          0          0
 44:          2197      5          0          0          00|00|00|00      0          0          0          0
 45:          2200      2          0          0          00|00|00|00      0          0          0          0
 46:          2202      0          0          0          00|00|00|00      0          0          0          0
 47:          2202      0          0          0          00|00|00|00      0          0          0          0
 48:          2201      1          0          0          00|00|00|00      0          0          0          0
 49:          2201      1          0          0          00|00|00|00      0          0          0          0
 50:          2201      1          0          0          00|00|00|00      0          0          0          0
 51:          2201      1          0          0          00|00|00|00      0          0          0          0
 52:          2195      7          0          0          00|00|00|00      0          0          0          0
 53:          2202      0          0          0          00|00|00|00      0          0          0          0
 54:          2200      2          0          0          00|00|00|00      0          0          0          0
 55:          2197      5          0          0          00|00|00|00      0          0          0          0
 56:          2202      0          0          0          00|00|00|00      0          0          0          0
 57:          2191     11          0          0          00|00|00|00      0          0          0          0
 58:          2202      0          0          0          00|00|00|00      0          0          0          0
 59:          2201      1          0          0          00|00|00|00      0          0          0          0
 60:          2202      0          0          0          00|00|00|00      0          0          0          0
 61:          2202      0          0          0          00|00|00|00      0          0          0          0
 62:          2202      0          0          0          00|00|00|00      0          0          0          0
 63:          2201      1          0          0          00|00|00|00      0          0          0          0
 64:          2201      1          0          0          00|00|00|00      0          0          0          0
 65:          2202      0          0          0          00|00|00|00      0          0          0          0
Variants detected: 2202
Known:
Unknown:
                                  CAGTGCTTGATATACTGGTCCCTCATTGCACTGTACTCCTCTTGACCTGCTGTGTGCAGAAATATCC [SEQ ID NO: 358]
   639   1767                                    ====A=======G====== [SEQ ID NO: 359]        MLTs: 4x7:3x4:2x12:1x5:0x3
3x96:3x47:2x56:1x38:0x49 2x196:2x132:1x53:0x94 1x135:1x130:0x50 0x181:0x181
     5     24                                    ==============C== [SEQ ID NO: 360]        MLTs: 3x1:2x1:0x1 2x1:2x1 0x16:0x16
     6     21                                    ====C============ [SEQ ID NO: 361]        MLTs: 3x1:3x1 2x1:2x1 1x1:1x1 0x10:0x10
     0     19                                    A=============== [SEQ ID NO: 362]        MLTs: 0x19:0x19
     0     17                                    ==========C====== [SEQ ID NO: 363]        MLTs: 1x2:1x2 0x13:0x13
     0     15                                    ==============C= [SEQ ID NO: 364]        MLTs: 1x1:1x1 0x13:0x13
     0     15                                    ==========C====== [SEQ ID NO: 365]        MLTs: 1x1:1x1 0x13:0x13
    10     15                                    ====A==-==G====== [SEQ ID NO: 366]        MLTs: 3x2:2x2:1x1:0x1 2x1:2x1 0x2:0x2
     0     12                                    =T=============== [SEQ ID NO: 367]        MLTs: 1x1:1x1 0x10:0x10
     0     12                                    ==========C====== [SEQ ID NO: 368]        MLTs: 2x1:1x1:0x1 0x9:0x9
     5      9                                    ===T=======T===== [SEQ ID NO: 369]        MLTs: 3x1:3x1 1x1:1x1 0x2:0x2
     0      9                                    ==========T====== [SEQ ID NO: 370]        MLTs: 2x1:2x1 0x6:0x6
(list truncated)
```

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2013/031014

A. CLASSIFICATION OF SUBJECT MATTER

C12Q 1/68(2006.01)i, C12N 15/11(2006.01)i

According to International Patent Classification (IPC) or to both national classification and IPC

B. FIELDS SEARCHED

Minimum documentation searched (classification system followed by classification symbols)
C12Q 1/68; C12P 19/34

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
Korean utility models and applications for utility models
Japanese utility models and applications for utility models

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)
eKOMPASS(KIPO internal) & Keywords: barcode, deep-sequencing, mutation, detection

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	DIEHL et al., 'Circulating mutant DNA to assess tumor dynamics' Nature Medicine, Vol.14, No.9, pp.985-990 (2007) See abstract and pages 985-986, 988.	1-3,10-12,15
A	VARLEY et al., 'Nested patch PCR enables highly multiplexed mutation discovery in candidate genes' Genome Research, Vol.18, No.11, pp.1844-1850 (2008) See abstract and figures 1-2.	1-3,10-12,15
A	JABARA et al., 'Accurate sampling and deep sequencing of the HIV-1 protease gene using a primer ID' PNAS, Vol.108, No.50, pp.20166-20171 (2011) See the whole document.	1-3,10-12,15
A	FLAHERTY et al., 'Ultrasensitive detection of rare mutations using next-generation targeted resequencing' Nucleic Acids Research, Vol.40, No.1, Article No.e2 (pages 1-12) (January 2012) See the whole document.	1-3,10-12,15
A	US 7622281 B2 (RONAGHI et al.) 24 November 2009 See the whole document.	1-3,10-12,15

☒ Further documents are listed in the continuation of Box C. ☒ See patent family annex.

* Special categories of cited documents:
"A" document defining the general state of the art which is not considered to be of particular relevance
"E" earlier application or patent but published on or after the international filing date
"L" document which may throw doubts on priority claim(s) or which is cited to establish the publication date of citation or other special reason (as specified)
"O" document referring to an oral disclosure, use, exhibition or other means
"P" document published prior to the international filing date but later than the priority date claimed
"T" later document published after the international filing date or priority date and not in conflict with the application but cited to understand the principle or theory underlying the invention
"X" document of particular relevance; the claimed invention cannot be considered novel or cannot be considered to involve an inventive step when the document is taken alone
"Y" document of particular relevance; the claimed invention cannot be considered to involve an inventive step when the document is combined with one or more other such documents,such combination being obvious to a person skilled in the art
"&" document member of the same patent family

Date of the actual completion of the international search	Date of mailing of the international search report
25 June 2013 (25.06.2013)	**25 June 2013 (25.06.2013)**
Name and mailing address of the ISA/KR Korean Intellectual Property Office 189 Cheongsa-ro, Seo-gu, Daejeon Metropolitan City, 302-701, Republic of Korea Facsimile No. 82-42-472-7140	Authorized officer KIM, Seung Beom Telephone No. 82-42-481-3371

Form PCT/ISA/210 (second sheet) (July 2009)

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2013/031014

C (Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
PX	SCHMITT et al., 'Detection of ultra-rare mutations by next-generation sequencing' Proceedings of the National Academy of Sciences, Vol.19, No.36, pp.14508-14513 (1 August 2012) See the whole document.	1-3,10-12,15
PX	NARAYAN et al., 'Ultrasensitive measurement of hotspot mutations in tumor DNA in blood using error-suppressed multiplexed deep sequencing' Cancer Research, Vol.72, No.14, pp.3492-3498 (10 May 2012) See the whole document.	1-3,10-12,15

IDT Exhibit 1020.5351

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2013/031014

Box No. II Observations where certain claims were found unsearchable (Continuation of item 2 of first sheet)

This international search report has not been established in respect of certain claims under Article 17(2)(a) for the following reasons:

1. ☐ Claims Nos.:
because they relate to subject matter not required to be searched by this Authority, namely:

2. ☒ Claims Nos.: 6
because they relate to parts of the international application that do not comply with the prescribed requirements to such an extent that no meaningful international search can be carried out, specifically:
Claim 6 does not comply with PCT Article 6, since it is referring to the multiple dependent claim which does not comply with PCT Rule 6.4(a).

