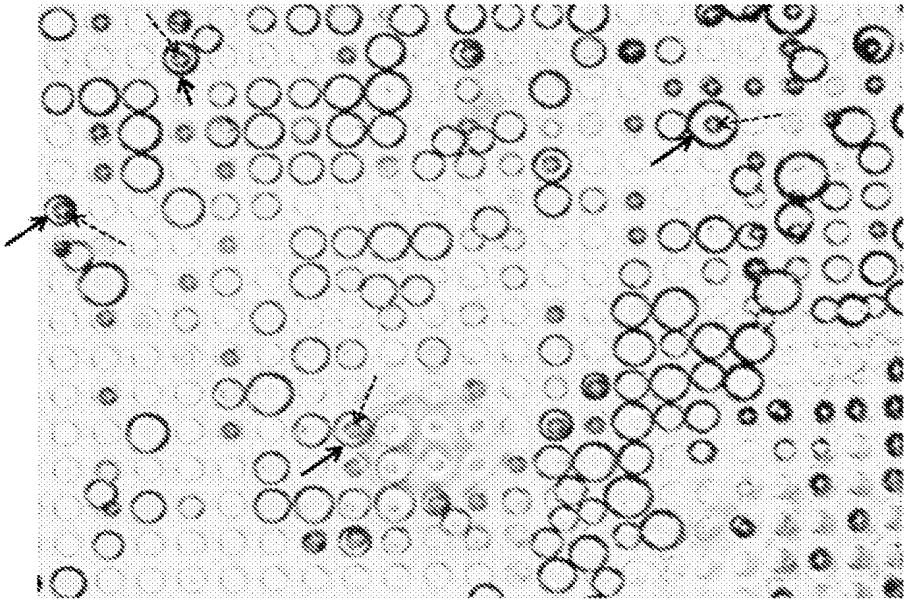


FIG. 8A

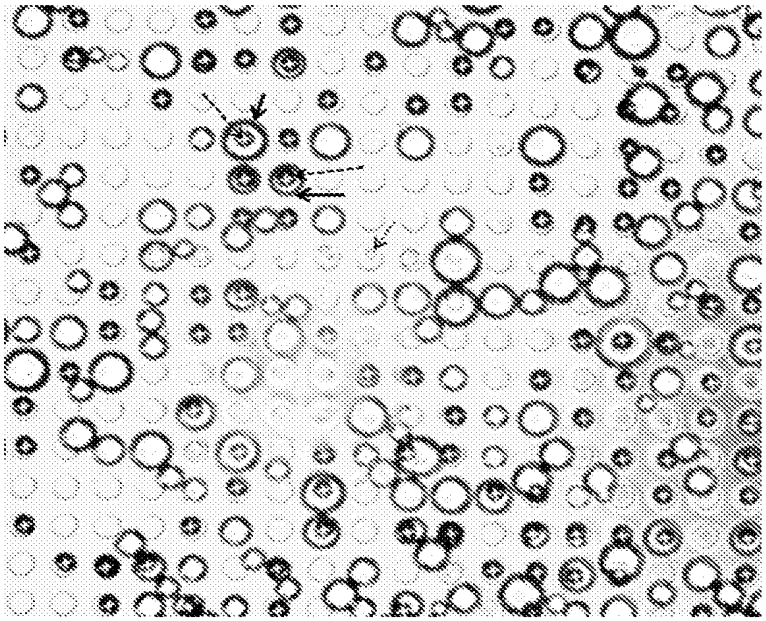


FIG. 8B

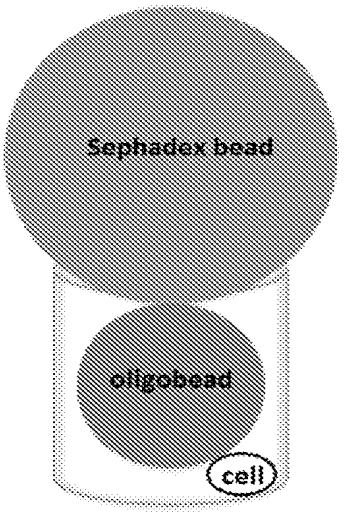


FIG. 8C

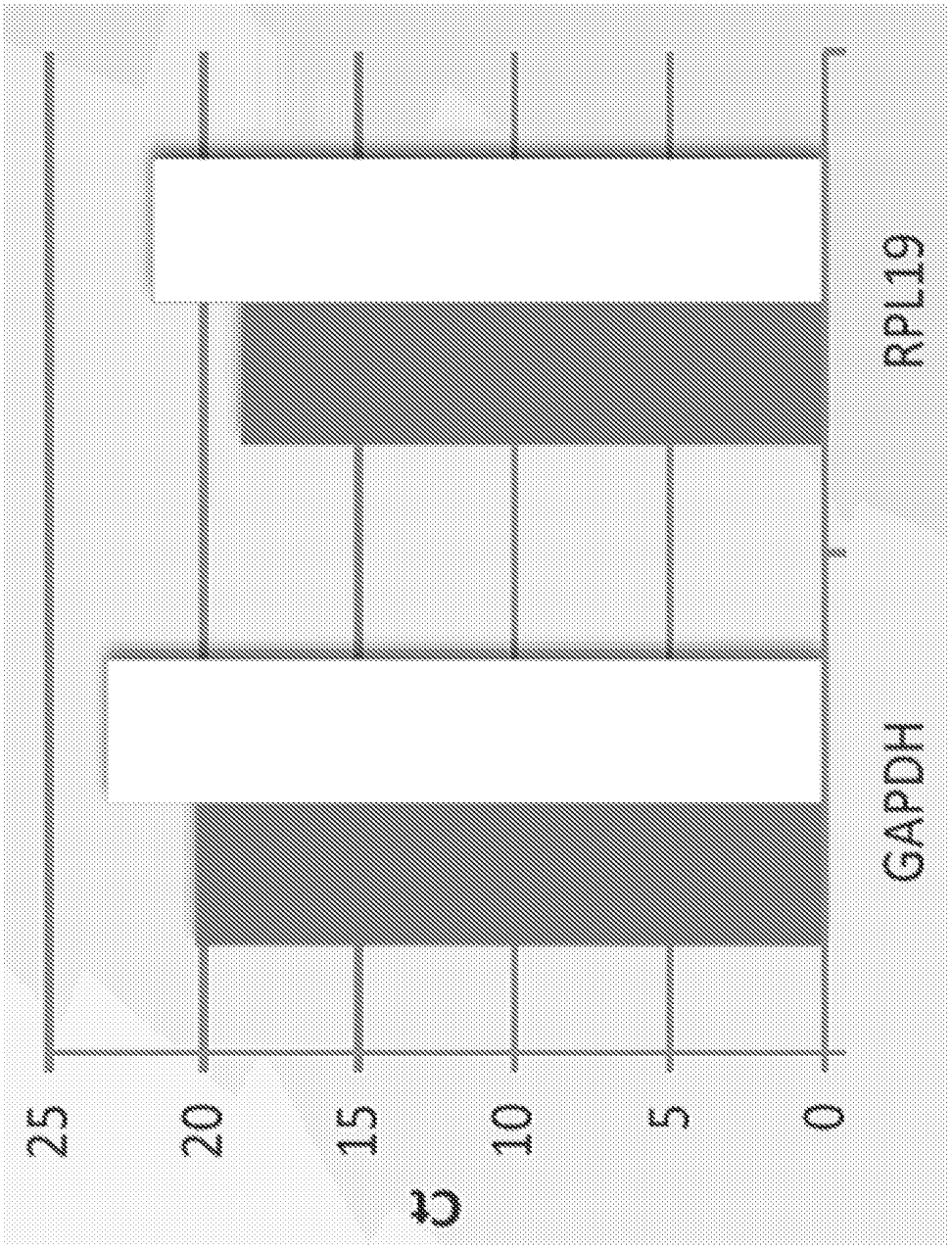


FIG. 9

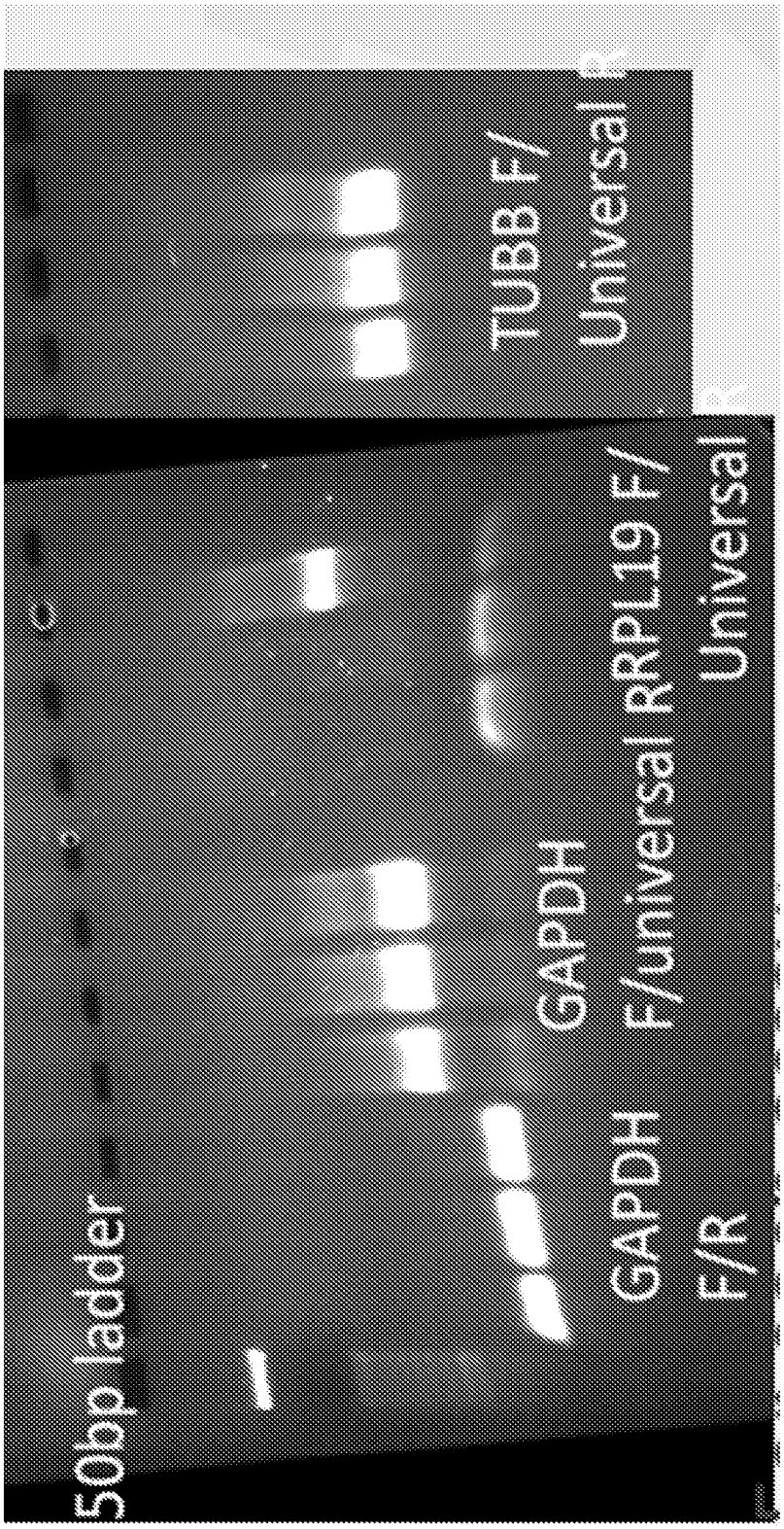

50bp ladder
GAPDH F/R
GAPDH F/universal
RRPL19 F/ Universal
TUBB F/ Universal

FIG. 10

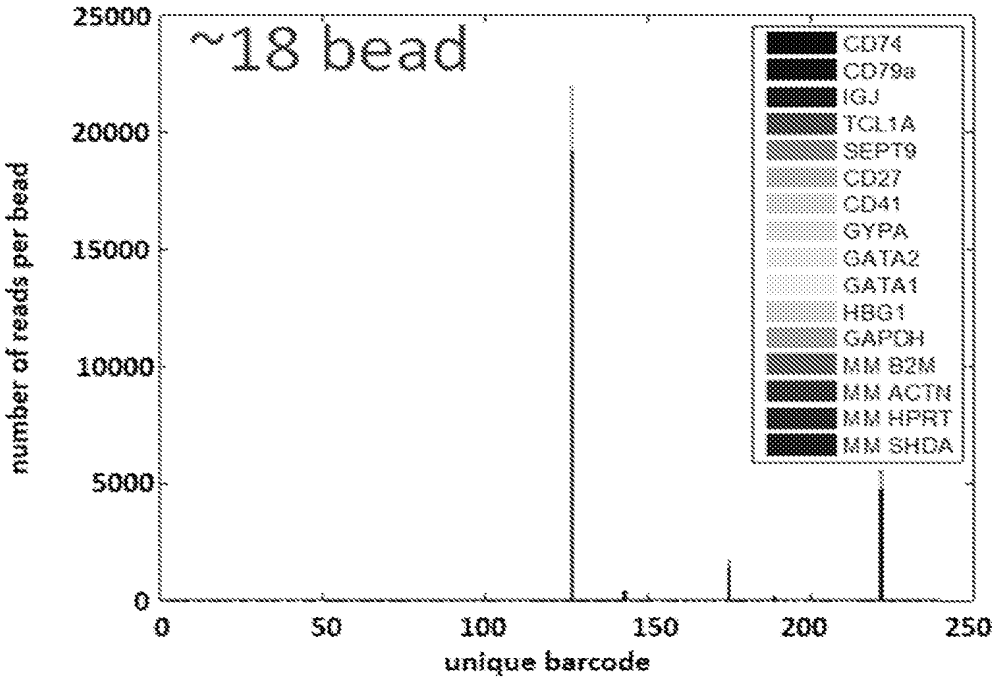


FIG. 11A

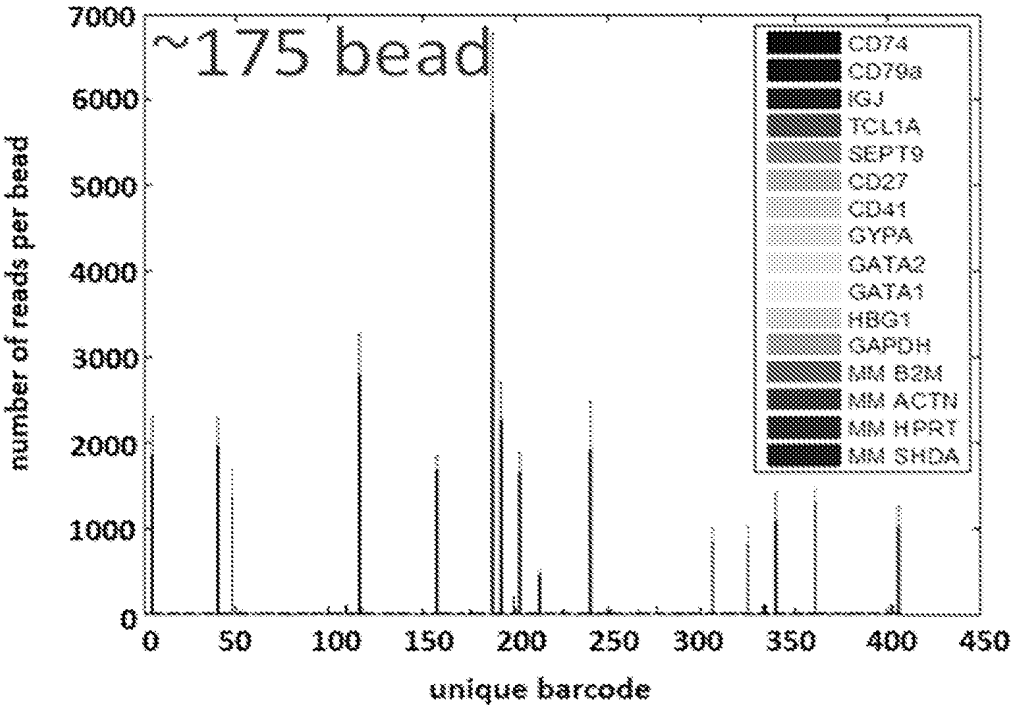


FIG. 11B

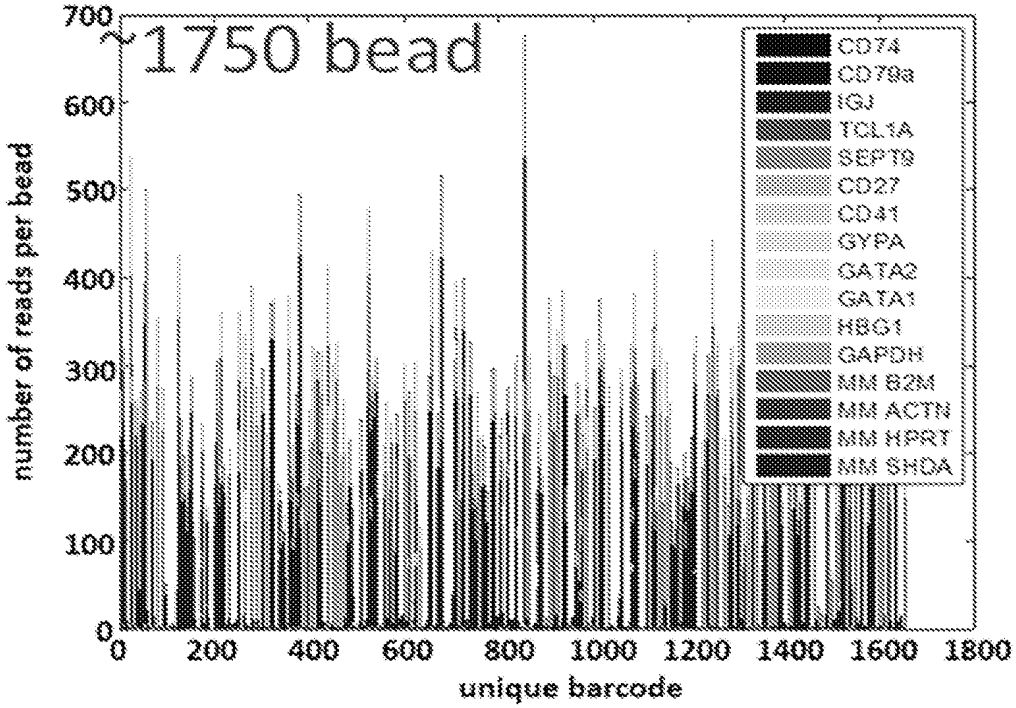


FIG. 11C

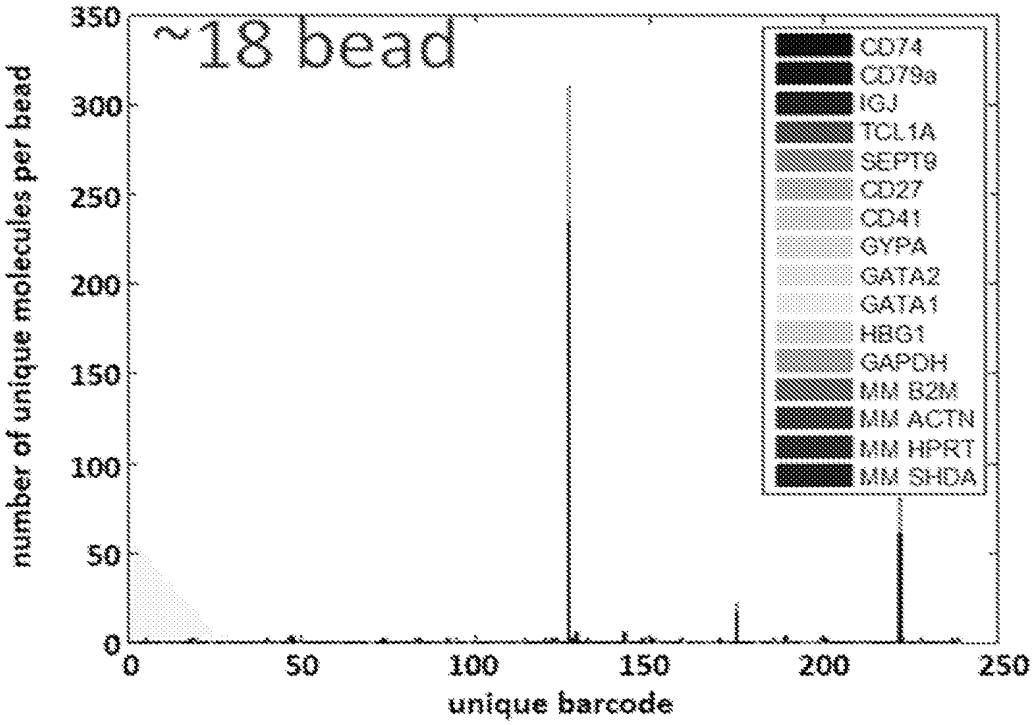


FIG. 11D

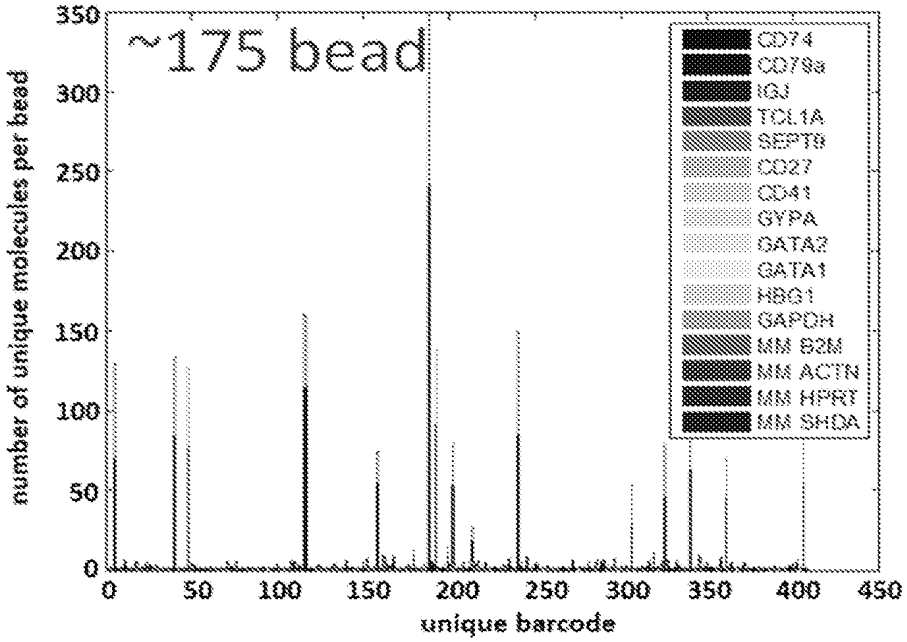


FIG. 11E

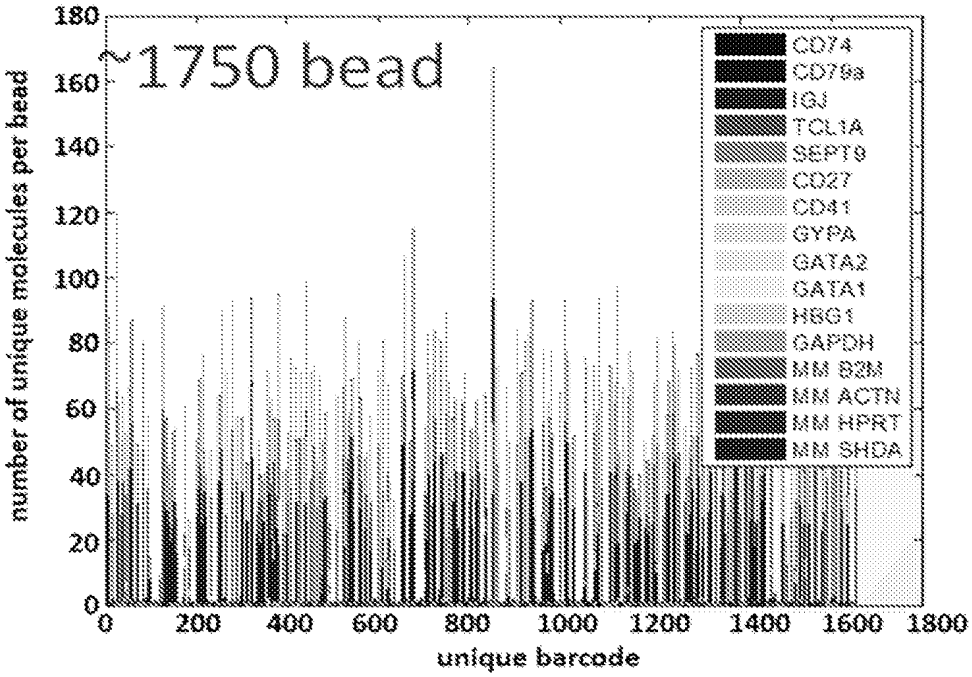


FIG. 11F

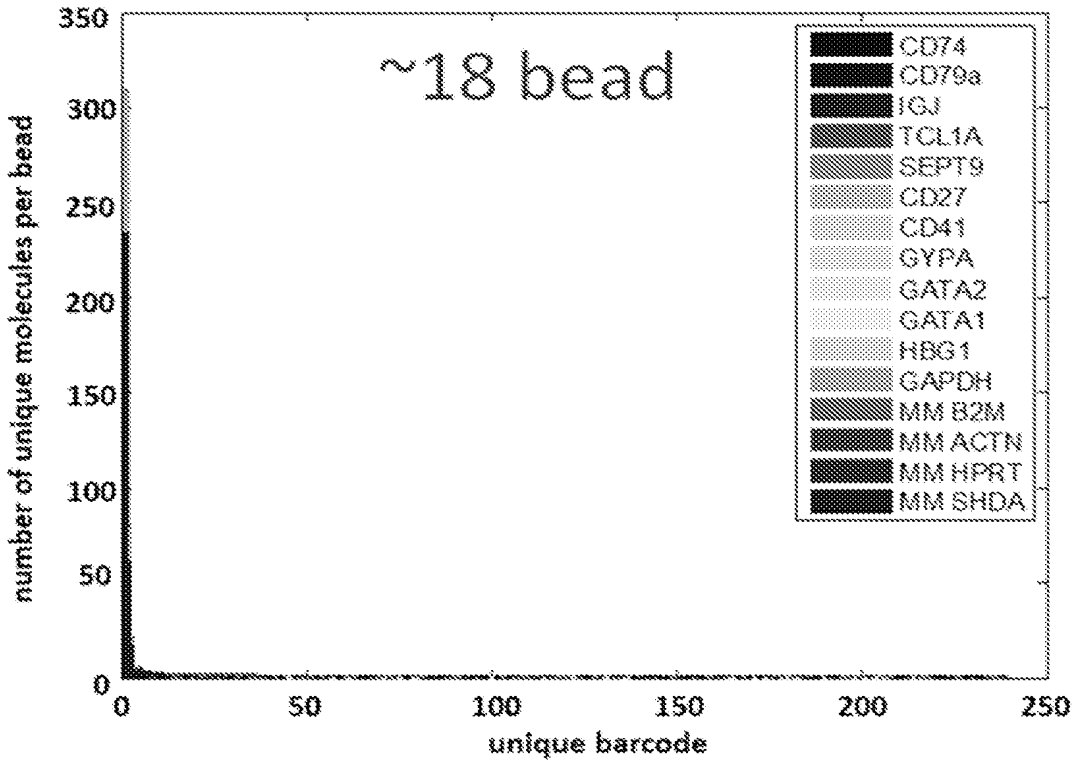


FIG. 11G

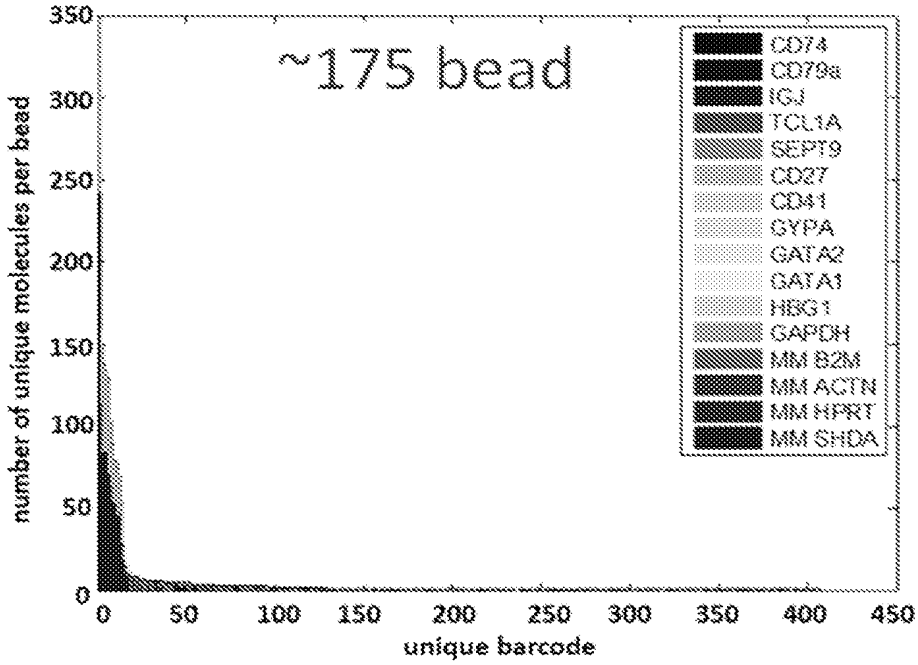


FIG. 11H

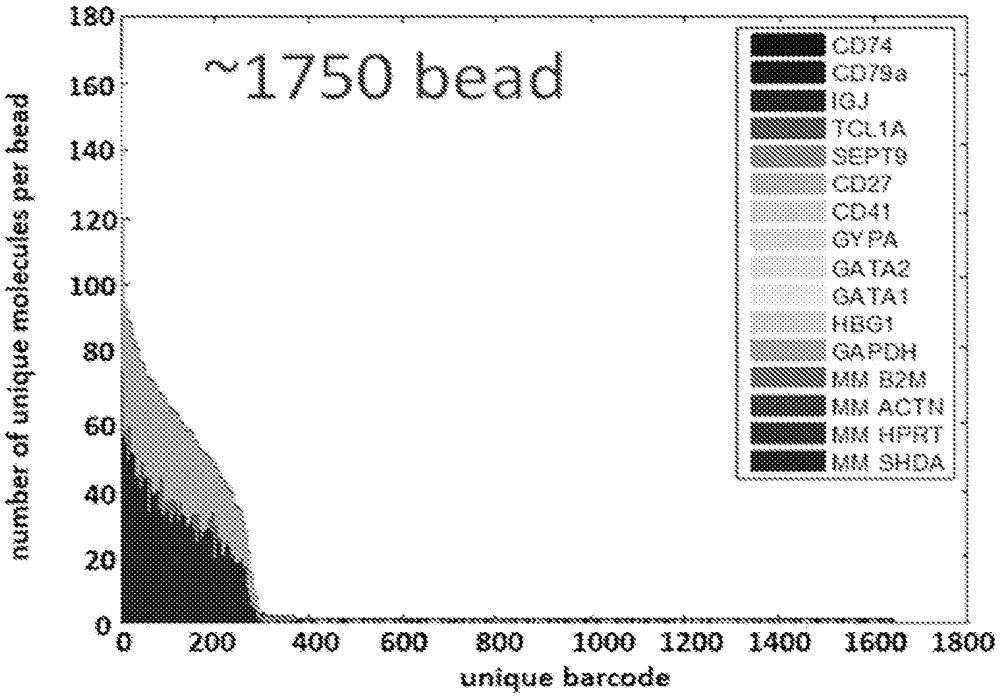


FIG. 11I

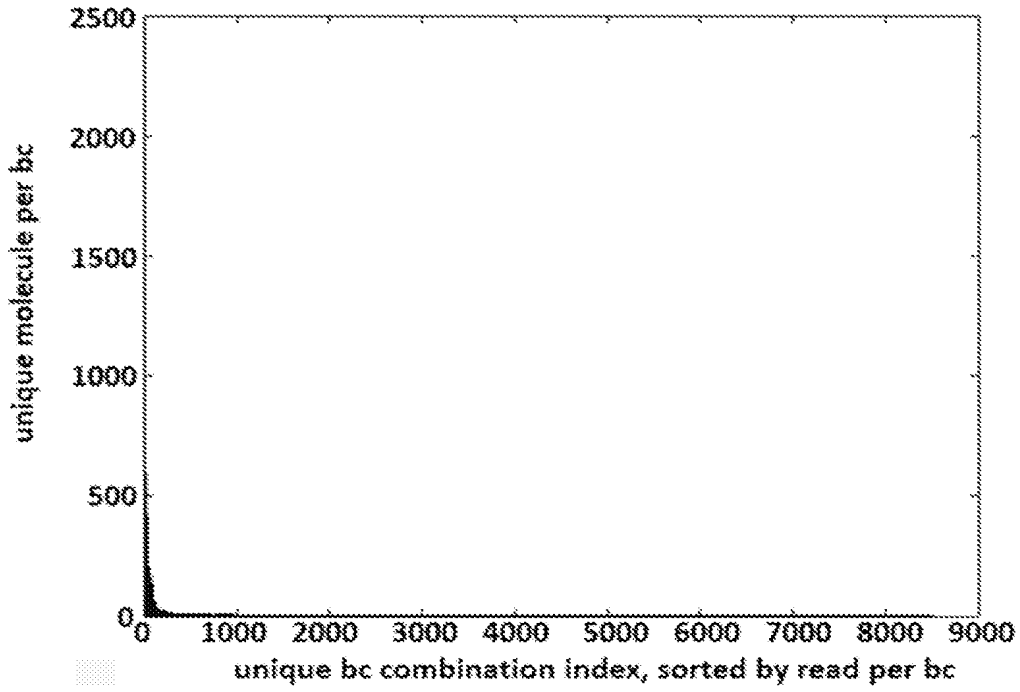


FIG. 12A

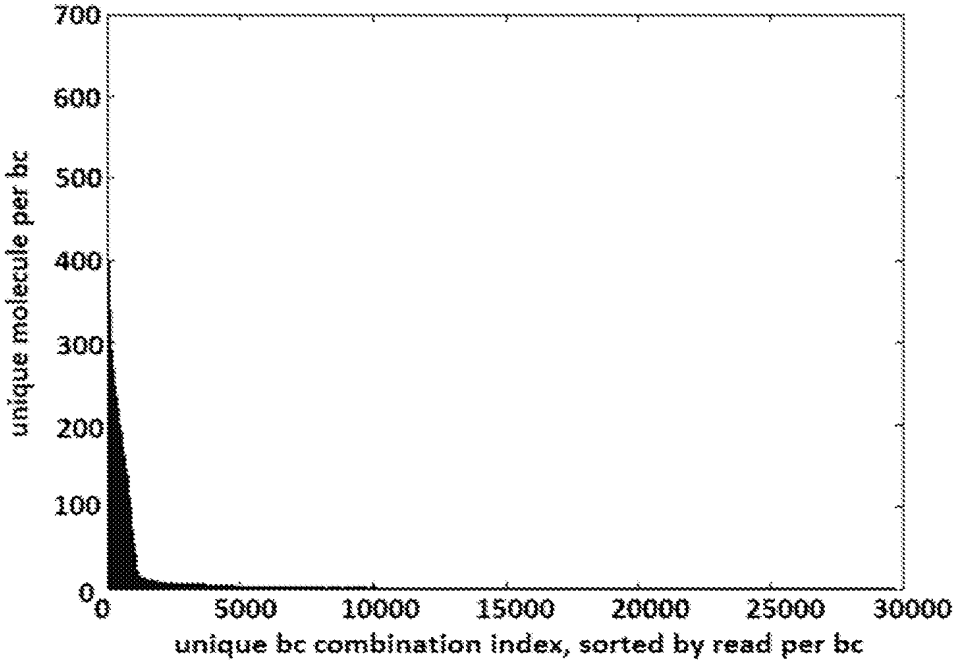


FIG. 12B

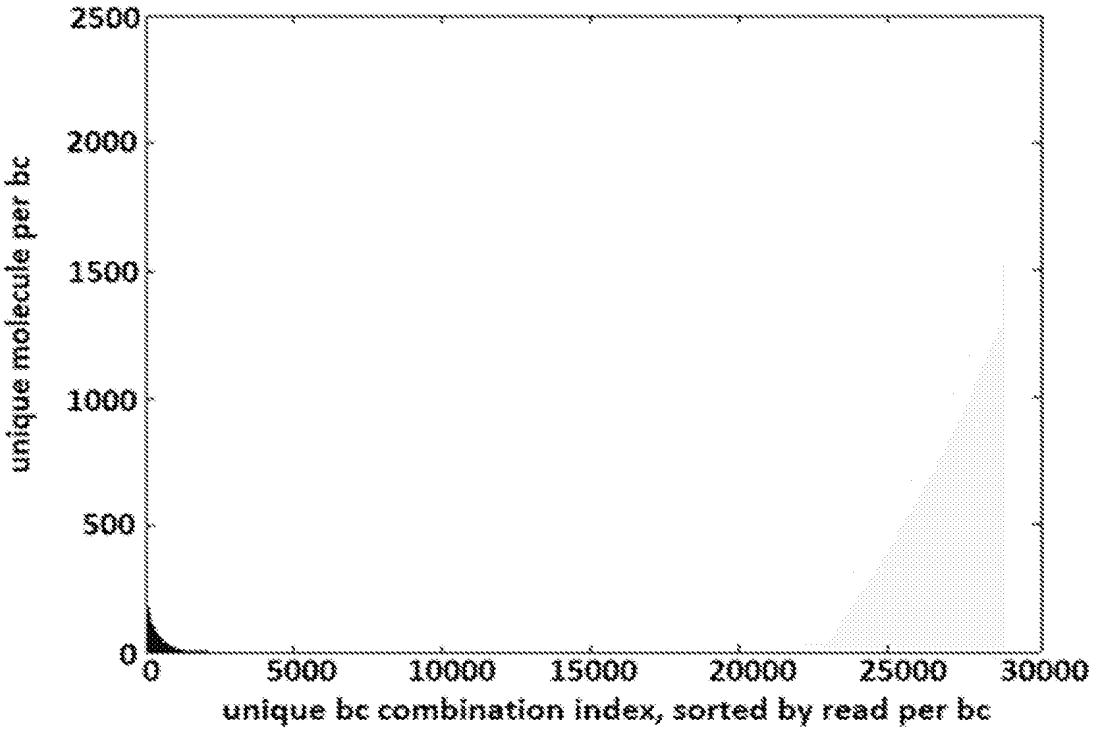


FIG. 12C