3. ☒ Claims Nos.: 4-5, 7-9, 13-14
because they are dependent claims and are not drafted in accordance with the second and third sentences of Rule 6.4(a).

Box No. III Observations where unity of invention is lacking (Continuation of item 3 of first sheet)

This International Searching Authority found multiple inventions in this international application, as follows:

1. ☐ As all required addtional search fees were timely paid by the applicant, this international search report covers all searchable claims.

2. ☐ As all searchable claims could be searched without effort justifying an additional fee, this Authority did not invite payment of any additional fee.

3. ☐ As only some of the required additional search fees were timely paid by the applicant, this international search report covers only those claims for which fees were paid, specifically claims Nos.:

4. ☐ No required additional search fees were timely paid by the applicant. Consequently, this international search report is restricted to the invention first mentioned in the claims; it is covered by claims Nos.:

Remark on Protest

☐ The additional search fees were accompanied by the applicant's protest and, where applicable, the payment of a protest fee.
☐ The additional search fees were accompanied by the applicant's protest but the applicable protest fee was not paid within the time limit specified in the invitation.
☐ No protest accompanied the payment of additional search fees.

IDT Exhibit 1020.5352

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No.
PCT/US2013/031014

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
US 7622281 B2	24.11.2009	US 2006-0008824 A1	12.01.2006

IDT Exhibit 1020.5353

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
27 December 2013 (27.12.2013)

WIPO | PCT

(10) International Publication Number
WO 2013/191775 A2

(51) International Patent Classification:
C40B 30/04 (2006.01)

(21) International Application Number:
PCT/US2013/032606

(22) International Filing Date:
15 March 2013 (15.03.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/661,293 18 June 2012 (18.06.2012) US

(71) Applicant *(for all designated States except US)*: **NUGEN TECHNOLOGIES, INC.** [US/US]; 201 Industrial Road, San Carlos, CA 94070 (US).

(72) Inventors; and
(71) Applicants *(for US only)*: **ARMOUR, Christopher** [US/US]; 13531 104th Place Ne, Kirkland, WA 98034 (US). **AMORESE, Doug** [US/US]; 357 South El Monte Avenue, Los Altos, CA 94022 (US). **LI, Bin** [CN/US]; 3723 Heron Way, Palo Alto, CA 94303 (US). **KURN, Nurith** [IL/US]; 2876 Ramona Street, Palo Alto, CA 94306 (US).

(74) Agents: **SKUBATCH, Maya** et al.; Wilson Sonsini Goodrich & Rosati, 650 Page Mill Road, Palo Alto, CA 94304-1050 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IS, JP, KE, KG, KM, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

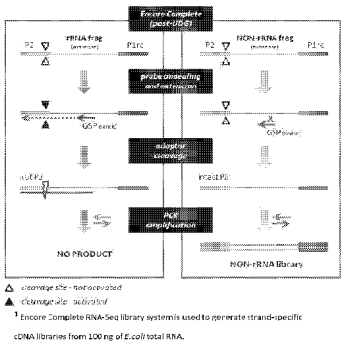
(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, ML, MR, NE, SN, TD, TG).

Published:

— *without international search report and to be republished upon receipt of that report (Rule 48.2(g))*

(54) **Title**: COMPOSITIONS AND METHODS FOR NEGATIVE SELECTION OF NON-DESIRED NUCLEIC ACID SEQUENCES

Figure 1. rRNA depletion from a nucleic acid library using insert-dependent adaptor cleavage (InDA-C).



(57) **Abstract**: The present invention provides methods, compositions and kits for the generation of next generation sequencing (NGS) libraries in which non-desired nucleic acid sequences have been depleted or substantially reduced. The methods, compositions and kits provided herein are useful, for example, for the production of libraries from total RNA with reduced ribosomal RNA and for the reduction of common mRNA species in expression profiling from mixed samples where the mRNAs of interest are present at low levels. The methods of the invention can be employed for the elimination of non-desired nucleic acid sequences in a sequence-specific manner, and consequently, for the enrichment of nucleic acid sequences of interest in a nucleic acid library.

WO 2013/191775 A2

COMPOSITIONS AND METHODS FOR NEGATIVE SELECTION OF NON-DESIRED NUCLEIC ACID SEQUENCES

CROSS REFERENCE

[0001] This claims the priority benefit of provisional patent application U.S. Serial No. 61/661,293, filed June 18, 2012, all of which is incorporated herein by reference in its entirety.

BACKGROUND OF THE INVENTION

[0002] Next generation sequencing (NGS) libraries are collections of DNA fragments whose nucleotide sequences will be determined. The sources of DNA for insertion into these libraries are typically genomic DNA that has been fragmented to a desired length, or copies of the transcriptome from a given cell population. Transcriptome libraries are generated by making a cDNA copy of an RNA population, creating a complement to each DNA strand, thereby generating double-stranded DNA, and then ligating the double-stranded DNAs to library-specific adaptors. The cDNA can be synthesized by using random primers, sequence-specific primers or primers containing oligo dT tails to prime a population of transcripts that are polyadenylated. Frequently, these fragment populations contain DNA that is not of interest to a particular study, and in some cases, these non-desired DNA sequences represent a very significant percentage of the overall DNA population. For example, in whole transcriptome studies, ribosomal RNA (rRNA) sequences comprise the majority (60-90%) of all fragments in a typical cDNA library, absent steps to remove rRNA from the samples. In another example, gene expression profiling from peripheral blood is primarily concerned with mRNA from peripheral blood mononuclear cells (PBMCs), which make up less than 0.1 % of the whole blood sample. Reduction of globin RNA from red blood cells, which make up majority of the cells in the blood sample, is desirable in such assays.

[0003] In the case of rRNA removal or depletion, three general methods have been described: 1) removal of rRNA from the starting population; 2) differential priming using oligo dT primers (i.e. priming polyadenylated transcripts only); and 3) differential priming where primers complementary to rRNA sequences are specifically eliminated (or under-represented) in a primer pool (Not-So-Random or NSR primer approach; see Armour et al., 2009). Priming a total RNA population with primers that only recognize poly(A)-sequences is problematic for two reasons. First, it cannot be used with prokaryotic organisms because prokaryotic mRNAs do not contain poly(A)- sequences at their 3' ends. Second, even with eukaryotic RNA samples, many biologically important elements, such as regulatory

IDT Exhibit 1020.5356

transcripts, are not polyadenylated and are therefore lost from the library with oligo dT priming. While NSR priming strategies can be effective when designed to specific organisms, NSR priming can cause distortions in the sample populations when a less optimized set of primers is employed across a broader range of sample types.

[0004] There is a need for improved methods for removal of specific non-desired DNA fragments from NGS libraries. Such methods would ideally enable starting with an unbiased template population and eliminating non-desired DNA fragments in a sequence-specific manner after the NGS library has been generated. The invention described herein fulfills this need.

SUMMARY OF THE INVENTION

[0005] The present invention provides novel methods, compositions and kits for construction of NGS libraries in which non-desired nucleic acid sequences have been depleted or substantially reduced. Specifically, an important aspect of this invention is the methods and compositions that allow for elimination or reduction of non-desired DNA sequences in a sequence-specific manner after the generation of a NGS library in which all sequences of the starting nucleic acid sequence population (for example, a transcriptome) are represented in an undistorted, unbiased manner. The methods of the invention can be employed for the elimination of non-desired nucleic acid sequences, such as ribosomal RNA, from a nucleic acid library, and consequently, for the enrichment of nucleic acid sequences of interest in the library.

[0006] In one aspect, the invention provides a method for the selective removal of non-desired nucleic acid sequences in a sequence-specific manner from a nucleic acid library with single-stranded DNA templates. In some embodiments, the method comprises: a) annealing a sequence-specific oligonucleotide primer or a set of sequence-specific oligonucleotide primers to the single-stranded DNA templates with adaptors of fixed orientation attached at each end, wherein the sequence-specific oligonucleotides are designed to be complementary to the non-desired nucleic acid sequences or regions adjacent to the non-desired nucleic acid sequences, and wherein one of the two adaptor sequences contains a recognition sequence for a restriction endonuclease specific for double-stranded DNA; b) extending the sequence-specific primers with a DNA polymerase beyond the adaptor-DNA template junction, thereby creating double-stranded DNA fragments where the oligonucleotide primers are complementary to the single-stranded DNA templates ; c) treating the population of DNA fragments (both single-stranded and double-stranded) with a restriction endonuclease specific for double-stranded DNA, thereby cleaving only double-stranded DNA fragments at the

IDT Exhibit 1020.5357

adaptor restriction endonuclease site and thus removing adaptor from one end of the fragments containing the non-desired nucleic acid sequences; and d) performing PCR using primers specific to each adaptor whereby exponential amplification only occurs when fragments have both PCR priming sites on the same template, thereby amplifying the desired nucleic acid sequences only.

[0007] In another aspect, the invention provides a method for the construction of a nucleic acid library from a sample of interest while retaining an unbiased nucleic acid template population in which all sequences of the starting nucleic acid sequence population are represented.

[0008] In some embodiments, the invention provides a method for the construction of a directional (i.e. strand-specific) nucleic acid library, the method comprising: a) reverse transcribing an RNA sample; b) generating double-stranded cDNA from a reverse transcribed RNA sample, wherein during the second strand cDNA synthesis at least one modified nucleotide is incorporated into the second strand of cDNA along the length of the strand; c) performing end repair on the double-stranded cDNA; d) ligating adaptors to the double-stranded cDNA, wherein one of the two adaptors has the modified nucleotide incorporated into the ligation strand of the adaptor; e) performing gap repair; f) selectively removing the second strand of cDNA by a suitable degrading agent; and g) removing the degradation products from the sample, thereby resulting in a library of single-stranded DNA templates with adaptors of fixed orientation attached to each end.

[0009] In a preferred embodiment, the modified nucleotide incorporated into the second strand of cDNA is deoxyuridine triphosphate (dUTP), and the degrading agent is the nuclease Uracil-N-Glycosylase (UNG).

[0010] In other embodiments, the nucleic acid library being constructed is not strand-specific.

[0011] In some embodiments, the nucleic acid sample of interest comprises total RNA. In some embodiments, the nucleic acid sample of interest is primed using a random primer population. In other embodiments, the nucleic acid sample of interest is primed using a partially selective primer population.

[0012] In various aspects the invention relates to methods for depleting undesired nucleic acids from a pool of nucleic acids. Libraries may be prepared with the remaining nucleic acids. Nucleic acid depletion and library generation may be performed in a strand specific manner. According to a first aspect, the invention relates to a method for depleting or reducing specific non-desired nucleic acid sequences from a nucleic acid library, the method

IDT Exhibit 1020.5358

comprising: (a) generating a nucleic acid library comprising single-stranded DNA fragments with adaptors of fixed orientation attached to each end of each DNA fragment; (b) annealing sequence-specific oligonucleotide probes to the single-stranded DNA fragments with adaptors of fixed orientation attached at each end, wherein the sequence-specific oligonucleotide probes are designed to be complementary to the non-desired nucleic acid sequences, and wherein at least one of the two adaptors comprises a recognition sequence for a restriction endonuclease specific for double-stranded DNA; (c) extending the sequence-specific oligonucleotide probes with a DNA polymerase, thereby creating double-stranded DNA fragments comprising at least a portion of the non-desired nucleic acid sequences; (d) treating the population of DNA fragments comprising double-stranded and single-stranded DNA with a restriction endonuclease specific for double-stranded DNA, thereby cleaving double-stranded DNA fragments at the restriction endonuclease site; and (e) performing PCR with a set of primers specific to the adaptor sequences, thereby amplifying the DNA fragments comprising the desired nucleic acid sequences. In some embodiments, the method further comprises an additional step of sequencing the amplified products. In some embodiments, the nucleic acid library originates from a population of sorted cells. In some embodiments, the nucleic acid library originates from a single cell. In some embodiments, the method further comprises sorting cells into a multiwell plate, microarray, microfluidic device, or slide and thereby generating the population of sorted cells. In some embodiments, the sorting is performed according to a cell surface marker. In some embodiments, the sorting is performed according to optical properties of the cells. In some embodiments, the sorting is performed according to cell size. In some embodiments, the non-desired nucleic acid sequences comprise bacterial ribosomal RNA, mitochondrial DNA, human globin mRNA, human cytoplasmic rRNA, human mitochondrial rRNA, grape cytoplasmic rRNA, grape mitochondrial rRNA, or grape chloroplast rRNA. In some embodiments, the restriction endonuclease of step d. is BspQI. In some embodiments, the DNA polymerase comprises a hot start polymerase. In some embodiments, the DNA polymerase is MyTaq polymerase. In some embodiments, step (a) comprises i. reverse transcribing a RNA sample; ii. generating double-stranded cDNA from the reverse transcribed RNA sample, wherein at least one of the four dNTPs dATP, dCTP, dGTP or dTTP is replaced by a non-canonical dNTP during second strand synthesis and incorporated into the second strand; iii. performing end repair on the double-stranded cDNA; iv. ligating adaptors to a 5' end of the double-stranded cDNA, wherein one of the adaptor strands has the non-canonical nucleotide incorporated into a ligation strand of the adaptor; v. performing gap repair; and iv. selectively removing the

IDT Exhibit 1020.5359

second strand by a cleavage agent. In some embodiments, the non-canonical nucleotide comprises uridine or inosine. In some embodiments, step vi comprises cleaving a base portion of the one or more non-canonical nucleotides, thereby forming an abasic site. In some embodiments, the cleavage agent comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the cleavage agent comprises a primary amine. In some embodiments, the cleavage agent comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the cleavage agent comprises a glycosylase and a polyamine. In some embodiments, the cleavage agent comprises endonuclease V.