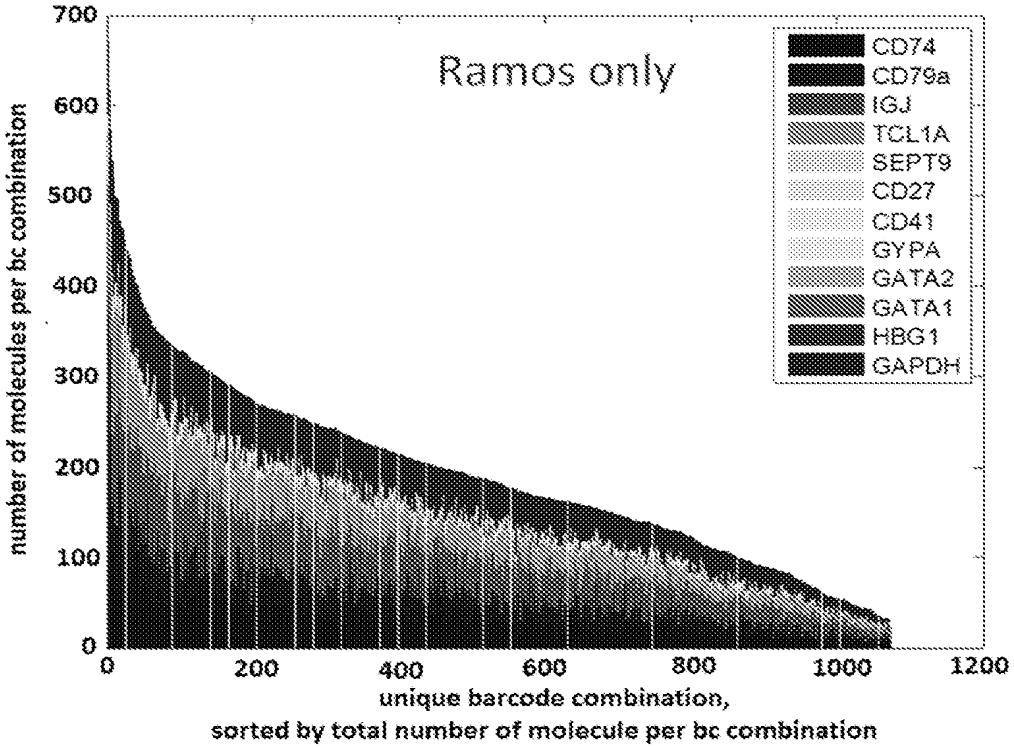


FIG. 12D

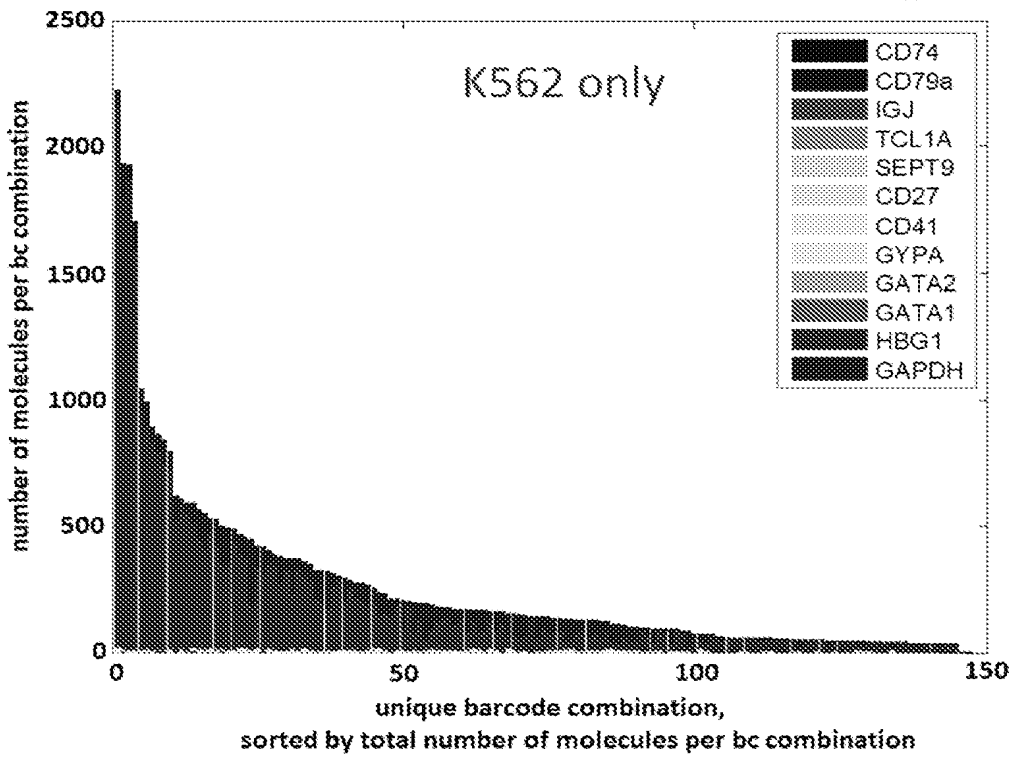


FIG. 12E

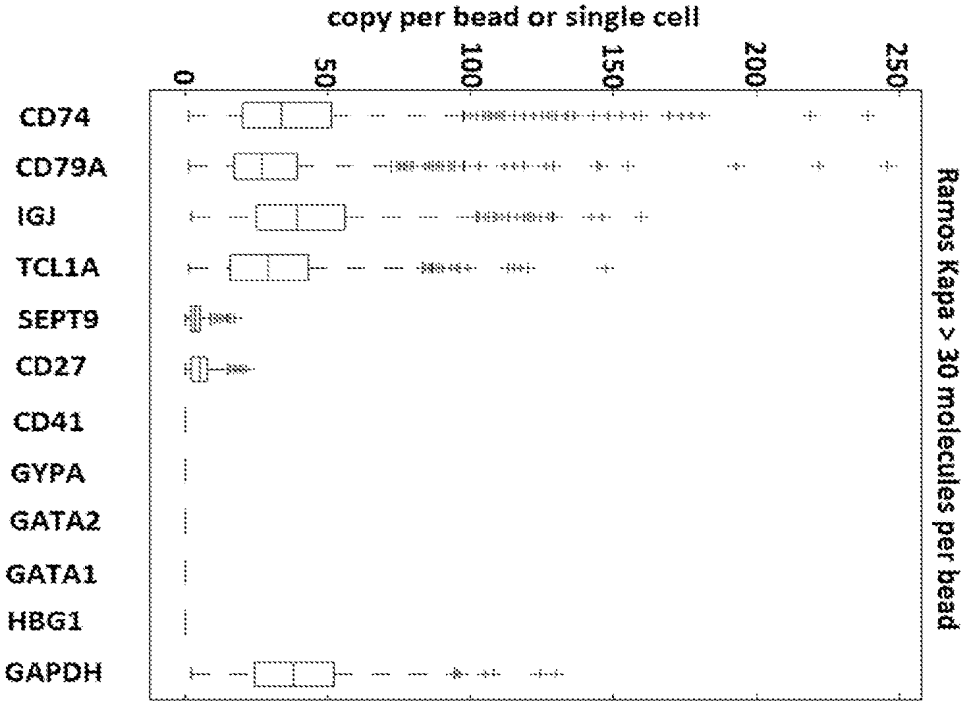


FIG. 12F

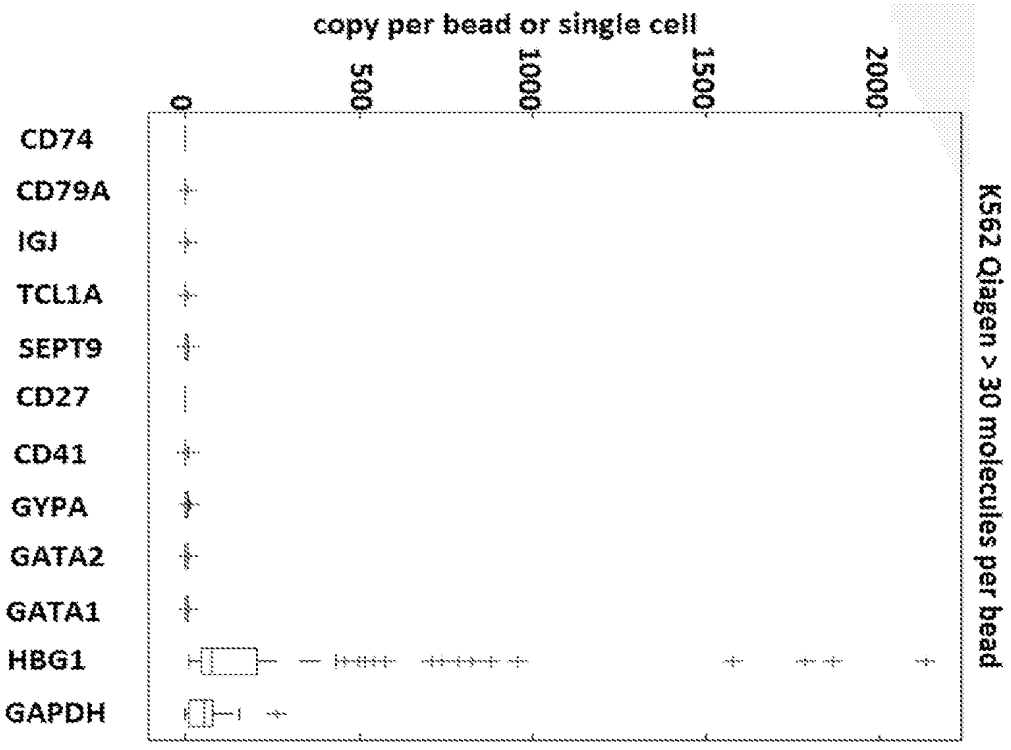


FIG. 12G

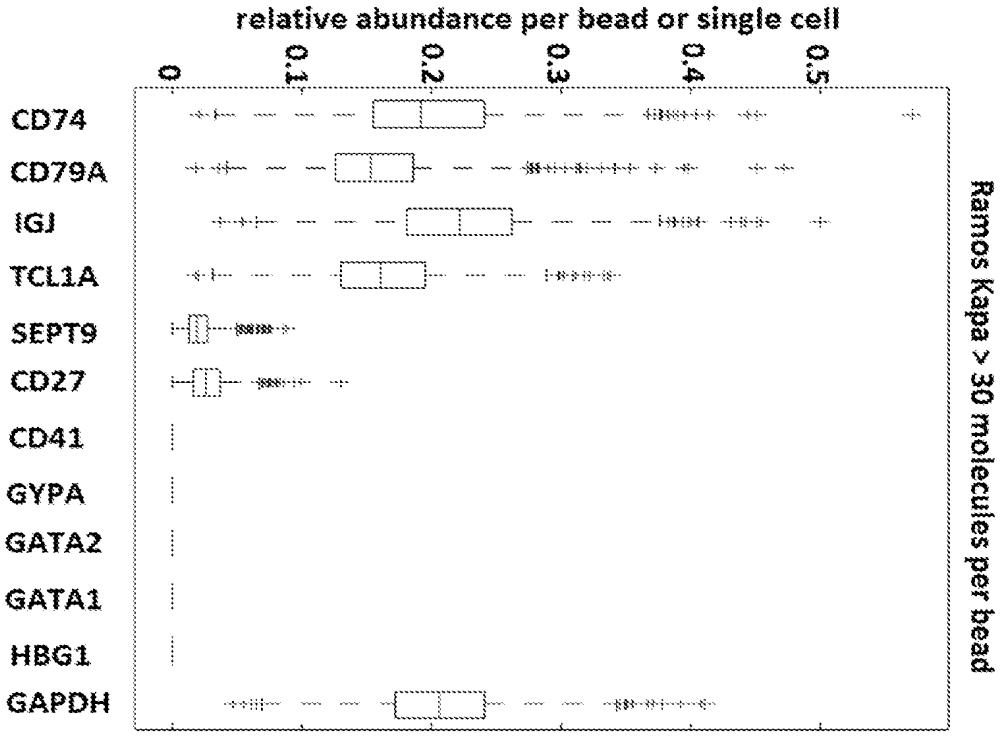


FIG. 12H

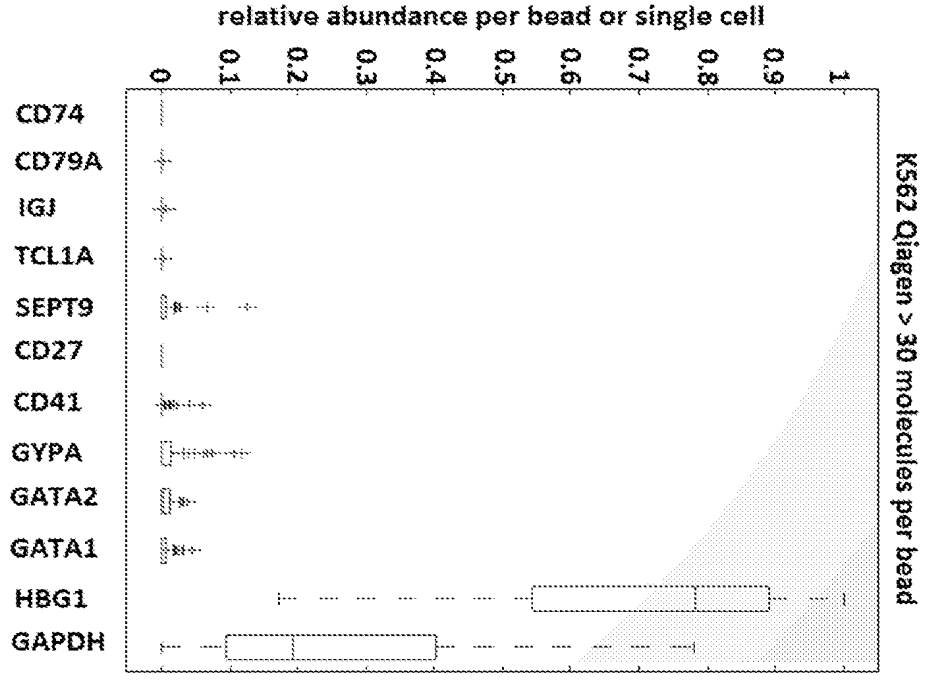


FIG. 12I

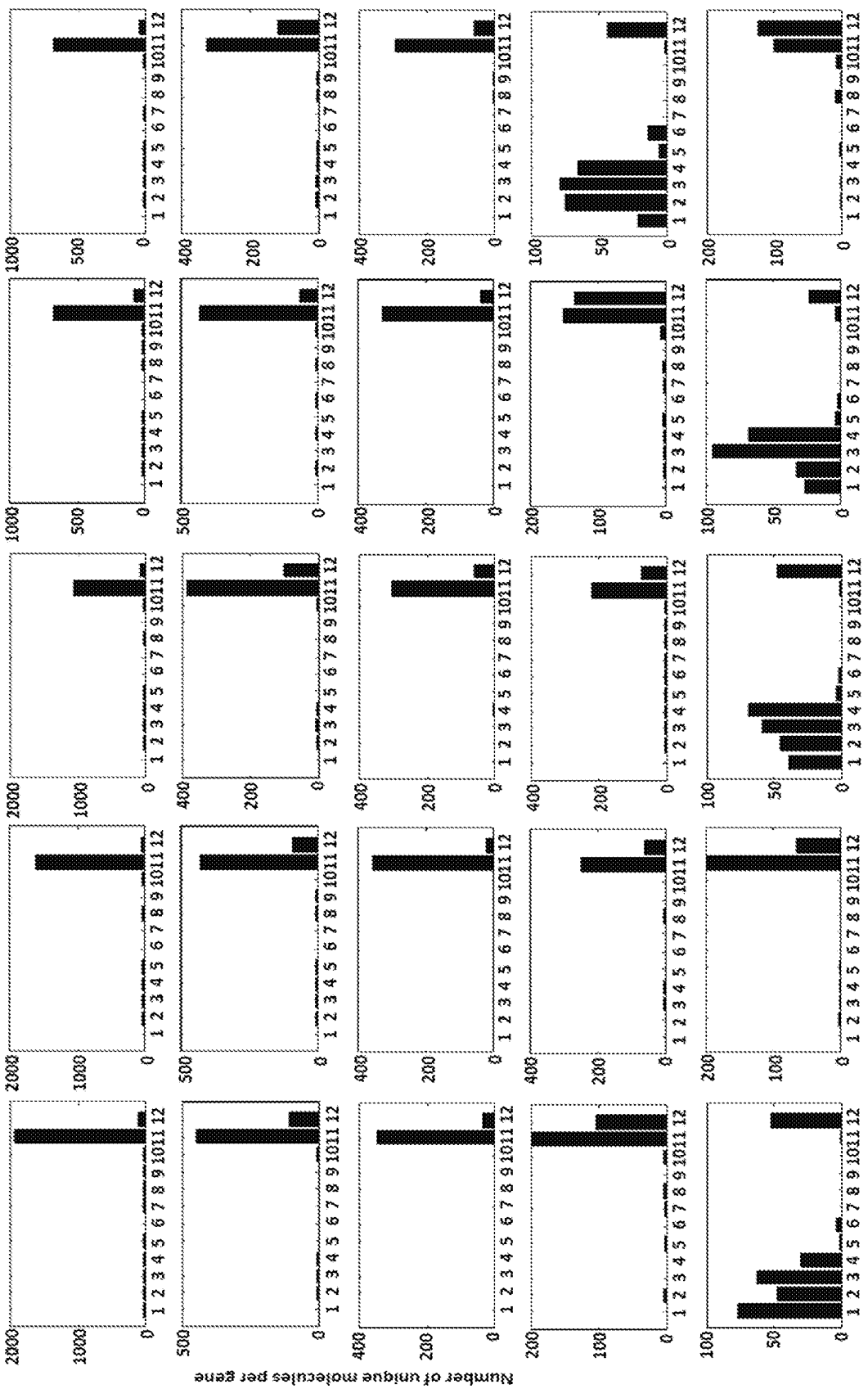


FIG. 12J

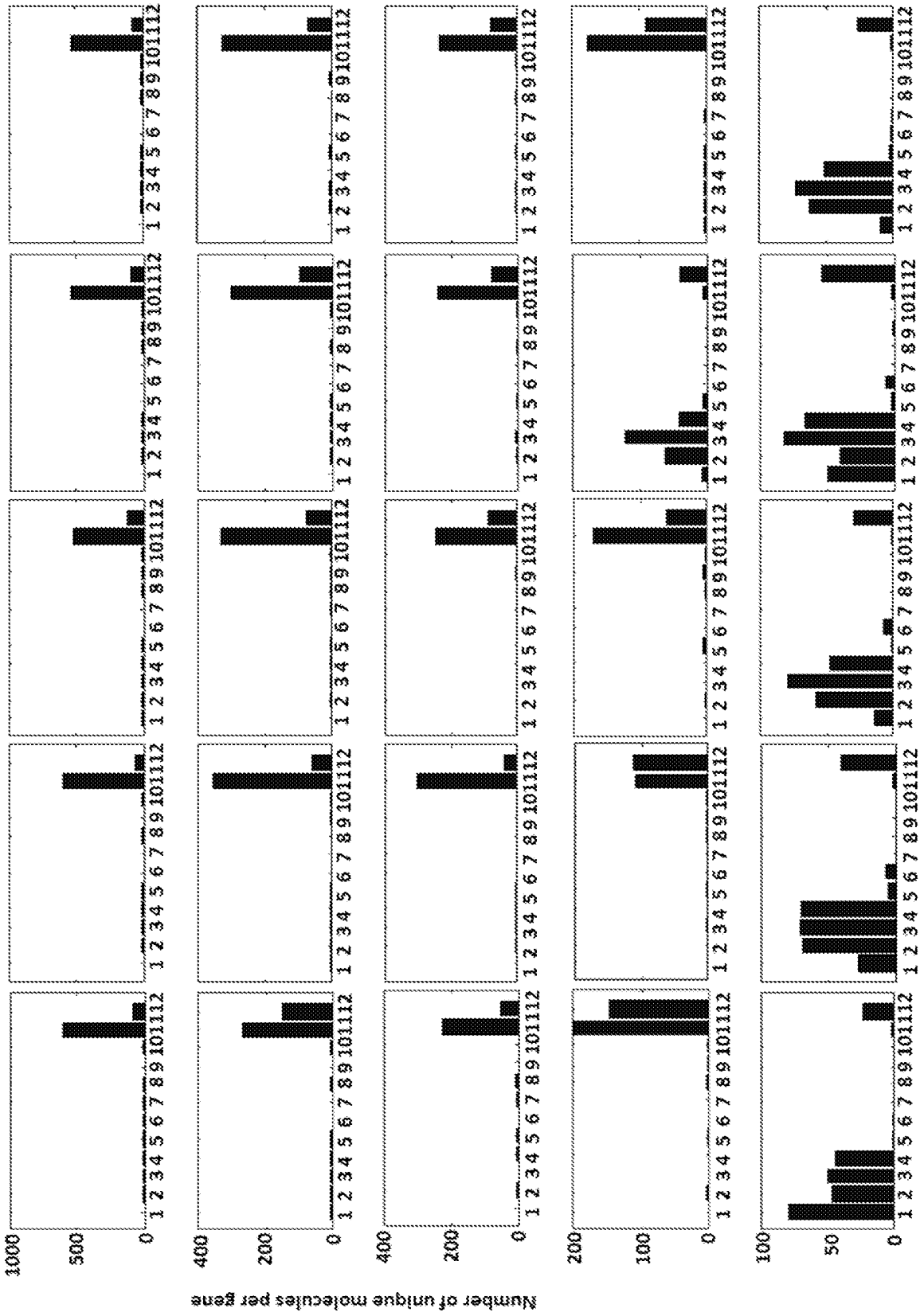


FIG. 12K

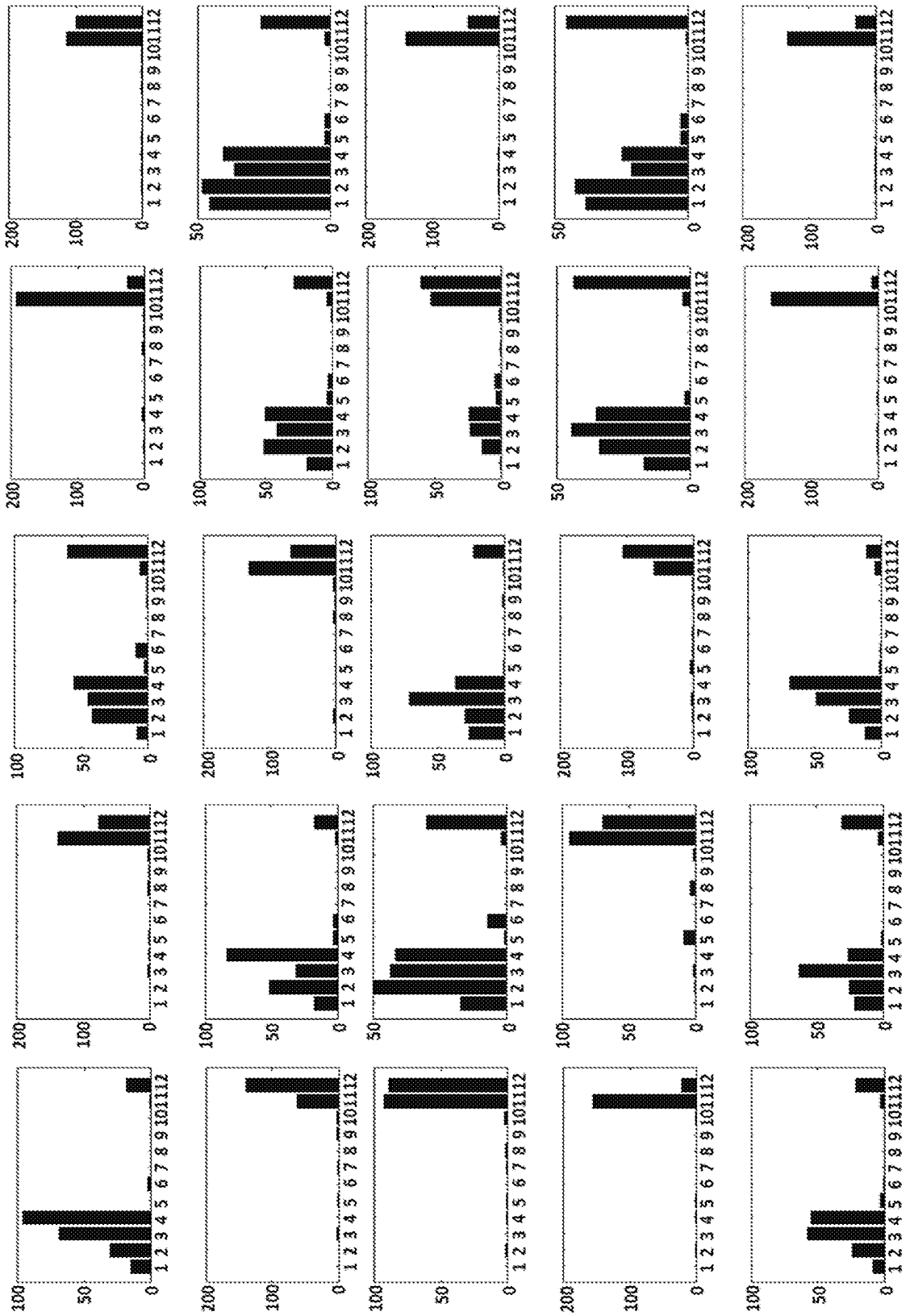


FIG. 12L

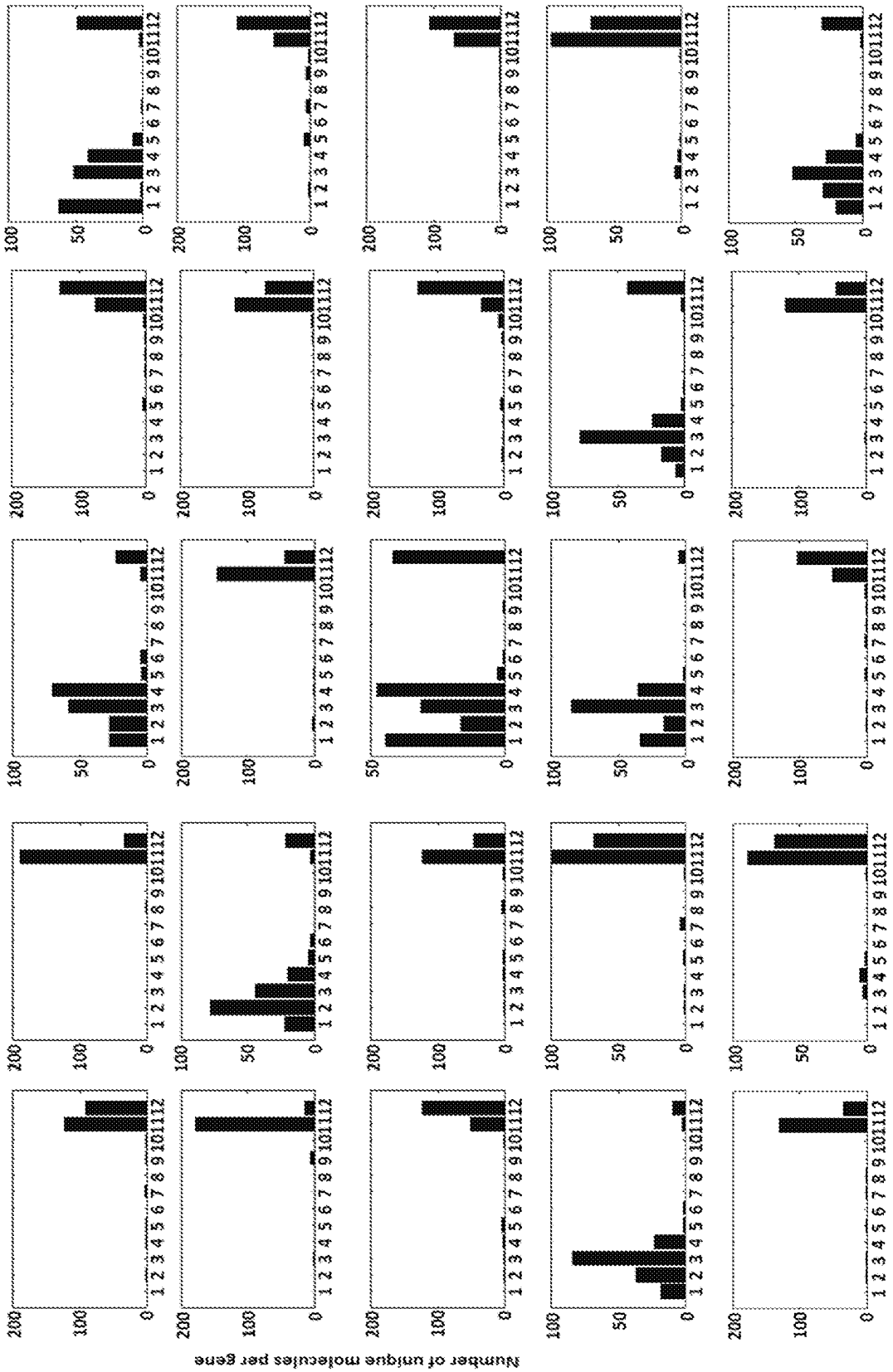


FIG. 12M

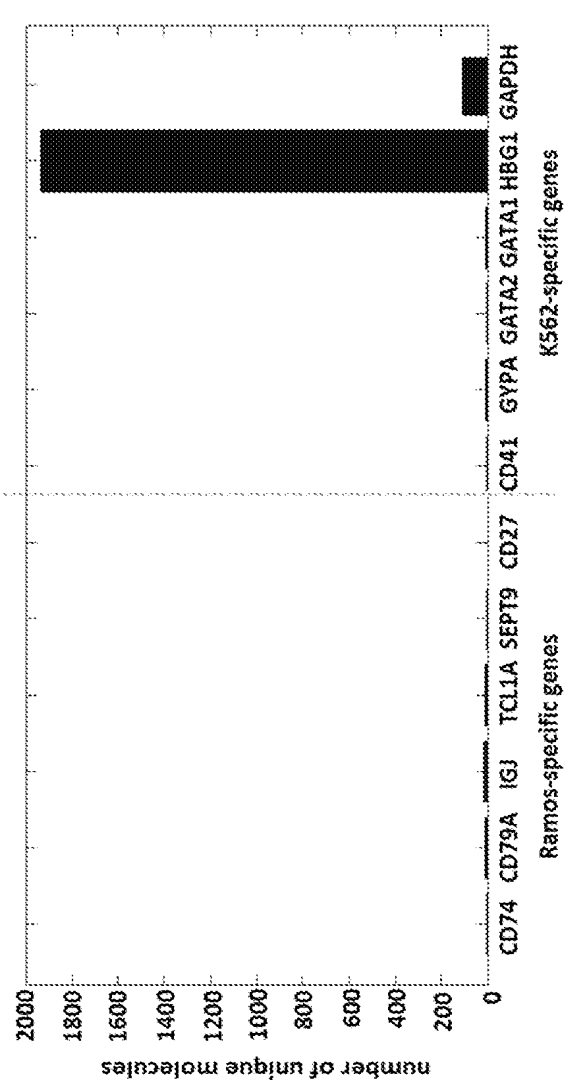


FIG. 12N

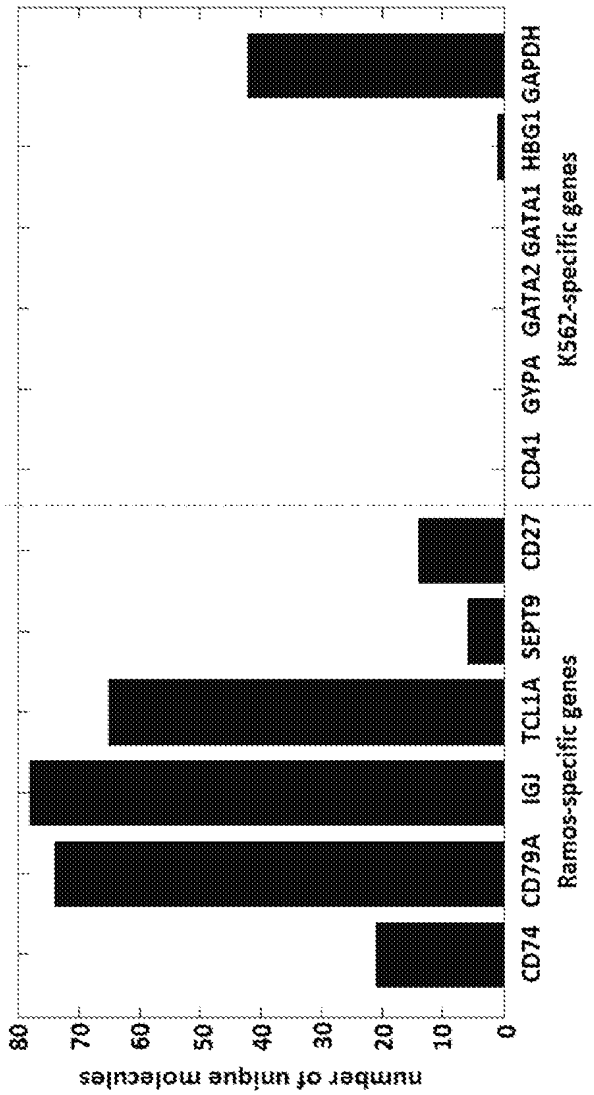


FIG. 12O

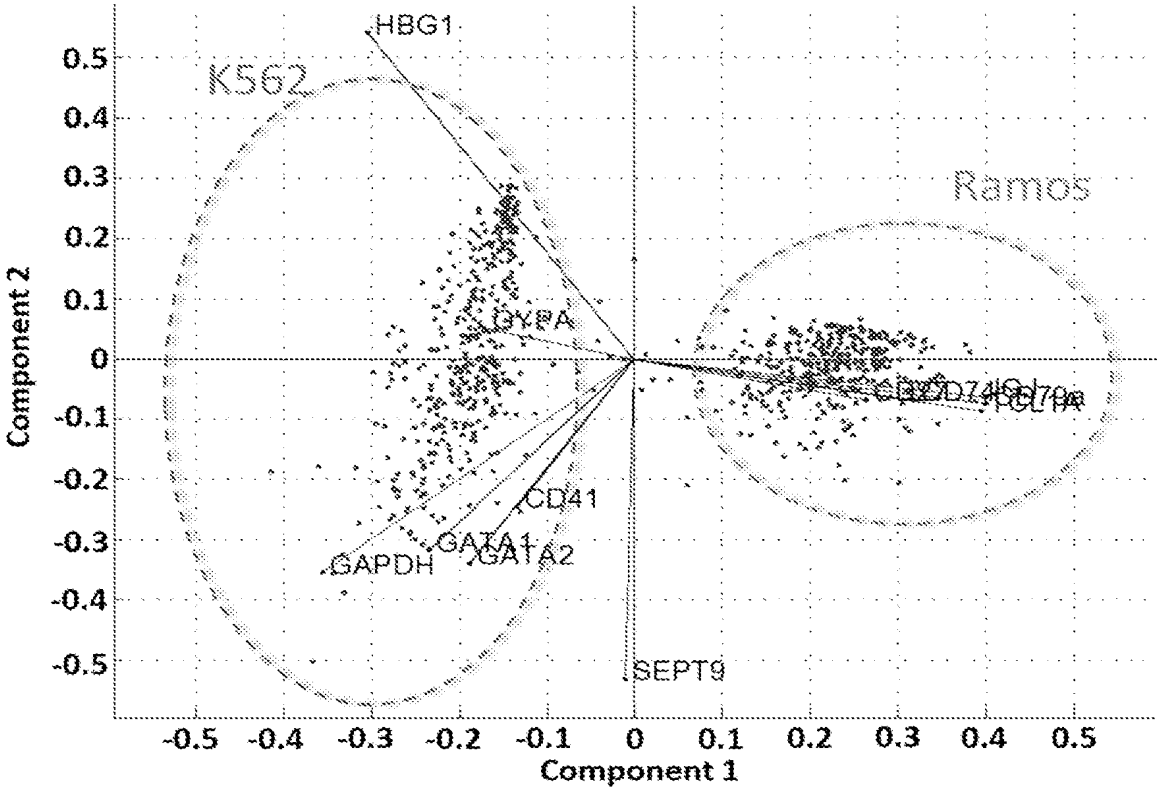


FIG. 12P

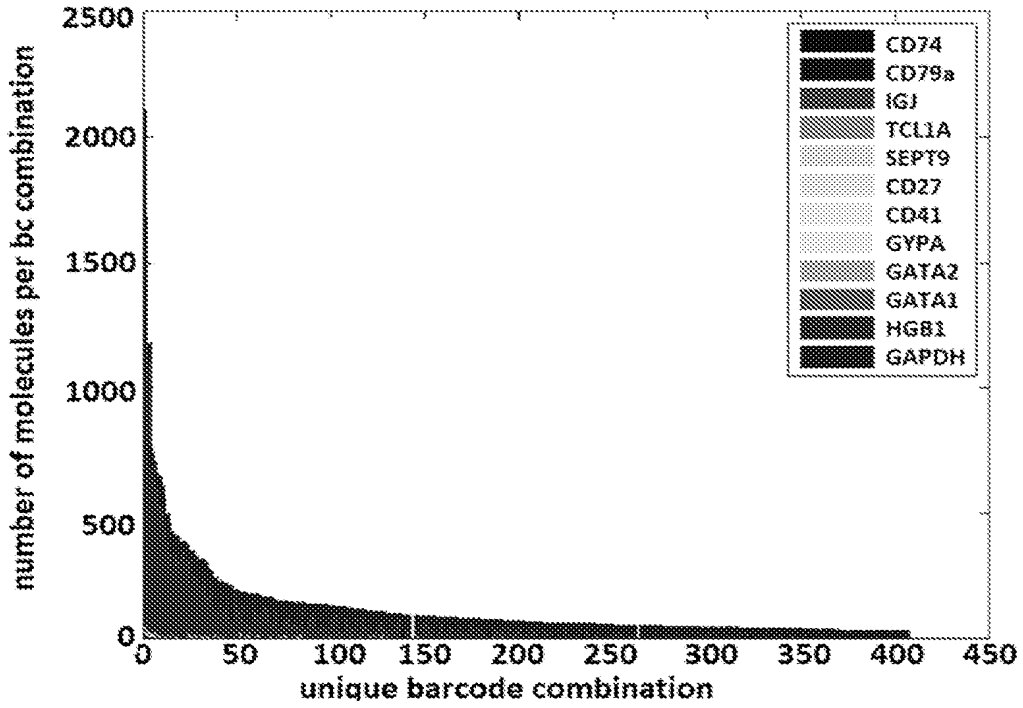


FIG. 12Q

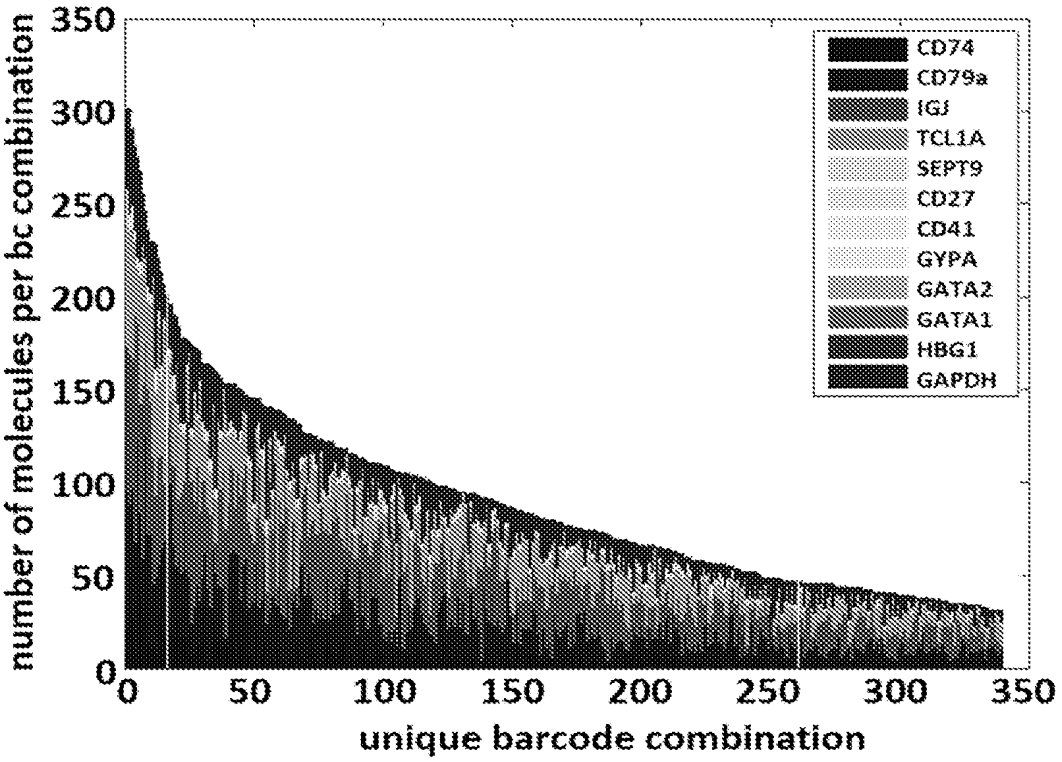


FIG. 12R

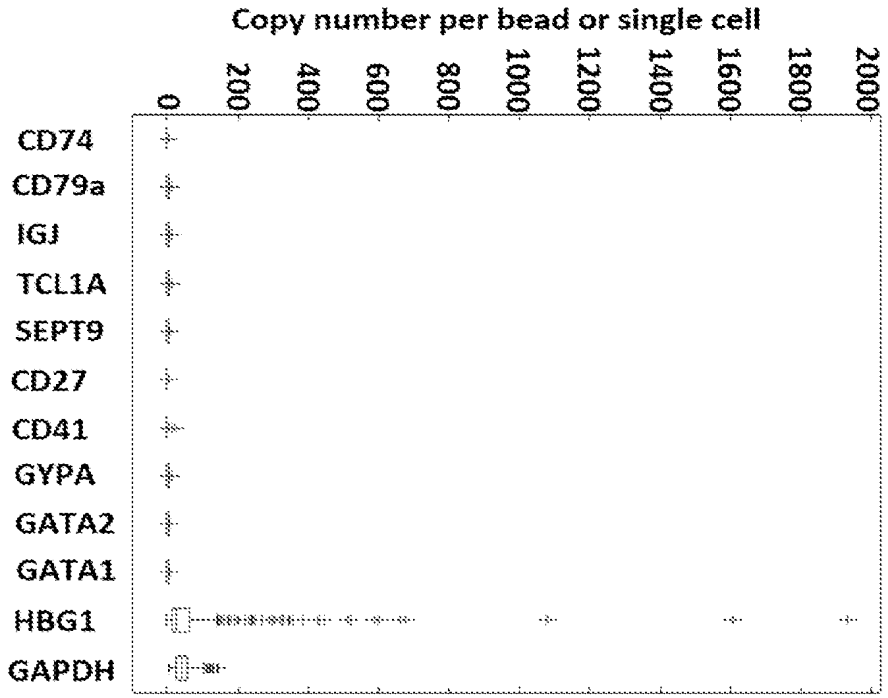


FIG. 12S

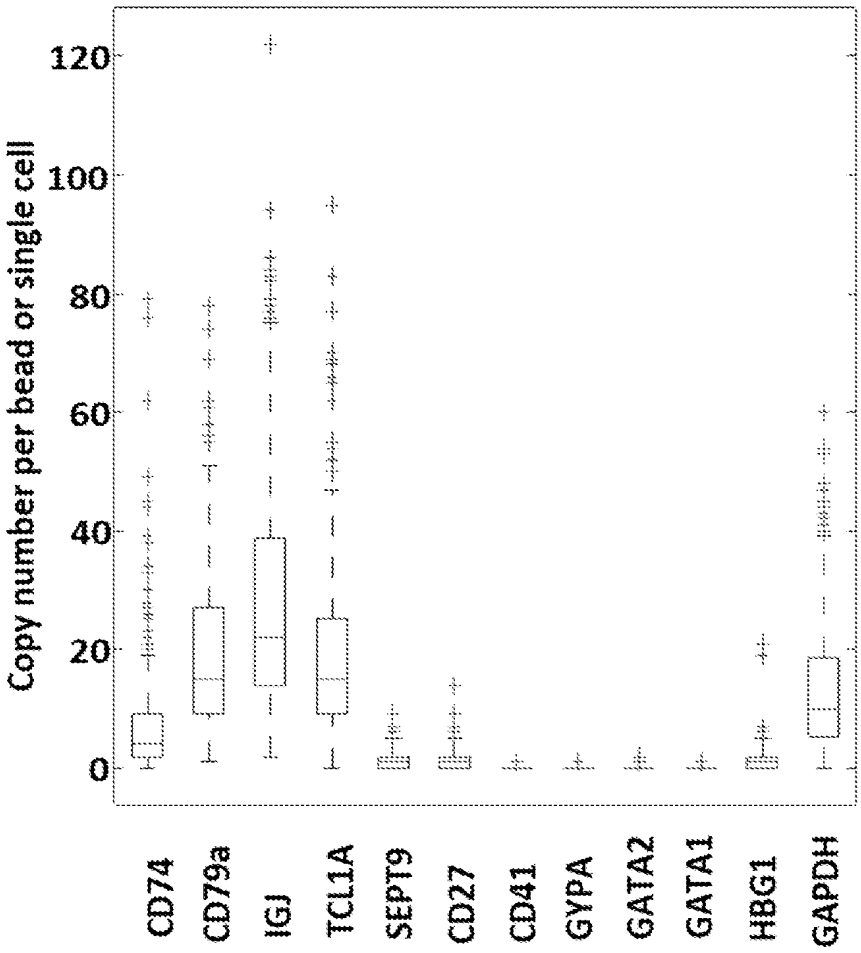

Copy number per bead or single cell
0
20
40
60
80
100
120
CD74
CD79a
IGJ
TCL1A
SEPT9
CD27
CD41
GYPA
GATA2
GATA1
HBG1
GAPDH