[0013] In a second aspect, the invention relates to a method of adapter ligation to a pool of nucleic acids, comprising: (a) ligating a nucleic acid comprising a first nucleic acid strand comprising a 5' phosphate, a second nucleic acid strand comprising a 5' phosphate and one or more non-canonical nucleotides, with at least a first adapter comprising a first adapter strand lacking a 5' phosphate and a second adapter strand lacking a 5' phosphate and one or more non-canonical nucleotides; (b) performing a 3' extension reaction; and (c) performing a cleavage reaction with an agent comprising one or more cleavage reagents, thereby cleaving at least one nucleic acid strand comprising one or more non-canonical nucleotides; wherein one of the one or more cleavage agents is specific for nucleic acid strands comprising the one or more non-canonical nucleotides. In some embodiments, the method comprises ligating the nucleic acid with a second adapter comprising a third adapter strand lacking a 5' phosphate and a fourth adapter strand lacking a 5' phosphate and one or more non-canonical nucleotides, wherein the first and the second adapters are different. In some embodiments, the nucleic acid is ligated with a first or second adapter at each end. In some embodiments, the non-canonical nucleotide is selected from uracil and inosine. In some embodiments, step c comprises cleaving a base portion of the one or more non-canonical nucleotides, thereby forming an abasic site. In some embodiments, the one or more cleavage reagents comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the one or more cleavage reagents comprises a primary amine. In some embodiments, the one or more cleavage reagents comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the one or more cleavage reagents comprises a glycosylase and a polyamine. In some embodiments, the one or more cleavage reagents comprises endonuclease V. In some embodiments, the method further comprises performing an amplification reaction comprising a first primer and second primer, wherein the first primer is hybridizable to the first adapter strand and the second primer is hybridizable to the fourth

IDT Exhibit 1020.5360

adapter strand, thereby generating amplified products. In some embodiments, the first adapter comprises a recognition sequence for a restriction endonuclease specific for double-stranded DNA. In some embodiments, the method further comprises (d) hybridizing a probe to a sequence on the first nucleic acid strand, (e) extending the probe with a DNA polymerase, thereby producing a partial duplex nucleic acid, and (f) treating the partial duplex nucleic acid with a restriction endonuclease specific for double-stranded DNA, thereby cleaving double-stranded DNA fragments at the recognition sequence. In some embodiments, the method further comprises performing PCR with a set of primers specific to the adaptor sequences, thereby amplifying at least a second nucleic acid in the pool of nucleic acids. In some embodiments, the second nucleic acid lacks the sequence in step d. In some embodiments, the method further comprises sequencing a portion of the second nucleic acid. In some embodiments, the nucleic acid is generated by i. performing a first strand synthesis on an RNA, thereby forming a first strand synthesis product; and ii. performing a second strand synthesis on the first strand in the presence of a non-canonical nucleotide, thereby forming a second strand synthesis product. In some embodiments, the method further comprises selectively cleaving the RNA. In some embodiments, selectively cleaving the RNA comprises treatment with RNAse H. In some embodiments, the method further comprises iii. fragmenting the first and second strand synthesis products, thereby generating fragmented first and second strand synthesis products; iv. performing end repair; and v. performing 5' phosphorylation. In some embodiments, the pool of nucleic acids originates from a population of sorted cells. In some embodiments, the pool of nucleic acids originates from a single cell. In some embodiments, the method further comprises sorting cells into a multiwell plate, microarray, microfluidic device, or slide and thereby generating the population of sorted cells. In some embodiments, the sorting is performed according to a cell surface marker. In some embodiments, the sorting is performed according to optical properties of the cells. In some embodiments, the sorting is performed according to cell size. In some embodiments, the pool of nucleic acids comprises bacterial ribosomal RNA, mitochondrial DNA, human globin mRNA, human cytoplasmic rRNA, human mitochondrial rRNA, grape cytoplasmic rRNA, grape mitochondrial rRNA, or grape chloroplast rRNA. In some embodiments, the restriction endonuclease is BspQI. In some embodiments, the 3' extension reaction is performed using a hot start polymerase. In some embodiments, the 3' extension reaction is performed using MyTaq polymerase.

[0014] In a third aspect, the invention relates to a method of adapter ligation to create a strand retained library of nucleic acids comprising desired and undesired nucleic acids,

IDT Exhibit 1020.5361

comprising: (a) mixing pool of templates comprising an undesired nucleic acid comprising one or more non-canonical nucleotides and a desired nucleic acid comprising one or more non-canonical nucleotides with a plurality of partial duplex primers each comprising a 3' overhang, (b) annealing to the templates the plurality of partial duplex primers; (c) performing primer extension reaction along the templates, thereby forming double-stranded nucleic acids each comprising a primer extension product, (d) ligating an adapter to at least one 5' end of the primer extension products; and (e) cleaving the templates from the double-stranded nucleic acids with a cleavage agent that is specific for nucleic acids comprising the one or more nucleotides. In some embodiments, the plurality of partial duplex primers comprise at least two partial duplex primers with dissimilar 3' overhang sequences. In some embodiments, the plurality of partial duplex primers comprise a shared sequence within a double-stranded portion. In some embodiments, the method further comprises a step comprising performing a primer extension reaction along the adapter. In some embodiments, step e comprises cleaving a base portion of the one or more non-canonical nucleotides, thereby forming an abasic site. In some embodiments, the cleavage agent comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the cleavage agent comprises a primary amine. In some embodiments, the cleavage agent comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the cleavage agents comprises a glycosylase and a polyamine. In some embodiments, the cleavage agent comprises endonuclease V. In some embodiments, the one or more non-canonical nucleotides comprise uracil or inosine. In some embodiments, the one or more non-canonical nucleotides comprise uracil and inosine. In some embodiments, the pool of templates is generated by i. performing a first strand synthesis on an RNA in the presence of the one or more non-canonical nucleotides, thereby forming a first strand synthesis product and ii. performing a fragmentation reaction. In some embodiments, the method further comprises selectively cleaving the RNA. In some embodiments, selectively cleaving the RNA comprises treatment with RNAse H. In some embodiments, the fragmentation reaction comprises utilizing a cleavage agent targeting the one or more non-canonical nucleotides. In some embodiments, fragmentation reaction comprises cleaving a base portion of the one or more non-canonical nucleotides, thereby forming an abasic site. In some embodiments, the cleavage agent comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the cleavage agent comprises a primary amine. In some embodiments, the cleavage agent comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the cleavage agents

IDT Exhibit 1020.5362

comprises a glycosylase and a polyamine. In some embodiments, the cleavage agent comprises endonuclease V. In some embodiments, the adapter comprises a recognition sequence for a restriction endonuclease specific for double-stranded DNA. In some embodiments, the method further comprises (f), hybridizing a probe to a sequence of the primer extension product; (g) extending the probe with a DNA polymerase, thereby producing a partial duplex nucleic acid, and (h) treating the partial duplex nucleic acid with a restriction endonuclease specific for double-stranded DNA, thereby cleaving double-stranded DNA fragments at the recognition sequence. In some embodiments, the method further comprises performing PCR with a primer that is hybridizable a sequence reverse complimentary to the adapter, thereby amplifying a desired nucleic acid in the pool of templates. In some embodiments, the method further comprises sequencing a portion of the desired nucleic acid. In some embodiments, the pool of templates originates from a population of sorted cells. In some embodiments, the pool of templates originates from a single cell. In some embodiments, the method further comprises sorting cells into a multiwell plate, microarray, microfluidic device, or slide and thereby generating the population of sorted cells. In some embodiments, the sorting is performed according to a cell surface marker. In some embodiments, the sorting is performed according to optical properties of the cells. In some embodiments, the sorting is performed according to cell size. In some embodiments, the pool of templates comprises bacterial ribosomal RNA, mitochondrial DNA, human globin mRNA, human cytoplasmic rRNA, human mitochondrial rRNA, grape cytoplasmic rRNA, grape mitochondrial rRNA, or grape chloroplast rRNA. In some embodiments, the restriction endonuclease is BspQI. In some embodiments, the primer extension reaction is performed using a hot start polymerase. In some embodiments, the primer extension reaction is performed using MyTaq polymerase.

[0015] In a fourth aspect, the invention relates to a method of adapter ligation to create a strand retained library of nucleic acids with desired and undesired nucleic acids, comprising: (a) mixing a pool of templates comprising an undesired nucleic acid and a desired nucleic acid with a plurality of partial duplex primers each comprising a 3' overhang; (b) annealing to the templates the plurality of partial duplex primers; (c) performing primer extension reaction along the templates, thereby forming double-stranded nucleic acids each comprising a primer extension product; (d) ligating an adapter to at least one 5' end of the primer extension product; and (e) cleaving the primer extension products from the double-stranded nucleic acids with a cleavage agent that is specific for nucleic acids comprising the one or more nucleotides. In some embodiments, the plurality of partial duplex primers comprise at

IDT Exhibit 1020.5363

least two partial duplex primers with dissimilar 3' overhang sequences. In some embodiments, the plurality of partial duplex primers comprise a shared sequence within a double-stranded portion. In some embodiments, the strand of the plurality of partial duplex primers with the 3' overhang lacks adenines in the shared sequence within the double-stranded portion. In some embodiments, the method further comprising a step comprising performing a primer extension reaction along the adapter. In some embodiments, the primer extension reaction is performed in the presence of one or more non-canonical nucleotides. In some embodiments, step e comprises cleaving a base portion of the one or more non-canonical nucleotides, thereby forming an abasic site. In some embodiments, the cleavage agent comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the cleavage agent comprises a primary amine. In some embodiments, the cleavage agent comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the cleavage agents comprises a glycosylase and a polyamine. In some embodiments, the cleavage agent comprises endonuclease V. In some embodiments, the one or more non-canonical nucleotides comprise uracil or inosine. In some embodiments, the pool of templates is generated by i. performing a first strand synthesis on an RNA in the presence of the one or more non-canonical nucleotides, thereby forming a first strand synthesis product; and ii. performing a fragmentation reaction. In some embodiments, the method further comprises selectively cleaving the RNA. In some embodiments, selectively cleaving the RNA comprises treatment with RNAse H. In some embodiments, the fragmentation reaction comprises utilizing a cleavage agent targeting the one or more non-canonical nucleotides. In some embodiments, the fragmentation reaction comprises cleaving a base portion of the one or more non-canonical nucleotides, thereby forming an abasic site. In some embodiments, the cleavage agent comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the cleavage agent comprises a primary amine. In some embodiments, the cleavage agent comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the cleavage agents comprises a glycosylase and a polyamine. In some embodiments, the cleavage agent comprises endonuclease V. In some embodiments, the adapter comprises a recognition sequence for a restriction endonuclease specific for double-stranded DNA. In some embodiments, the method further comprises, (f) hybridizing a probe to a sequence of the undesired nucleic acid; (g) extending the probe with a DNA polymerase, thereby producing a partial duplex nucleic acid, and (h) treating the partial duplex nucleic acid with a restriction endonuclease specific for double-stranded DNA, thereby cleaving double-stranded DNA

IDT Exhibit 1020.5364

fragments at the recognition sequence. In some embodiments, the method further comprises performing PCR with a set of primers that are hybridizable to a sequence reverse complimentary to the adapter and the shared sequence of the partial duplex primers opposite the 3' overhang, thereby amplifying a desired nucleic acid in the pool of templates. In some embodiments, the method further comprises sequencing a portion of the desired nucleic acid. In some embodiments, the pool of templates originates from a population of sorted cells. In some embodiments, the pool of templates originates from a single cell. In some embodiments, the method further comprises sorting cells into a multiwell plate, microarray, microfluidic device, or slide and thereby generating the population of sorted cells. In some embodiments, the sorting is performed according to a cell surface marker. In some embodiments, the sorting is performed according to optical properties of the cells. In some embodiments, the sorting is performed according to cell size. In some embodiments, the pool of templates comprises bacterial ribosomal RNA, mitochondrial DNA, human globin mRNA, human cytoplasmic rRNA, human mitochondrial rRNA, grape cytoplasmic rRNA, grape mitochondrial rRNA, or grape chloroplast rRNA. In some embodiments, the restriction endonuclease is BspQI. In some embodiments, the primer extension reaction is performed using a hot start polymerase. In some embodiments, the primer extension reaction is performed using MyTaq polymerase.

[0016] According to any of the aspects, the invention relates to partially, substantially, or completely depleting undesired sequences, wherein at least 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 65, 70, 75, 80, 85, 90, 95, 98, 99, 99.5, 99.8, 99.9, 99.99 percent or more of some or all of the undesired sequences is depleted and, optionally, partially substantially, or completely retaining desired sequences, wherein less than 98, 95, 90, 85, 80, 75, 70, 65, 60, 55, 50, 45, 40, 35, 30, 25, 20, 15, 10, 5, 2, 1, 0.5, 0.2, 0.1, 0.05, 0.01 percent or less of some or all of the desired sequences is depleted. In some embodiments, the methods in any of the aspects deplete some or all of the undesired sequences completely. In some embodiments, the methods in any of the aspects retain some or all of the desired sequences completely. In any of the aspects, the methods described herein may decrease the abundance ratio between an undesired and a desired nucleic acid sequence, between some undesired and some desired nucleic acid sequences, or between all undesired and all desired nucleic acid sequences by 1.1, 1.2, 1.5, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 50, 60, 75, 100, 200, 500, 1000, 5000, 10000, 100000, 1000000 fold or more.

[0017] In various aspects, the invention relates to kits. In a first aspect, the invention relates to a kit comprising a restriction endonuclease, a first adapter comprising one or more non-

IDT Exhibit 1020.5365

canonical nucleotides on one strand and lacking 5' phosphates, a second adapter lacking said one or more non-canonical nucleotides and lacking 5' phosphates, a ligase, a polymerase, a cleavage agent, a library of probes, a set of primers specific to the adaptor sequences; wherein the second adapter comprises a recognition sequence for the restriction endonuclease.

[0018] In a second aspect, the invention relates to a kit comprising a restriction endonuclease, a first adapter lacking 5' phosphates, a plurality of partial duplex primers each comprising a 3' overhang and comprising a shared sequence within a double-stranded portion, a ligase, a polymerase, a cleavage agent, a library of probes capable of acting as a primer for a primer extension reaction; and a primer that is hybridizable a sequence reverse complimentary to the adapter; wherein the first adapter comprises a recognition sequence for the restriction endonuclease; and wherein the plurality of partial duplex primers comprise at least two partial duplex primers with dissimilar 3' overhang sequences.

[0019] In a third aspect, the invention relates to a kit comprising a restriction endonuclease, a first adapter lacking 5' phosphates, a plurality of partial duplex primers each comprising a 3' overhang, comprising a shared sequence within a double-stranded portion, and the strand of the plurality of partial duplex primers with the 3' overhang lacking adenines in the shared sequence within the double-stranded portion, a ligase, a polymerase, a cleavage agent, a library of probes capable of acting as a primer for a primer extension reaction; and a set of primers that are hybridizable to a sequence reverse complimentary to the adapter and the shared sequence of the partial duplex primers opposite the 3' overhang; wherein the first adapter comprises a recognition sequence for the restriction endonuclease; and wherein the plurality of partial duplex primers comprise at least two partial duplex primers with dissimilar 3' overhang sequences.

[0020] In any of the aspects: The restriction endonuclease may be BspQI. In some embodiments, the polymerase is a hot start polymerase, e.g. MyTaq. In some embodiments, the kit further comprises one or more non-canonical nucleotides. In some embodiments, the one or more non-canonical nucleotides comprise uracil or inosine. In some embodiments, the cleavage agent comprises a glycosylase. In some embodiments, the glycosylase is UNG or UDG. In some embodiments, the cleavage agent comprises a primary amine. In some embodiments, the cleavage agent comprises a polyamine. In some embodiments, the polyamine is DMED. In some embodiments, the cleavage agent comprises a glycosylase and a polyamine. In some embodiments, the cleavage agent comprises endonuclease V.