FIG. 12T

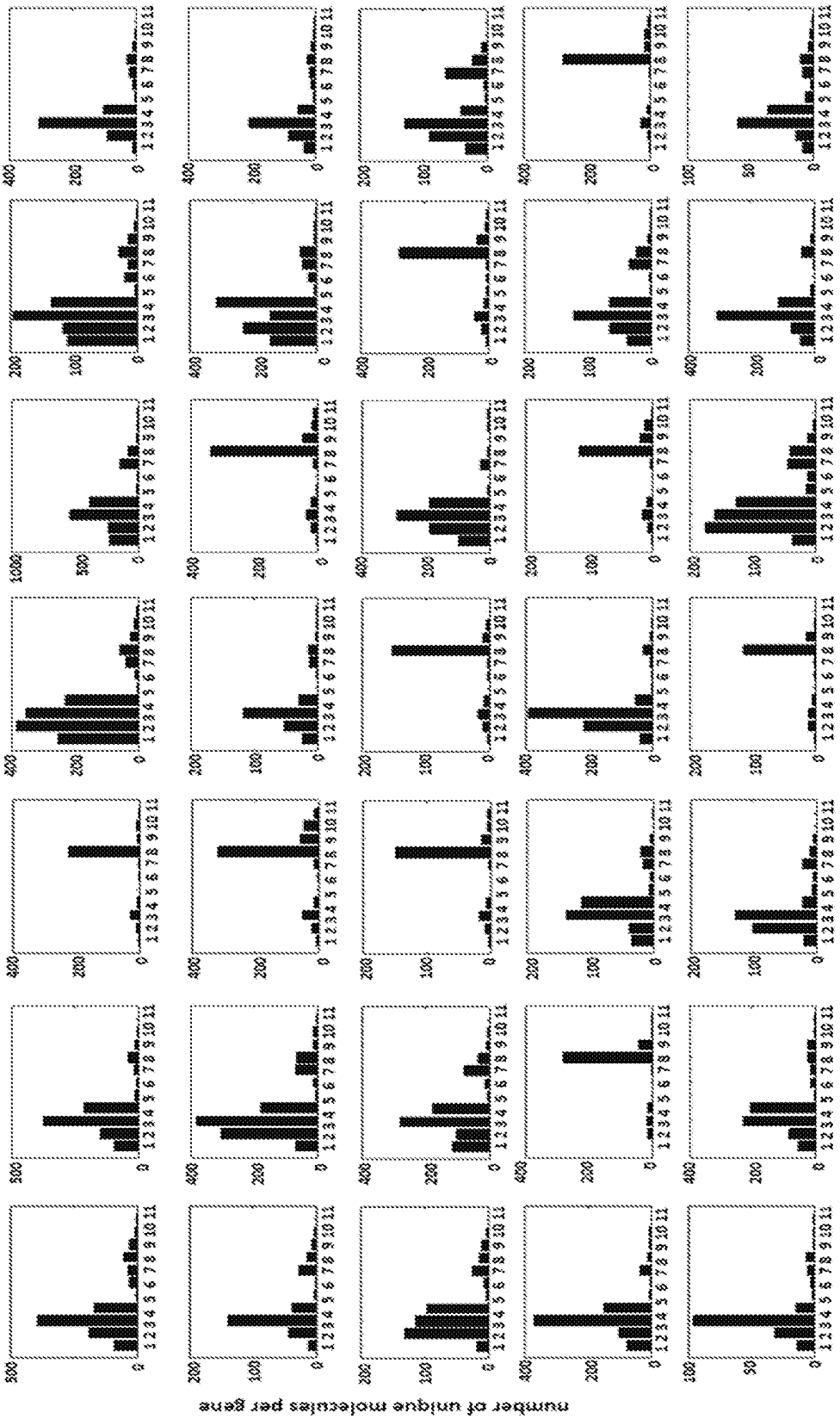


FIG. 13A

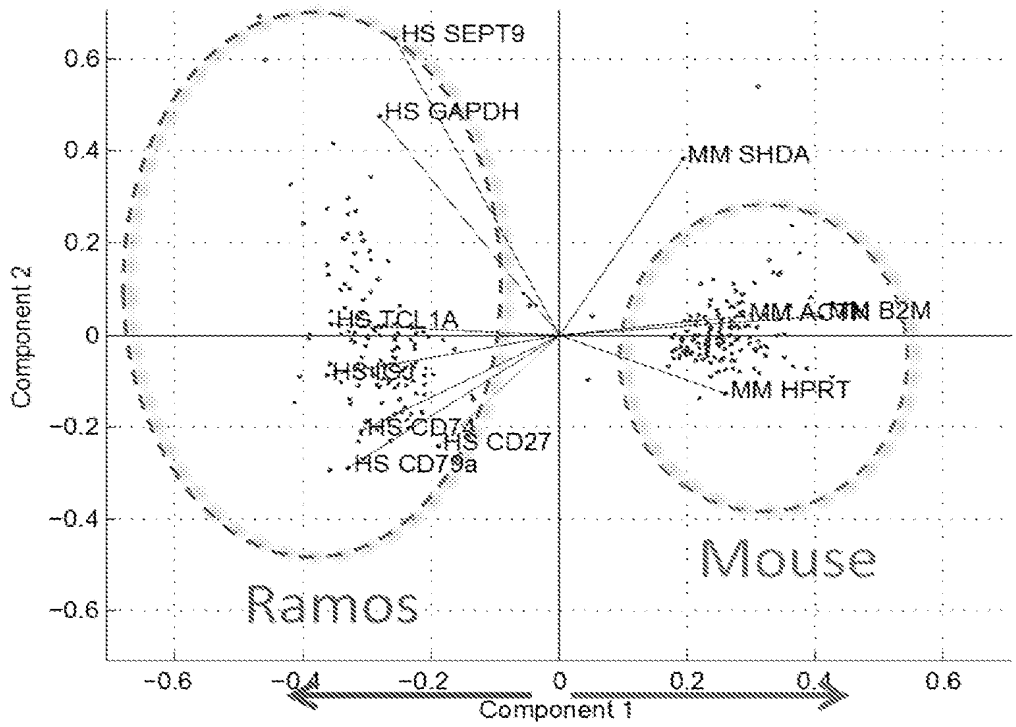


FIG. 13B

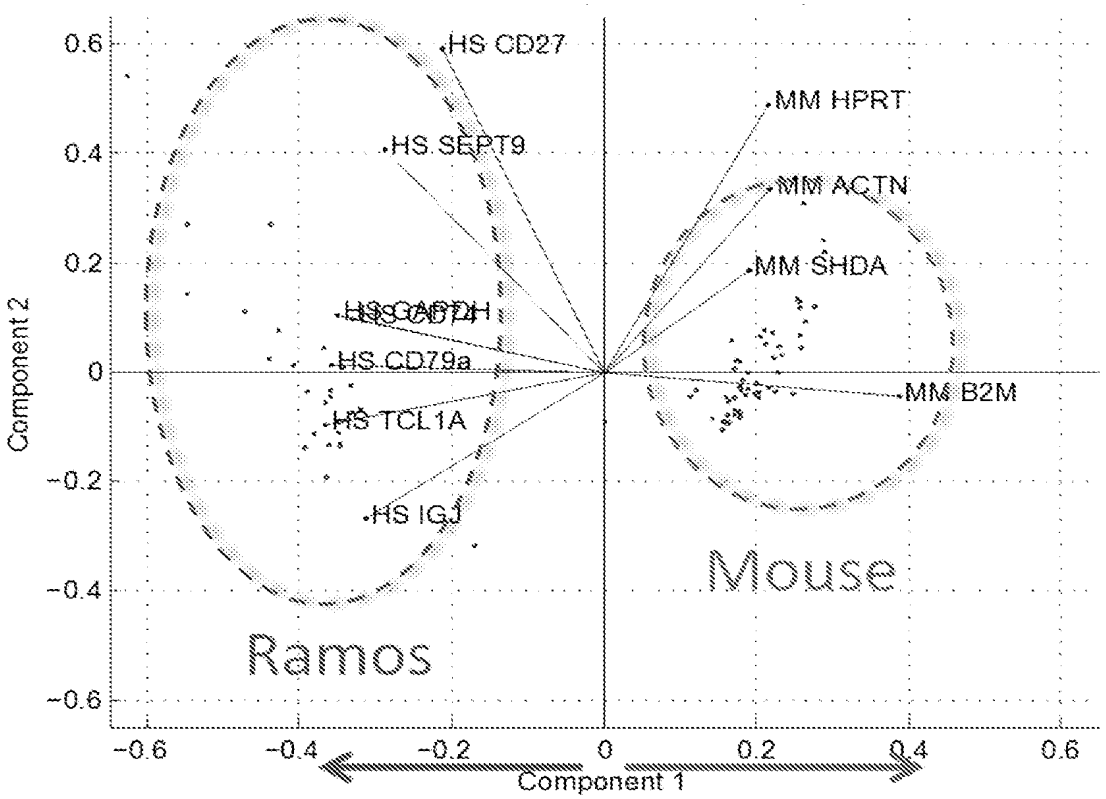


FIG. 13C

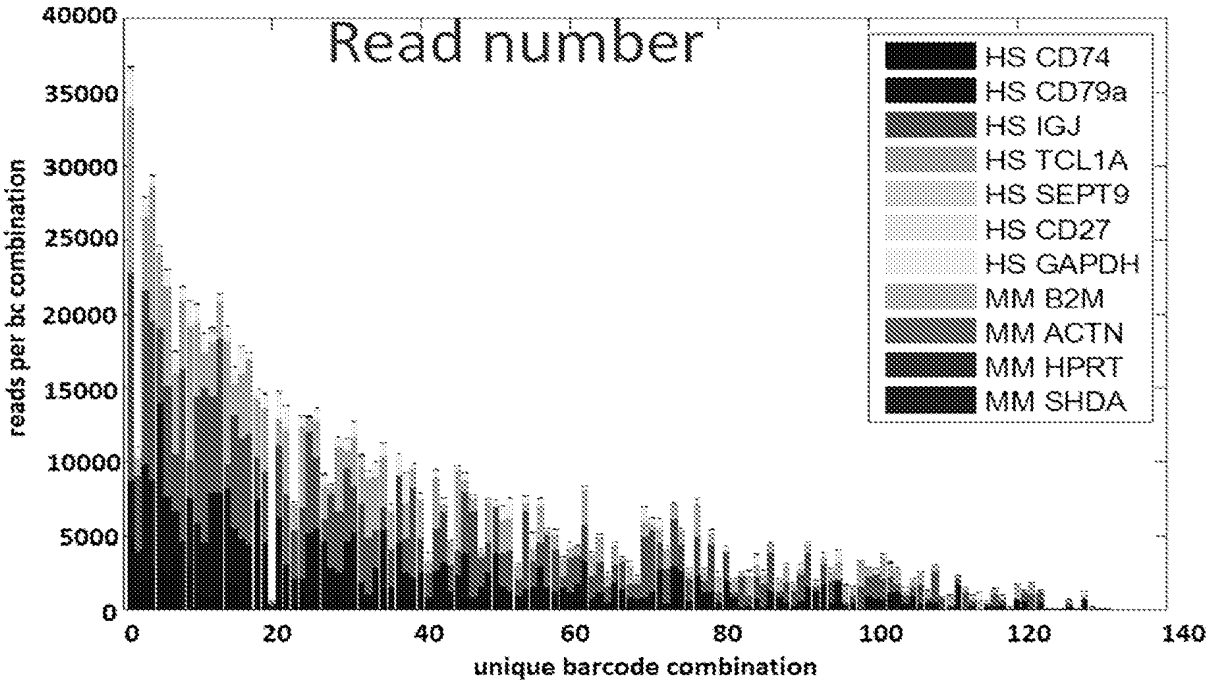


FIG. 13D

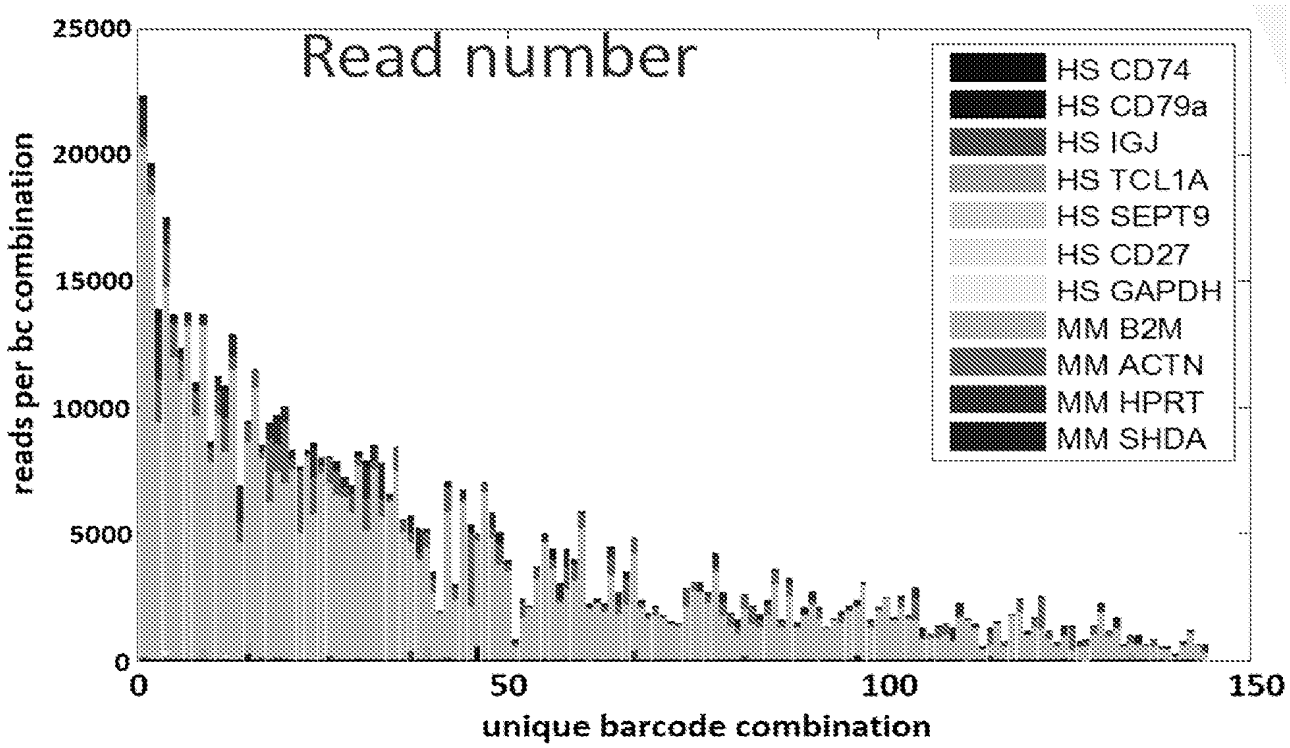


FIG. 13E

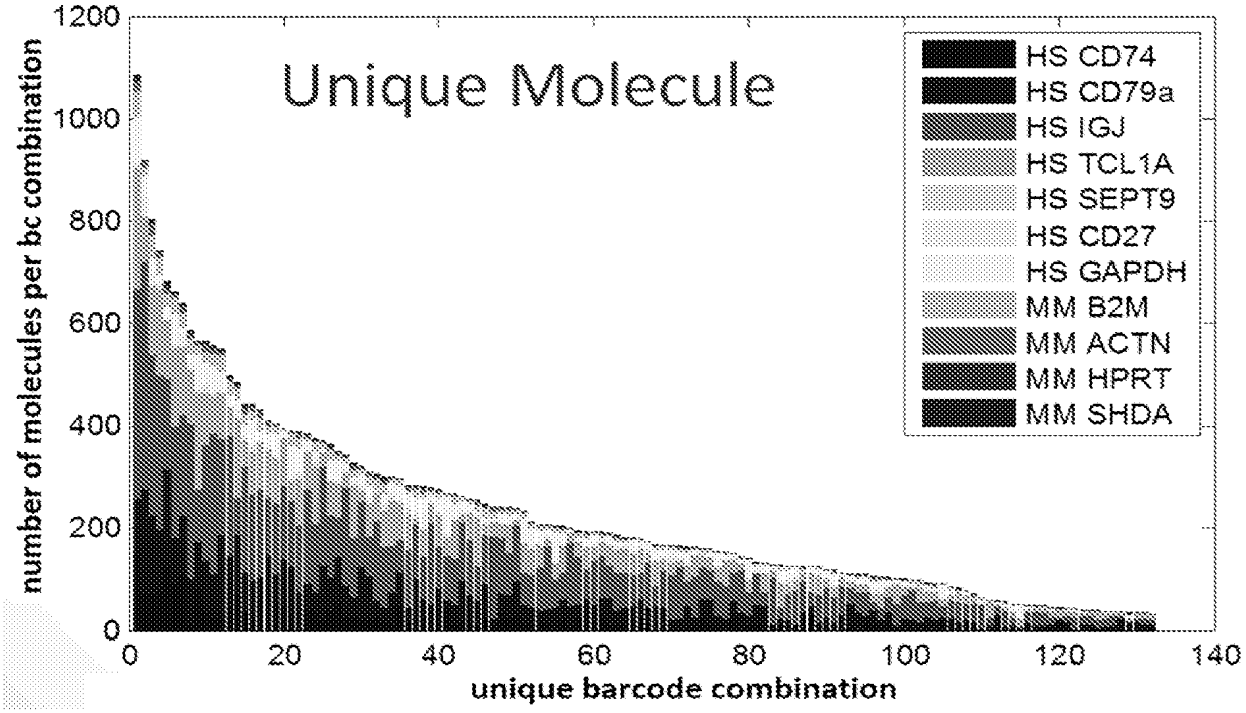


FIG. 13F

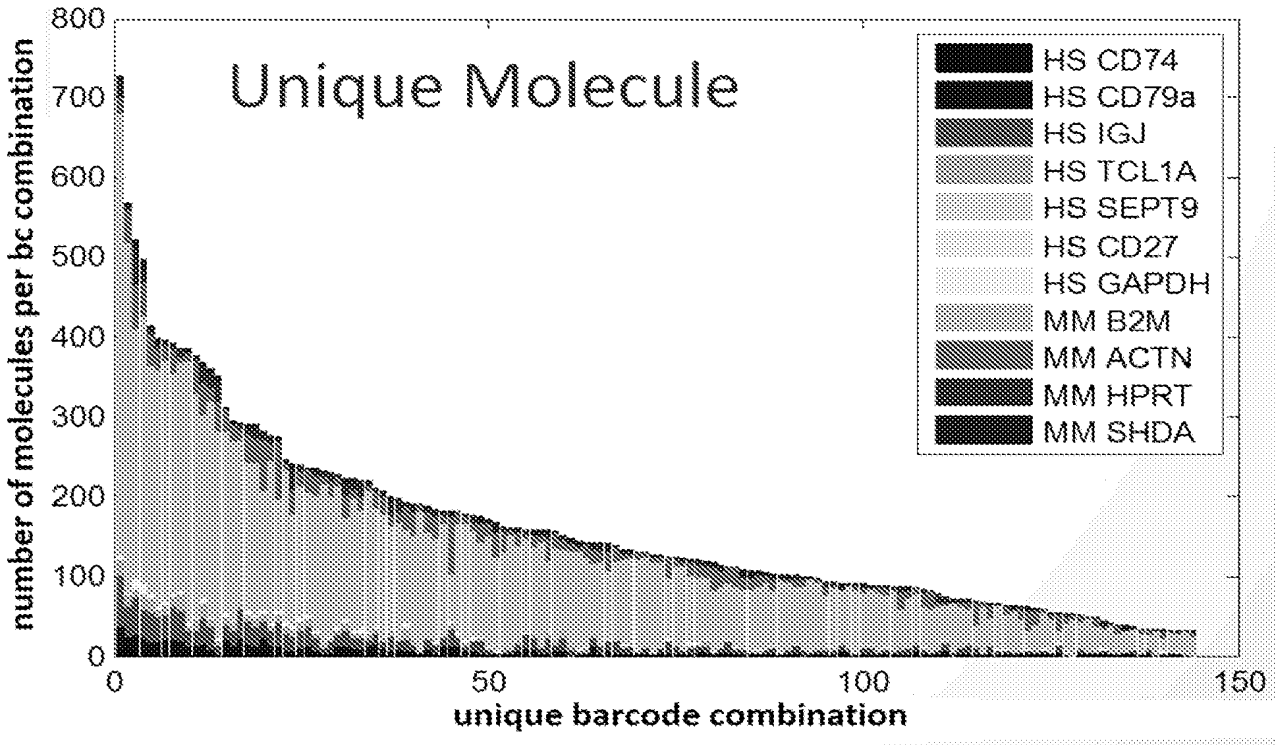


FIG. 13G

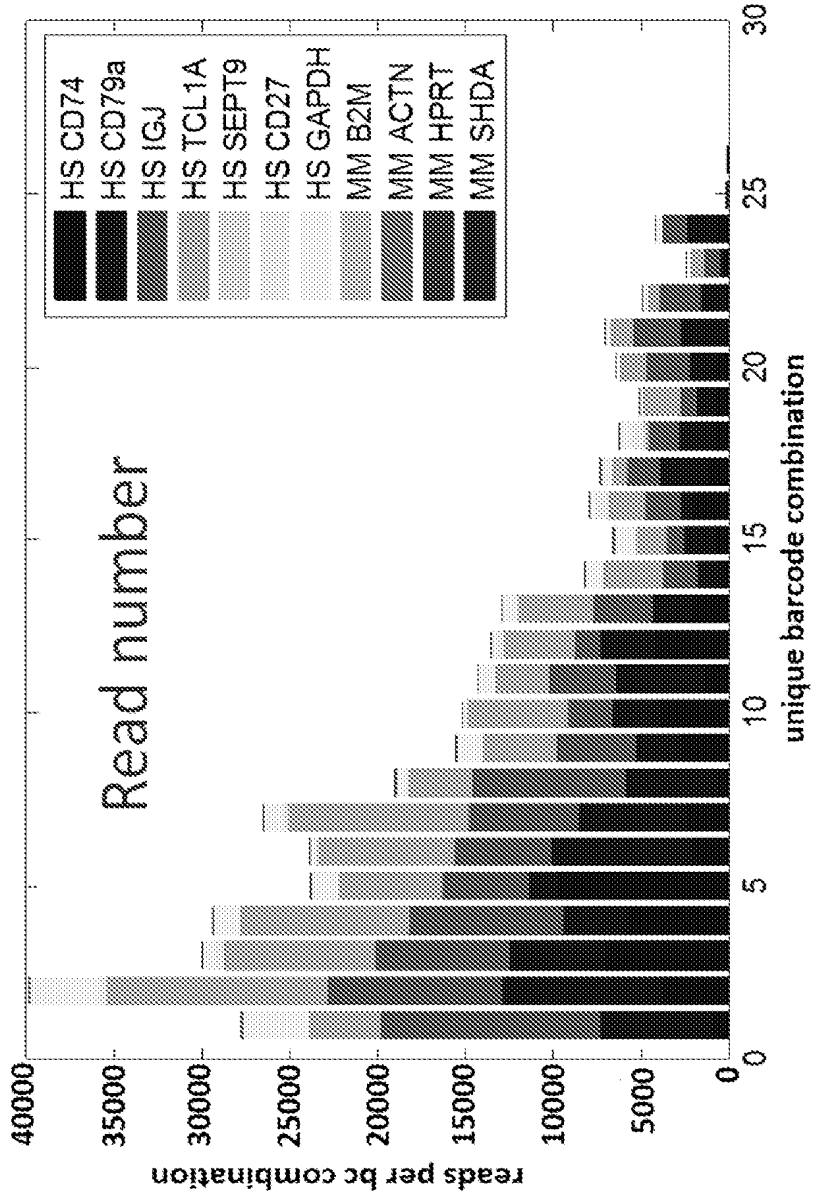

Read number
HS CD74
HS CD79a
HS IGJ
HS TCL1A
HS SEPT9
HS CD27
HS GAPDH
MM B2M
MM ACTN
MM HPRT
MM SHDA
reads per bc combination
unique barcode combination
0
5000
10000
15000
20000
25000
30000
35000
40000
0
5
10
15
20
25
30

FIG. 13H

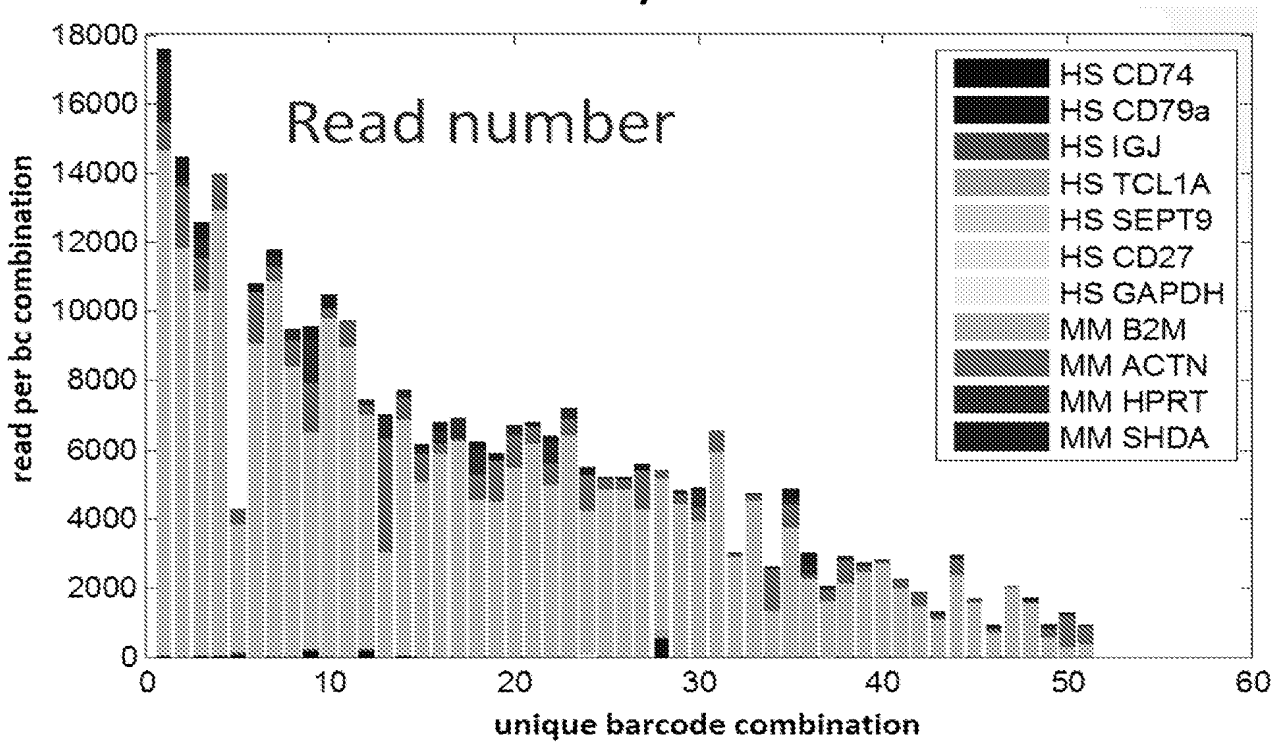


FIG. 13I

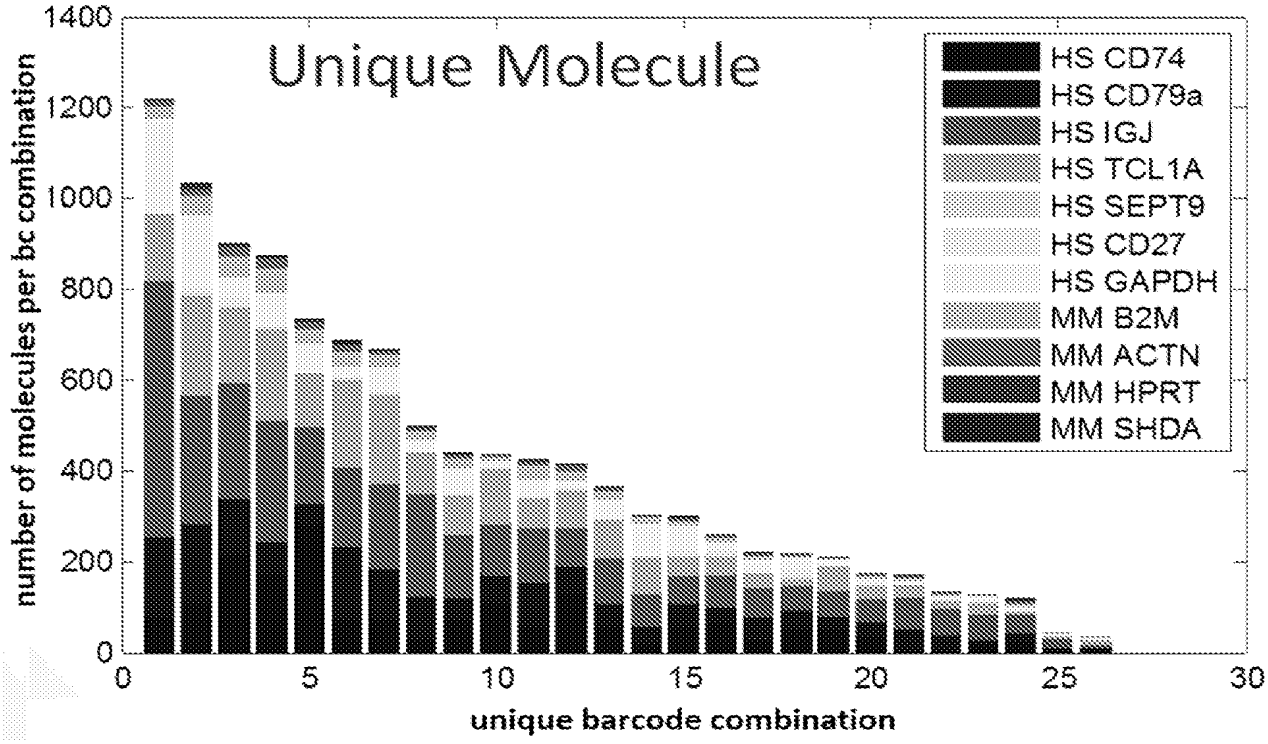


FIG. 13J

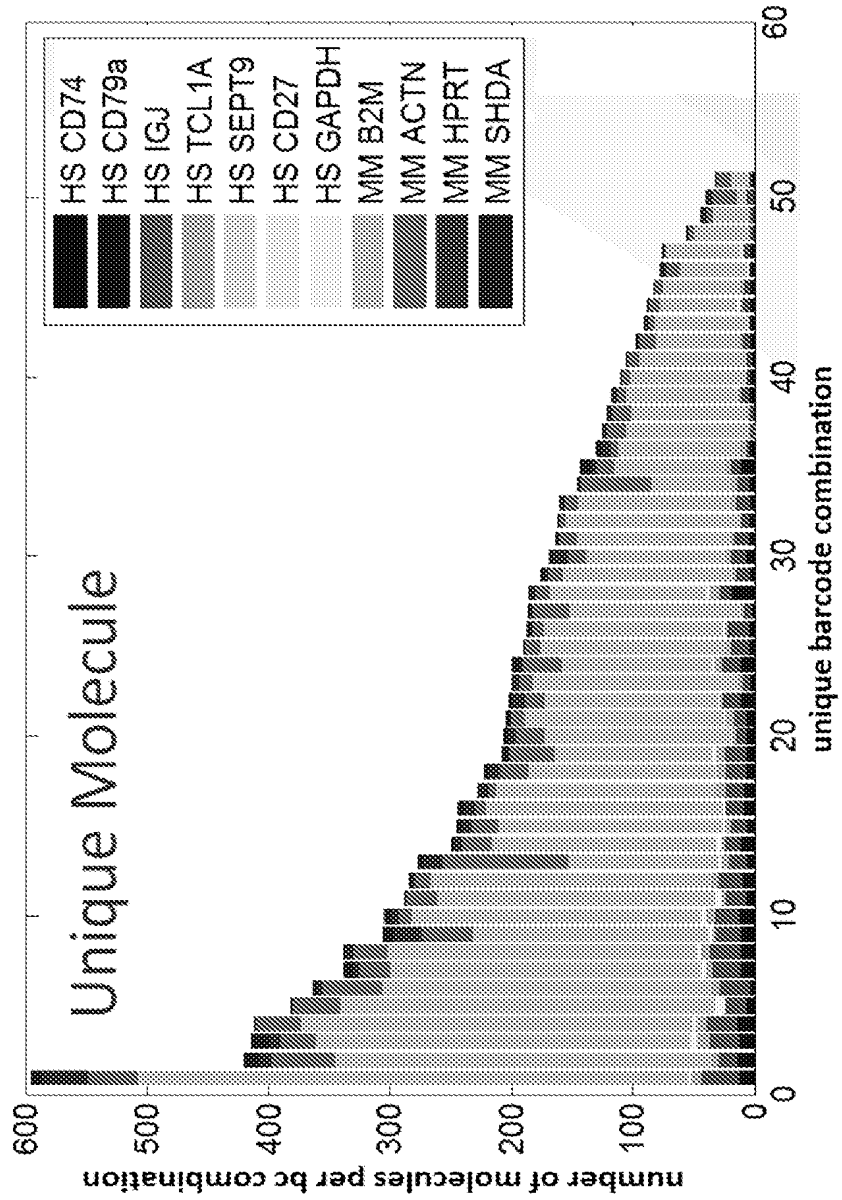

Unique Molecule
HS CD74
HS CD79a
HS IGJ
HS TCL1A
HS SEPT9
HS CD27
HS GAPDH
MM B2M
MM ACTN
MM HPRT
MM SHDA
number of molecules per bc combination
unique barcode combination
0
100
200
300
400
500
600
0
10
20
30
40
50
60

FIG. 13K

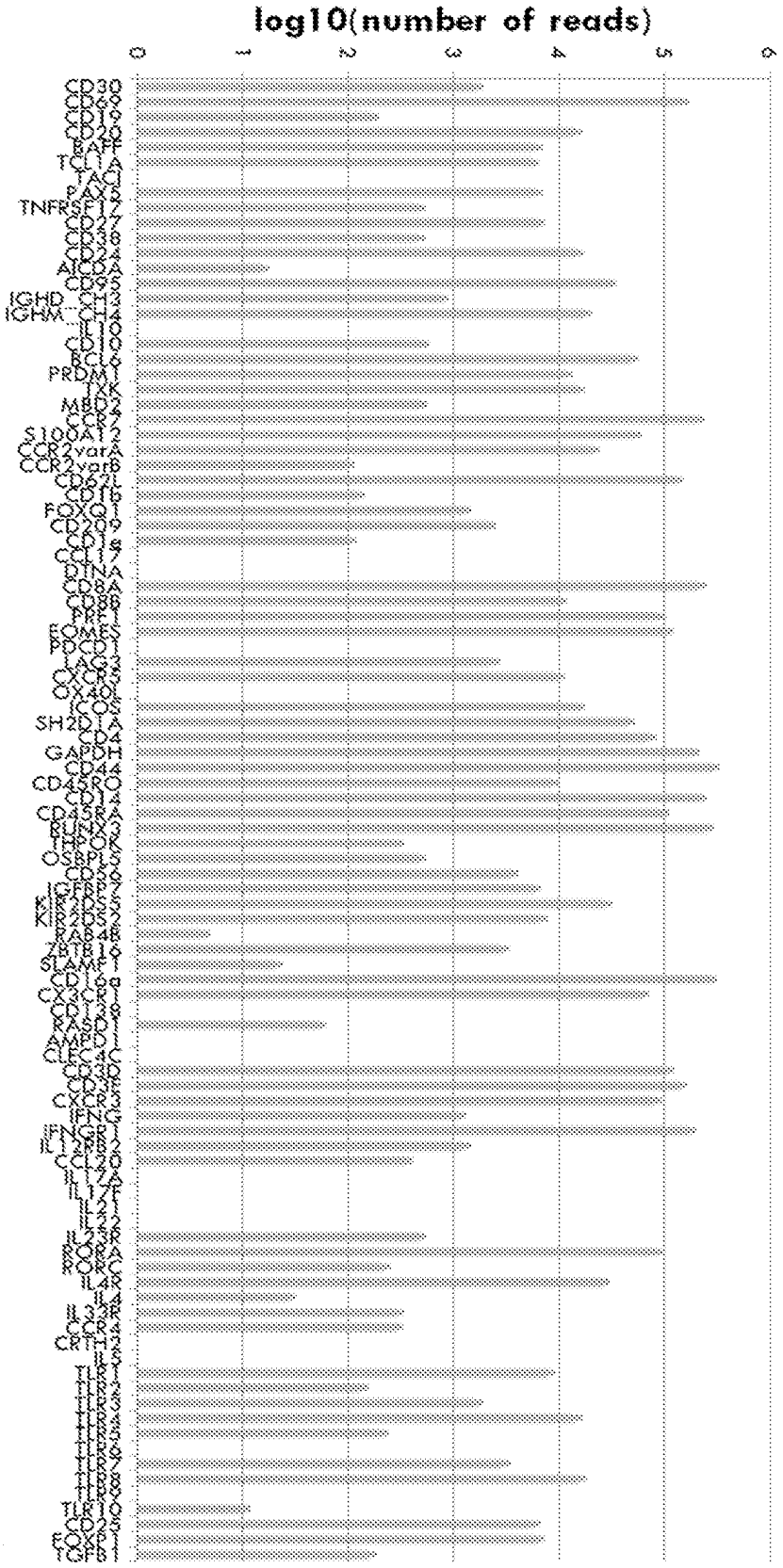

log10(number of reads)
0
1
2
3
4
5
6
CD30
CD69
CD19
CD20
BAFF
TCL1A
TACI
PAX5
TNFRSF17
CD27
CD38
CD24
AICDA
CD95
IGHD_CH3
IGHM_CH1
IL10
CD10
BCL6
PRDM1
IXK
MBD2
CCR7
S100A12
CCR2varA
CCR2varB
CD62L
CD16
FOXQ1
CD209
CD68
CCL17
DNA
CD8A
CD8B
PRF1
EOMES
PDCD1
LAG3
CXCR5
OX40L
ICOS
SH2D1A
CD4
GAPDH
CD44
CD45RO
CD14
CD45RA
RUNX3
THPOK
OS8PL5
CD56
IGFBP2
KIR2DS5
KIR2DS2
RAB48
ZBTB16
SLAMF1
CD16a
CX3CR1
CD138
EAS5D1
AMPD1
CLEC4C
CD303
CD304
CXCR3
IFNG
IFNGR1
IL12RB2
CCL4/3/0
IL17A
IL17F
IL22
IL26
IL23R
RORA
RORC
IL4
IL4R
IL13
CCR4
CRTH2
IL5
TLR1
TLR2
TLR3
TLR4
TLR5
TLR6
TLR7
TLR8
TLR9
TLR10
CD207
FOXP3
TCF7

FIG. 14

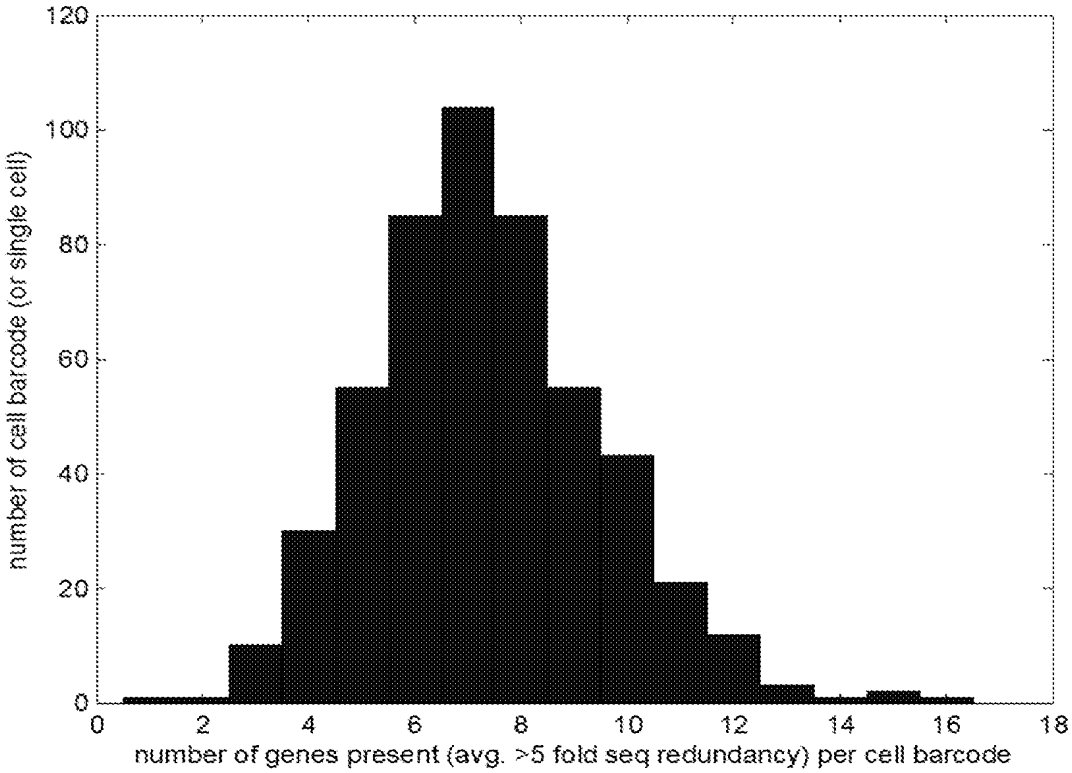


FIG. 15A

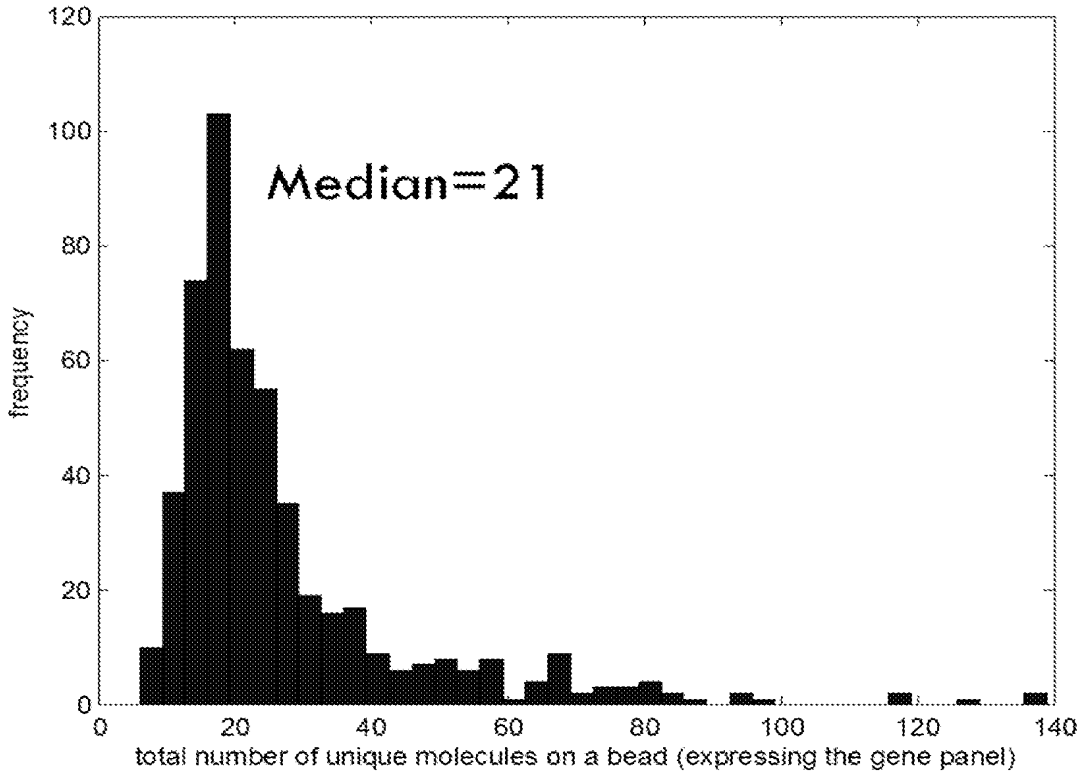


FIG. 15B

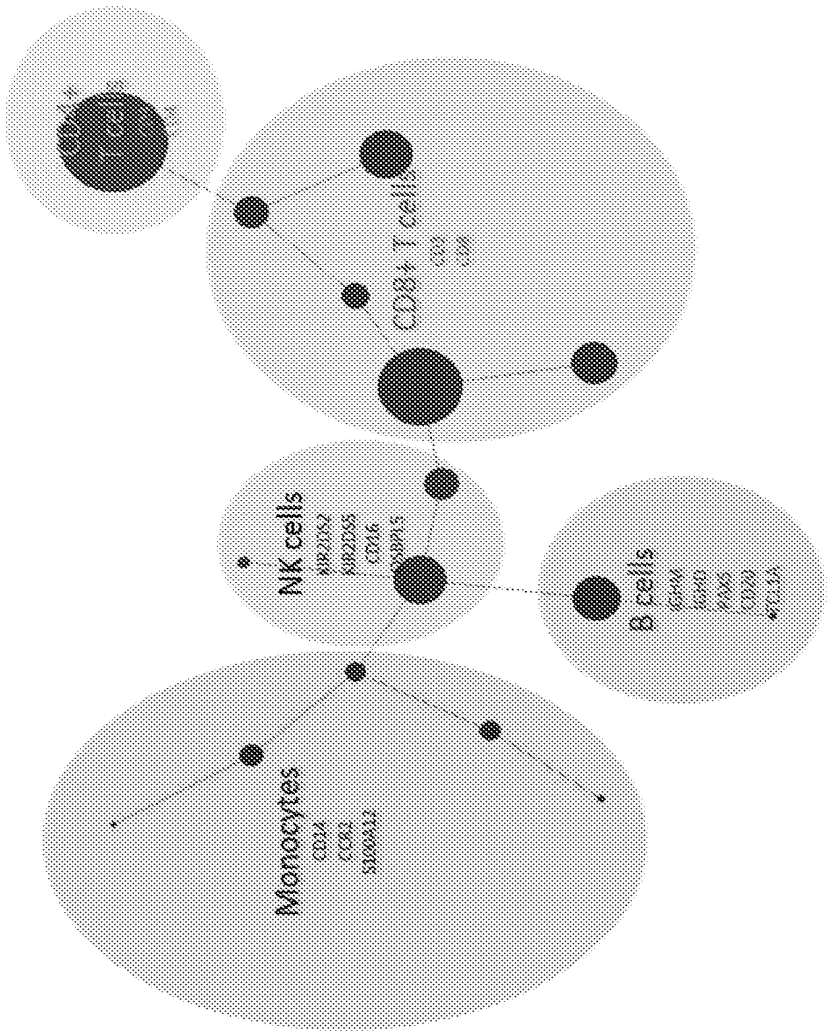

CD8+ T cells
CD3
CD8
NK cells
KIR2DL2
KIR2DS5
CD16
B cells
IgM
IgD
Monocytes
CD14
CCR2
SIGNR3