IDT Exhibit 1020.5366

[0021] Kits for performing any of the methods described herein are another feature of the invention. Such kits may include reagents, enzymes and platforms for selective enrichment, amplification, and sequencing of nucleic acids. In one embodiment, a kit is provided comprising: a) an adaptor or several adaptors, b) one or more of oligonucleotide primers, and c) reagents for amplification. In another embodiment, the kit further comprises a sequence and dsDNA-specific nucleic acid modifying enzyme, such as a restriction enzyme. In yet another embodiment, the kit further comprises reagents for sequencing. A kit will preferably include instructions for employing the kit components as well as the use of any other reagent not included in the kit.

[0022] In any of the aspects, the methods, compositions and kits of the invention relate to primer probes that are capable of acting as primers targeting sequences on undesired nucleic acids.

[0023] In any of the aspects, the invention relates to a set of primer probes that are generated by, compiling undesired nucleic acid sequences, optionally compiling desired nucleic acid sequences, choosing one or more strands in each undesired nucleic acid sequence, computationally fragmenting each undesired nucleic acid sequence to stretches of a selected length, for example 40-200, 50-180, 60-150, 70-120, 80-110, 90-100 bases long, selecting a target melting temperature range, for example 40-90, 45-85, 50-80, 55-75, 60-70, 55-65° C etc., for amplification primers, optionally selecting a target length range, for example 10-80, 11-70, 12-65, 13-60, 14-55, 15-50, 16-45, 17-40, 18-35, 19-30, 10-30, 11-28, 12-26, 13-24, 14-22, 15-20, 10-20, 11-19, 12-18, 13-17, 14-16 nucleotides long, etc., for amplification primers, designing qualified amplification primers targeting a portion in the one or more strands in each undesired nucleic acid sequence, with a predicted melting temperature within the target temperature range and, optionally, with a length within the target length range, optionally determining whether one or more of the designed amplification primers are hybridizable to one or more of the desired nucleic acid sequences and optionally removing any such designed amplification primers from a list of qualified amplification primers, and synthesizing oligonucleotides from the list of qualified amplification primers. Those of skill in the art appreciate that the target melting temperature range and the target length range may fall within any range bounded by any of these values (e.g., 45-55° C or 1-12 nucleotides long etc.). In some embodiments, the oligonucleotide synthesis is performed using standard phosporamidite chemistries known in the art. In some embodiments, the synthesized oligonucleotides are pooled.

IDT Exhibit 1020.5367

[0024] In any of the aspects the set of primer probes targeting undesired nucleic acid sequences may comprise between 50-10000, 55-5000, 60-1000, 70-500, 80-250, 90-200, 100-180, 110-170, 120-180, 130-170, 140-160, 100-150, 250-1000 distinct oligonucleotides. Thus, a composition of the invention comprises, consists essentially of, or consists of, a set of at least 50, 100, 150, 200, 250 different oligonucleotides wherein said oligonucleotides selectively hybridize to mitochondrial RNA, mitochondtrial DNA, human rRNA, mitochondrial rRNA, bacterial ribosomal RNA, mitochondrial DNA, human globin mRNA, human cytoplasmic rRNA, human mitochondrial rRNA, grape cytoplasmic rRNA, grape mitochondrial rRNA, or grape chloroplast rRNA. Such composition can be isolated in a vial (e.g., a reagent). Each of the oligonucleotides can have the properties (e.g., size and Tm) described herein. Those of skill in the art appreciate that the target melting temperature range and the target length range may fall within any range bounded by any of these values (e.g., 60-70, 160-200, or 150-250 etc.).

INCORPORATION BY REFERENCE

[0025] All publications, patents, and patent applications mentioned in this specification are herein incorporated by reference to the same extent as if each individual publication, patent, or patent application was specifically and individually indicated to be incorporated by reference.

BRIEF DESCRIPTION OF THE DRAWINGS

[0026] The novel features of the invention are set forth with particularity in the appended claims. A better understanding of the features and advantages of the present invention will be obtained by reference to the following description that sets forth illustrative embodiments, in which the principles of the invention are utilized, and the accompanying drawings of which:

[0027] Figure 1 depicts the elimination of non-desired nucleic acid sequences from a nucleic acid library of single-stranded DNA fragments using insert-dependent adaptor cleavage (InDA-C). The gene-specific primer (GSP) anneals to its complementary sequence only, creating a population of double-stranded or partially double-stranded molecules following polymerase-based extension. Subsequent treatment with an adaptor-specific restriction endonuclease cleaves only fragments which were activated by the GSP extension reaction, thereby removing one of the PCR priming sites from the non-desired fragments. PCR amplification produces a library that is enriched for the nucleic acid sequences of interest.

[0028] Figure 2 depicts a summary of the results from an experiment depleting bacterial rRNA fragments from strand-specific whole transcriptome cDNA libraries, as outlined in Example 1.

IDT Exhibit 1020.5368

[0029] Figure 3 depicts a comparison of the expression profiles from the four test libraries described in Example 1.

[0030] Figure 4 depicts targeted depletion of 16S rRNA sites by universal prokaryotic InDA-C probes in Example 1.

[0031] Figure 5 depicts a method of directional library construction.

[0032] Figure 6 depicts a method of nucleic acid depletion using InDA-C probes, comprising double cDNA hydrolysis.

[0033] Figure 7 depicts another method of nucleic acid depletion using InDA-C probes.

[0034] Figure 8 depicts designs for two of the partial-duplex primers.

[0035] Figure 9 depicts a method for of undesired nucleic acids by universal prokaryotic InDA-C probes.

DETAILED DESCRIPTION OF THE INVENTION

General

[0036] The methods of the invention can be used for the generation of next generation sequencing (NGS) libraries in which non-desired nucleic acid sequences have been depleted or substantially reduced. Such methods are useful, for example, for the production of sequencing libraries with reduced ribosomal RNA representation, and for the enrichment of nucleic acid sequences of interest in a nucleic acid library. Altogether, the methods of the present invention provide an improvement over the existing methods for creating NGS libraries which are depleted from non-desired nucleic acid sequences because the elimination of non-desired nucleic acid sequences occurs after the generation of the nucleic acid library, thereby enabling starting with a non-distorted, unbiased nucleic acid template population.

[0037] The methods and compositions of the invention can be used for directional library construction. The methods of the invention can further be used to generate adaptor ligated single stranded DNA samples, wherein the orientation of the adaptor is fixed.

[0038] As used herein, unless otherwise indicated, some inventive embodiments herein contemplate numerical ranges. A variety of aspects of this invention can be presented in a range format. It should be understood that the description in range format is merely for convenience and brevity and should not be construed as an inflexible limitation on the scope of the invention. Accordingly, the description of a range should be considered to have specifically disclosed all the possible subranges as well as individual numerical values within that range as if explicitly written out. For example, description of a range such as from 1 to 6 should be considered to have specifically disclosed subranges such as from 1 to 3, from 1 to 4, from 1 to 5, from 2 to 4, from 2 to 6, from 3 to 6 etc., as well as individual numbers within

IDT Exhibit 1020.5369

that range, for example, 1, 2, 3, 4, 5, and 6. This applies regardless of the breadth of the range. When ranges are present, the ranges include the range endpoints.