FIG. 16

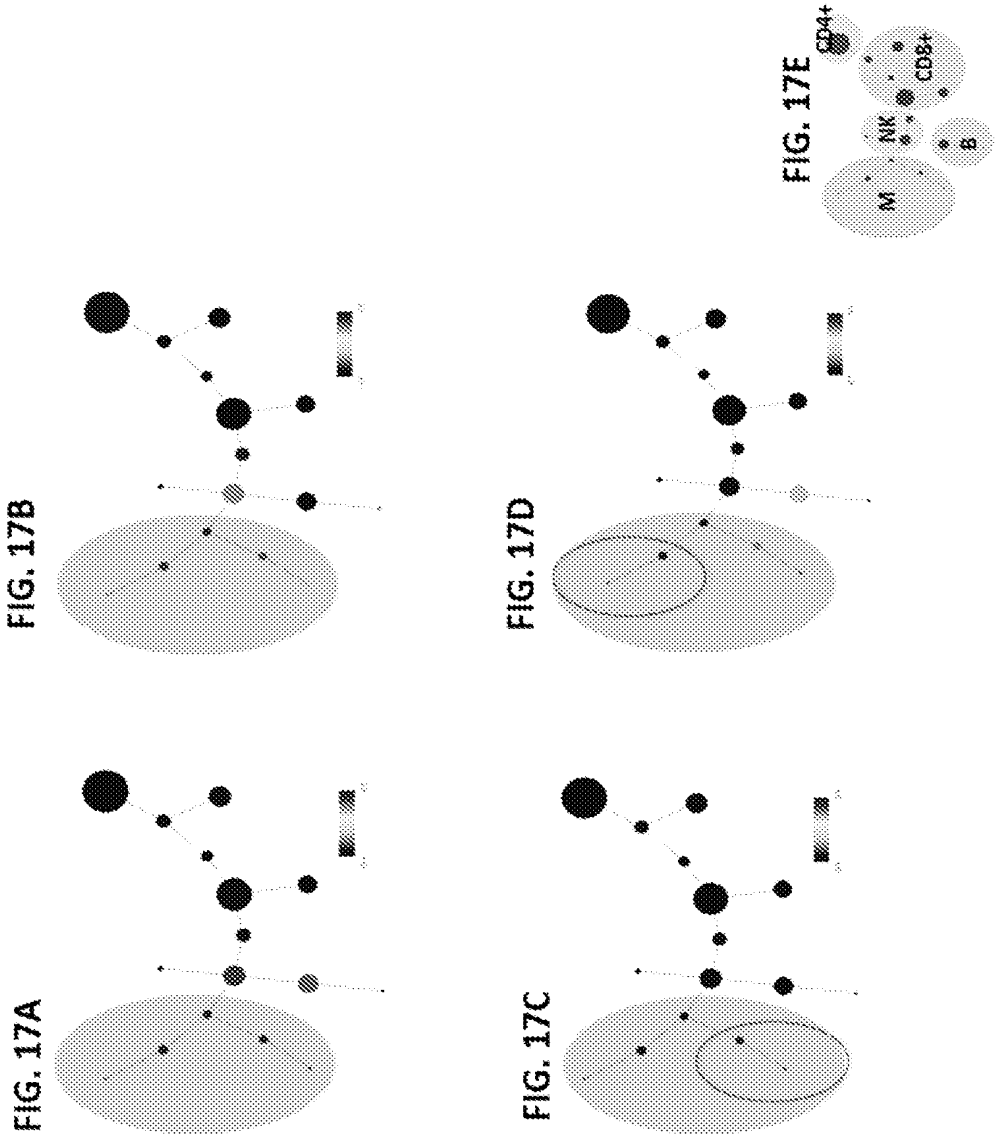


FIG. 17A-E

FIG. 18A

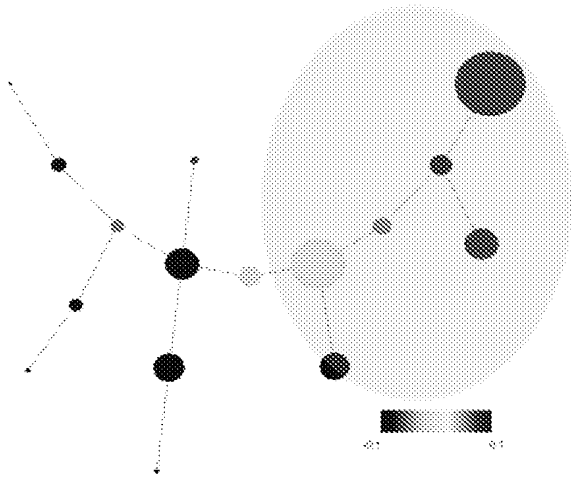


FIG. 18B

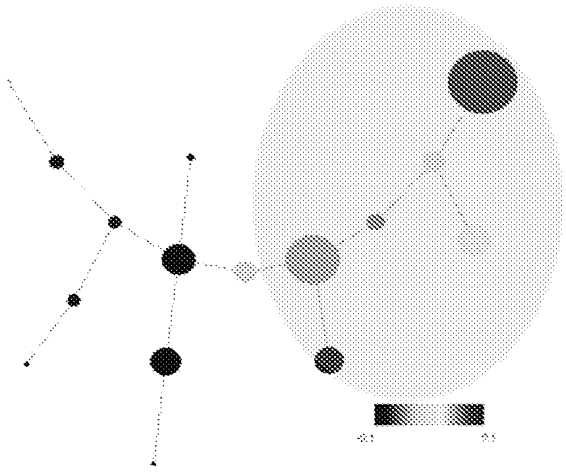


FIG. 18C

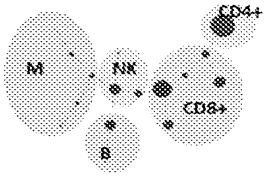


FIG. 18A-C

FIG. 19A

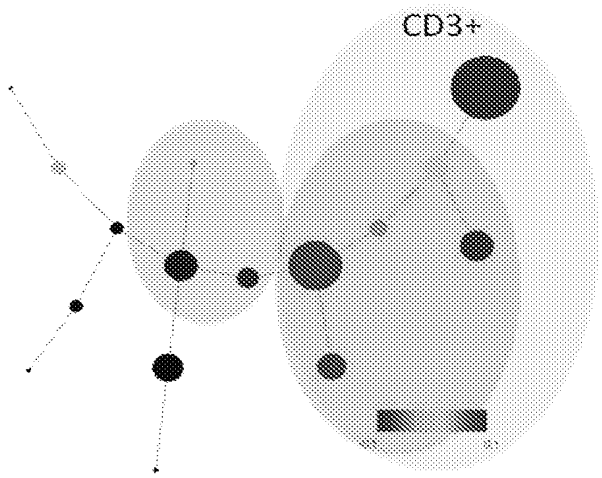


FIG. 19B

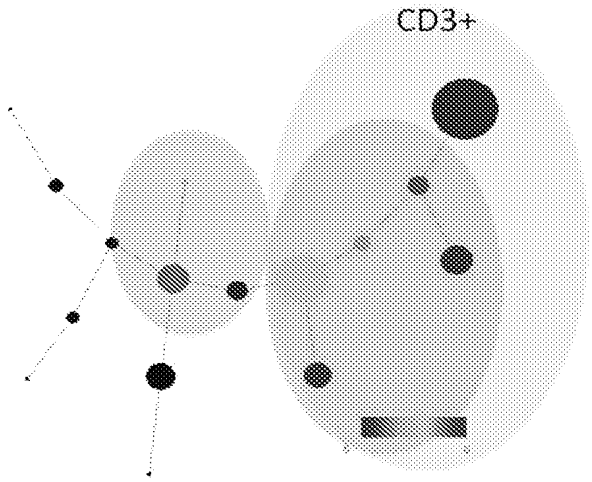


FIG. 19C

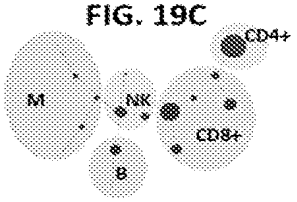


FIG. 19A-C

FIG. 20A

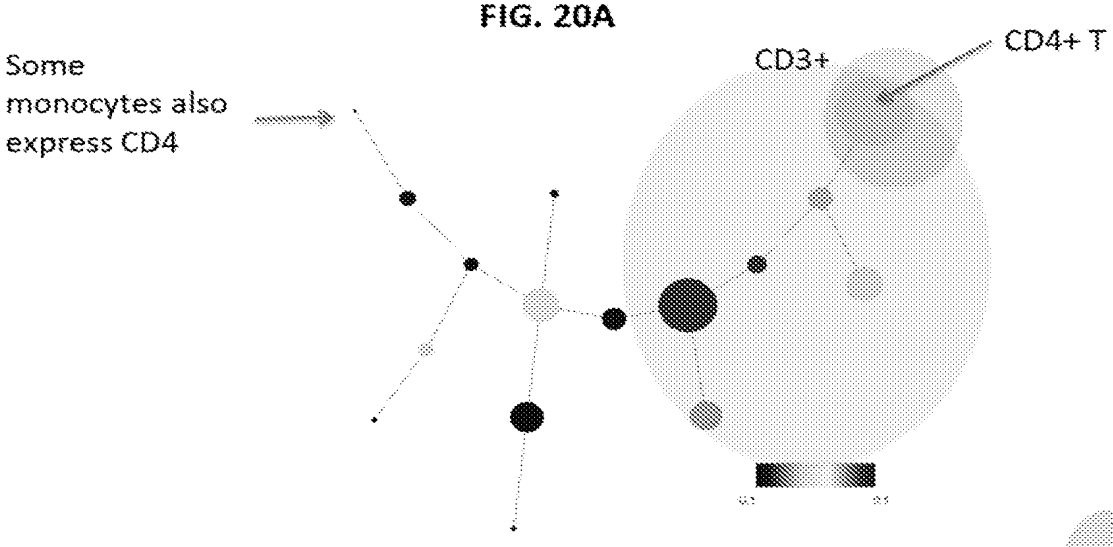


FIG. 20B

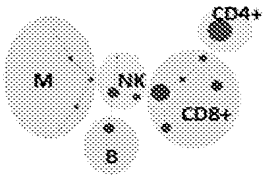


FIG. 20A-B

FIG. 21A

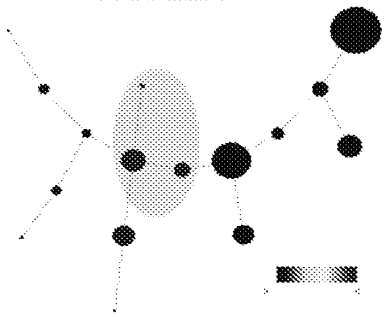


FIG. 21B

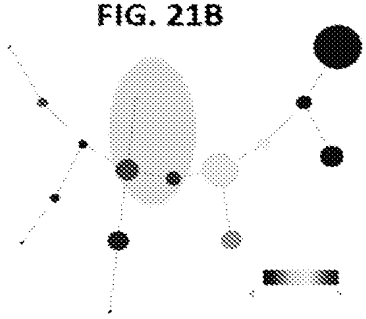


FIG. 21C

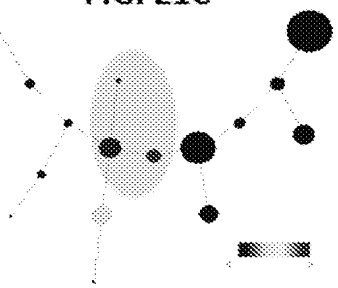


FIG. 21D

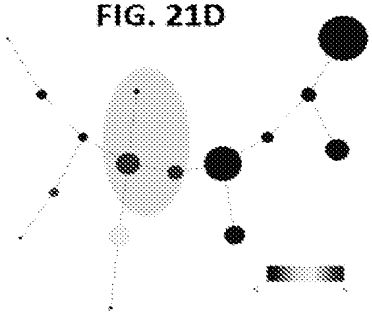


FIG. 21E

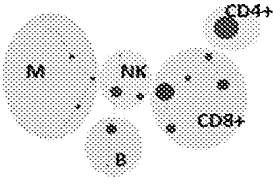


FIG. 21A-E

FIG. 22A

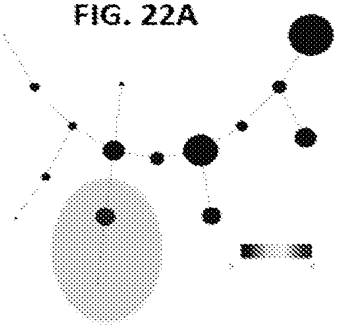


FIG. 22B

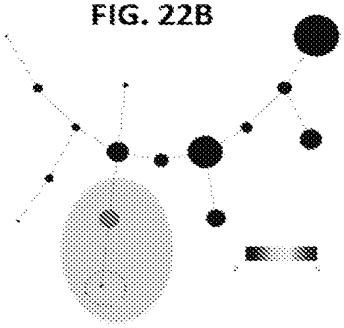


FIG. 22C

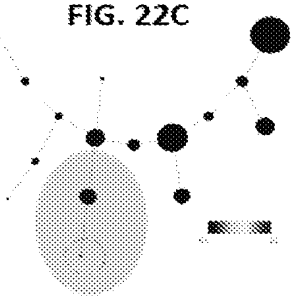


FIG. 22D

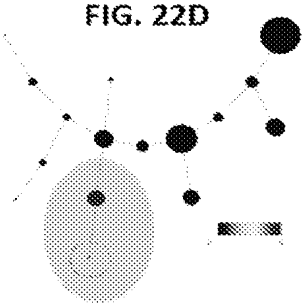


FIG. 22E

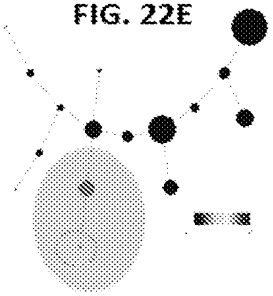


FIG. 22F

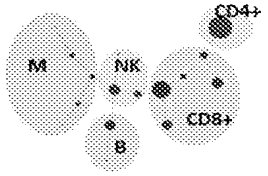


FIG. 22A-F

FIG. 23A

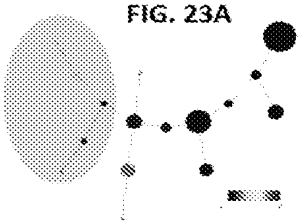


FIG. 23B

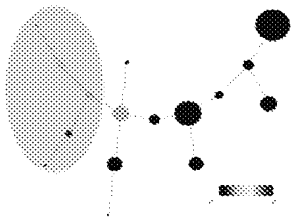


FIG. 23C

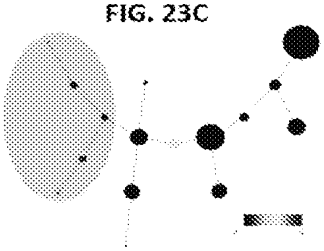


FIG. 23D

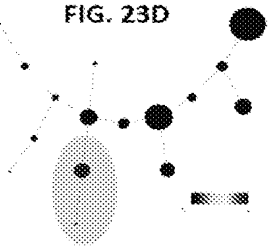


FIG. 23E

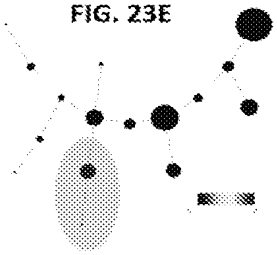


FIG. 23F

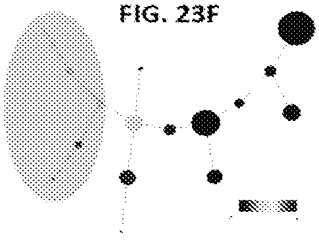


FIG. 23G

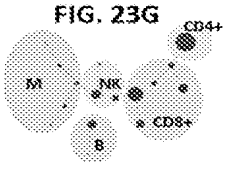


FIG. 23A-G



FIG. 24

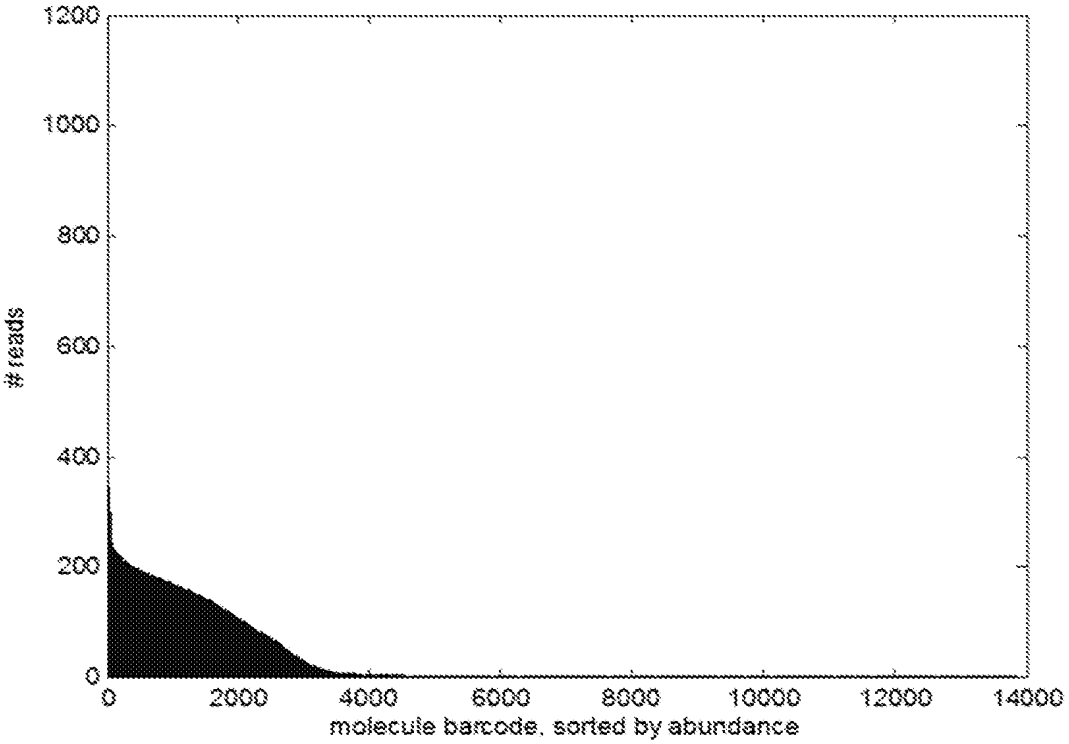


FIG. 25A

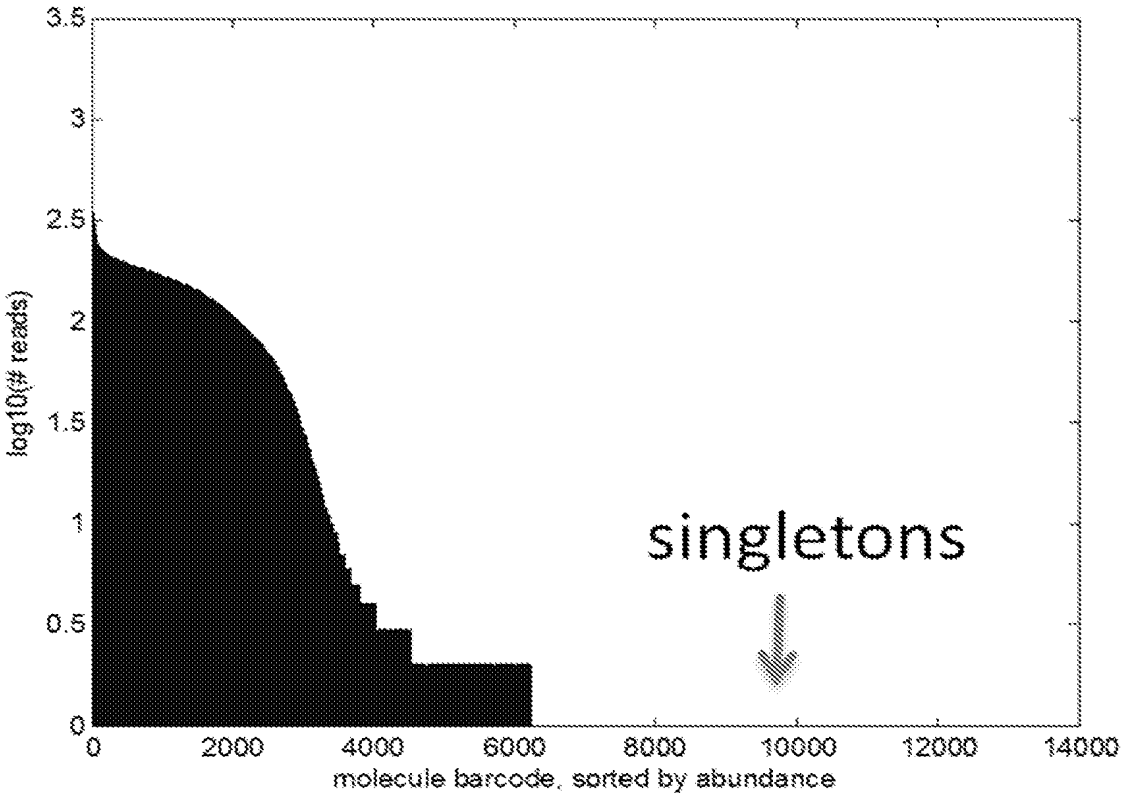


FIG. 25B

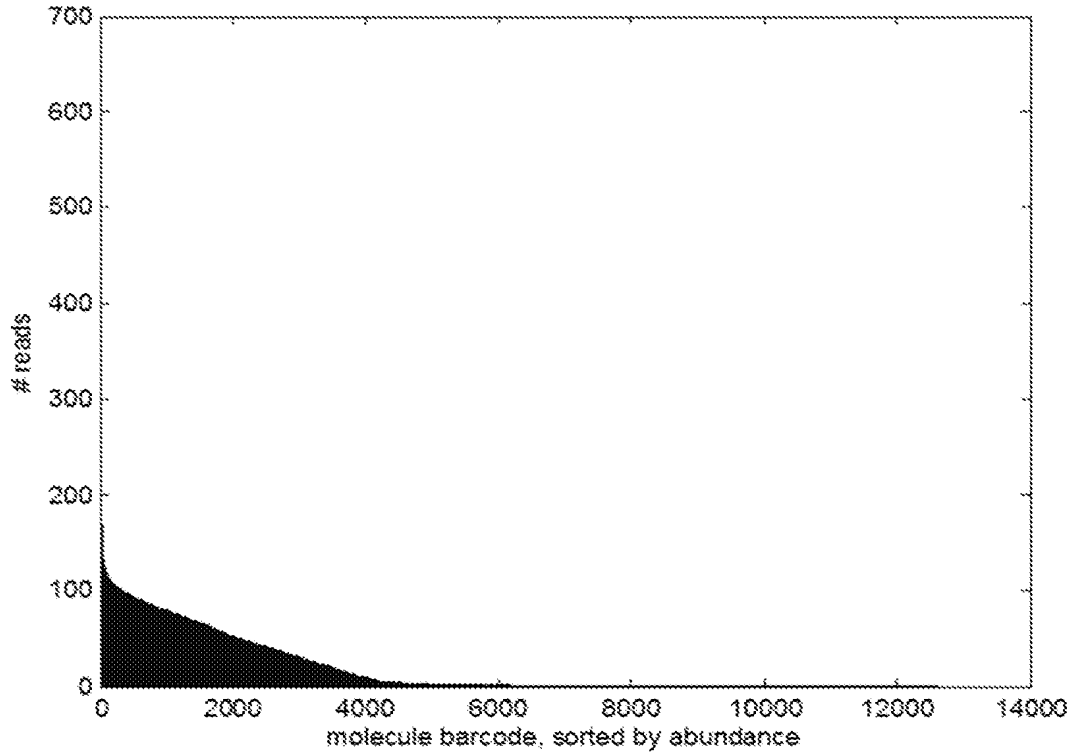


FIG. 25C

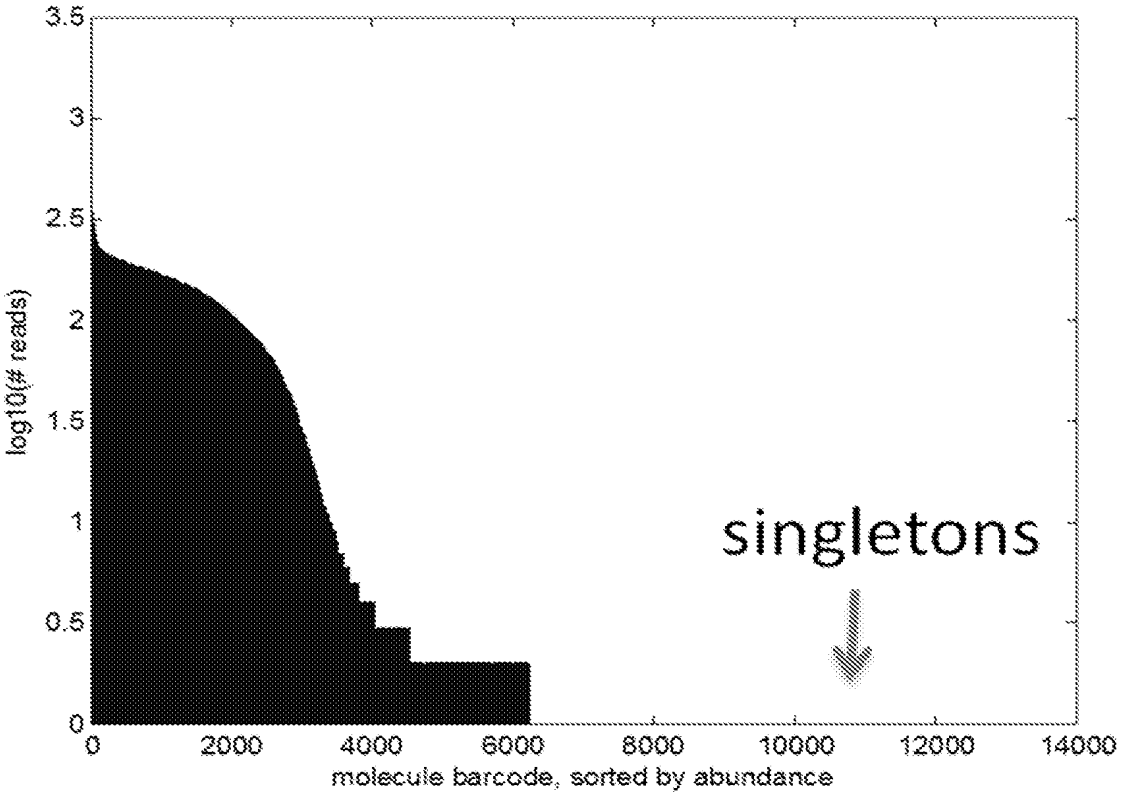


FIG. 25D

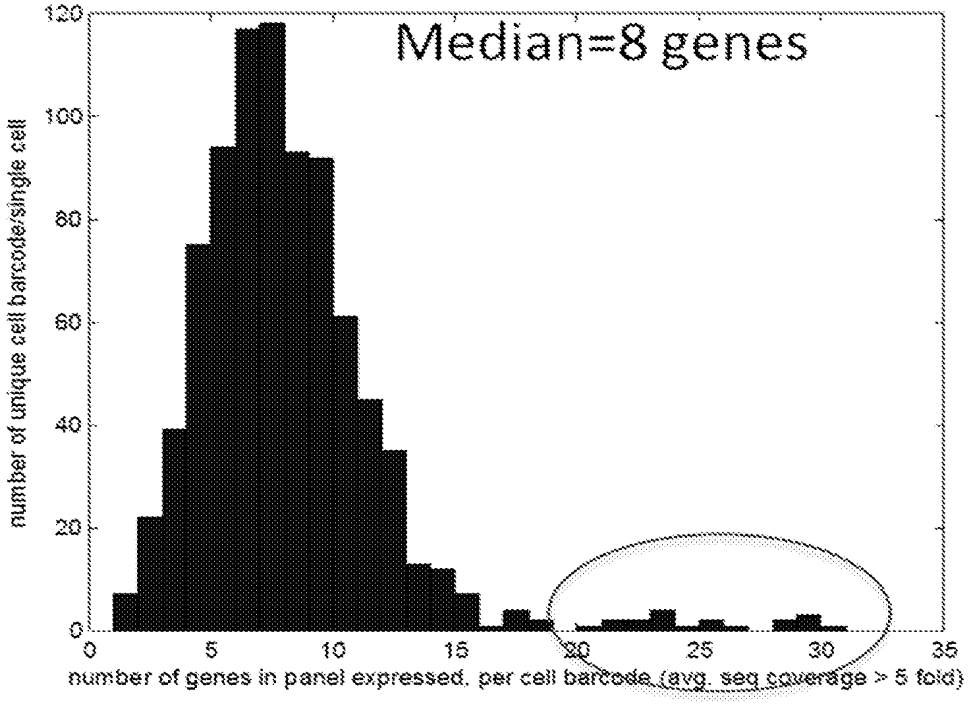


FIG. 26A

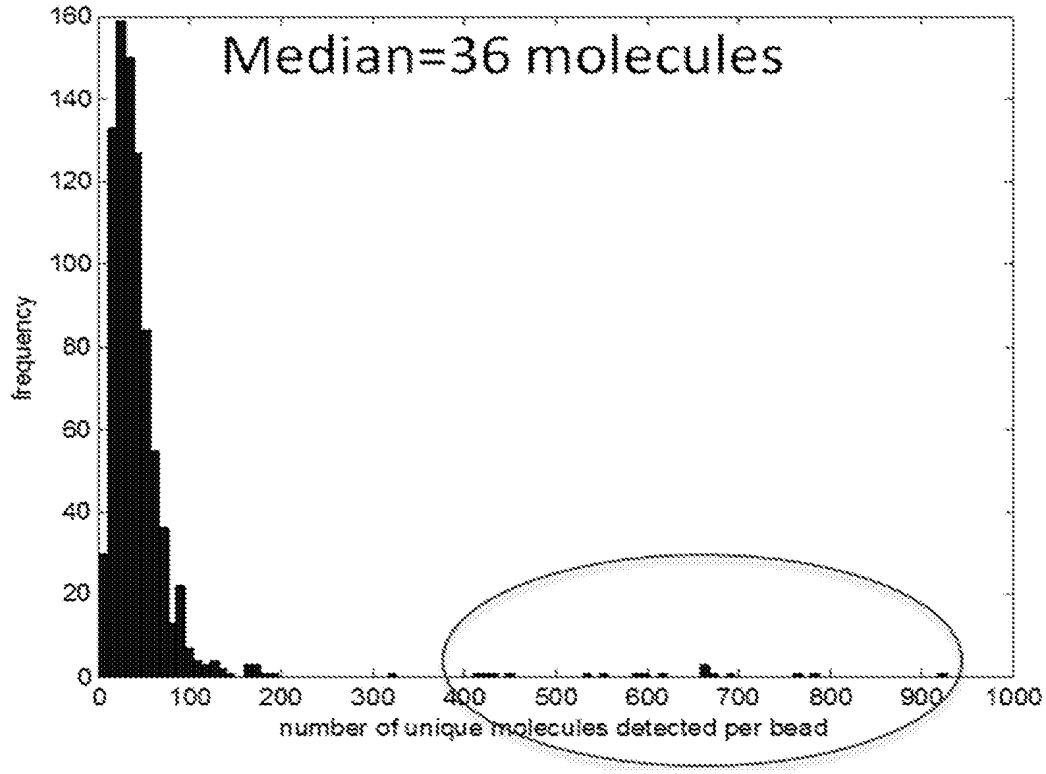


FIG. 26B

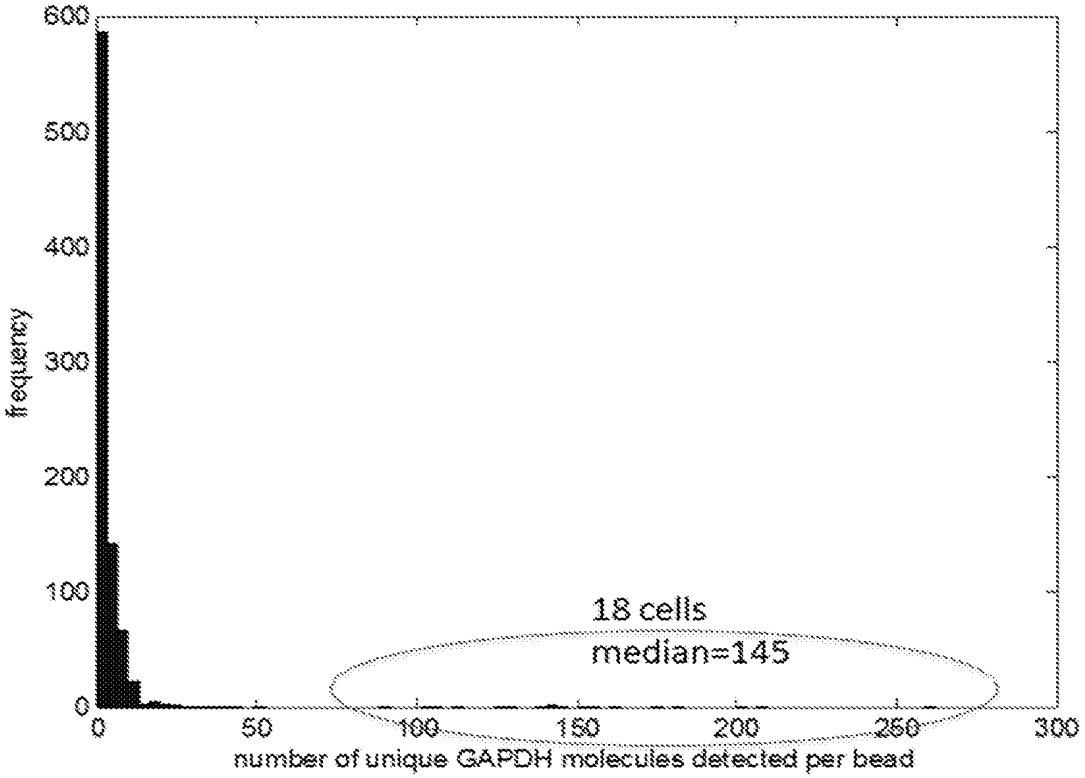


FIG. 26C

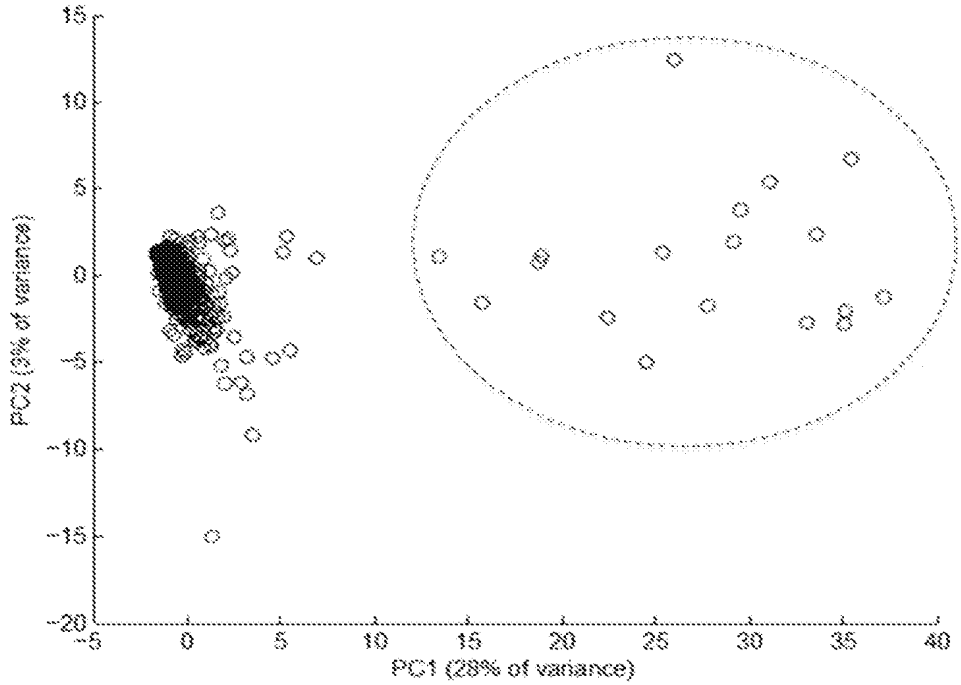


FIG. 27

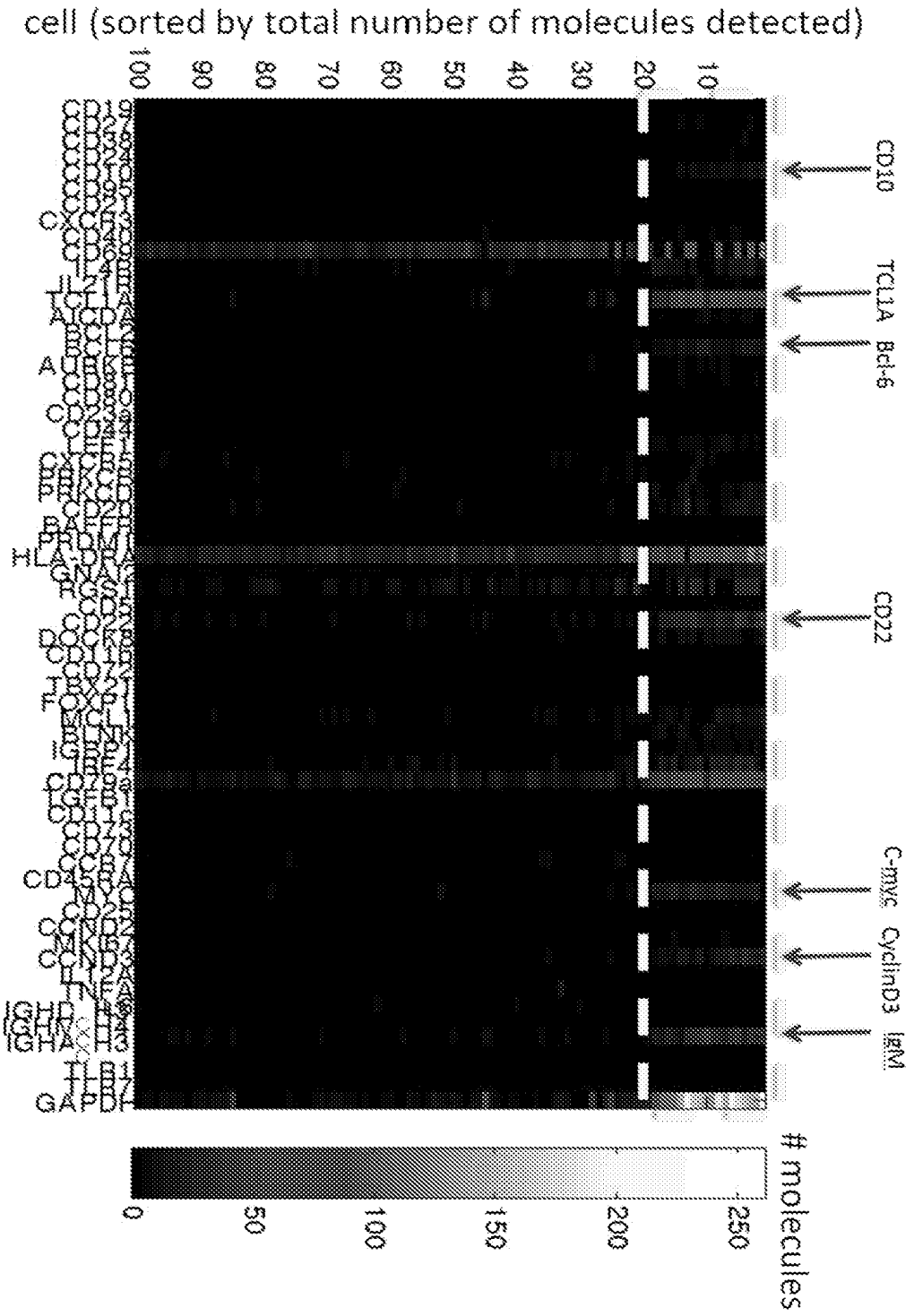

cell (sorted by total number of molecules detected)
100
90
80
70
60
50
40
30
20
10
CD10
TCL1A
Bcl-6
CD22
C-myc
CyclinD3
IgM
0
50
100
150
200
250
molecules

FIG. 28

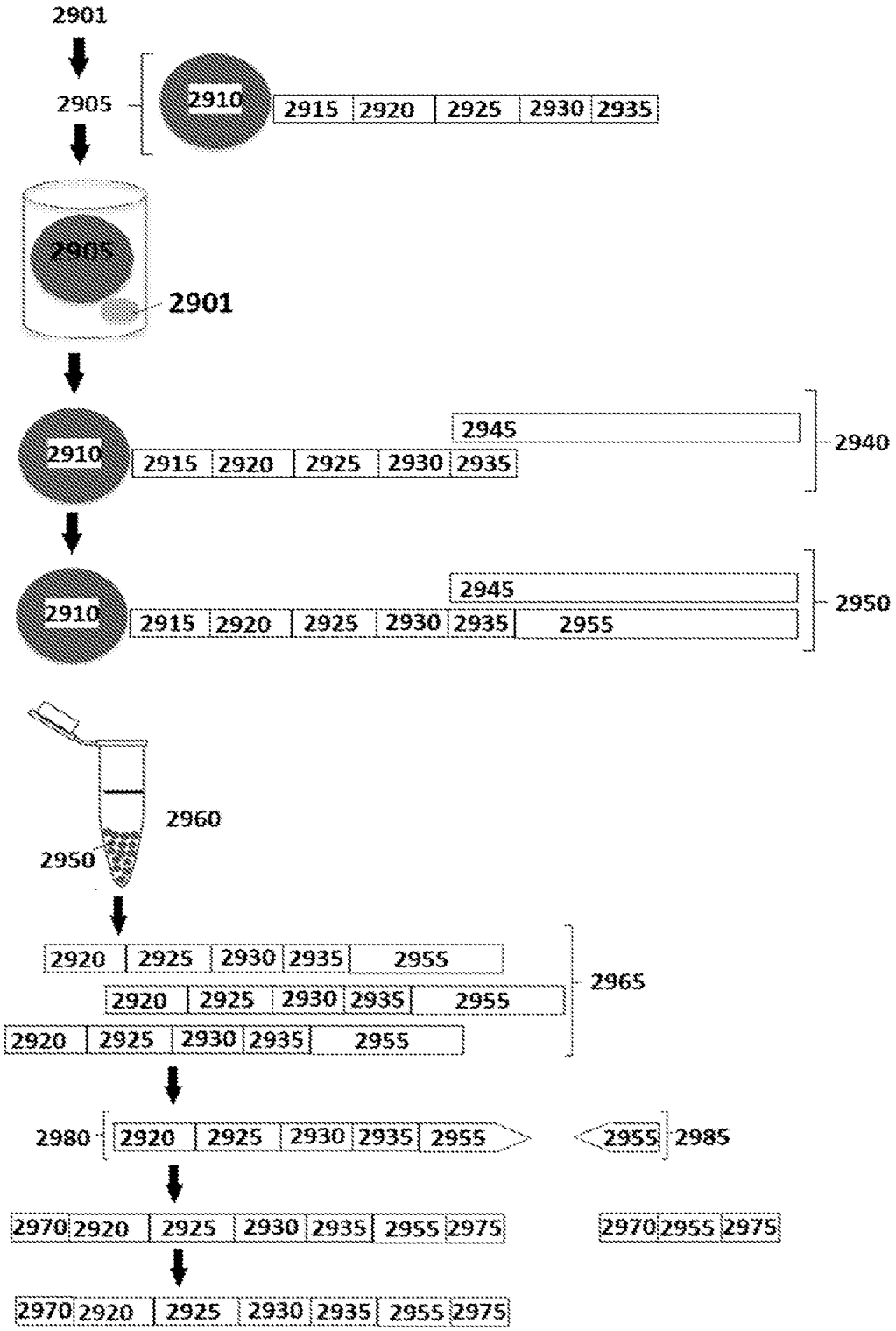

2901
2905
2910
2915 2920 2925 2930 2935
2905
2901
2910
2945
2915 2920 2925 2930 2935
2940
2910
2945
2915 2920 2925 2930 2935 2955
2950
2960
2950
2920 2925 2930 2935 2955
2920 2925 2930 2935 2955
2920 2925 2930 2935 2955
2965
2980
2920 2925 2930 2935 2955
2955
2985
2970 2920 2925 2930 2935 2955 2975
2970 2955 2975
2970 2920 2925 2930 2935 2955 2975

FIG. 29

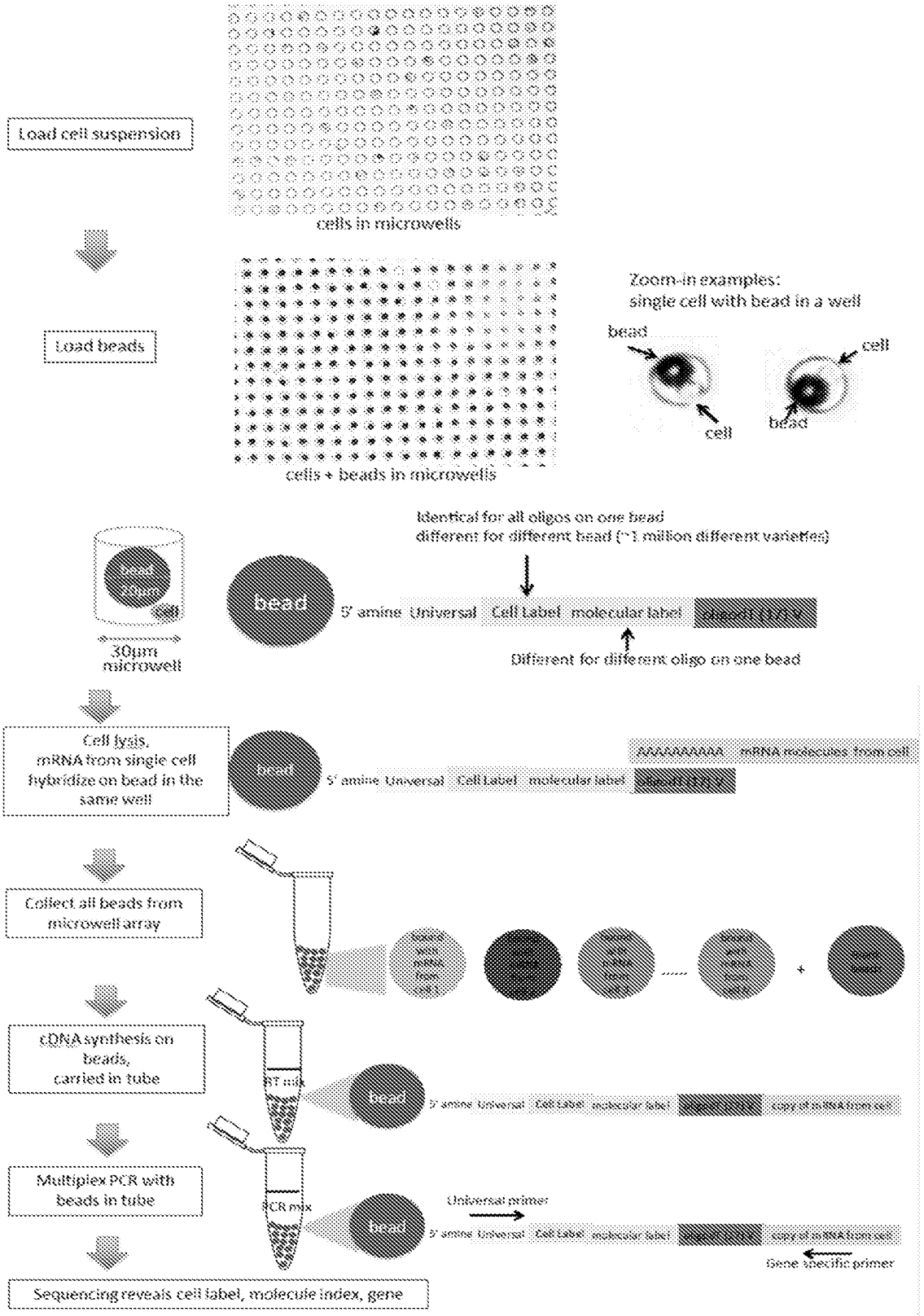

Load cell suspension
cells in microwells
Load beads
cells + beads in microwells
Zoom-in examples:
single cell with bead in a well
bead
cell
cell
bead
Identical for all oligos on one bead
different for different bead (~1 million different varieties)
bead
5' amine
Universal
Cell Label
molecular label
Different for different oligo on one bead
30μm microwell
Cell lysis,
mRNA from single cell hybridize on bead in the same well
AAAAAAAAA
mRNA molecules from cell
Collect all beads from microwell array
cDNA synthesis on beads,
carried in tube
RT mix
Multiplex PCR with beads in tube
PCR mix
Universal primer
Gene specific primer
Sequencing reveals cell label, molecule index, gene

FIG. 30

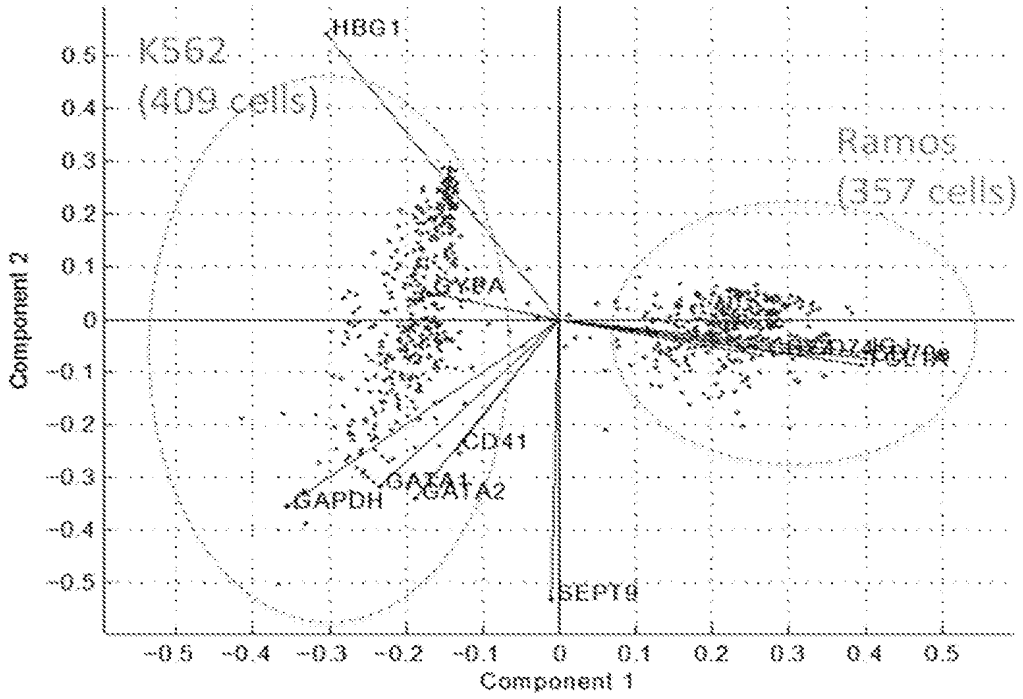


FIG. 31A

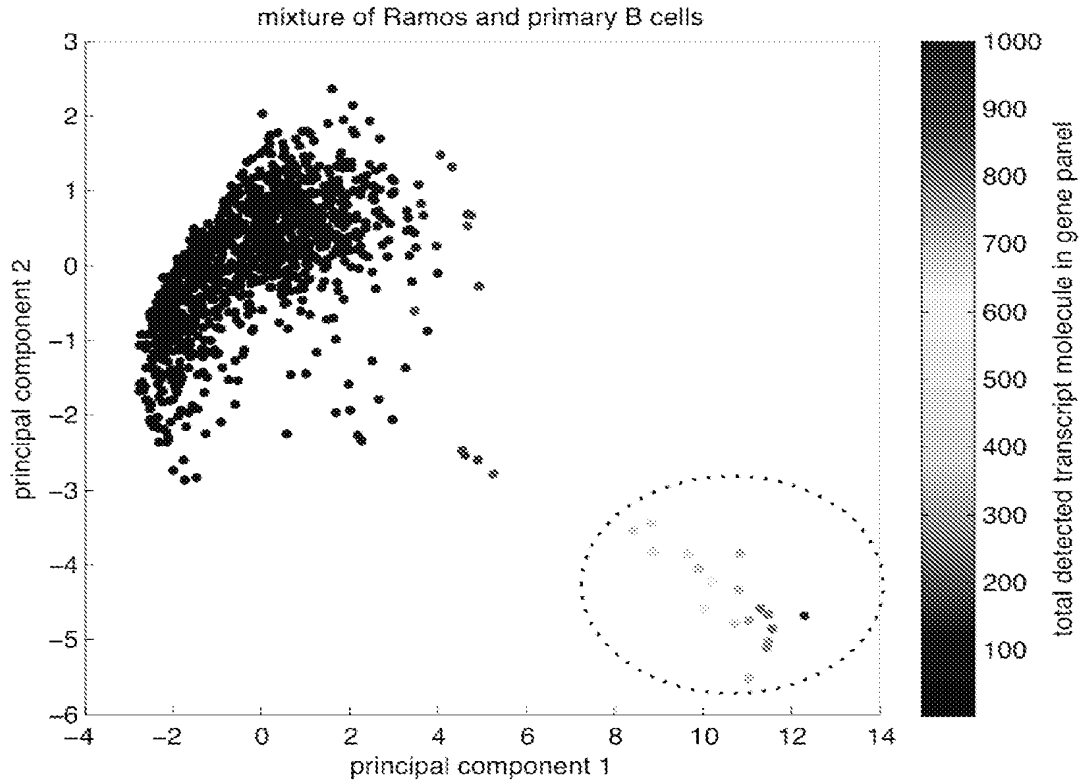


FIG. 31B

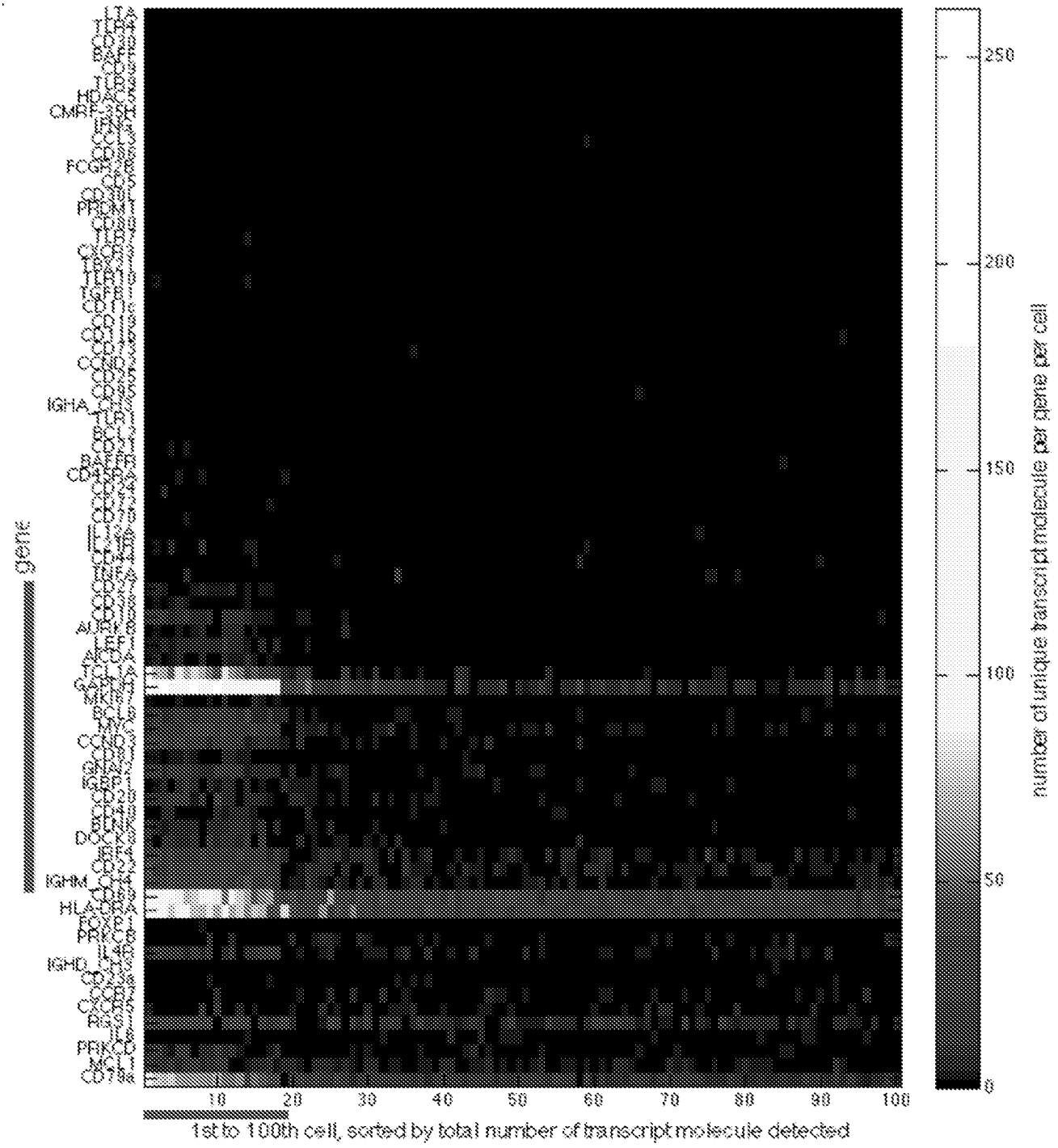

IGHA
IGHM
HLA-DRA
IGHD
MS4A1
CD79A
gene
10
20
30
40
50
60
70
80
90
100
1st to 100th cell, sorted by total number of transcript molecule detected
250
200
150
100
50
0
number of unique transcript molecule per gene per cell

FIG. 31C

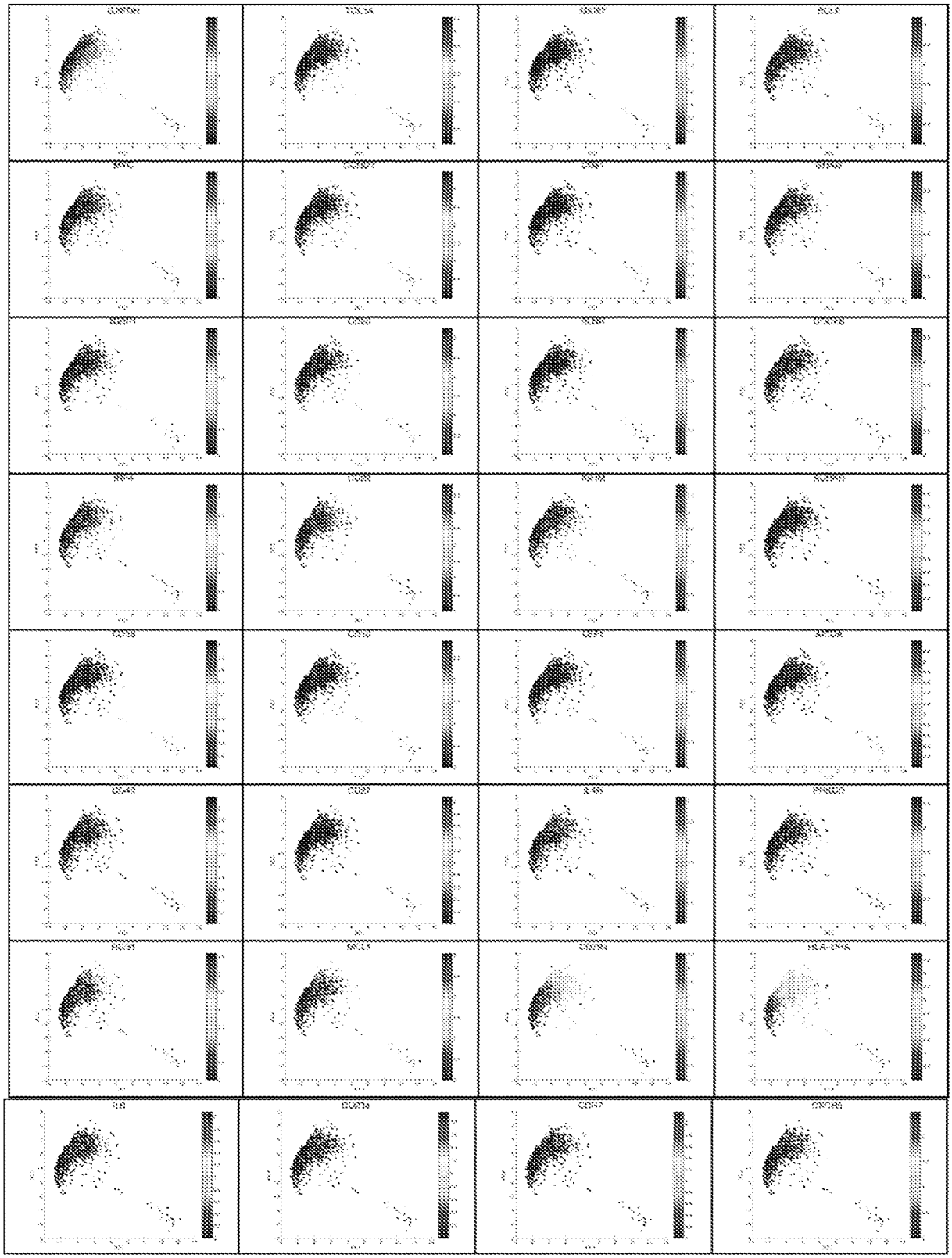


FIG. 31D

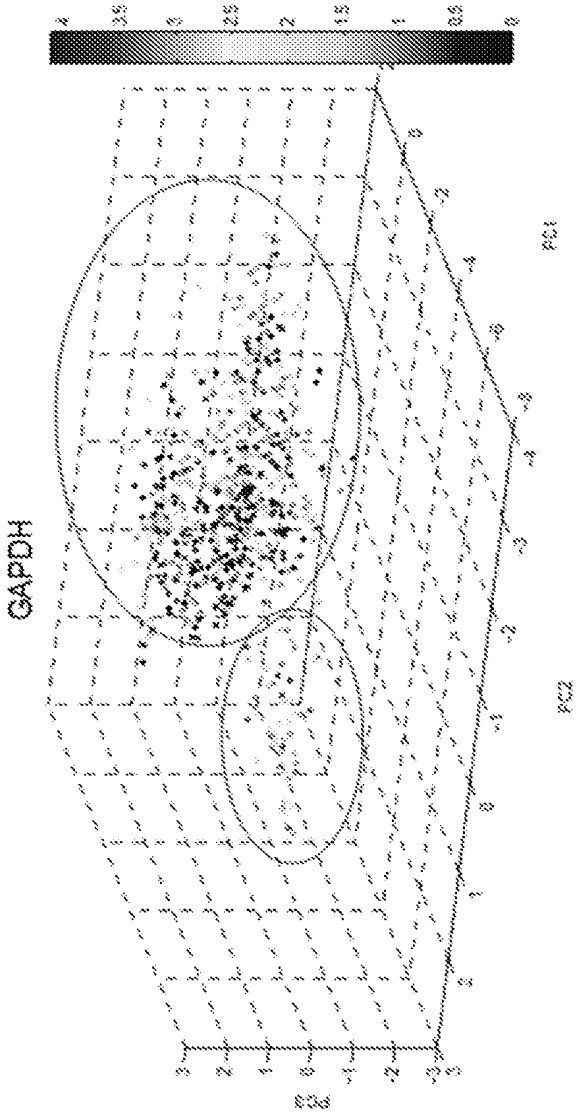
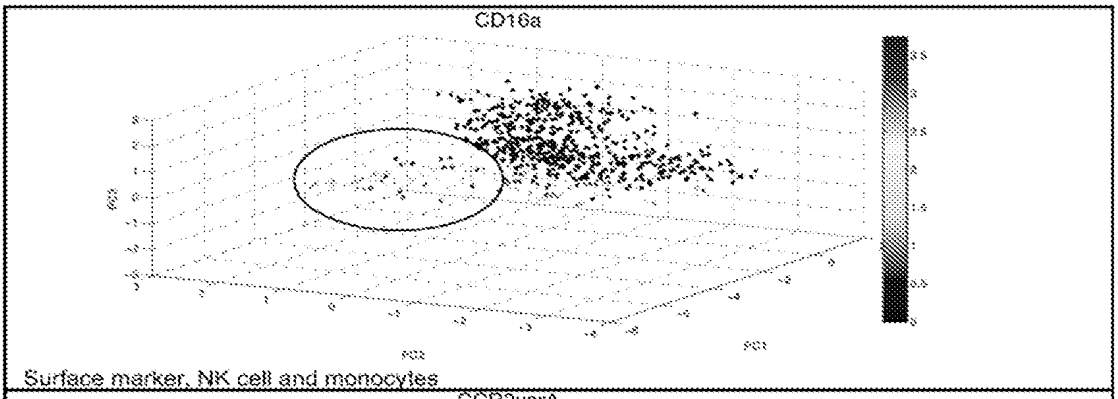
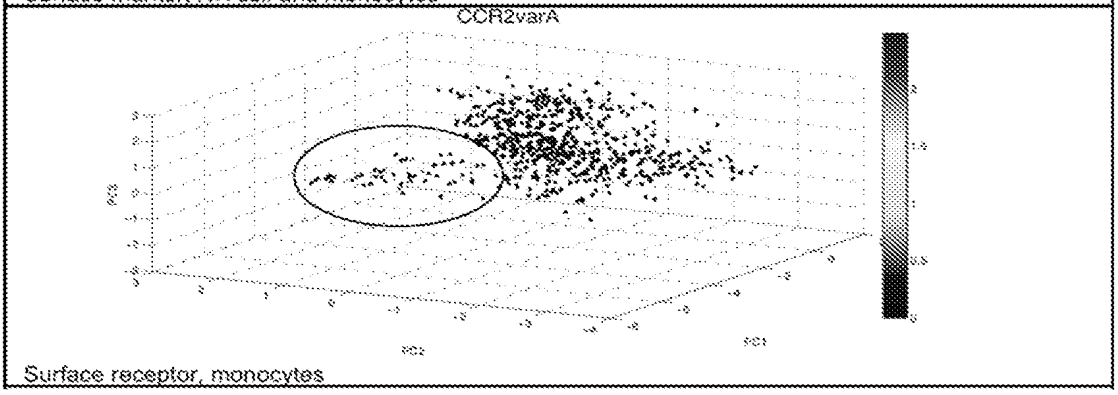