[0039] Reference will now be made in detail to exemplary embodiments of the invention. While the disclosed methods and compositions will be described in conjunction with the exemplary embodiments, it will be understood that these exemplary embodiments are not intended to limit the invention. On the contrary, the invention is intended to encompass alternatives, modifications and equivalents, which may be included in the spirit and scope of the invention.

[0040] Unless otherwise specified, terms and symbols of genetics, molecular biology, biochemistry and nucleic acid used herein follow those of standard treatises and texts in the field, e.g. Kornberg and Baker, *DNA Replication*, Second Edition (W.H. Freeman, New York, 1992); Lehninger, *Biochemistry*, Second Edition (Worth Publishers, New York, 1975); Strachan and Read, Human *Molecular Genetics*, Second Edition (Wiley-Liss, New York, 1999); Eckstein, editor, *Oligonucleotides and Analogs: A Practical Approach* (Oxford University Press, New York, 1991); Gait, editor, *Oligonucleotide Synthesis: A Practical Approach* (IRL Press, Oxford, 1984); and the like.

Oligonucleotides of the invention

[0041] As used within the invention, the term "oligonucleotide" refers to a polynucleotide chain, typically less than 200 residues long, most typically between 15 and 100 nucleotides long, but also intended to encompass longer polynucleotide chains. Oligonucleotides may be single-or double-stranded. The terms "oligonucleotide probe" or "probe", as used in this invention, refer to an oligonucleotide capable of hybridizing to a complementary nucleotide sequence. As used in this invention, the term "oligonucleotide" may be used interchangeably with the terms "primer", "adaptor" and "probe".

[0042] As used herein, the terms "hybridization"/ "hybridizing" and "annealing" are used interchangeably and refer to the pairing of complementary nucleic acids.

[0043] The term "primer", as used herein, refers to an oligonucleotide, generally with a free 3' hydroxyl group that is capable of hybridizing with a template (such as a target polynucleotide, target DNA, target RNA or a primer extension product) and is also capable of promoting polymerization of a polynucleotide complementary to the template. A primer may contain a non-hybridizing sequence that constitutes a tail of the primer. A primer may still be hybridizing to a target even though its sequences are not fully complementary to the target.

[0044] The primers of the invention are generally oligonucleotides that are employed in an extension reaction by a polymerase along a polynucleotide template, such as in PCR or

cDNA synthesis, for example. The oligonucleotide primer is often a synthetic polynucleotide that is single stranded, containing a sequence at its 3′-end that is capable of hybridizing with a sequence of the target polynucleotide. Normally, the 3′ region of the primer that hybridizes with the target nucleic acid has at least 80%, preferably 90%, more preferably 95%, most preferably 100%, complementarity to a sequence or primer binding site.

[0045] "Complementary", as used herein, refers to complementarity to all or only to a portion of a sequence. The number of nucleotides in the hybridizable sequence of a specific oligonucleotide primer should be such that stringency conditions used to hybridize the oligonucleotide primer will prevent excessive random non-specific hybridization. Usually, the number of nucleotides in the hybridizing portion of the oligonucleotide primer will be at least as great as the defined sequence on the target polynucleotide that the oligonucleotide primer hybridizes to, namely, at least 5, at least 6, at least 7, at least 8, at least 9, at least 10, at least 11, at least 12, at least 13, at least 14, at least 15, at least about 20, and generally from about 6 to about 10 or 6 to about 12 of 12 to about 200 nucleotides, usually about 10 to about 50 nucleotides. In general, the target polynucleotide is larger than the oligonucleotide primer or primers as described previously.

[0046] Complementary may generally refer to the capacity for precise pairing between two nucleotides. I.e., if a nucleotide at a given position of a nucleic acid is capable of hydrogen bonding with a nucleotide of another nucleic acid, then the two nucleic acids are considered to be complementary to one another at that position. A "complement" may be an exactly or partially complementary sequence. Complementarity between two single-stranded nucleic acid molecules may be "partial," in which only some of the nucleotides bind, or it may be complete when total complementarity exists between the single-stranded molecules. The degree of complementarity between nucleic acid strands has significant effects on the efficiency and strength of hybridization between nucleic acid strands. Two sequences that are partially complementary may have, for example, at least 90% identity, or at least 95%, 96%, 97%. 98%, or 99% identity sequence over a sequence of at least 7 nucleotides, more typically in the range of 10-30 nucleotides, and often over a sequence of at least 14-25 nucleotides. It will be understood that the 3′ base of a primer sequence will desirably be perfectly complementary to corresponding bases of the target nucleic acid sequence to allow priming to occur.

[0047] "Specific hybridization" refers to the binding of a nucleic acid to a target nucleotide sequence in the absence of substantial binding to other nucleotide sequences present in the hybridization mixture under defined stringency conditions. Those of skill in the art recognize

IDT Exhibit 1020.5371

that relaxing the stringency of the hybridization conditions allows sequence mismatches to be tolerated. In particular embodiments, hybridizations are carried out under stringent hybridization conditions.

[0048] "Tm" refers to "melting temperature", which is the temperature at which a population of double-stranded nucleic acid molecules becomes half-dissociated into single strands. The Tm of a single stranded oligonucleotide, as used herein, refers to the Tm of a double stranded molecule comprising the oligonucleotide and its exact complement. Tm may be determined by calculation. Specifically, the Tm of an oligonucleotide may be a calculated Tm according to the equation: "Tm (° C.)=4(G+C)+2(A+T)" (Thein and Wallace, 1986, in Human genetic disorders, p 33-50, IRL Press, Oxford UK, incorporated herein by reference).

[0049] In some cases, the identity of the investigated target polynucleotide sequence is known, and hybridizable primers can be synthesized precisely according to the antisense sequence of the aforesaid target polynucleotide sequence. In other cases, when the target polynucleotide sequence is unknown, the hybridizable sequence of an oligonucleotide primer is a random sequence. Oligonucleotide primers comprising random sequences may be referred to as "random primers", as described below. In yet other cases, an oligonucleotide primer such as a first primer or a second primer comprises a set of primers such as for example a set of first primers or a set of second primers. In some cases, the set of first or second primers may comprise a mixture of primers designed to hybridize to a plurality (e.g. 2, 3, 4, about 6, 8, 10, 20, 40, 80, 100, 125, 150, 200, 250, 300, 400, 500, 600, 800, 1000, 1500, 2000, 2500, 3000, 4000, 5000, 6000, 7000, 8000, 10,000, 20,000, 25,000 or more) of target sequences. In some cases, the plurality of target sequences may comprise a group of related sequences, random sequences, a whole transcriptome or fraction (e.g. substantial fraction) thereof, or any group of sequences such as mRNA.

[0050] In some embodiments of the invention, random priming is used. A "random primer", as used herein, is a primer that generally comprises a sequence that is not designed based on a particular or specific sequence in a sample, but rather is based on a statistical expectation (or an empirical observation) that a sequence of the random primer is hybridizable, under a given set of conditions, to one or more sequences in a sample. A random primer will generally be an oligonucleotide or a population of oligonucleotides comprising a random sequence(s) in which the nucleotides at a given position on the oligonucleotide can be any of the four nucleotides A, T, G, C or any of their analogs. A random primer may comprise a 5' or 3' region that is a specific, non-random sequence. In some embodiments of the invention, the random primers comprise tailed primers with a 3' random sequence region and a 5' non-

IDT Exhibit 1020.5372

hybridizing region that comprises a specific, common adaptor sequence. The sequence of a random primer, or its complement, may or may not be naturally occurring, and may or may not be present in a pool of sequences in a sample of interest. A "random primer" can also refer to a primer that is a member of a population of primers (a plurality of random primers) which are collectively designed to hybridize to a desired target sequence or sequences.