FIG. 32

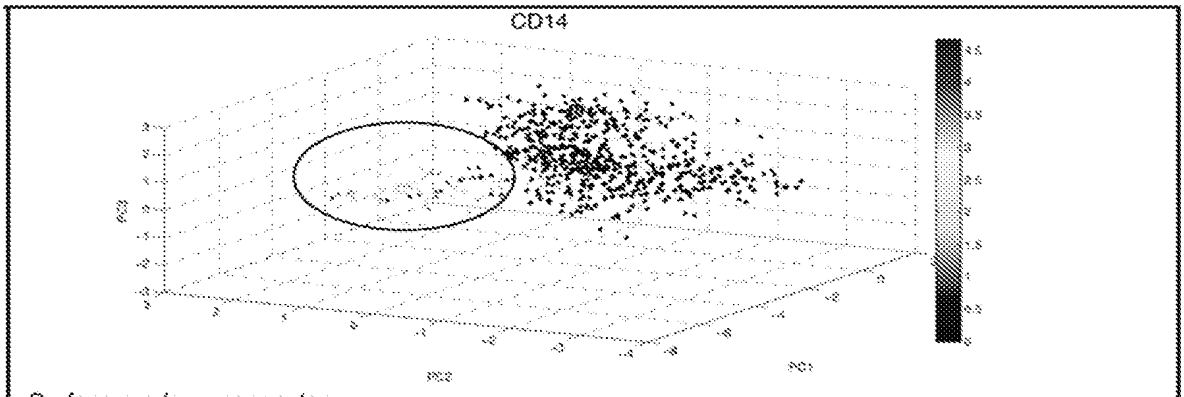
33A



33B



33C



33D

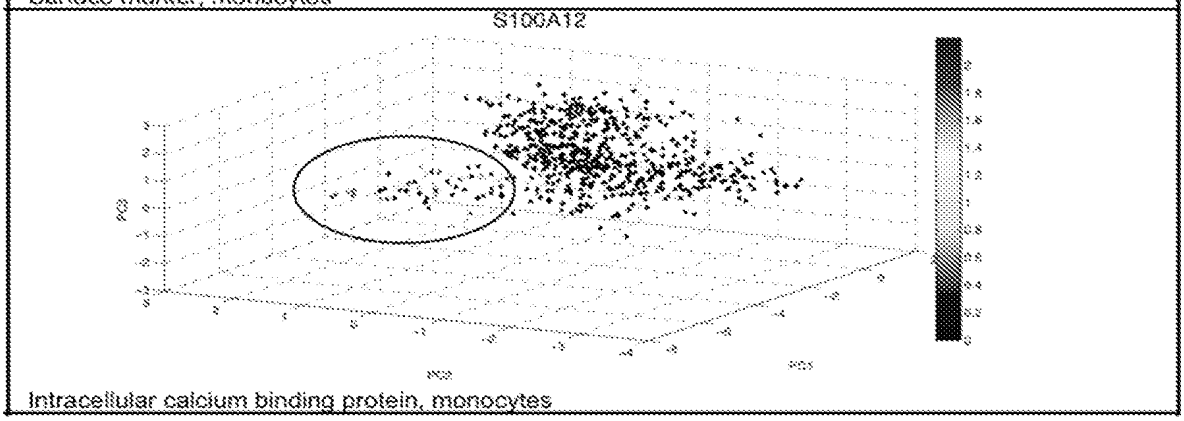


FIG. 33A-D

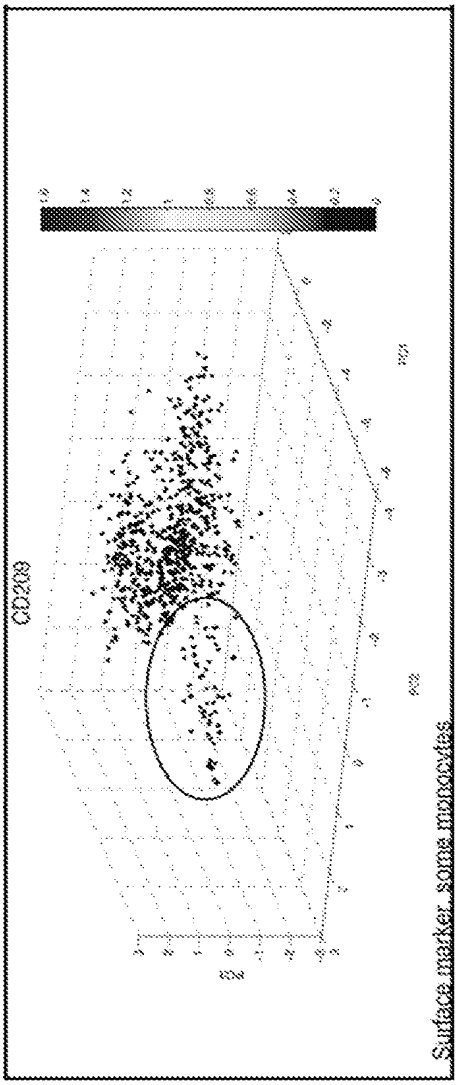


FIG. 33E

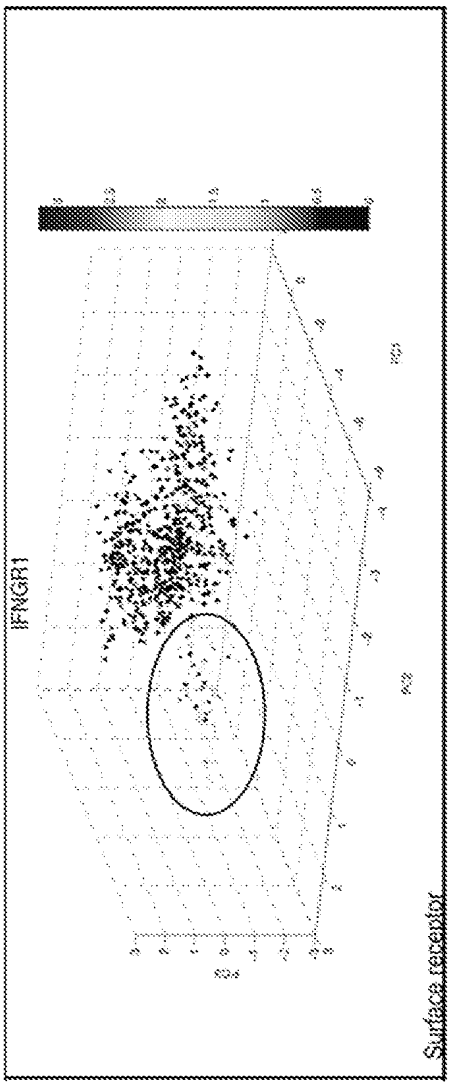
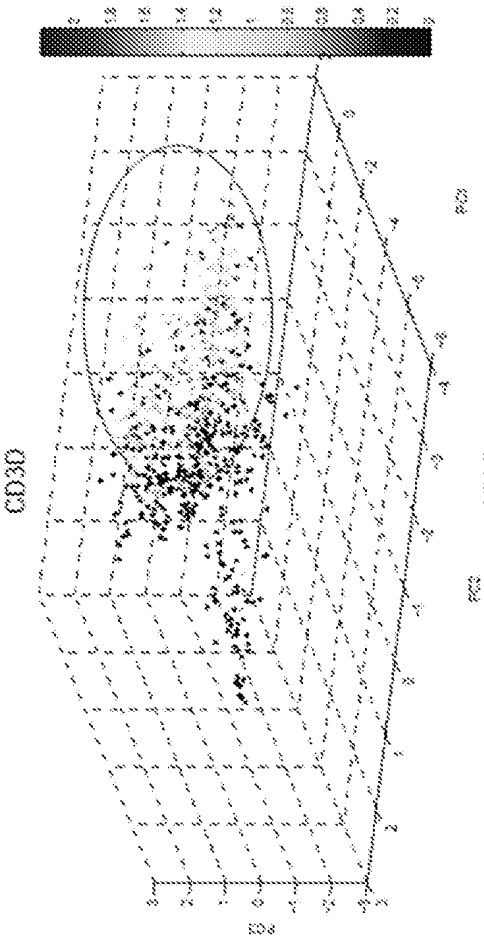
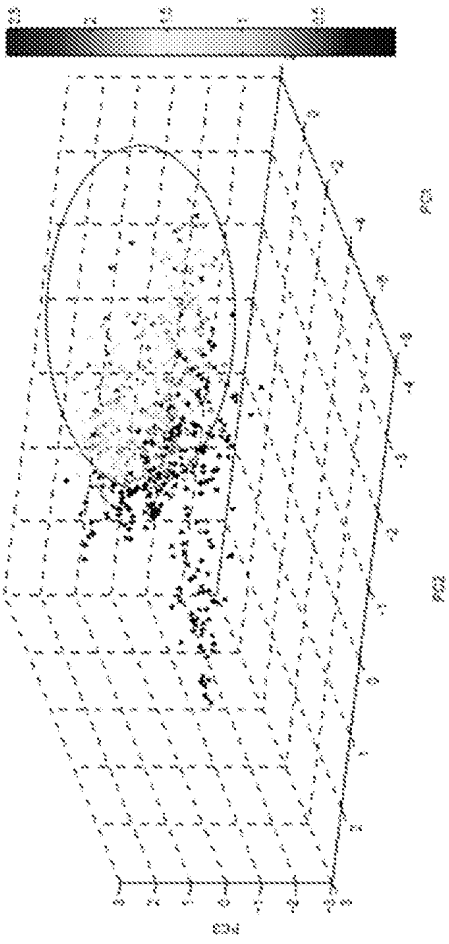


FIG. 33F



34A



34B

FIG. 34A-B

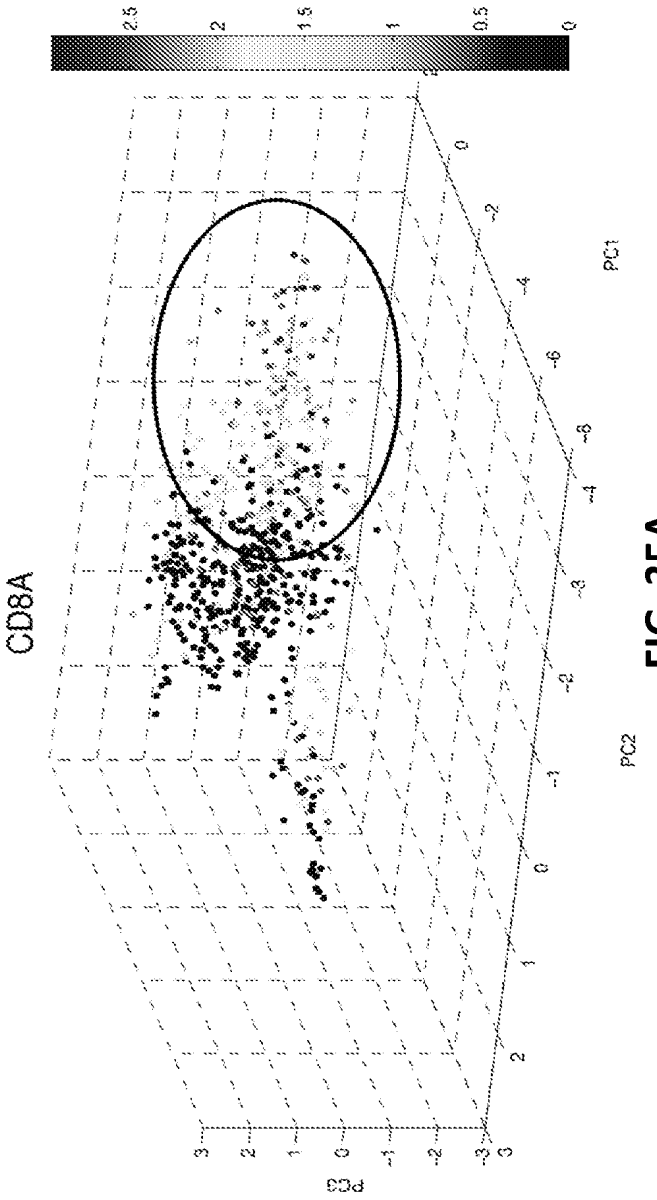


FIG. 35A

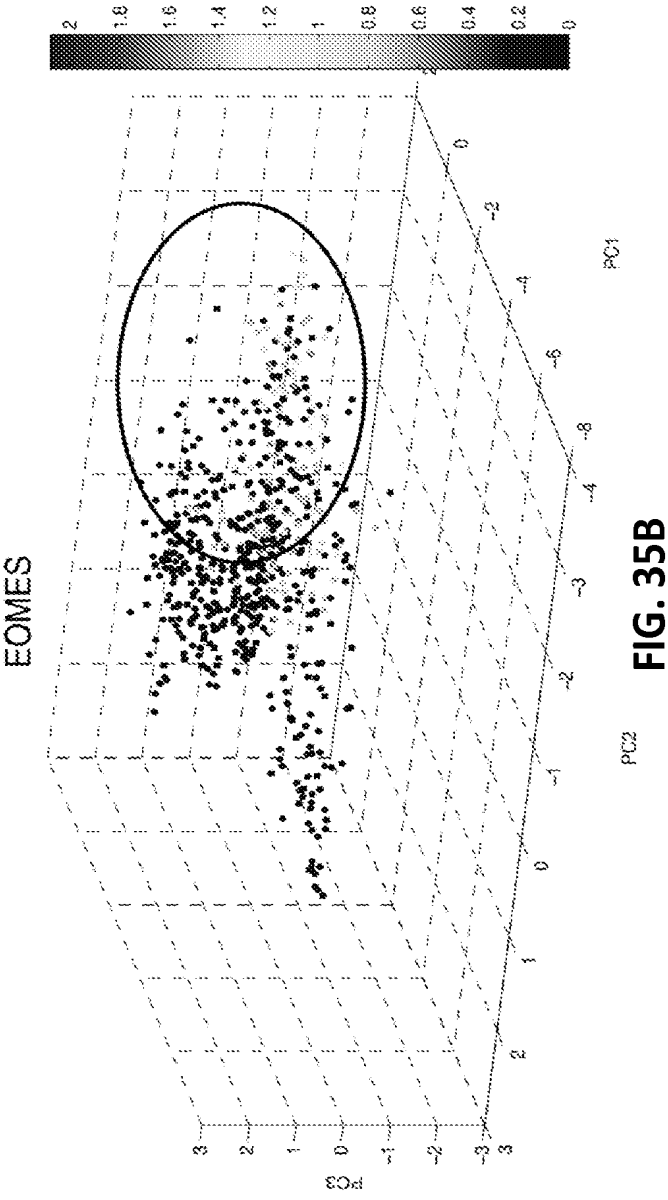


FIG. 35B

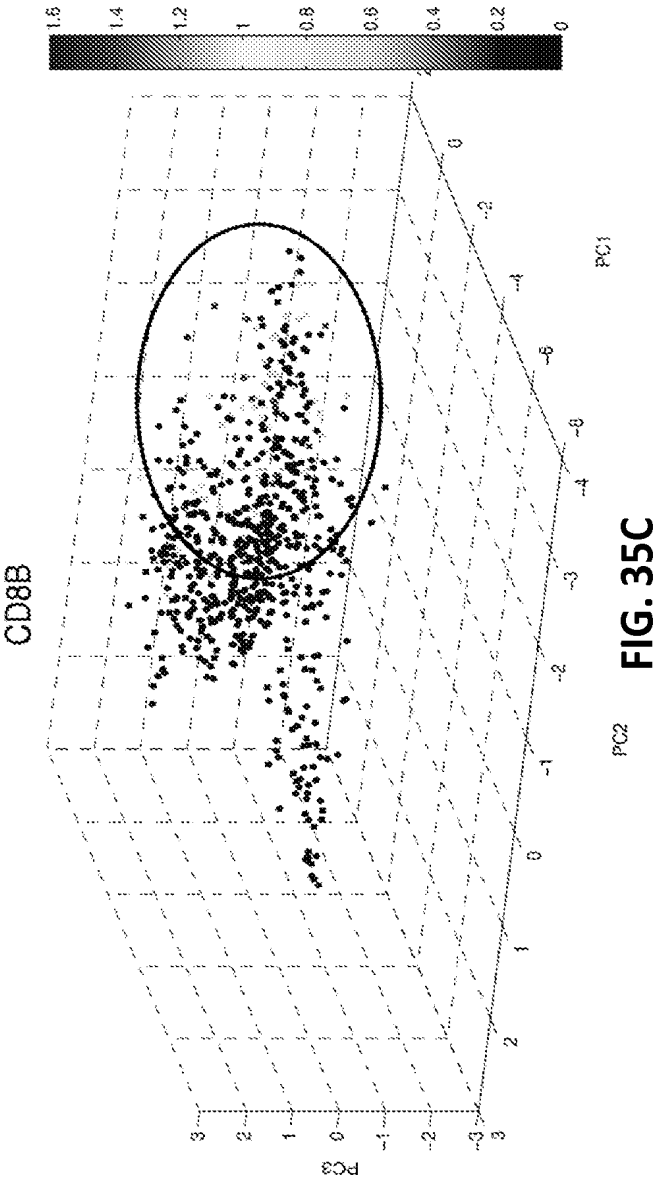


FIG. 35C

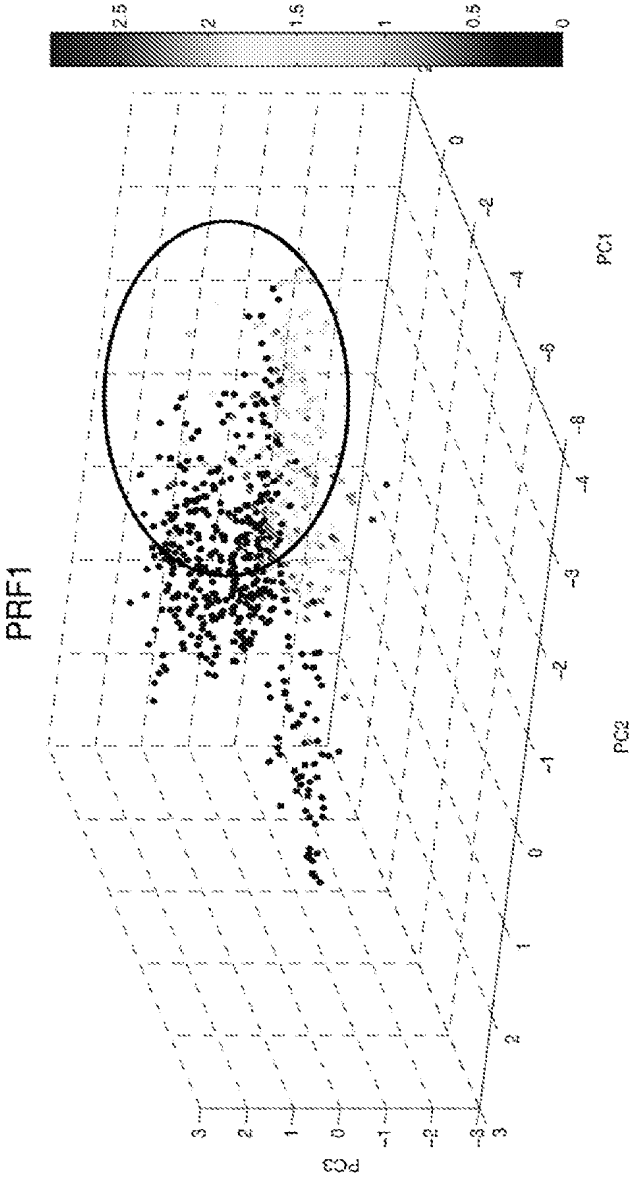


FIG. 35D

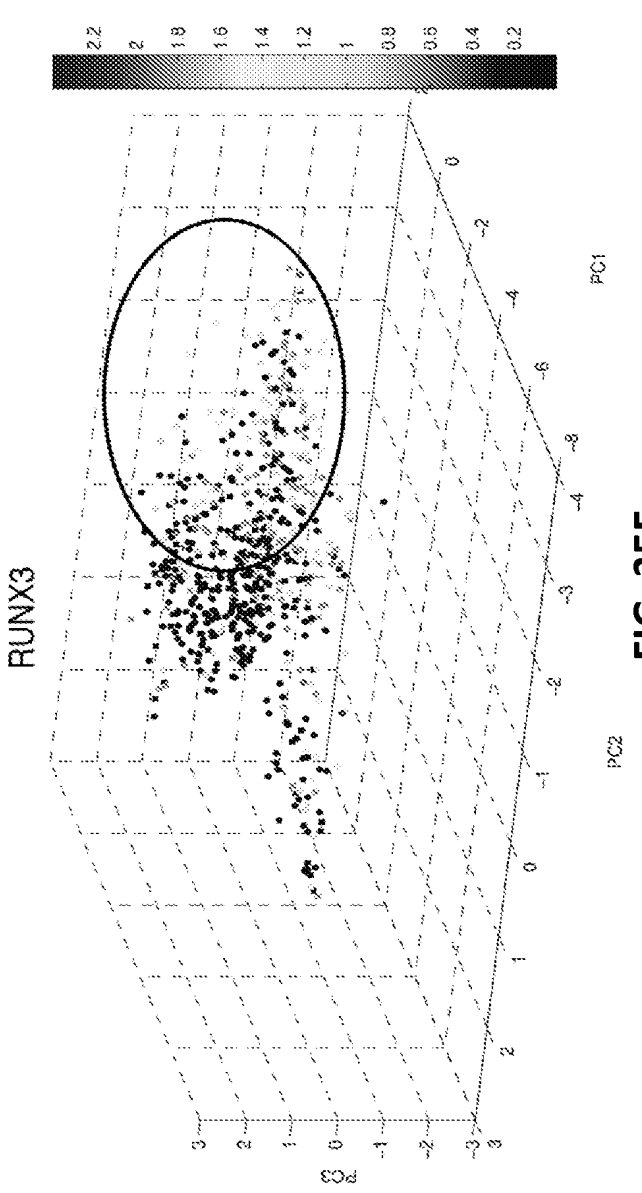


FIG. 35E

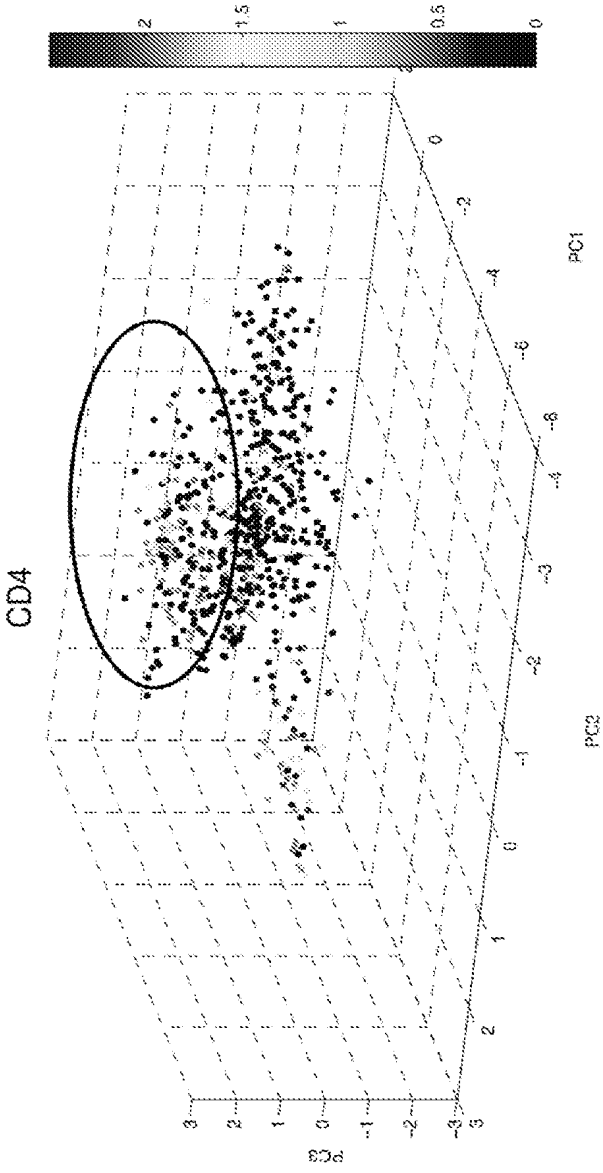


FIG. 36A

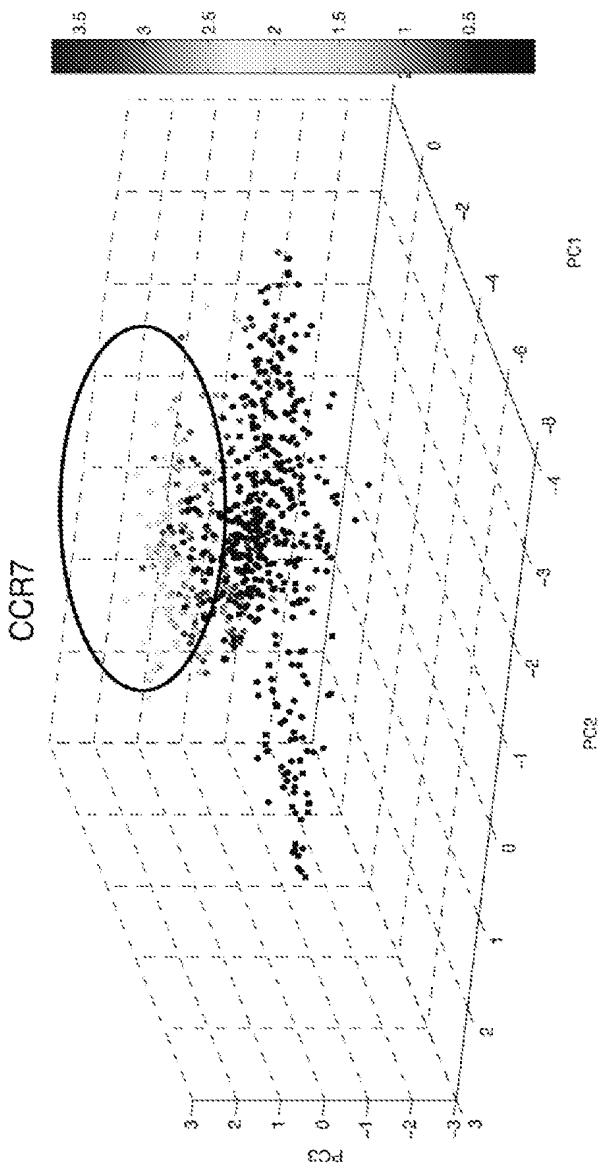


FIG. 36B

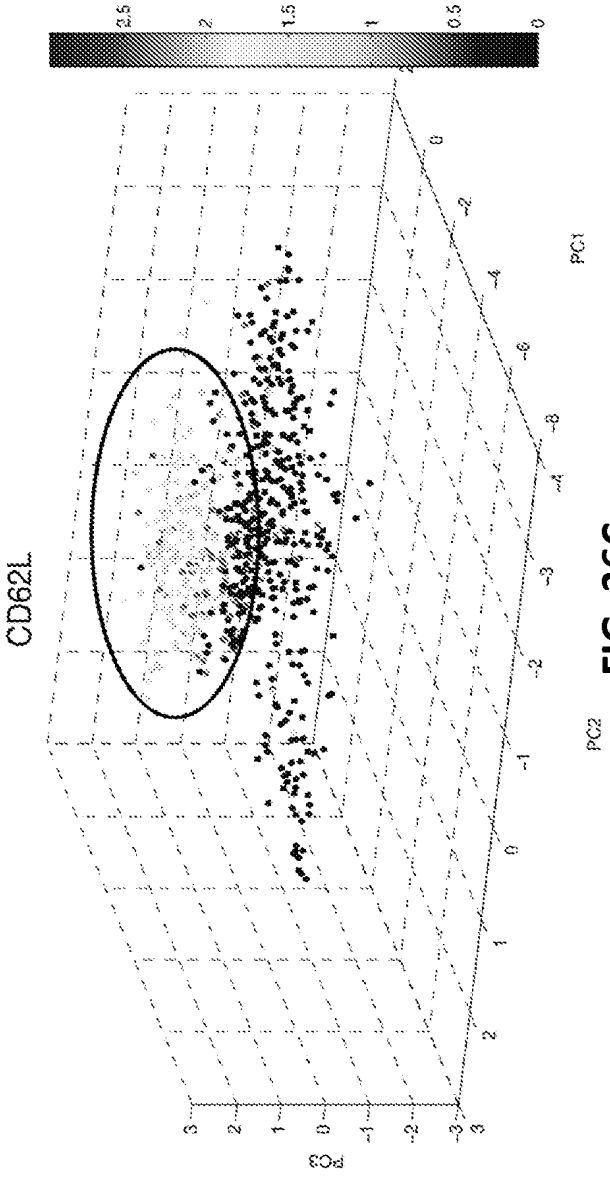


FIG. 36C

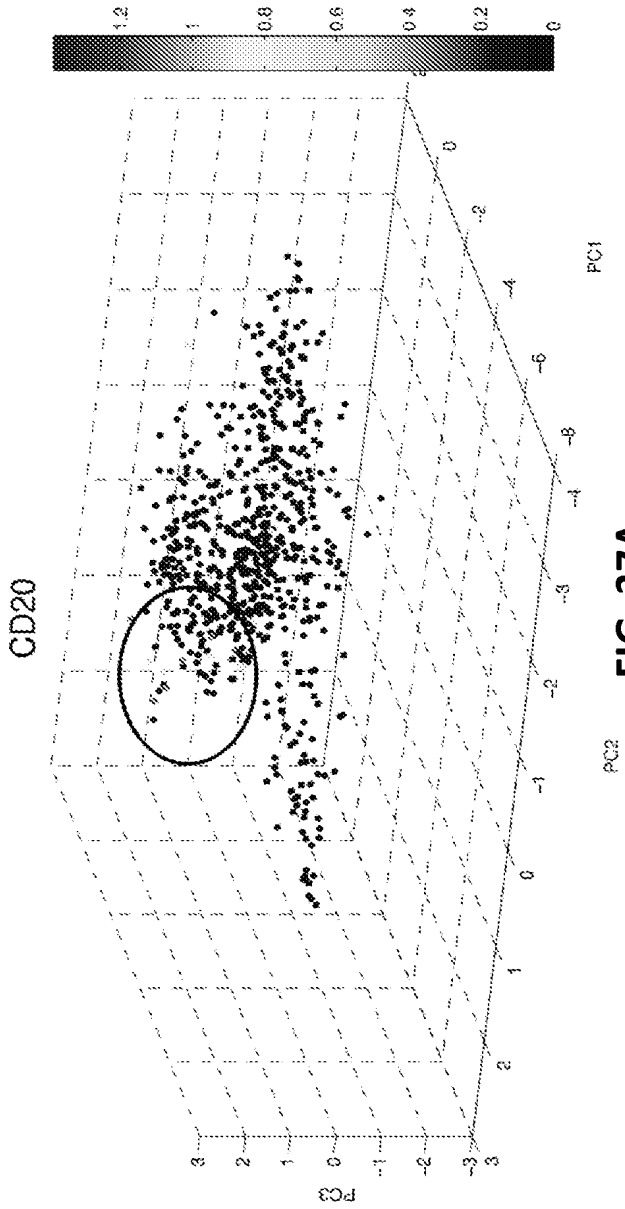


FIG. 37A

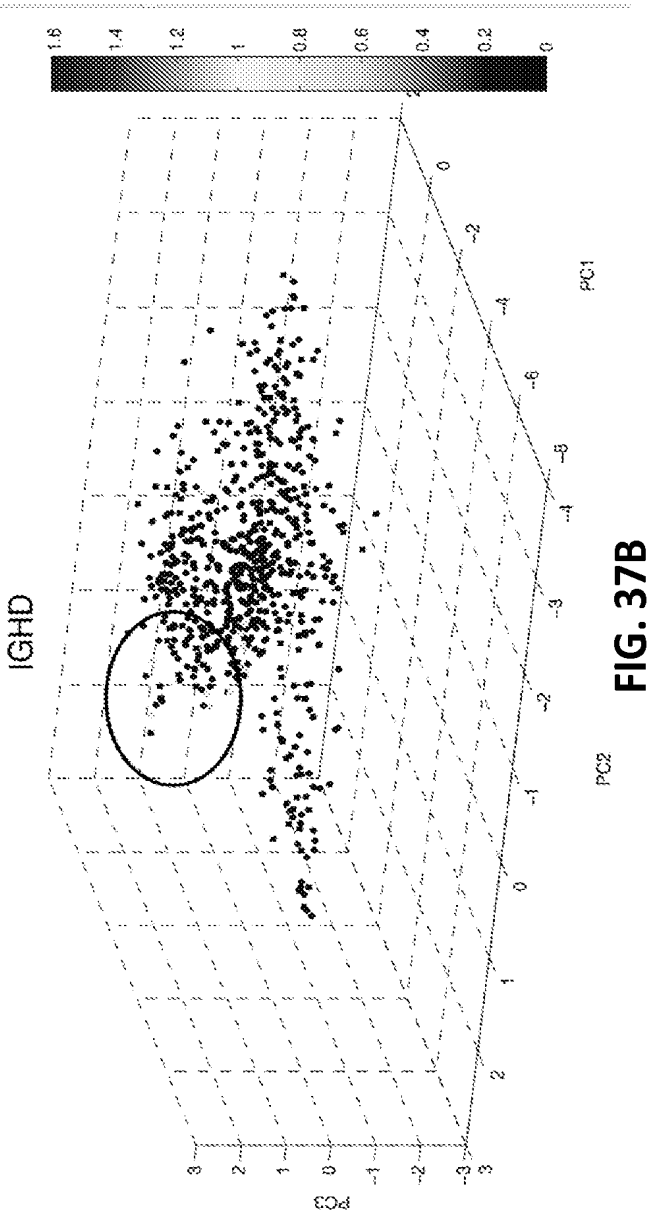


FIG. 37B

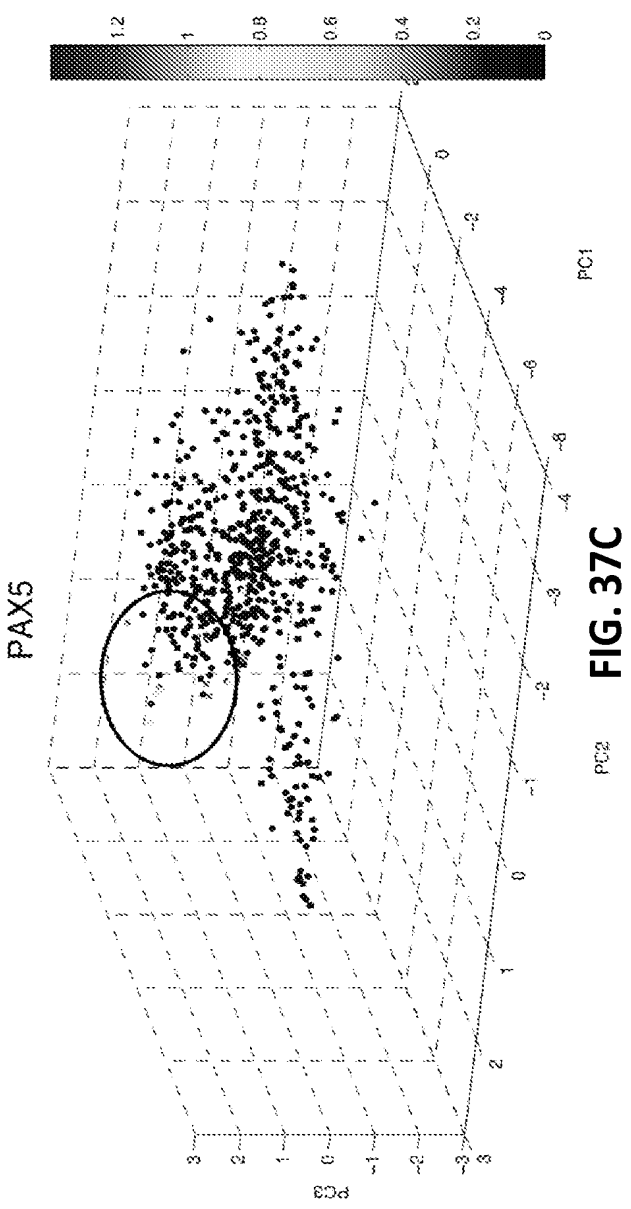


FIG. 37C

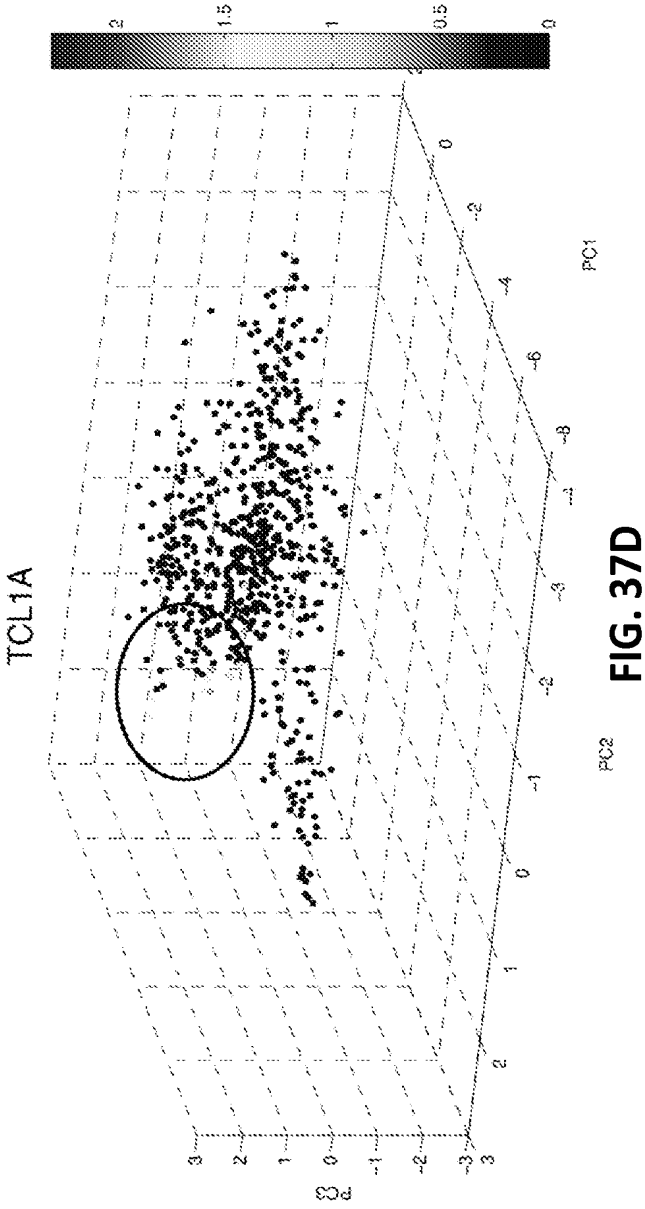


FIG. 37D

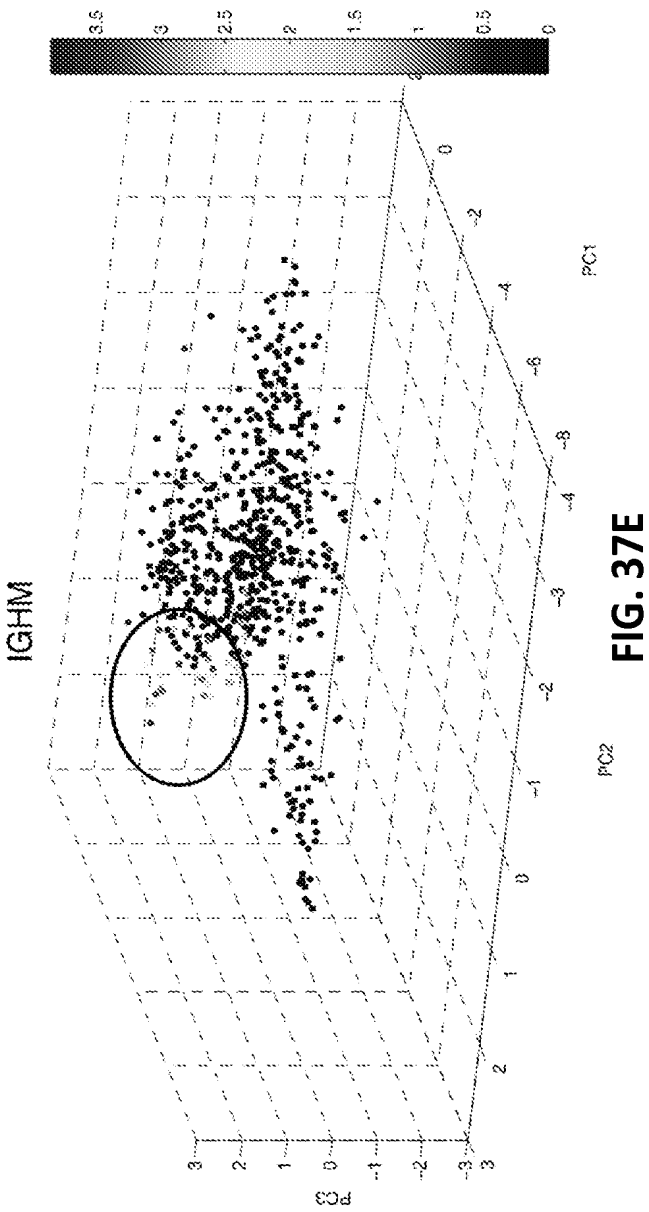


FIG. 37E

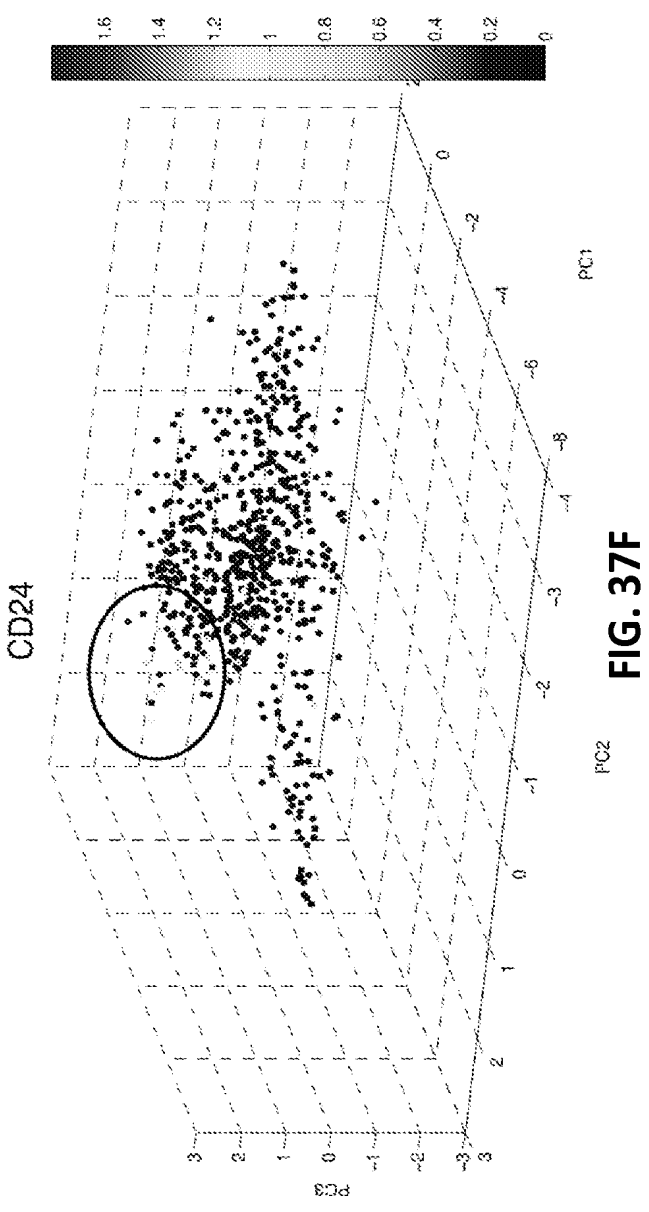


FIG. 37F

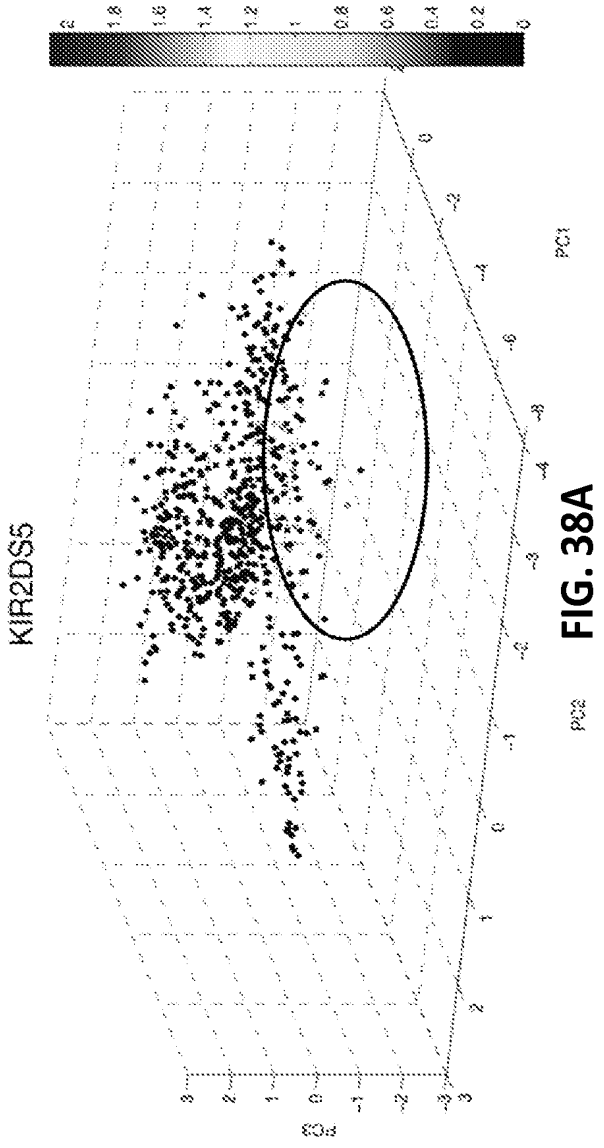


FIG. 38A

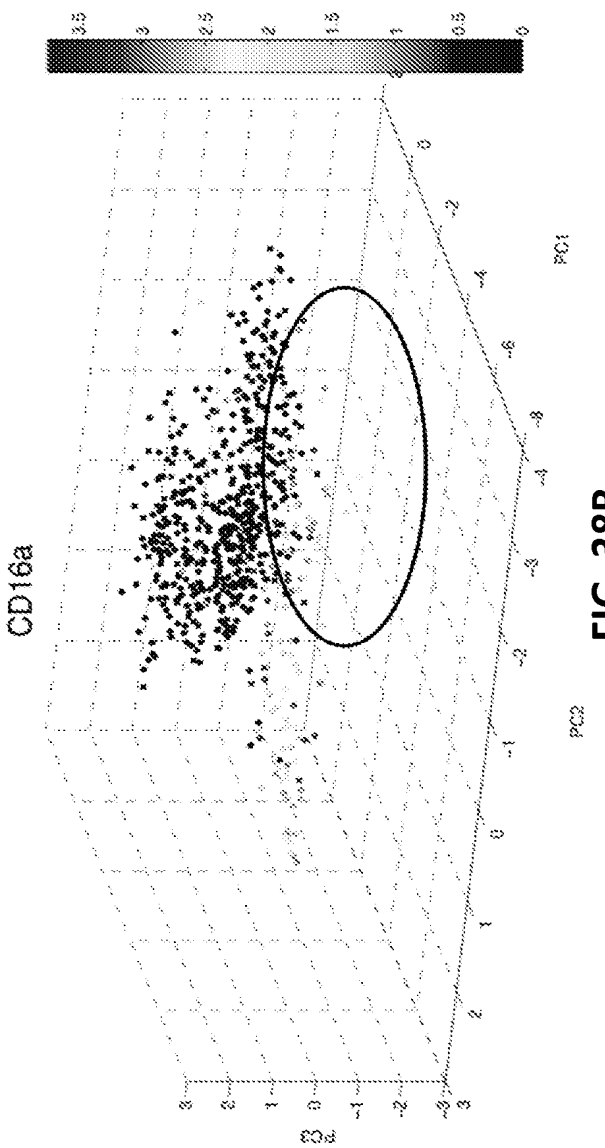


FIG. 38B

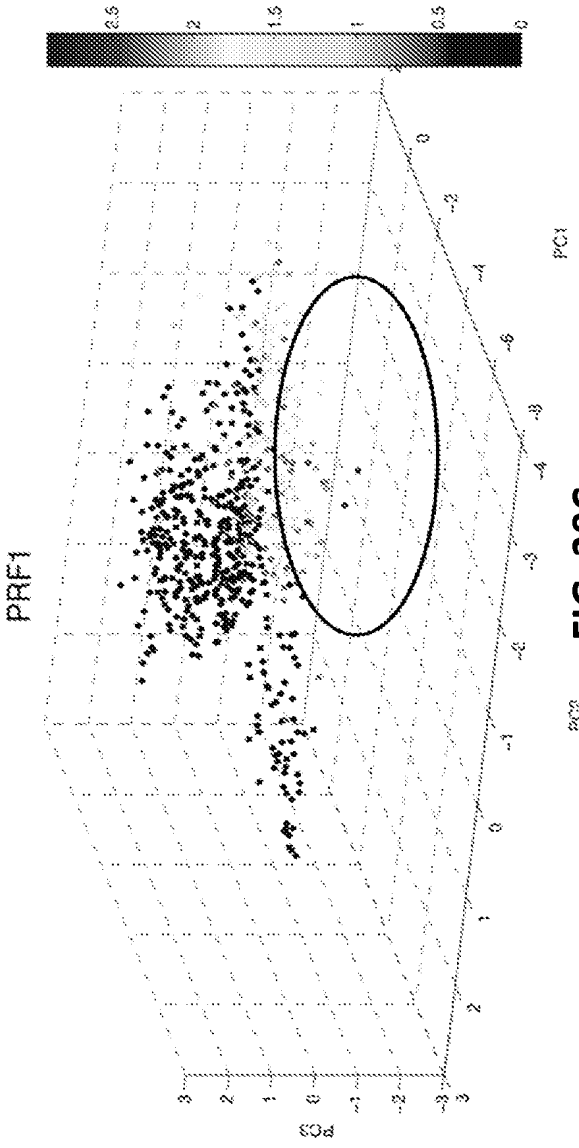


FIG. 38C

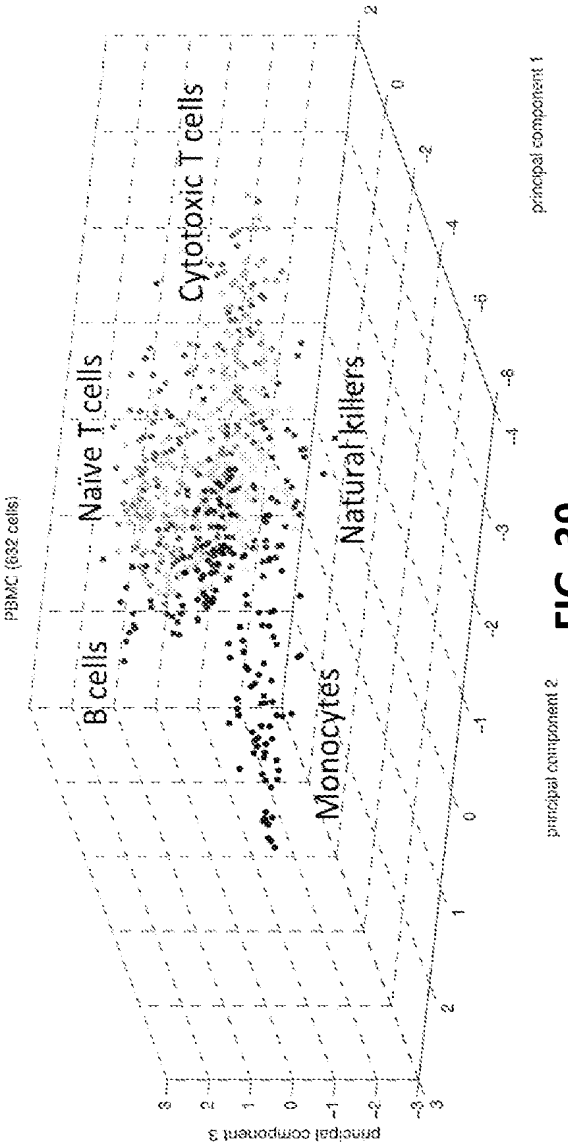


FIG. 39

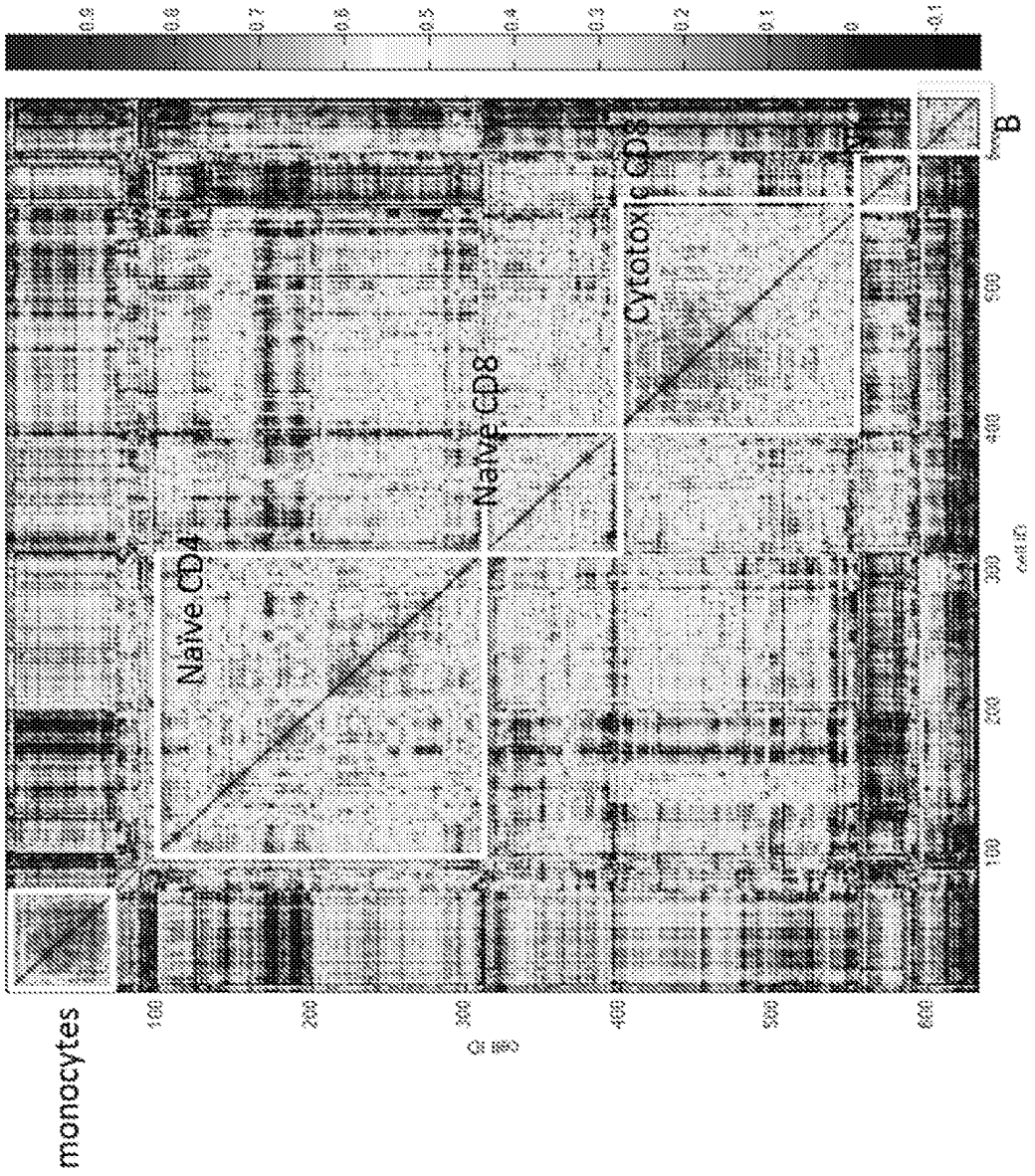

monocytes
B
Naïve CD4
Naïve CD8
Cytotoxic CD8
CD45RO
CD45RA
100
200
300
400
500
600