[0051] The term "adaptor", as used herein, refers to an oligonucleotide of known sequence, the ligation or incorporation of which to a target polynucleotide or a target polynucleotide strand of interest enables the generation of amplification-ready products of the target polynucleotide or the target polynucleotide strand of interest. Various adaptor designs are envisioned. Various ligation processes and reagents are known in the art and can be useful for carrying out the methods of the invention. For example, blunt ligation can be employed. Similarly, a single dA nucleotide can be added to the 3′-end of the double-stranded DNA product, by a polymerase lacking 3′-exonuclease activity and can anneal to an adaptor comprising a dT overhang (or the reverse). This design allows the hybridized components to be subsequently ligated (e.g., by T4 DNA ligase). Other ligation strategies and the corresponding reagents and known in the art and kits and reagents for carrying out efficient ligation reactions are commercially available (e.g., from New England Biolabs, Roche).

[0052] The term "insert-dependent adaptor cleavage" (InDA-C), as used herein, refers to a multi-step process for depleting or removing specific nucleotide sequences from a nucleotide library. The first step comprises annealing sequence-specific oligonucleotides, designed to be complementary to non-desired nucleic acid sequences or sequences directly adjacent to regions of non-desired sequence, to single-stranded nucleic acid templates with adaptors of fixed orientation attached at each end. The adaptors at the 5' ends of each fragment contain a recognition sequence for a restriction endonuclease specific for double-stranded DNA. Following the annealing of the sequence-specific oligonucleotides, primer extension is performed, thereby creating double-stranded DNA fragments in the regions where the oligonucleotides are complementary to the single-stranded nucleic acid templates. The resulting nucleic acid library, containing both single-stranded and double-stranded fragments, is treated with the restriction endonuclease, resulting in cleavage at the restriction endonuclease site of the double-stranded fragments only, and thus, the removal of the adaptor at one end of the fragments containing the non-desired nucleic acid sequences. Following adaptor cleavage, PCR may be performed using primers specific to each adaptor, resulting in amplification of the desired nucleic acid fragments only (i.e. amplification of the

IDT Exhibit 1020.5373

fragments containing both PCR priming sites on the same template) . Insert-dependent adaptor cleavage is depicted in Figure 1.

[0053] Methods for designing oligonucleotides of various lengths and melting temperatures that are capable of hybridizing or that are excluded from hybridizing to a selected list of sequences are well known in the art and are described in further detail in EP 1957645B1, which is incorporated herein by reference in its entirety.

Nucleic acid modifying enzymes

[0054] The methods of the invention employ the use of nucleic acid (NA) modifying enzymes. The nucleic acid modifying enzyme can be DNA-specific modifying enzyme. The NA-modifying enzyme can be selected for specificity for double-stranded DNA. The enzyme can be a duplex-specific endonuclease, a blunt-end frequent cutter restriction enzyme, or other restriction enzyme. Examples of blunt-end cutters include DraI or SmaI. The NA-modifying enzyme can be an enzyme provided by New England Biolabs. The NA-modifying enzyme can be a homing endonuclease (a homing endonuclease can be an endonuclease that does not have a stringently-defined recognition sequence). The NA-modifying enzyme can be a high fidelity endonuclease (a high fidelity endonuclease can be an engineered endonuclease that has less "star activity" than the wild-type version of the endonuclease).

[0055] In some embodiments , the NA-modifying enzyme is a sequence- and duplex-specific DNA-modifying restriction endonuclease. In a preferred embodiment, the NA-acid modifying enzyme is the enzyme BspQI, a type IIS restriction endonuclease.

Attachment of Adaptors

Ligation

[0056] The terms "joining" and "ligation" as used herein, with respect to two polynucleotides, such as a stem-loop adaptor/primer oligonucleotide and a target polynucleotide, refers to the covalent attachment of two separate polynucleotides to produce a single larger polynucleotide with a contiguous backbone. Methods for joining two polynucleotides are known in the art, and include without limitation, enzymatic and non-enzymatic (e.g. chemical) methods. Examples of ligation reactions that are non-enzymatic include the non-enzymatic ligation techniques described in U.S. Pat. Nos. 5,780,613 and 5,476,930, which are herein incorporated by reference. In some embodiments, an adaptor oligonucleotide is joined to a target polynucleotide by a ligase, for example a DNA ligase or RNA ligase. Multiple ligases, each having characterized reaction conditions, are known in the art, and include, without limitation NAD^+-dependent ligases including tRNA ligase, Taq DNA ligase, *Thermus filiformis* DNA ligase, Escherichia coli DNA ligase, Tth DNA ligase,

IDT Exhibit 1020.5374

Thermus scotoductus DNA ligase (I and II), thermostable ligase, Ampligase thermostable DNA ligase, VanC-type ligase, 9° N DNA Ligase, Tsp DNA ligase, and novel ligases discovered by bioprospecting; ATP-dependent ligases including T4 RNA ligase, T4 DNA ligase, T3 DNA ligase, T7 DNA ligase, Pfu DNA ligase, DNA ligase 1, DNA ligase III, DNA ligase IV, and novel ligases discovered by bioprospecting; and wild-type, mutant isoforms, and genetically engineered variants thereof. Ligation can be between polynucleotides having hybridizable sequences, such as complementary overhangs. Ligation can also be between two blunt ends. Generally, a 5' phosphate is utilized in a ligation reaction. The 5' phosphate can be provided by the target polynucleotide, the adaptor oligonucleotide, or both. 5' phosphates can be added to or removed from polynucleotides to be joined, as needed. Methods for the addition or removal of 5' phosphates are known in the art, and include without limitation enzymatic and chemical processes. Enzymes useful in the addition and/or removal of 5' phosphates include kinases, phosphatases, and polymerases. In some embodiments, both of the two ends joined in a ligation reaction (e.g. an adaptor end and a target polynucleotide end) provide a 5' phosphate, such that two covalent linkages are made in joining the two ends. In some embodiments, only one of the two ends joined in a ligation reaction (e.g. only one of an adaptor end and a target polynucleotide end) provides a 5' phosphate, such that only one covalent linkage is made in joining the two ends. In some embodiments, only one strand at one or both ends of a target polynucleotide is joined to an adaptor oligonucleotide. In some embodiments, both strands at one or both ends of a target polynucleotide are joined to an adaptor oligonucleotide. In some embodiments, 3' phosphates are removed prior to ligation. In some embodiments, an adaptor oligonucleotide is added to both ends of a target polynucleotide, wherein one or both strands at each end are joined to one or more adaptor oligonucleotides. When both strands at both ends are joined to an adaptor oligonucleotide, joining can be followed by a cleavage reaction that leaves a 5' overhang that can serve as a template for the extension of the corresponding 3' end, which 3' end may or may not include one or more nucleotides derived from the adaptor oligonucleotide. In some embodiments, a target polynucleotide is joined to a first adaptor oligonucleotide on one end and a second adaptor oligonucleotide on the other end. In some embodiments, the target polynucleotide and the adaptor to which it is joined comprise blunt ends. In some embodiments, separate ligation reactions are carried out for each sample, using a different first adaptor oligonucleotide comprising at least one barcode sequence for each sample, such that no barcode sequence is joined to the target polynucleotides of more

IDT Exhibit 1020.5375