FIG. 40A

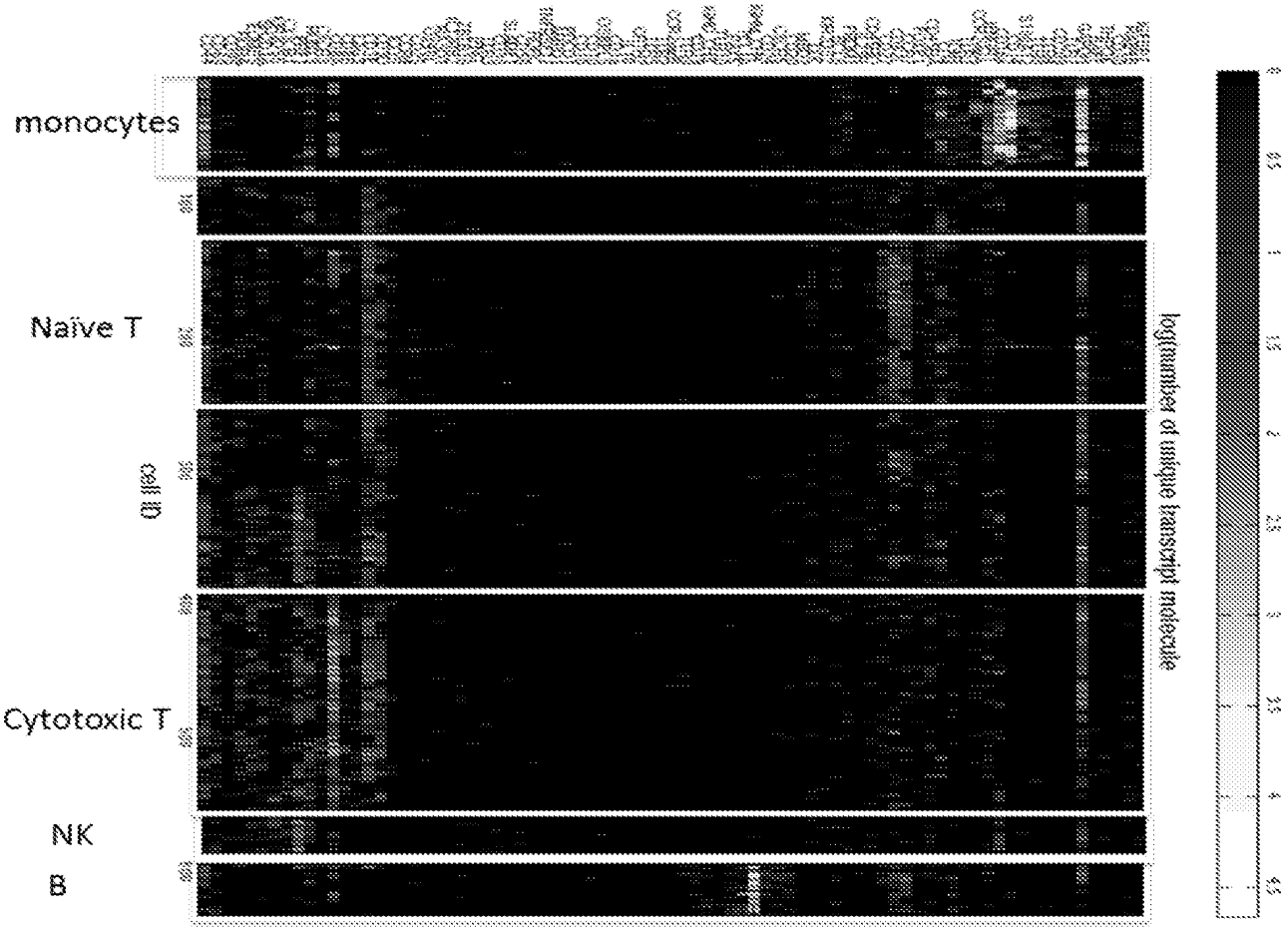


FIG. 40B

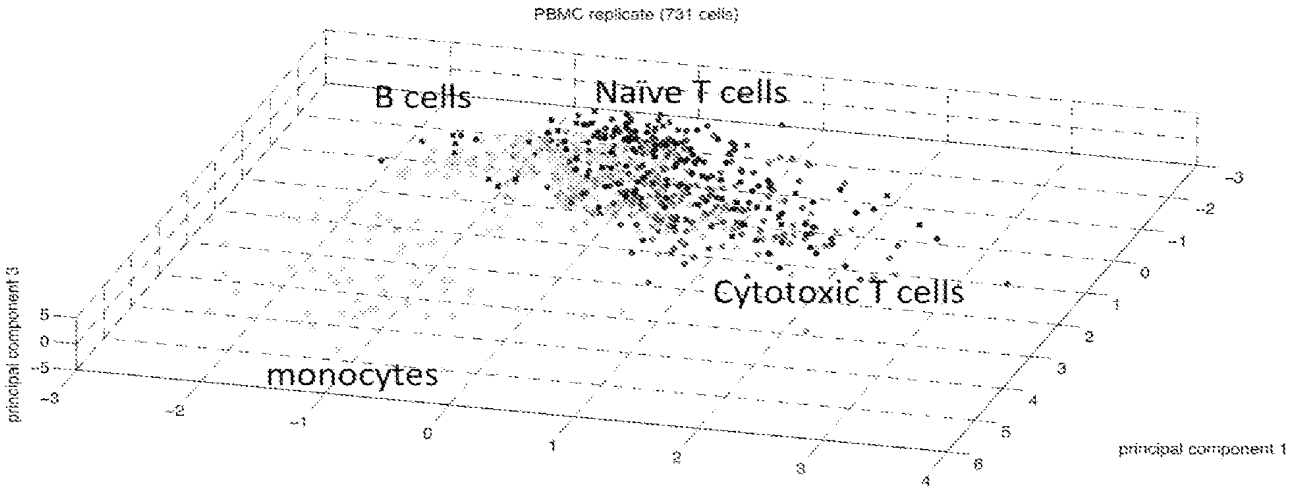


FIG. 41

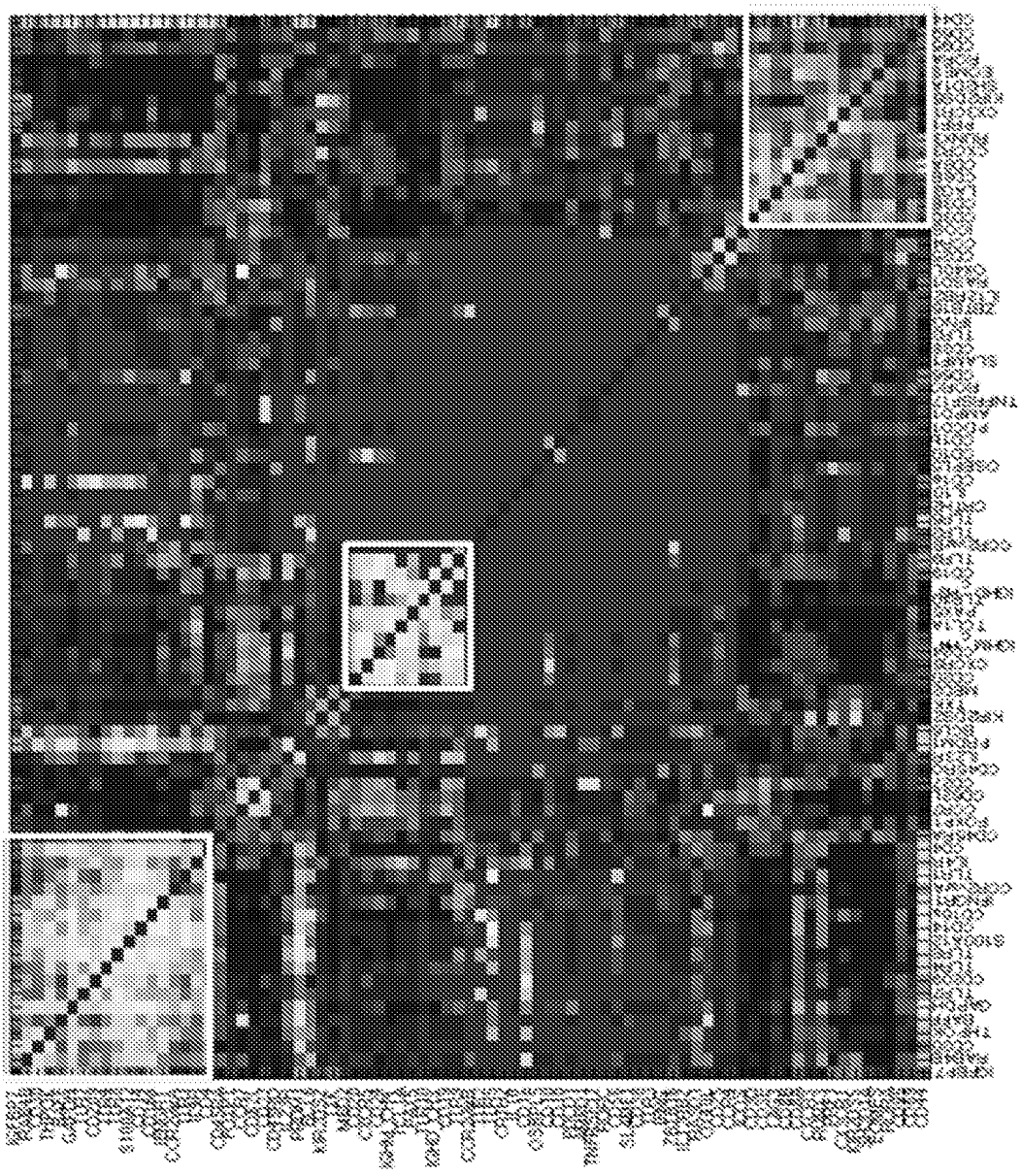


FIG. 42

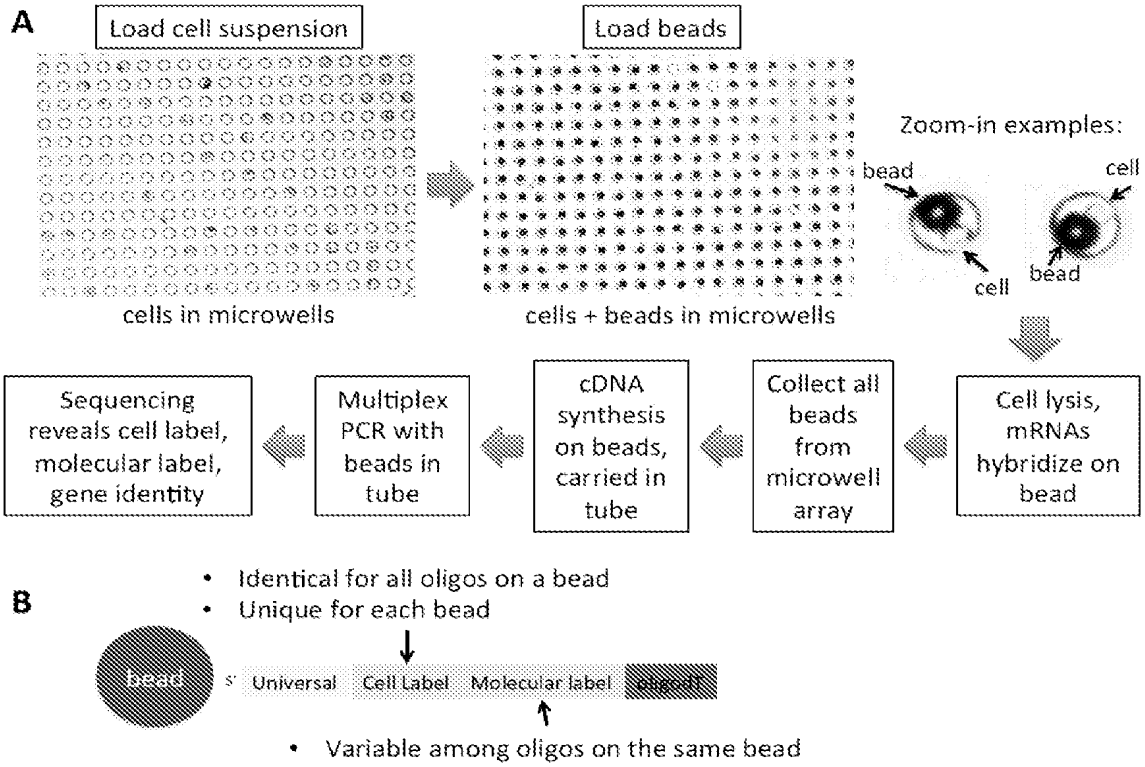


FIG. 43

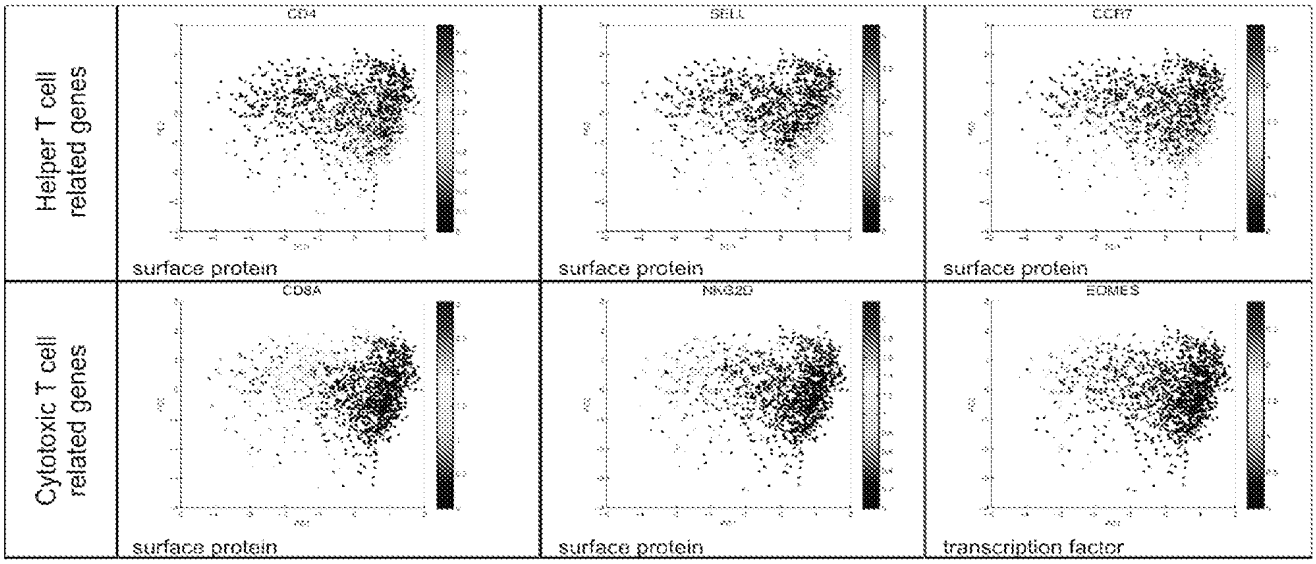


FIG. 44A

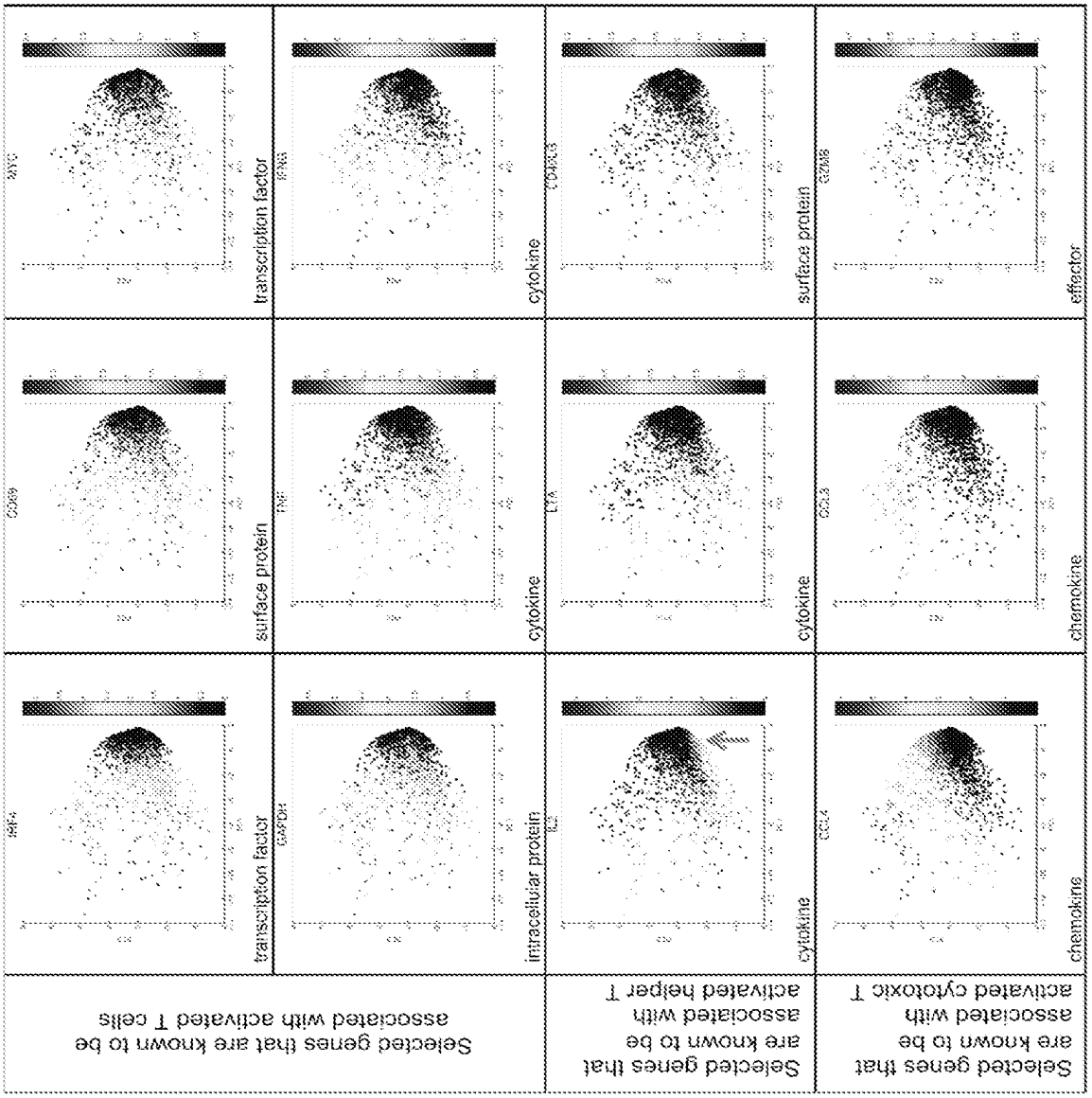

Selected genes that are known to be associated with activated cytotoxic T
Selected genes that are known to be associated with activated helper T
Selected genes that are known to be associated with activated T cells
effector
surface protein
cytokine
transcription factor
chemokine
cytokine
surface protein
cytokine
transcription factor
chemokine
cytokine
intracellular protein
cytokine
surface protein
transcription factor

FIG. 44B

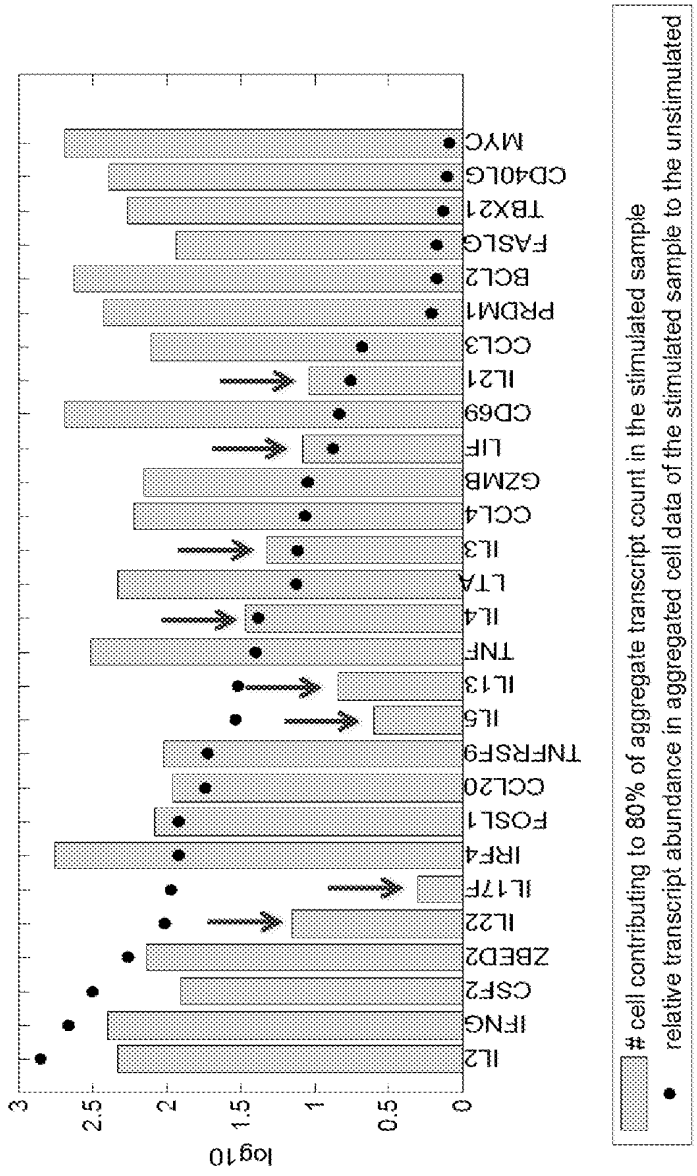

IL2
IFNG
CSF2
ZBED2
IL22
IL17F
IRF4
FOSL1
CCL20
TNFRSF9
IL5
IL13
TNF
IL4
LTA
IL3
CCL4
GZMB
LIF
CD69
IL21
CCL3
PRDM1
BCL2
FASLG
TBX21
CD40LG
MYC
log10
3
2.5
2
1.5
1
0.5
0
cell contributing to 80% of aggregate transcript count in the stimulated sample
relative transcript abundance in aggregated cell data of the stimulated sample to the unstimulated

FIG. 44C

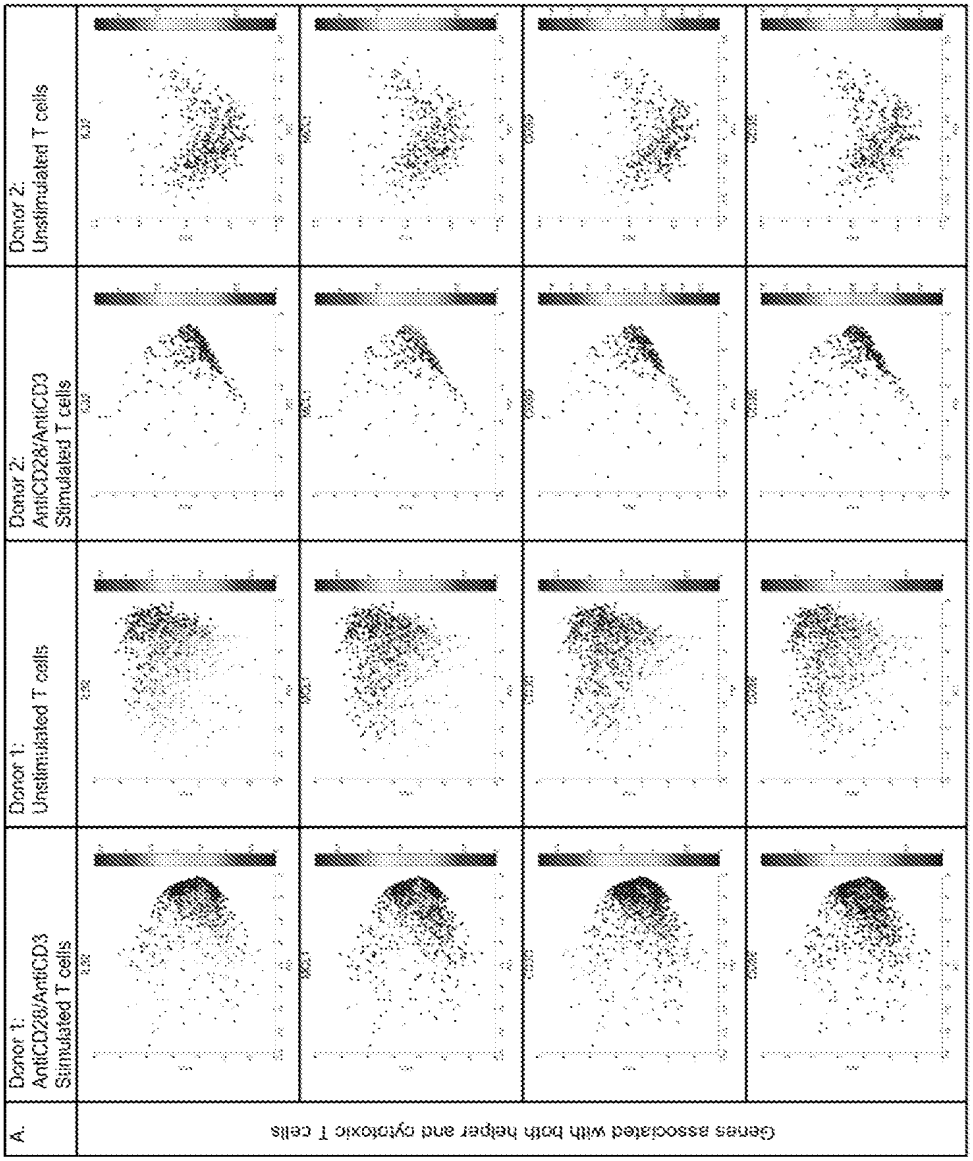

A.
Genes associated with both helper and cytotoxic T cells
Donor 1:
Anti-CD28/Anti-CD3
Stimulated T cells
Donor 1:
Unstimulated T cells
Donor 2:
Anti-CD28/Anti-CD3
Stimulated T cells
Donor 2:
Unstimulated T cells

FIG. 45A

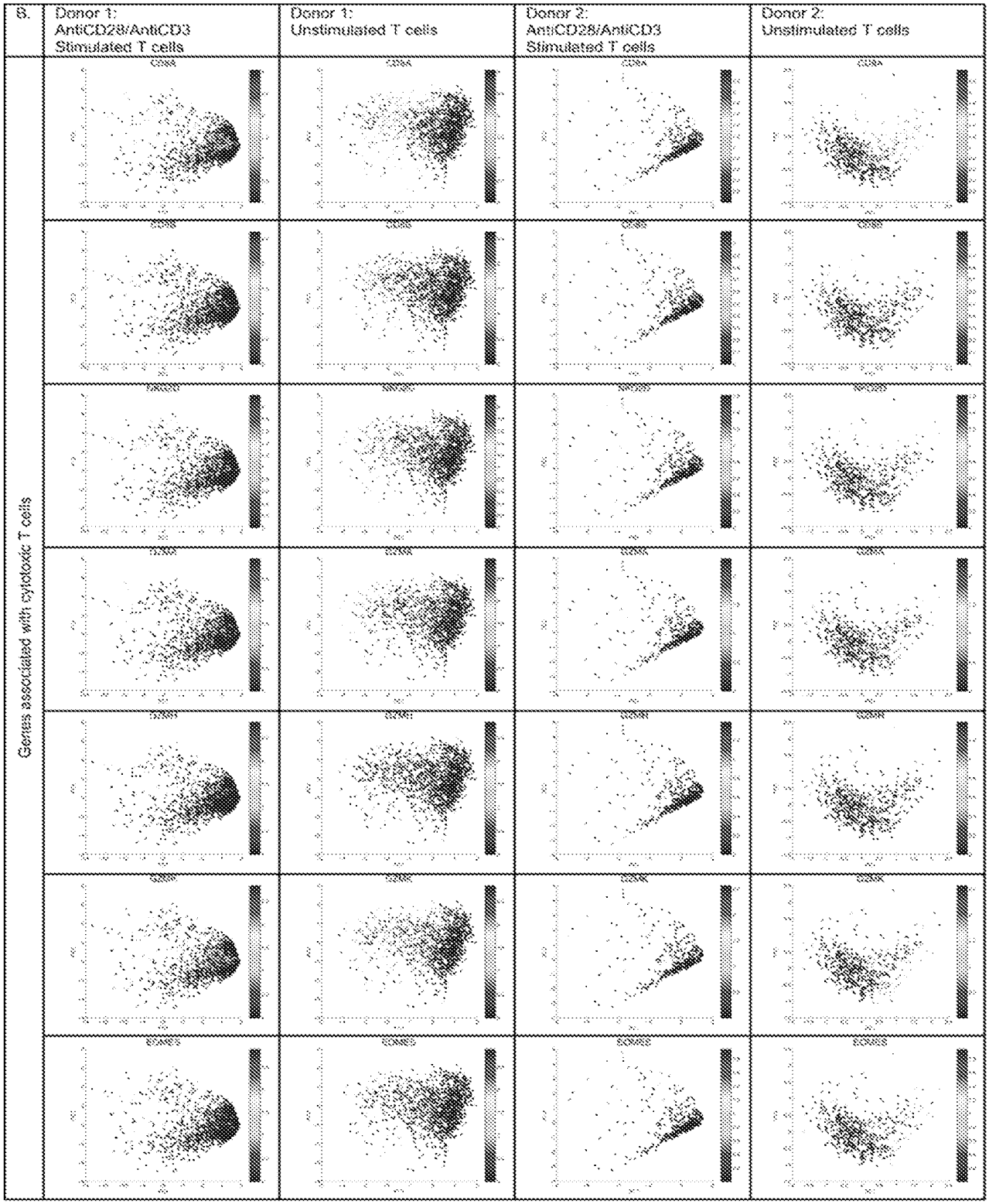

B.
Donor 1:
AntiCD28/AntiCD3
Stimulated T cells
Donor 1:
Unstimulated T cells
Donor 2:
AntiCD28/AntiCD3
Stimulated T cells
Donor 2:
Unstimulated T cells
Genes associated with cytotoxic T cells

FIG. 45B

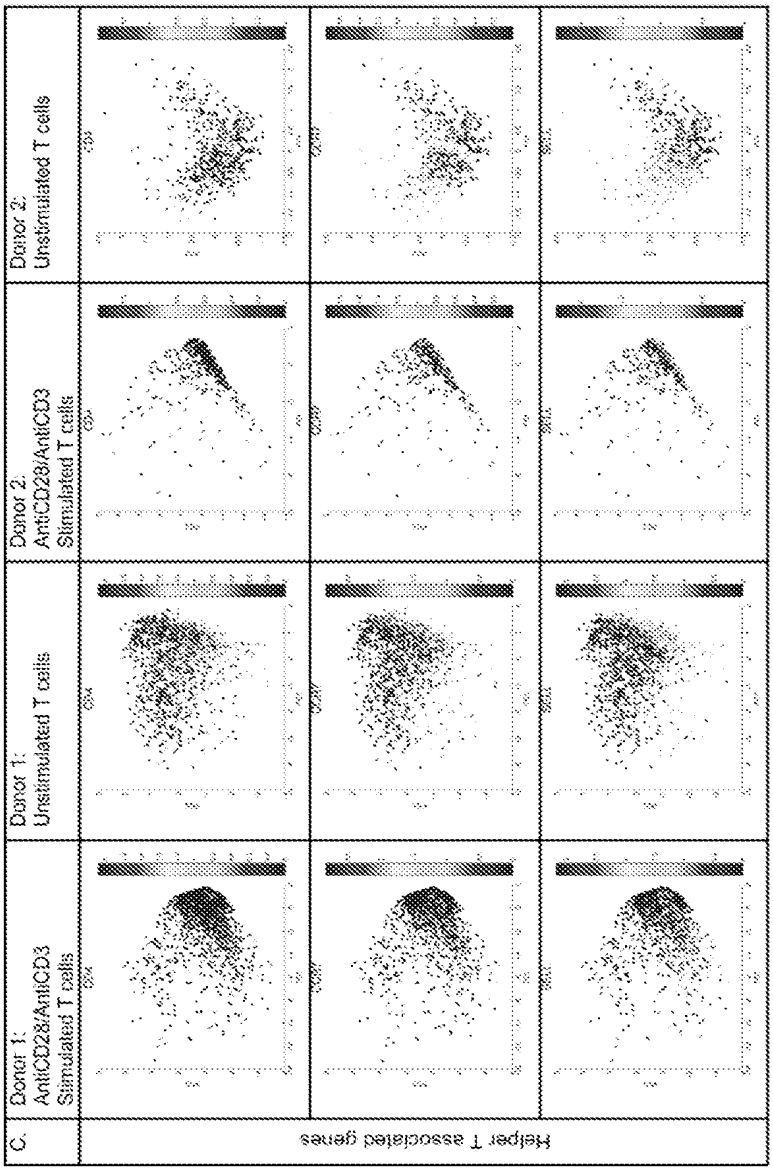

C:
Donor 1:
Anti-CD28/Anti-CD3
Stimulated T cells
Donor 1:
Unstimulated T cells
Donor 2:
Anti-CD28/Anti-CD3
Stimulated T cells
Donor 2:
Unstimulated T cells
Helper T associated genes

FIG. 45C

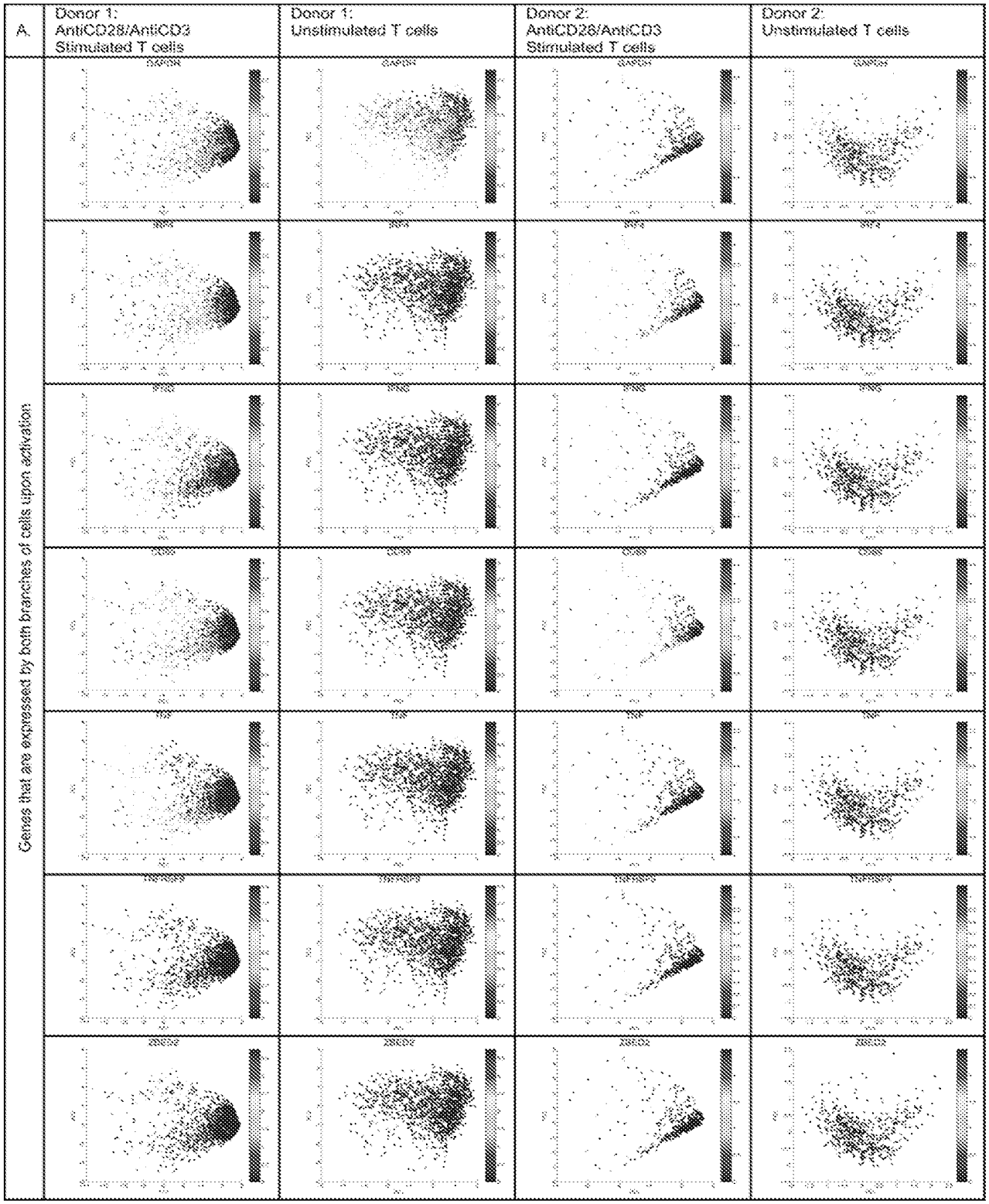

A.
Donor 1:
AntiCD28/AntiCD3
Stimulated T cells
Donor 1:
Unstimulated T cells
Donor 2:
AntiCD28/AntiCD3
Stimulated T cells
Donor 2:
Unstimulated T cells
Genes that are expressed by both branches of cells upon activation

FIG. 46A

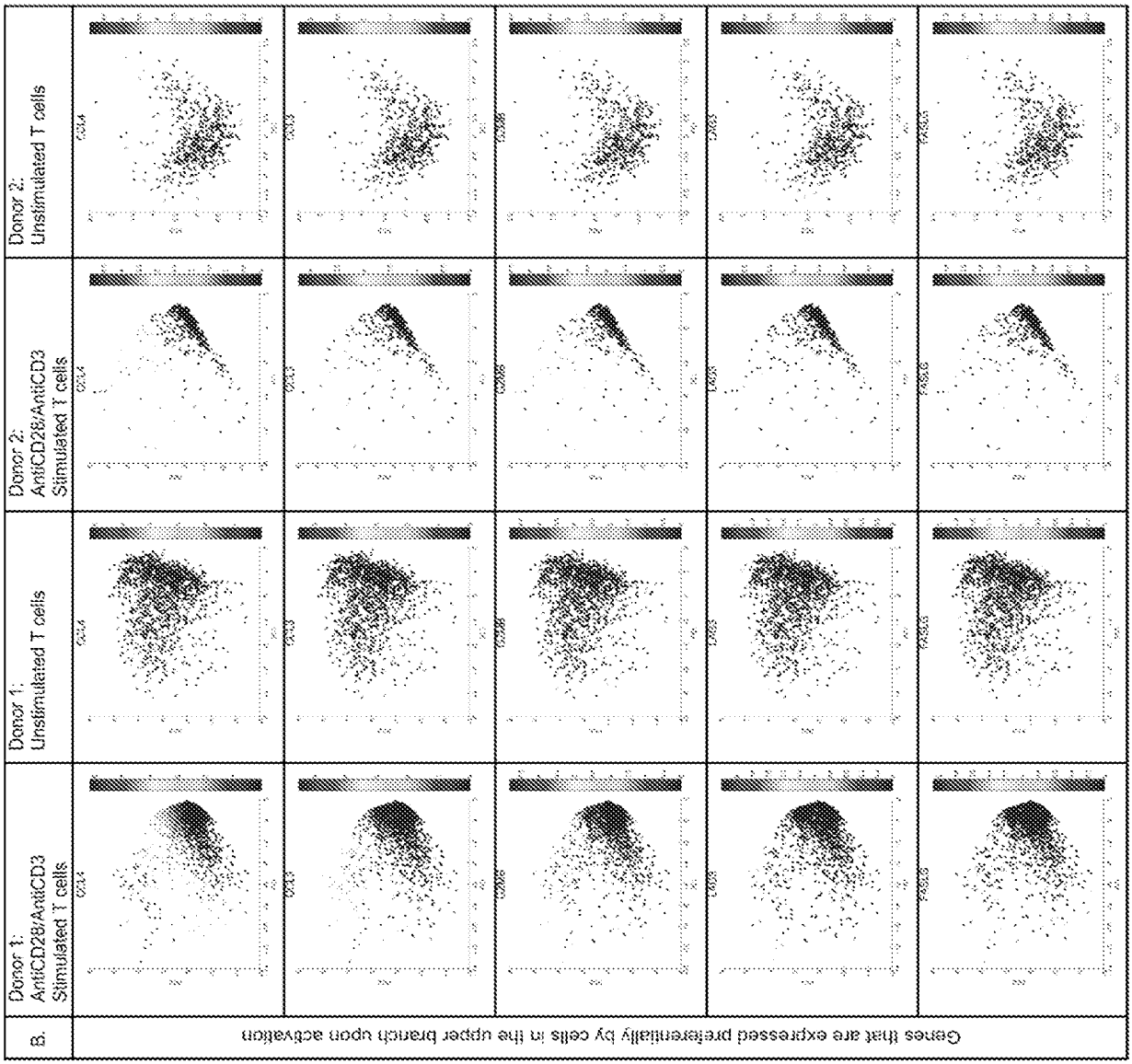

B.
Genes that are expressed preferentially by cells in the upper branch upon activation
Donor 1: Anti-CD28/Anti-CD3 Stimulated T cells
Donor 1: Unstimulated T cells
Donor 2: Anti-CD28/Anti-CD3 Stimulated T cells
Donor 2: Unstimulated T cells

FIG. 46B

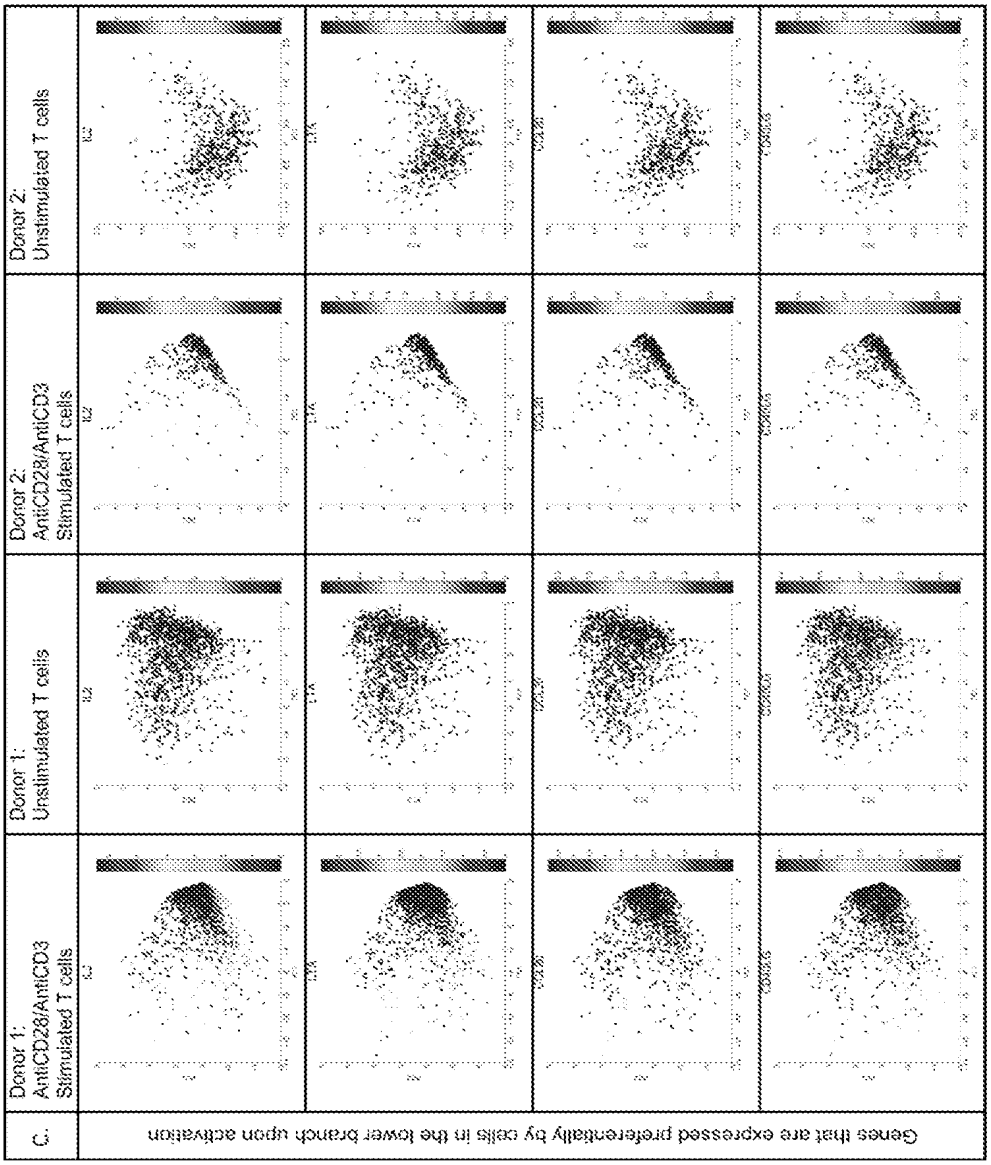

c.
Genes that are expressed preferentially by cells in the lower branch upon activation
Donor 1:
Anti-CD28/Anti-CD3
Stimulated T cells
Donor 1:
Unstimulated T cells
Donor 2:
Anti-CD28/Anti-CD3
Stimulated T cells
Donor 2:
Unstimulated T cells

FIG. 46C

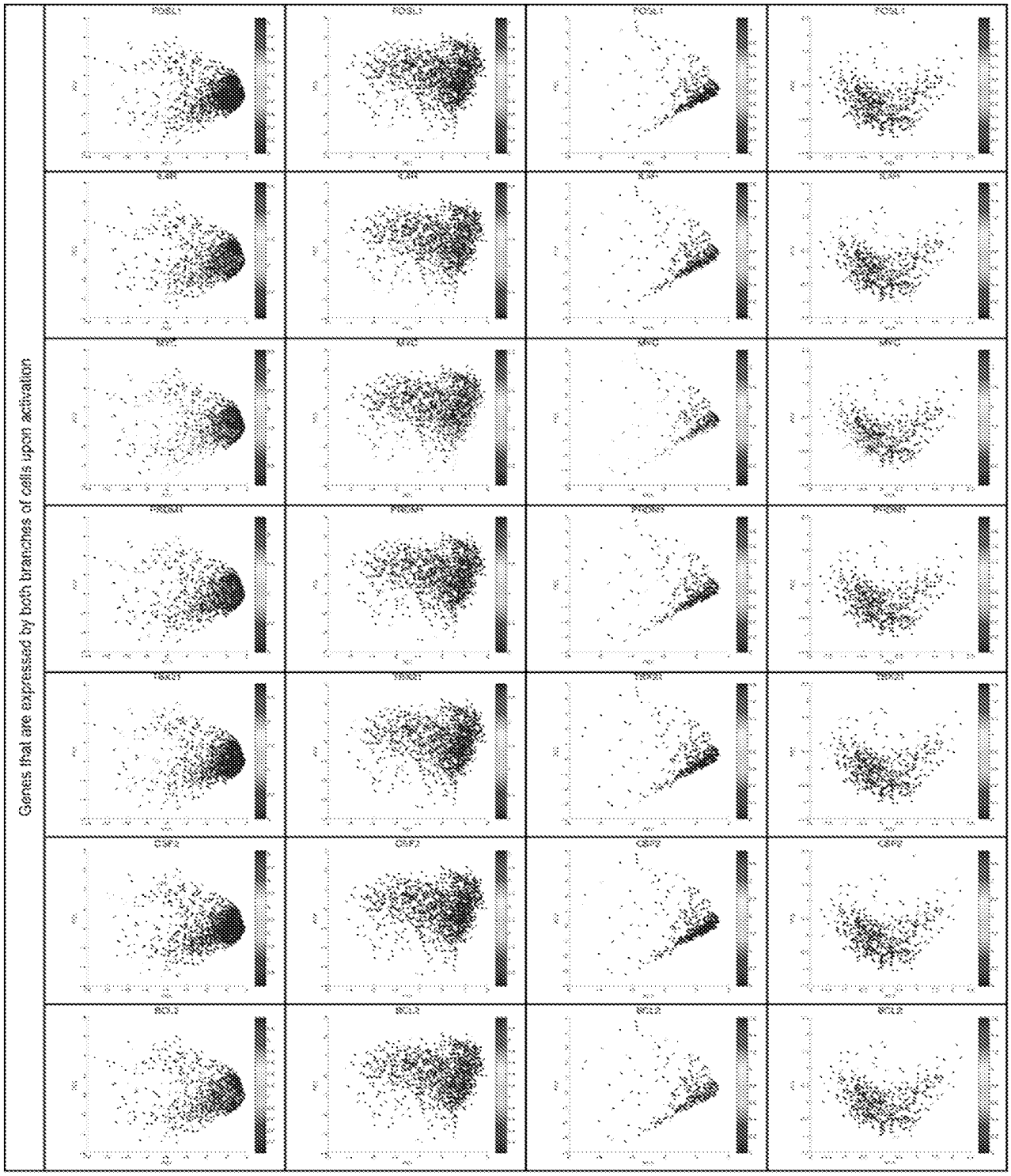

Genes that are expressed by both branches of cells upon activation

FIG. 46D

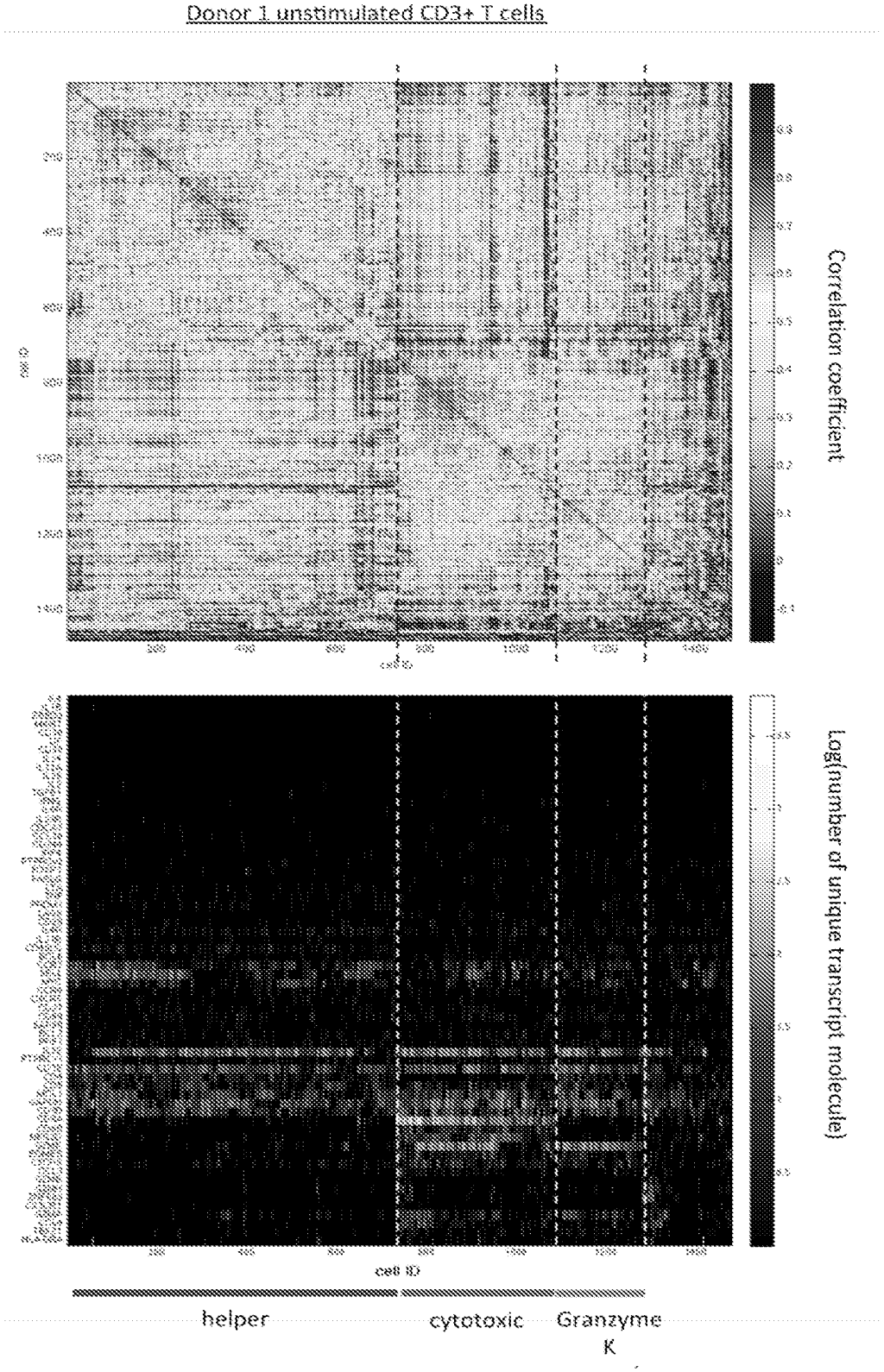

Donor 1 unstimulated CD3+ T cells
Correlation coefficient
Log(number of unique transcript molecule)
cell ID
helper
cytotoxic
Granzyme K

FIG. 47

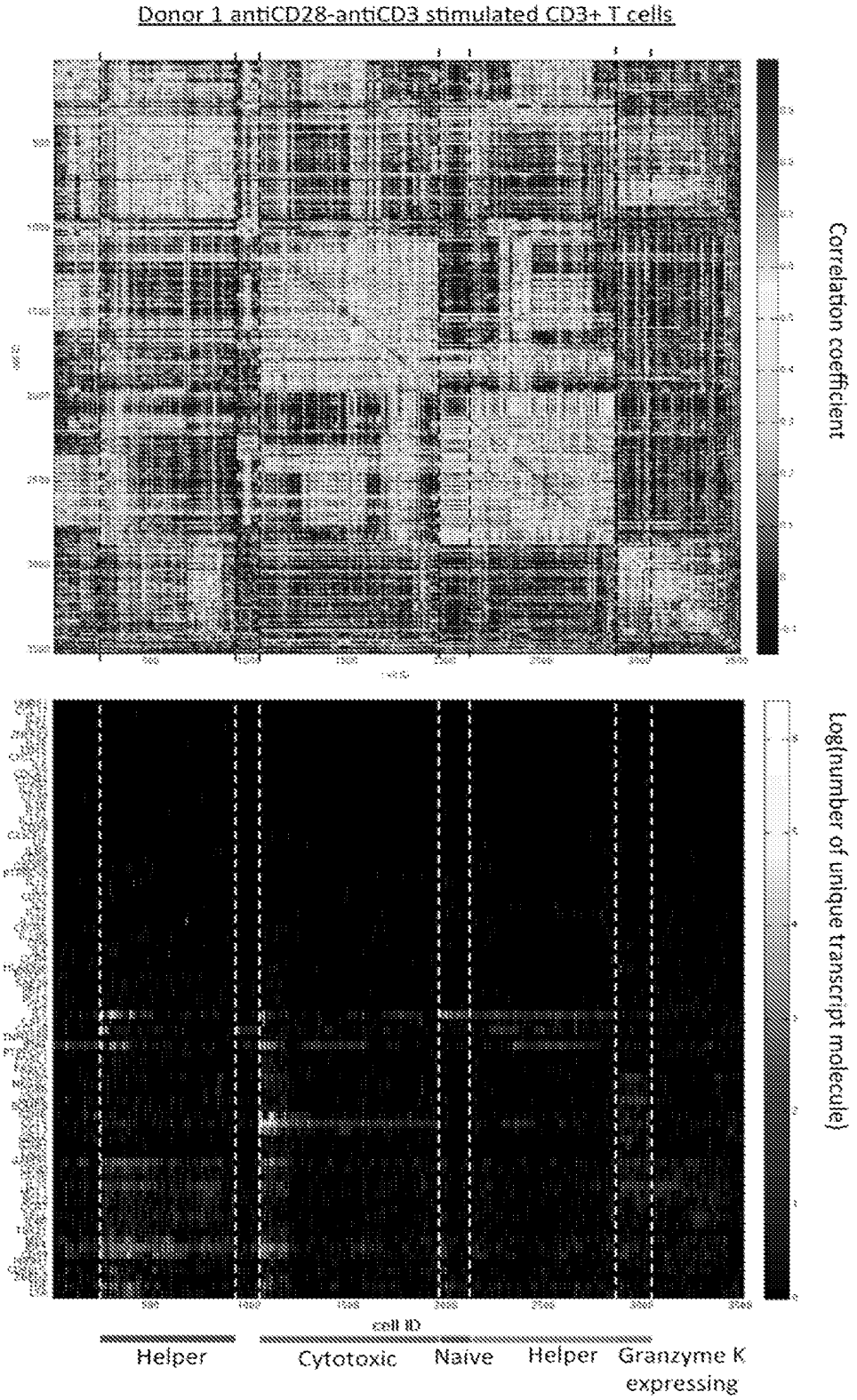

Donor 1 antiCD28-antiCD3 stimulated CD3+ T cells
Correlation coefficient
Log(number of unique transcript molecule)
cell ID
Helper
Cytotoxic
Naive
Helper
Granzyme K expressing

FIG. 48

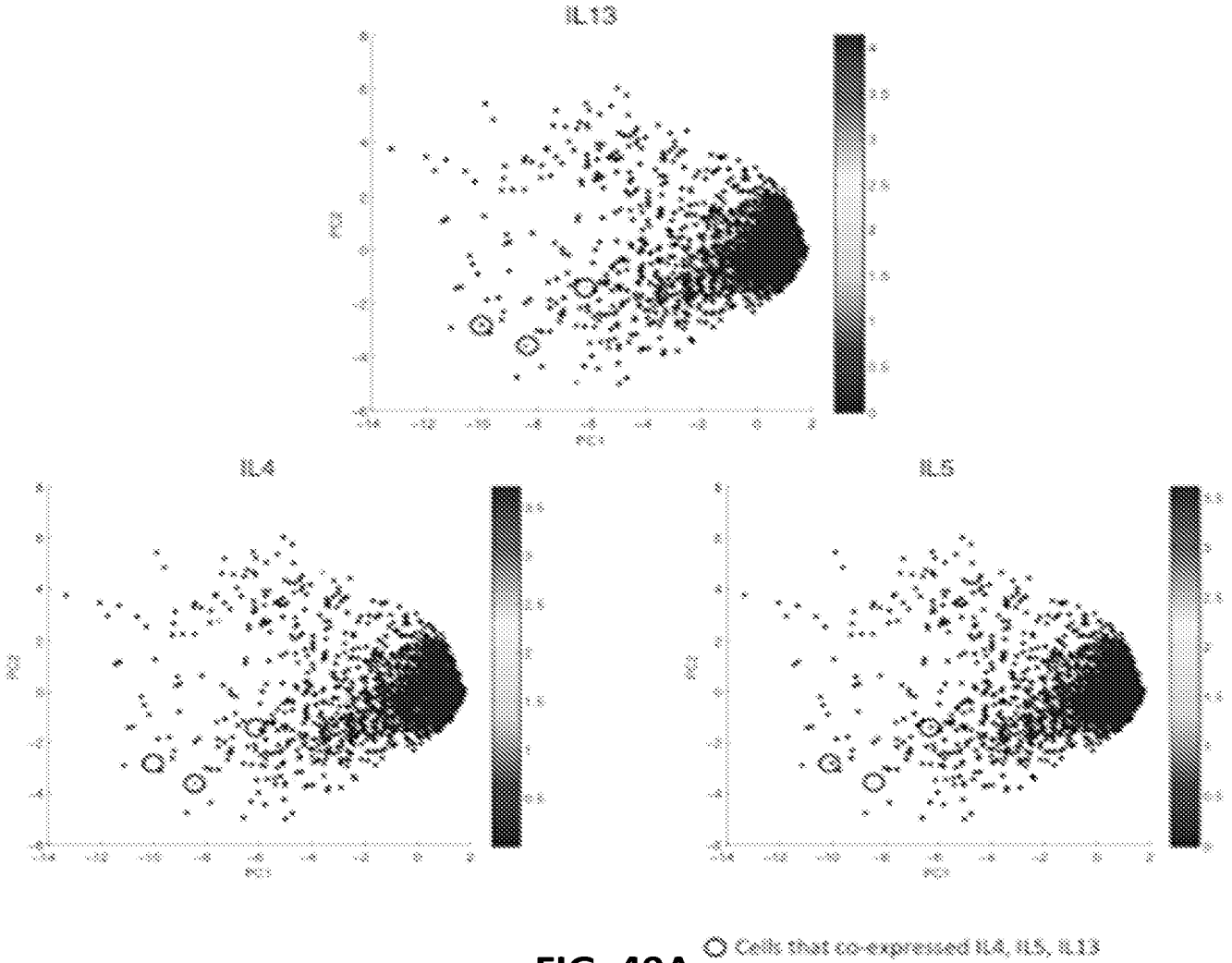


FIG. 49A

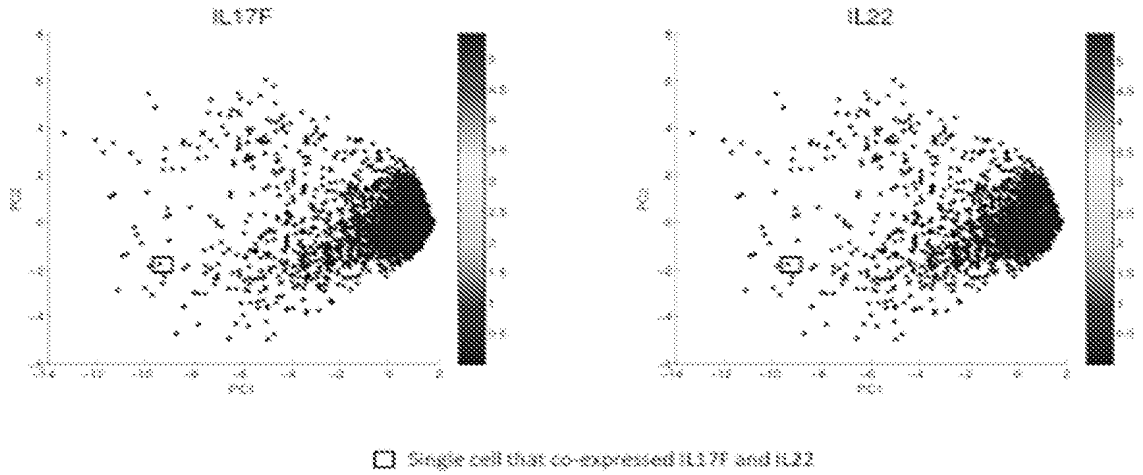


FIG. 49B

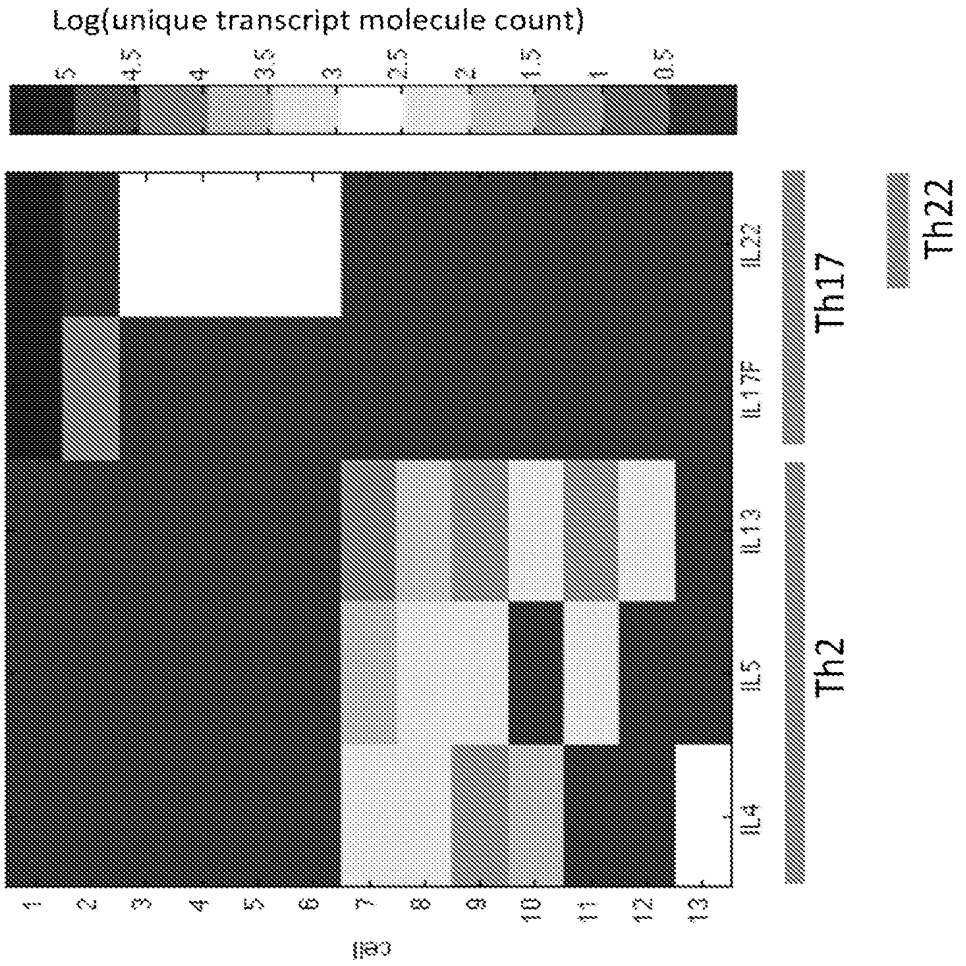

Log(unique transcript molecule count)
5
4.5
4
3.5
3
2.5
2
1.5
1
0.5
IL22
IL17F
IL13
IL5
IL4
Th17
Th22
Th2
1
2
3
4
5
6
7
8
9
10
11
12
13

FIG. 49C

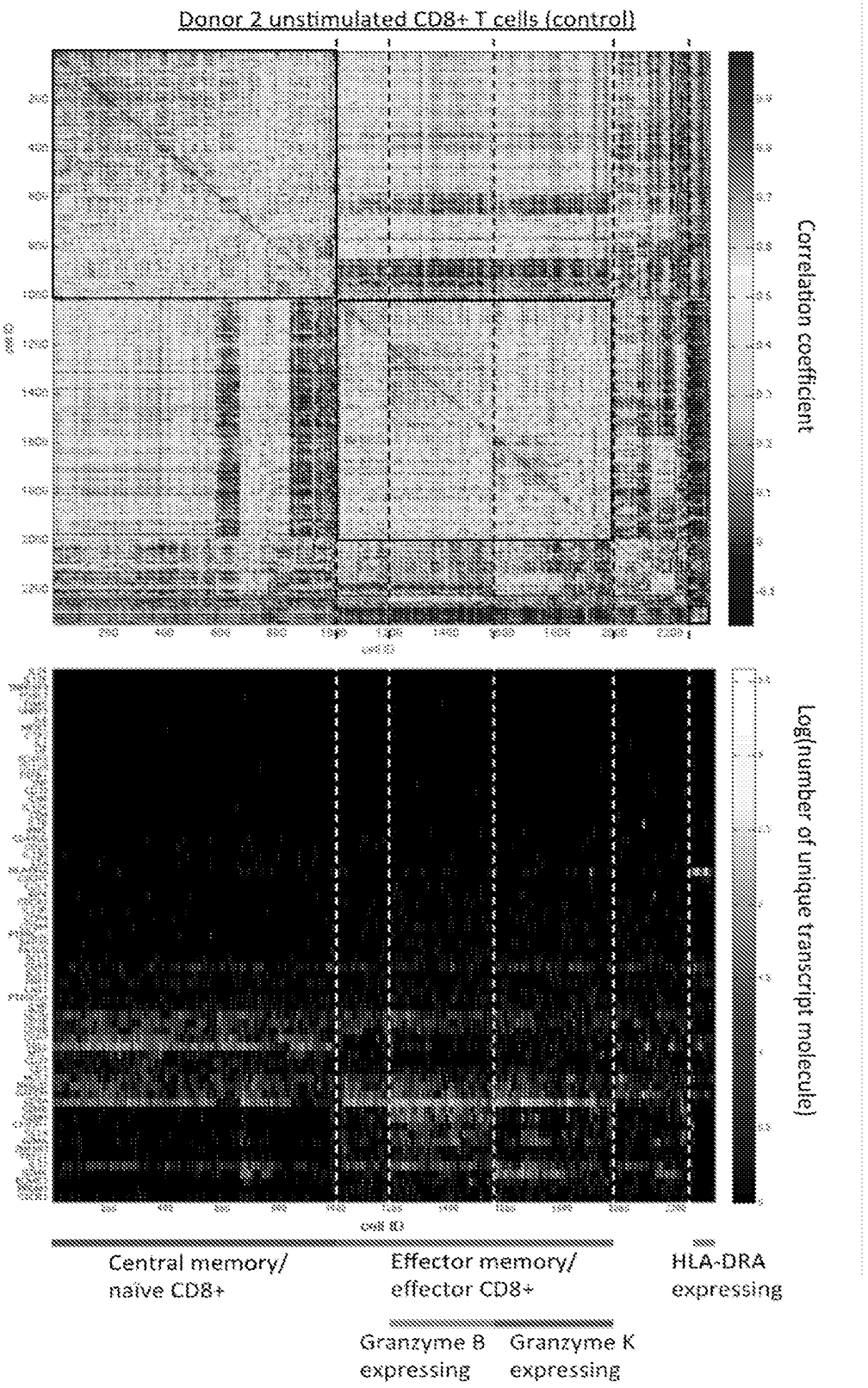

Donor 2 unstimulated CD8+ T cells (control)
Correlation coefficient
Log(number of unique transcript molecule)
Central memory/
naïve CD8+
Effector memory/
effector CD8+
HLA-DRA
expressing
Granzyme B
expressing
Granzyme K
expressing

FIG. 50A

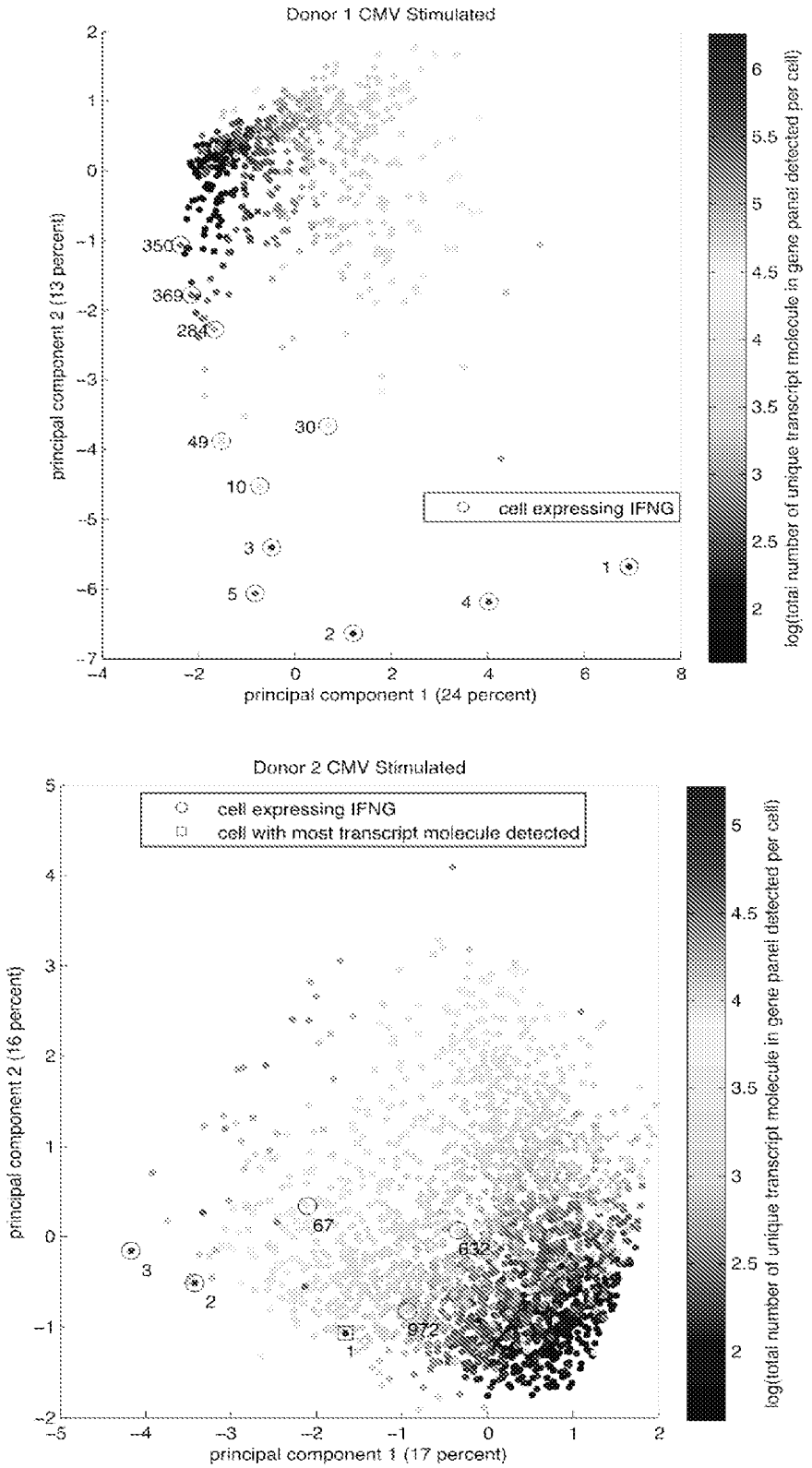

Donor 1 CMV Stimulated
principal component 2 (13 percent)
principal component 1 (24 percent)
cell expressing IFNG
log(total number of unique transcript molecule in gene panel detected per cell)
Donor 2 CMV Stimulated
cell expressing IFNG
cell with most transcript molecule detected
principal component 2 (16 percent)
principal component 1 (17 percent)
log(total number of unique transcript molecule in gene panel detected per cell)

FIG. 50B

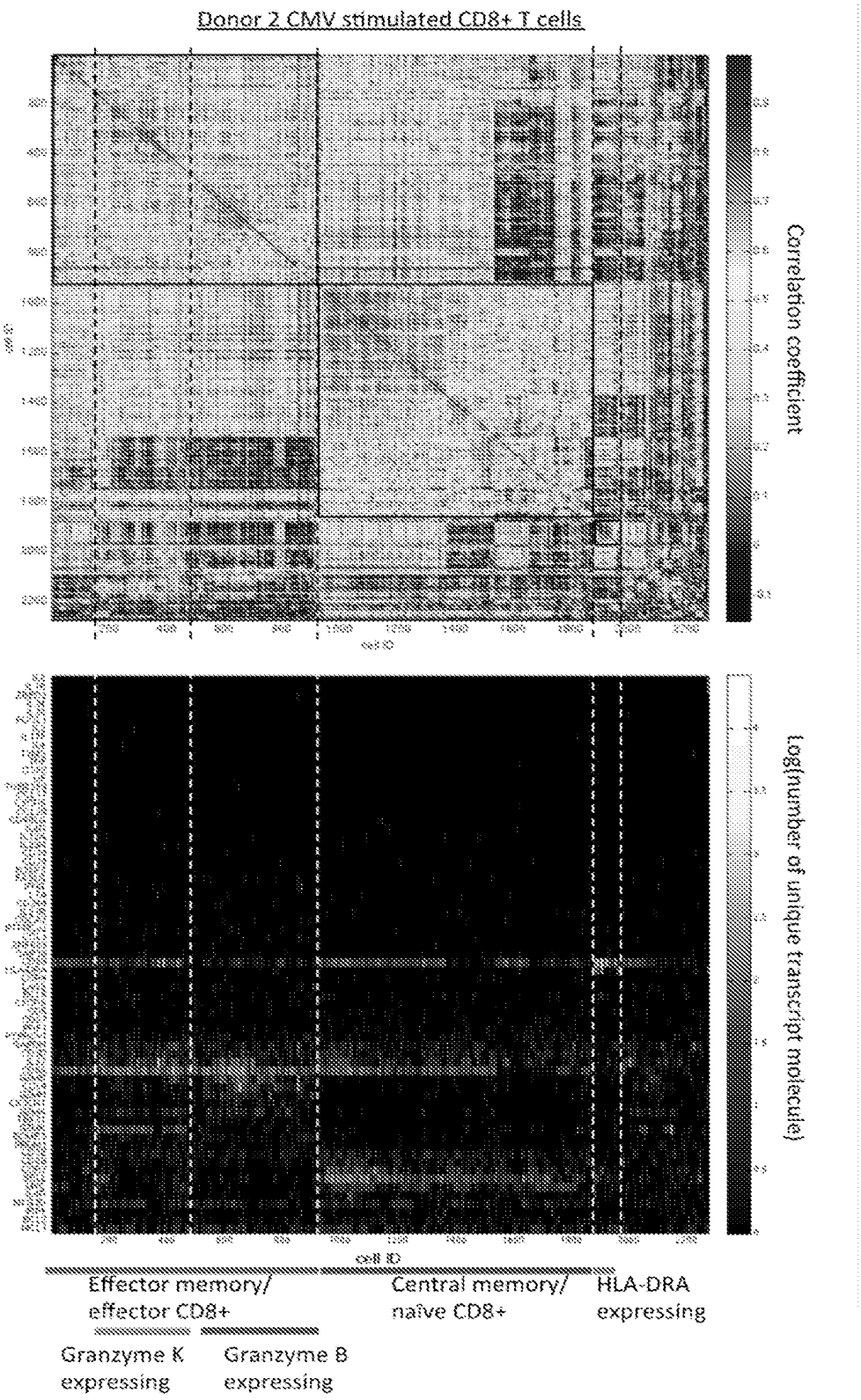

Donor 2 CMV stimulated CD8+ T cells
Correlation coefficient
Log(number of unique transcript molecule)
Effector memory/
effector CD8+
Central memory/
naïve CD8+
HLA-DRA
expressing
Granzyme K
expressing
Granzyme B
expressing

FIG. 51

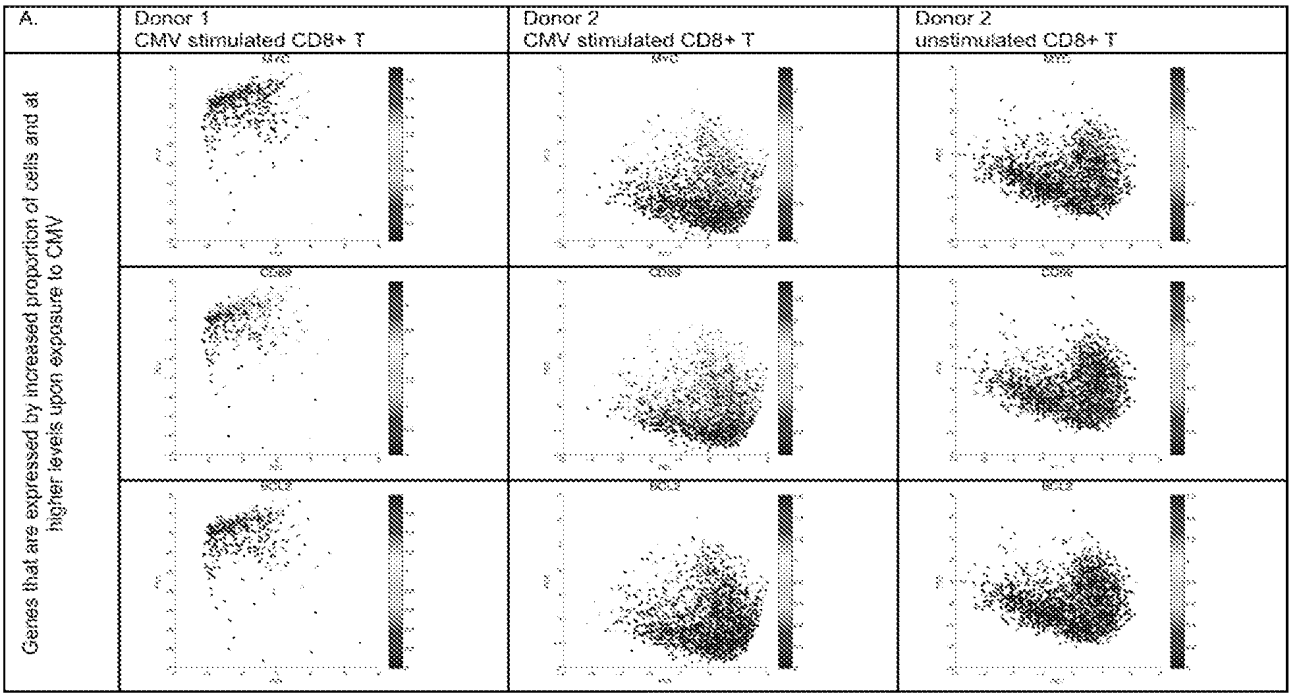


FIG. 52A

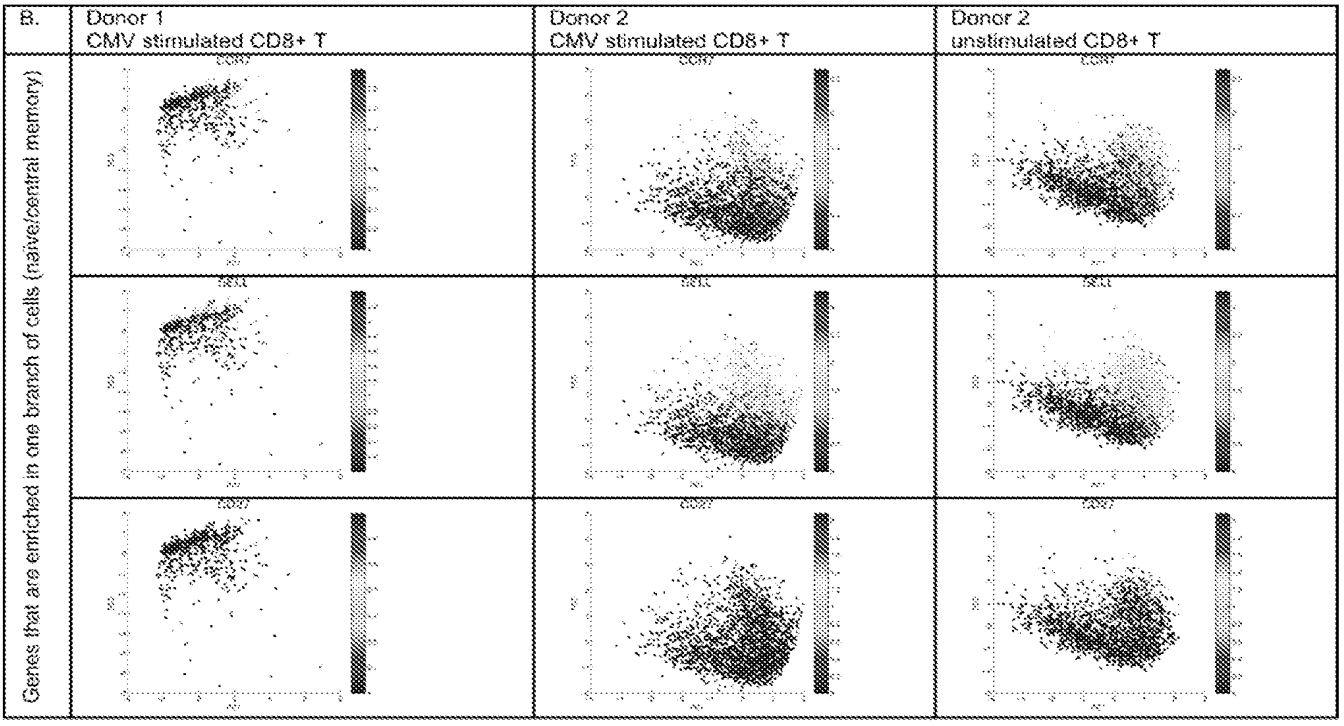


FIG. 52B

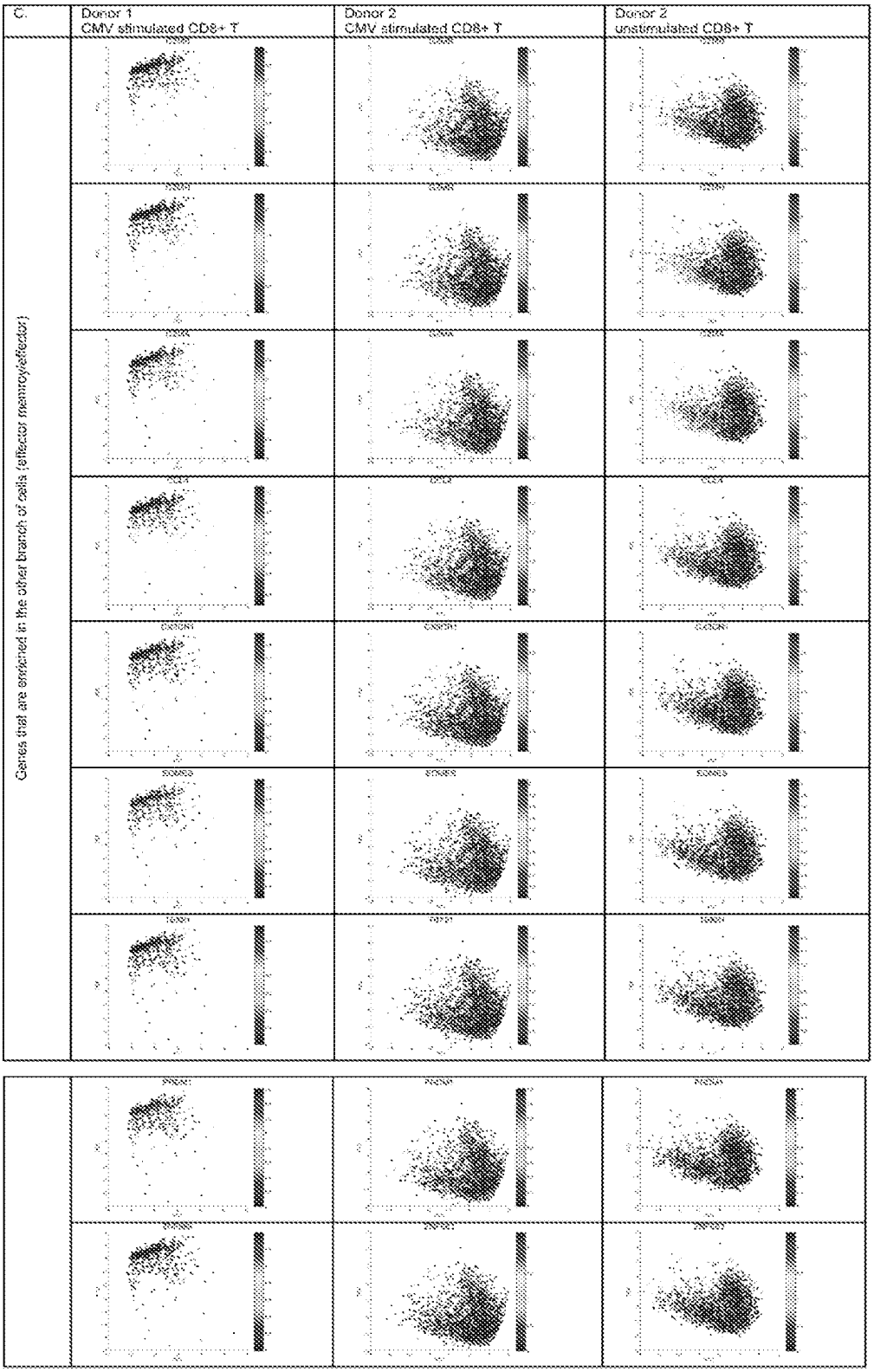

C.
Donor 1
CMV stimulated CD8+ T
Donor 2
CMV stimulated CD8+ T
Donor 2
unstimulated CD8+ T
Genes that are enriched in the other branch of cells (effector memory/effector)

FIG. 52C

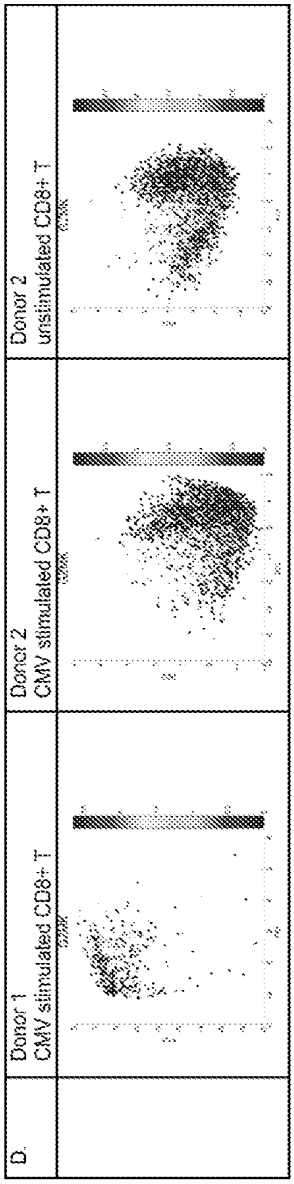


FIG. 52D

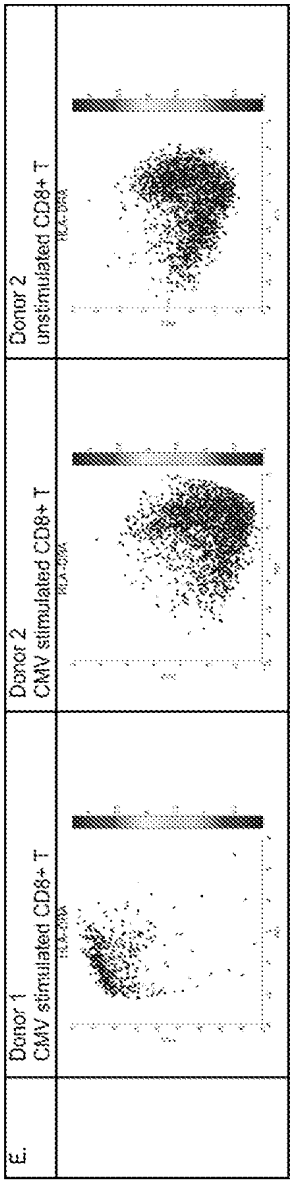


FIG. 52E

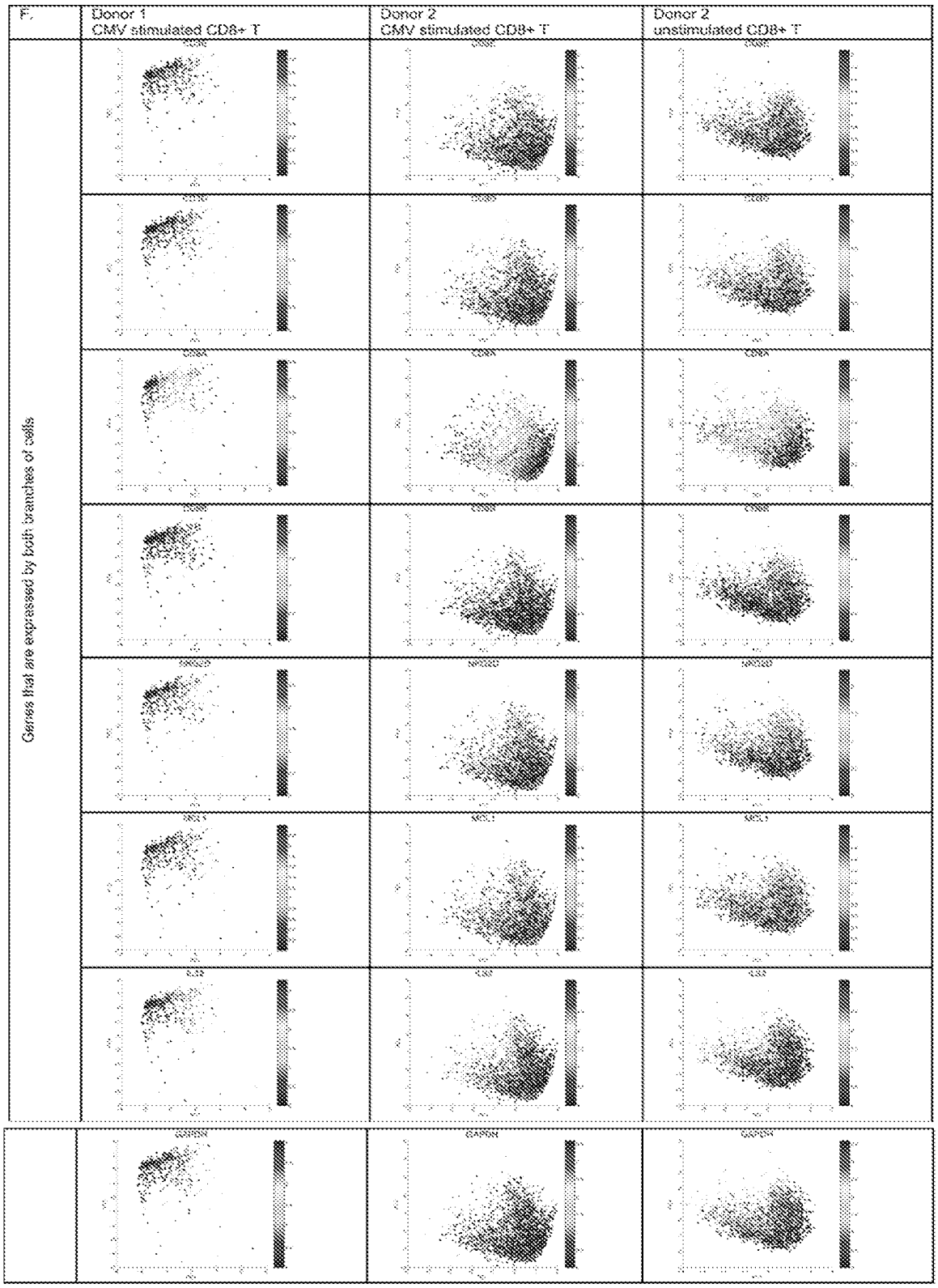

F.
Donor 1
CMV stimulated CD8+ T
Donor 2
CMV stimulated CD8+ T
Donor 2
unstimulated CD8+ T
Genes that are expressed by both branches of cells

FIG. 52F

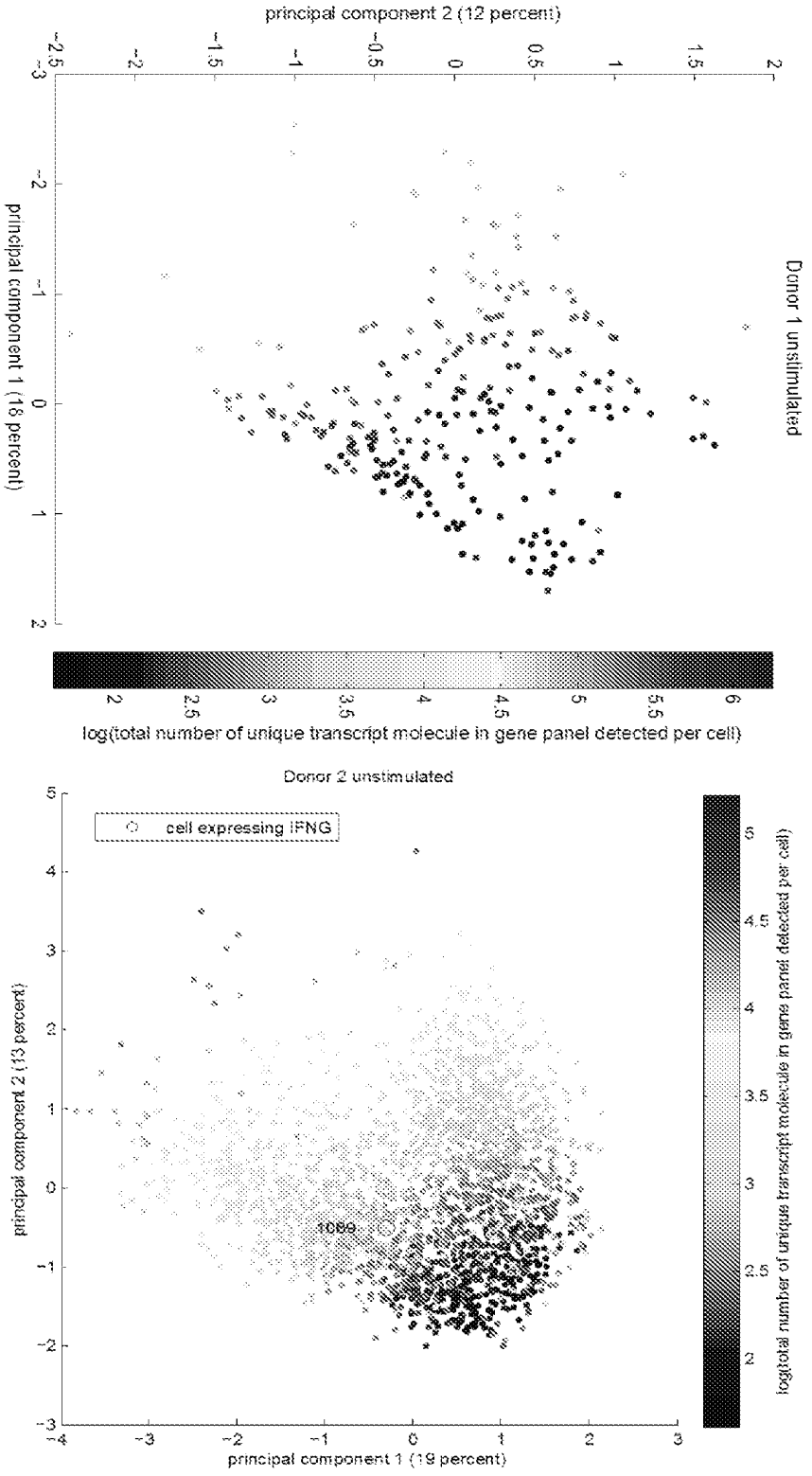

principal component 2 (12 percent)
principal component 1 (18 percent)
Donor 1 unstimulated
log(total number of unique transcript molecule in gene panel detected per cell)
Donor 2 unstimulated
cell expressing IFNG
1089
principal component 2 (13 percent)
principal component 1 (19 percent)
log(total number of unique transcript molecule in gene panel detected per cell)

FIG. 53

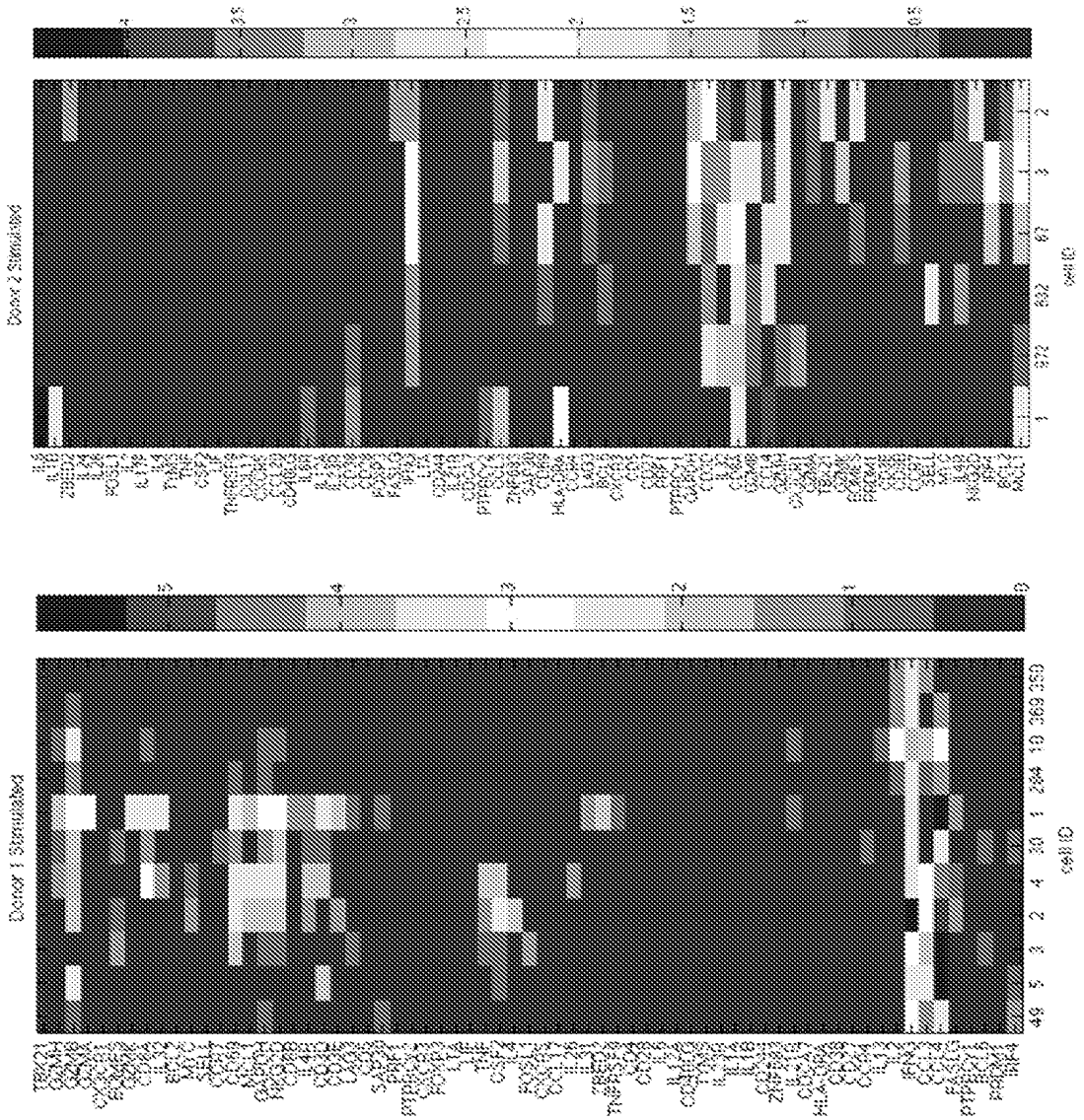


FIG. 54

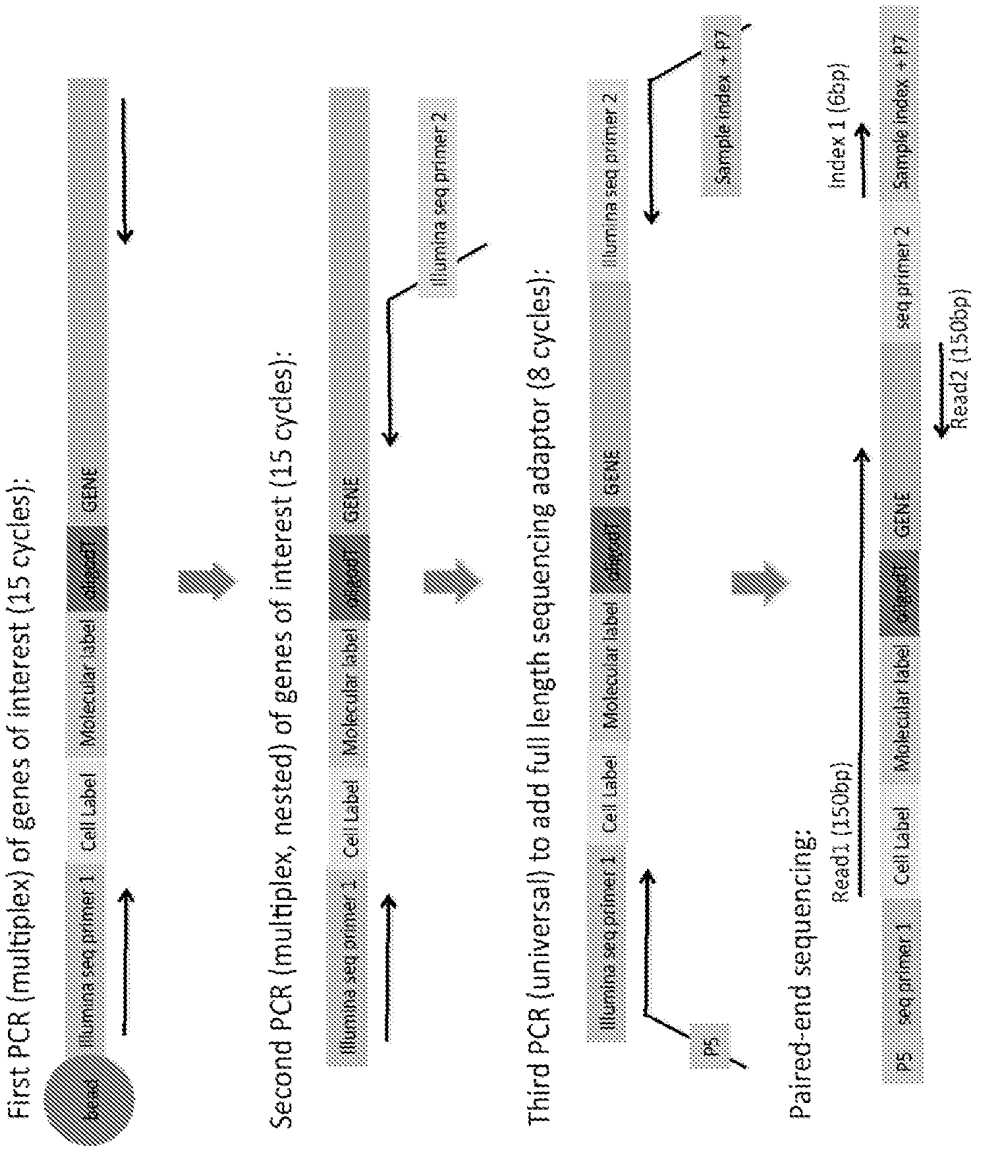

First PCR (multiplex) of genes of interest (15 cycles):
Read1
Illumina seq primer 1
Cell Label
Molecular label
Barcode
GENE
Second PCR (multiplex, nested) of genes of interest (15 cycles):
Illumina seq primer 1
Cell Label
Molecular label
Barcode
GENE
Illumina seq primer 2
Third PCR (universal) to add full length sequencing adaptor (8 cycles):
Illumina seq primer 1
Cell Label
Molecular label
Barcode
GENE
Illumina seq primer 2
P5
Sample index + P7
Paired-end sequencing:
Read1 (150bp)
Index 1 (6bp)
P5
seq primer 1
Cell Label
Molecular label
Barcode
GENE
seq primer 2
Sample index + P7
Read2 (150bp)

FIG. 55

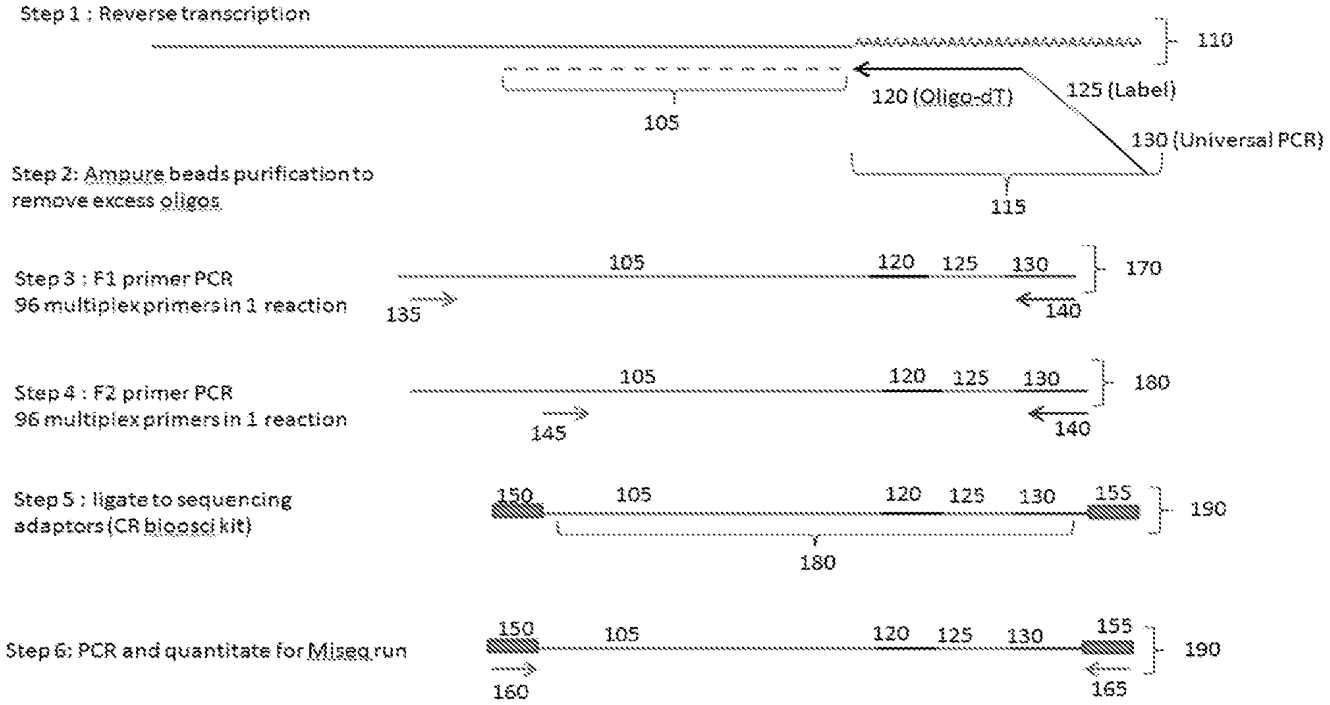

Step 1 : Reverse transcription
110
105
120 (Oligo-dT)
125 (Label)
130 (Universal PCR)
115
Step 2: Ampure beads purification to remove excess oligos
Step 3 : F1 primer PCR
96 multiplex primers in 1 reaction
105
120
125
130
170
135
140
Step 4 : F2 primer PCR
96 multiplex primers in 1 reaction
105
120
125
130
180
145
140
Step 5 : ligate to sequencing adaptors (CR bioosci kit)
150
105
120
125
130
155
190
180
Step 6: PCR and quantitate for Miseq run
150
105
120
125
130
155
190
160
165

FIG. 56

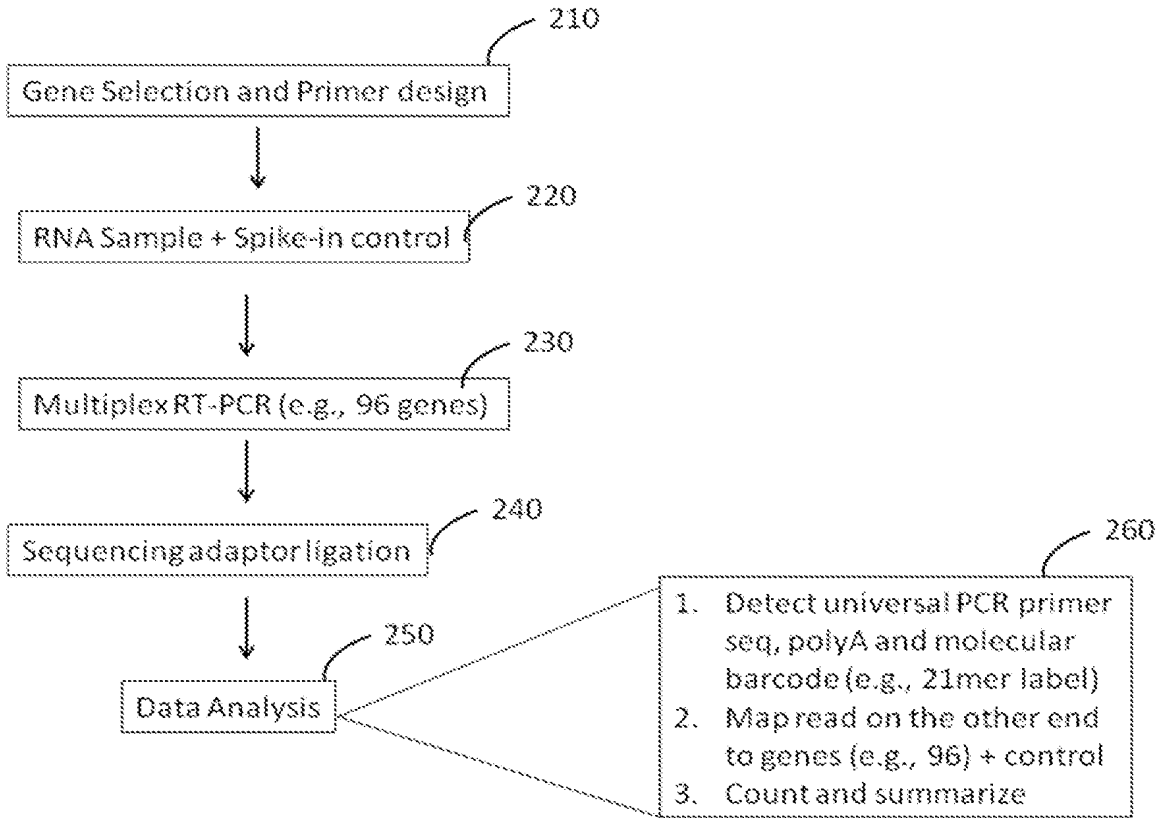

210
Gene Selection and Primer design
220
RNA Sample + Spike-in control
230
Multiplex RT-PCR (e.g., 96 genes)
240
Sequencing adaptor ligation
250
Data Analysis
260
1. Detect universal PCR primer seq, polyA and molecular barcode (e.g., 21mer label)
2. Map read on the other end to genes (e.g., 96) + control
3. Count and summarize

FIG. 57

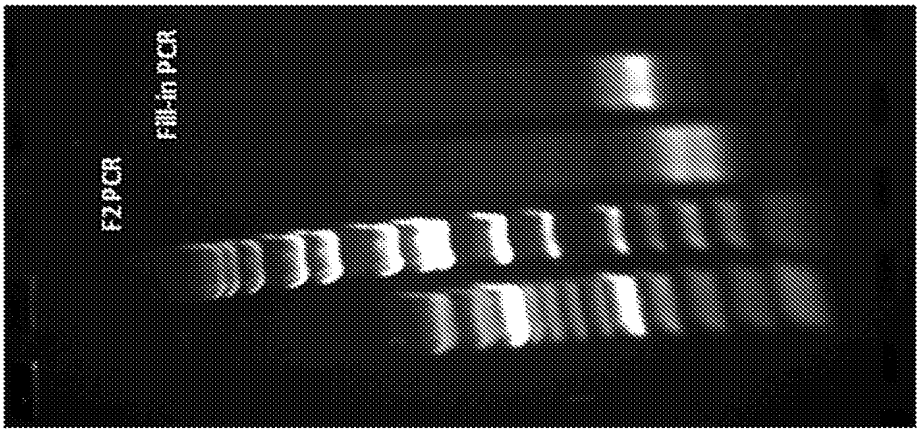


FIG. 58B

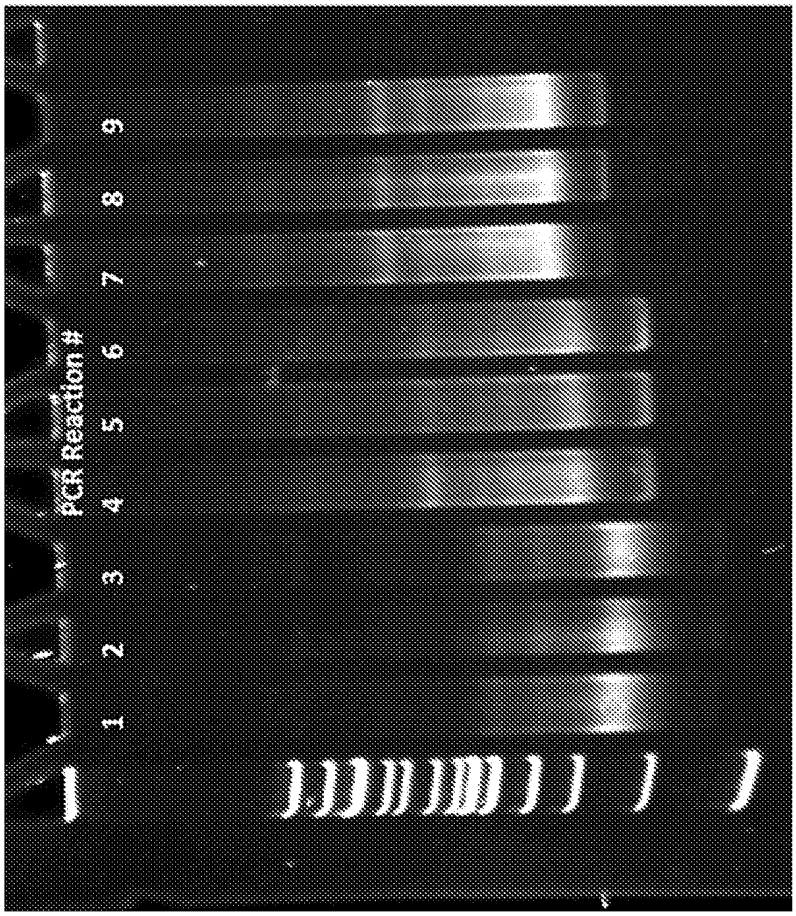


FIG. 58A

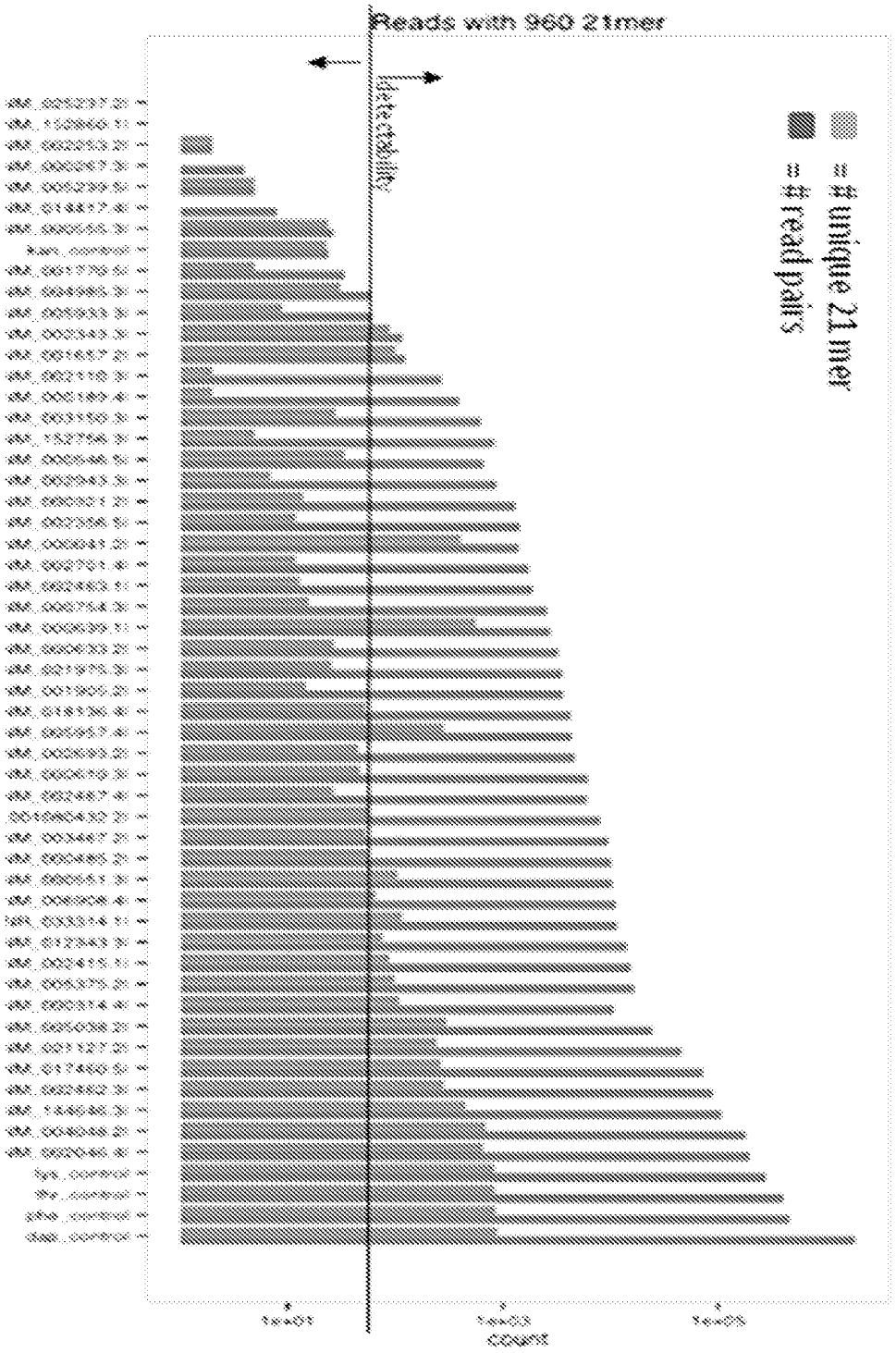


FIG. 59

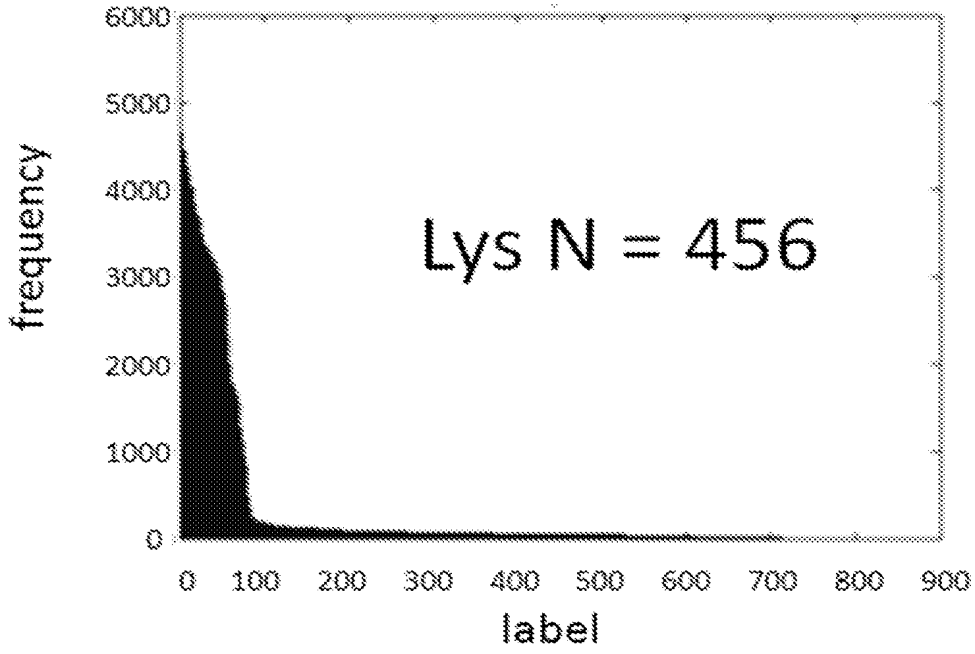


FIG. 60A

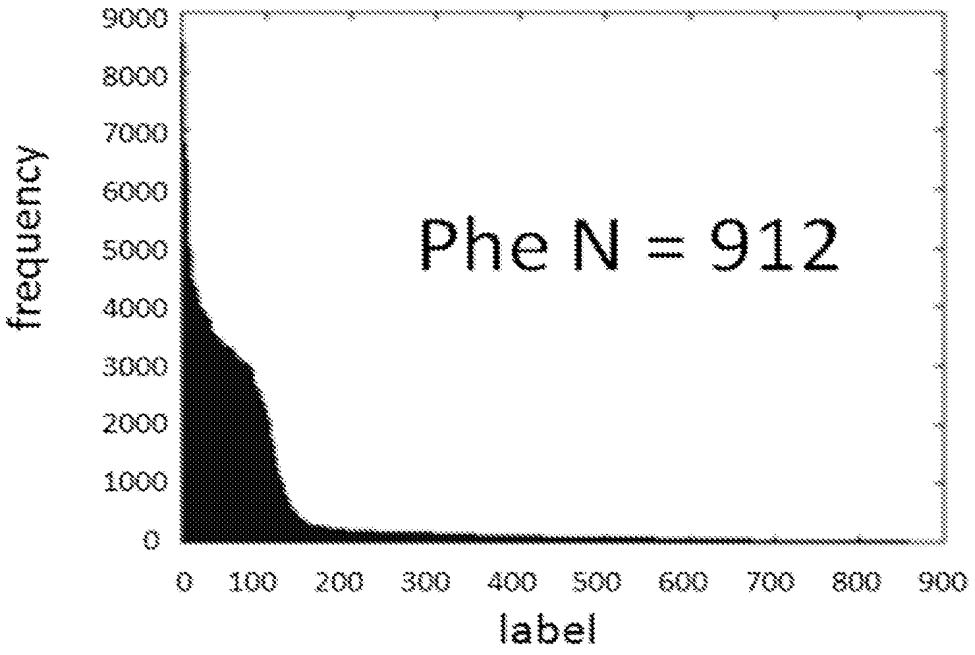


FIG. 60B

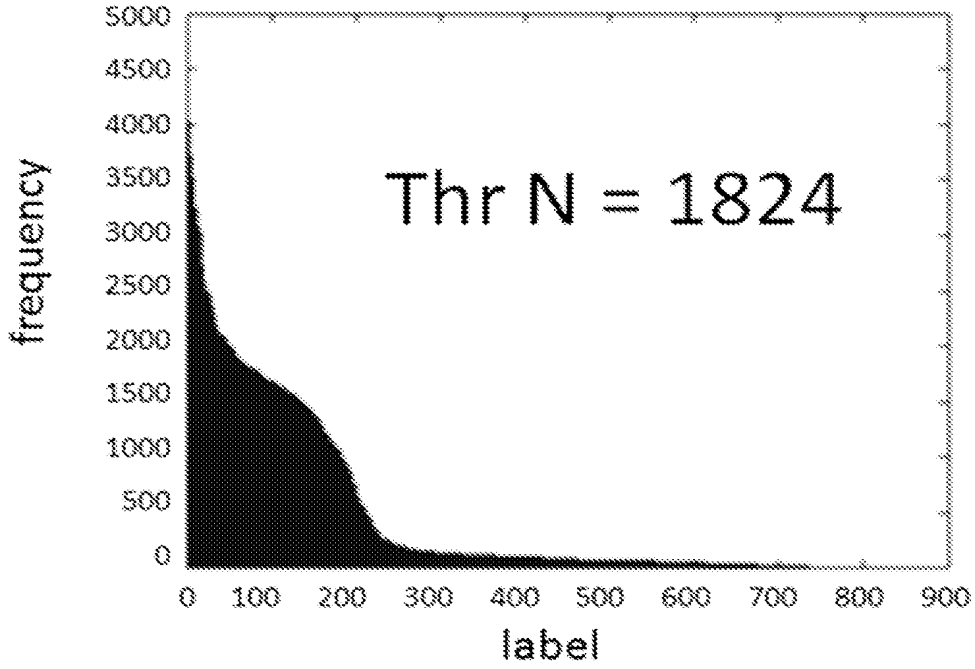


FIG. 60C

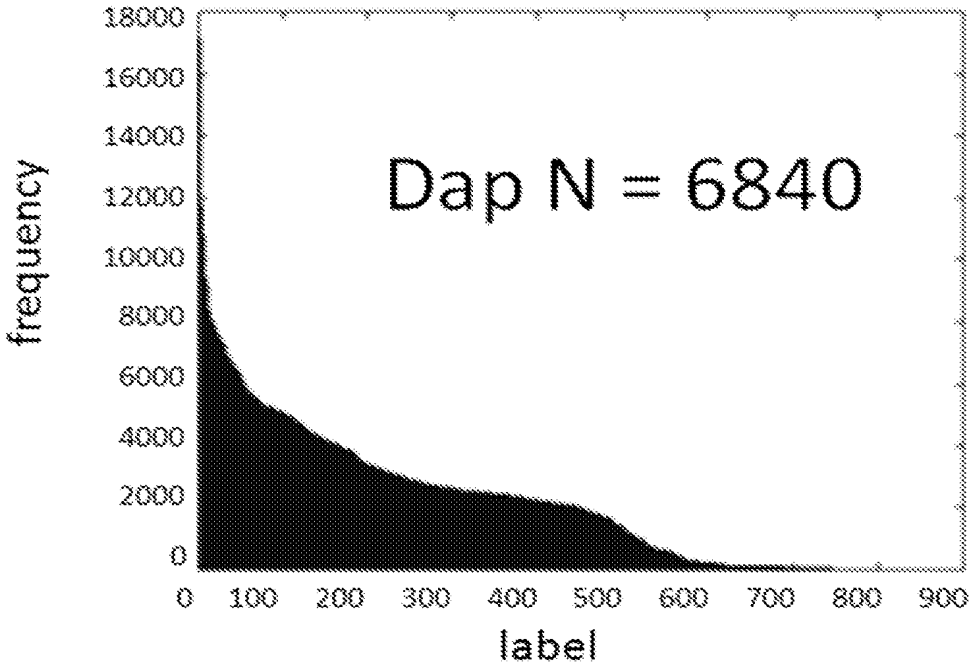


FIG. 60D

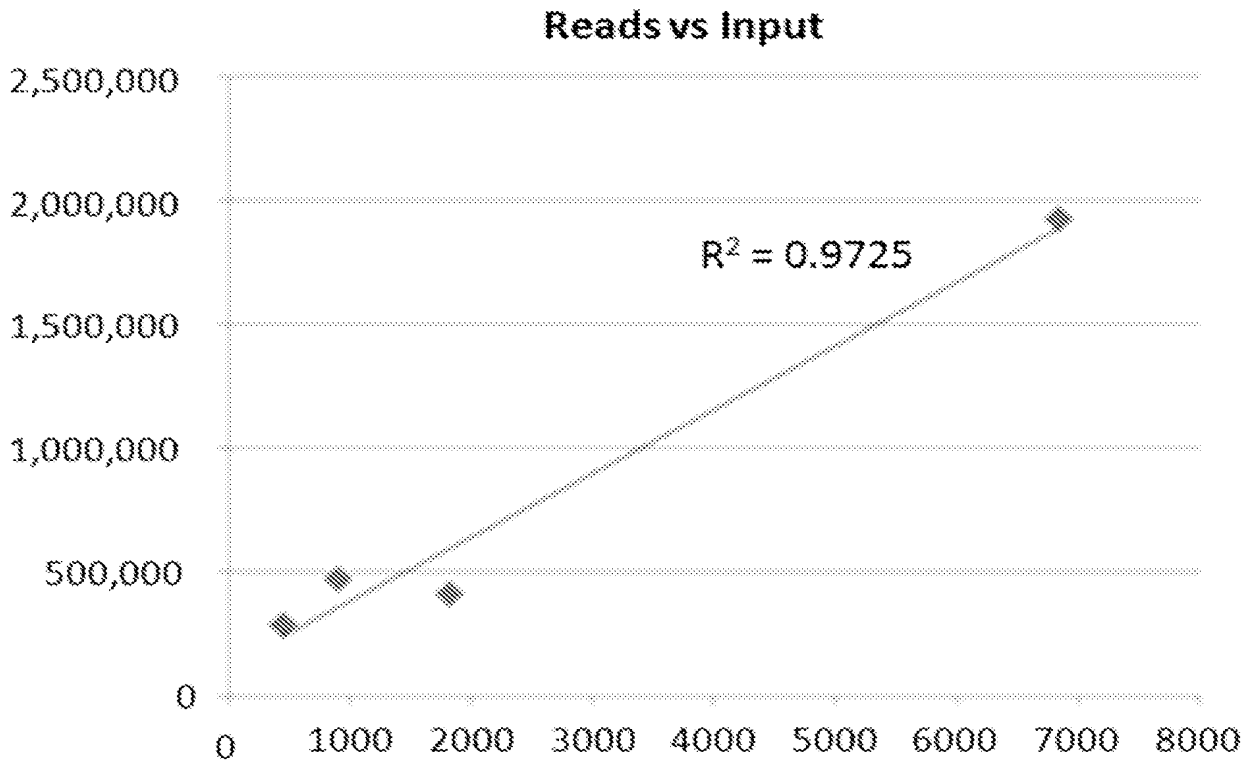


FIG. 60E

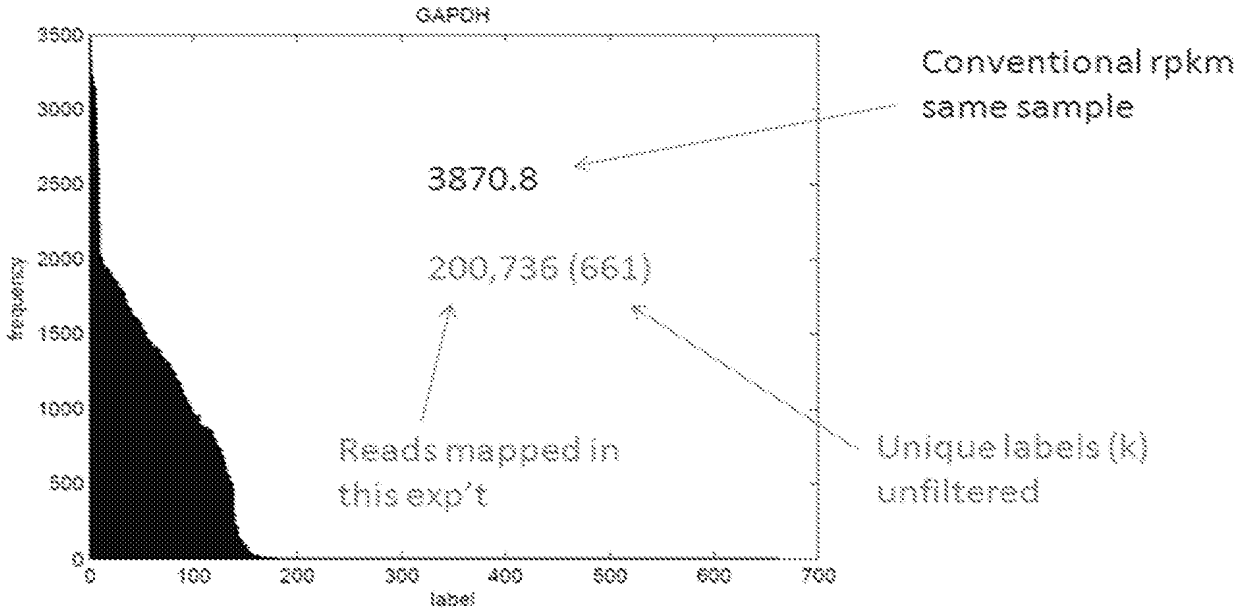


FIG. 61

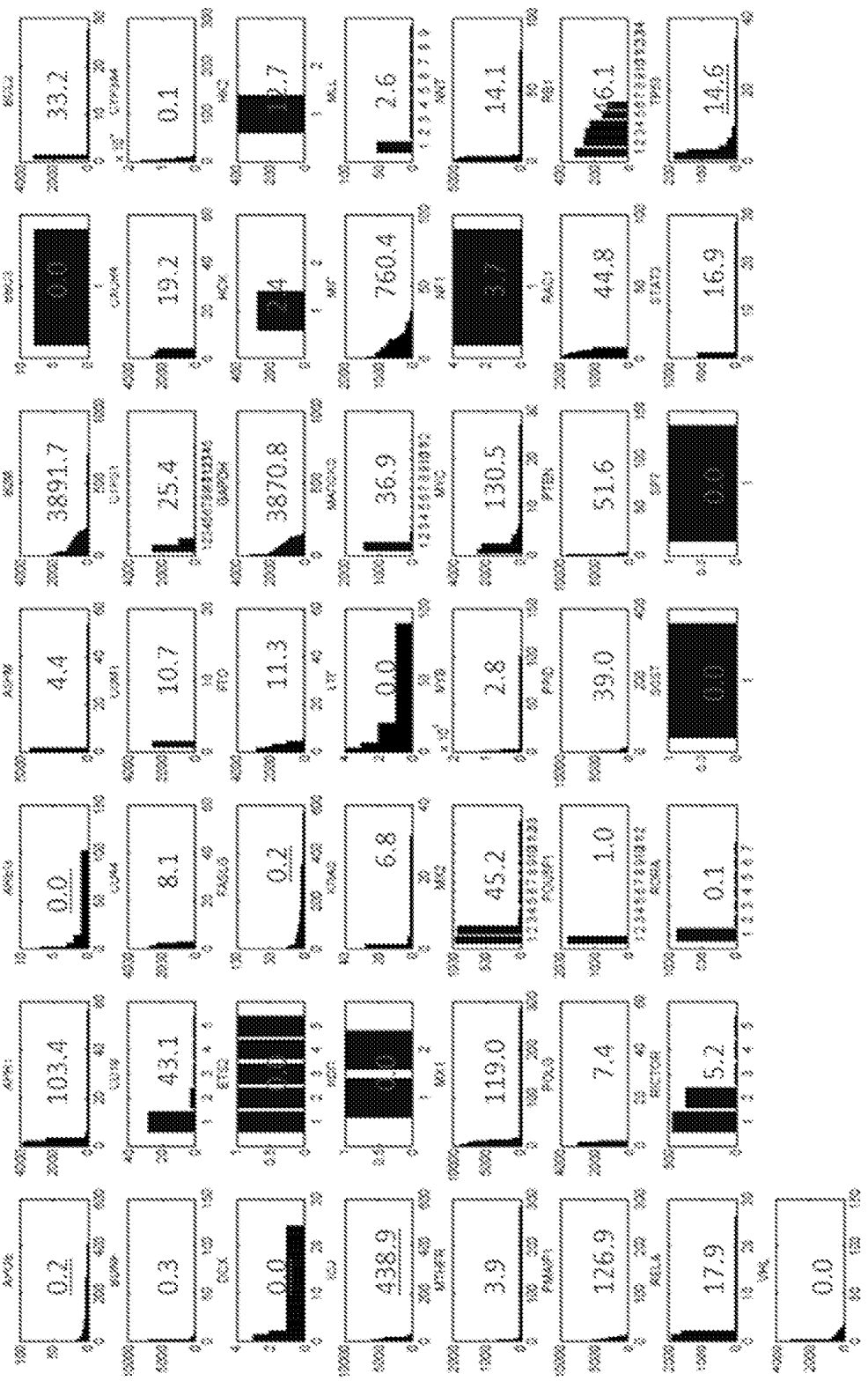


FIG. 62

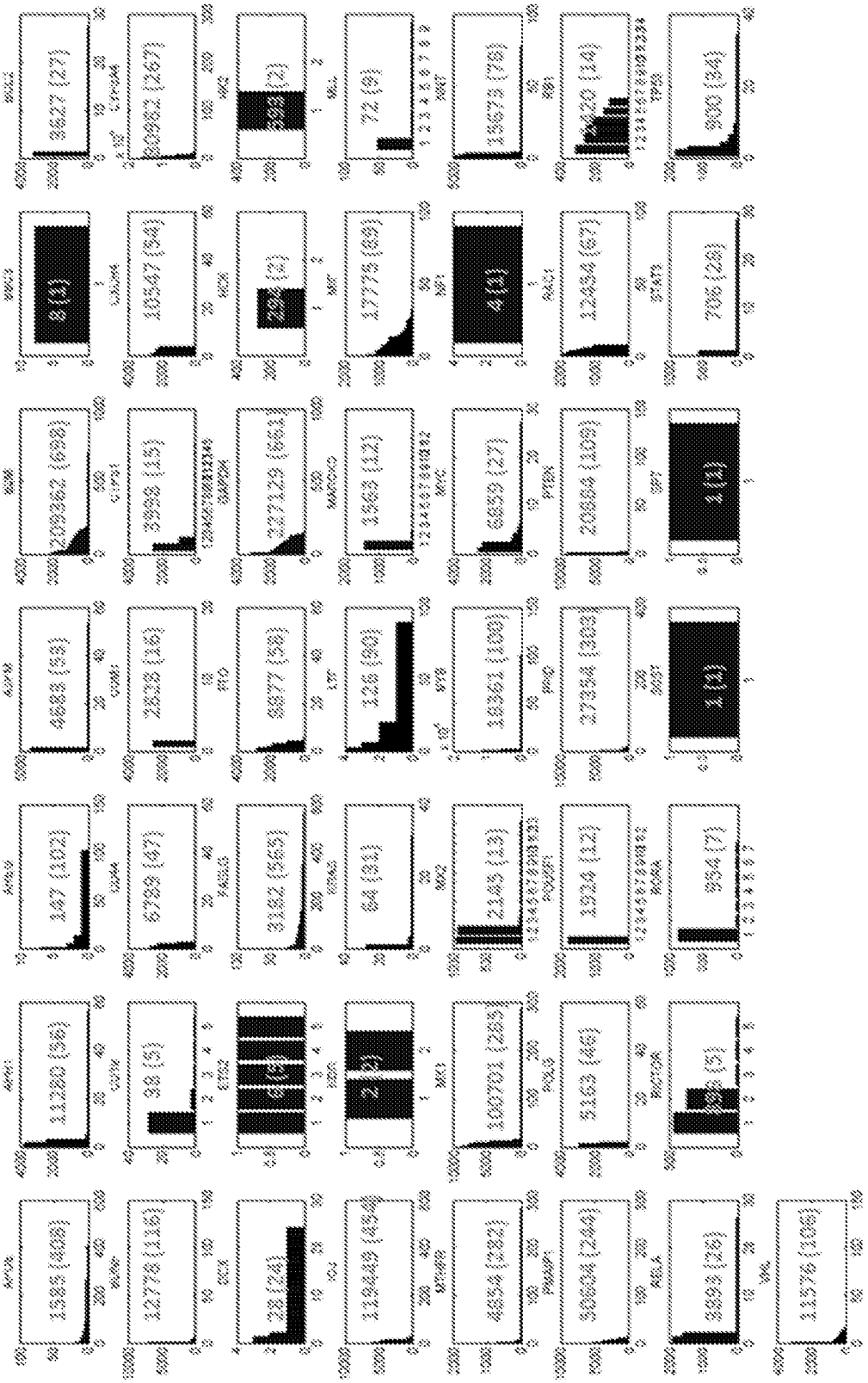


FIG. 63

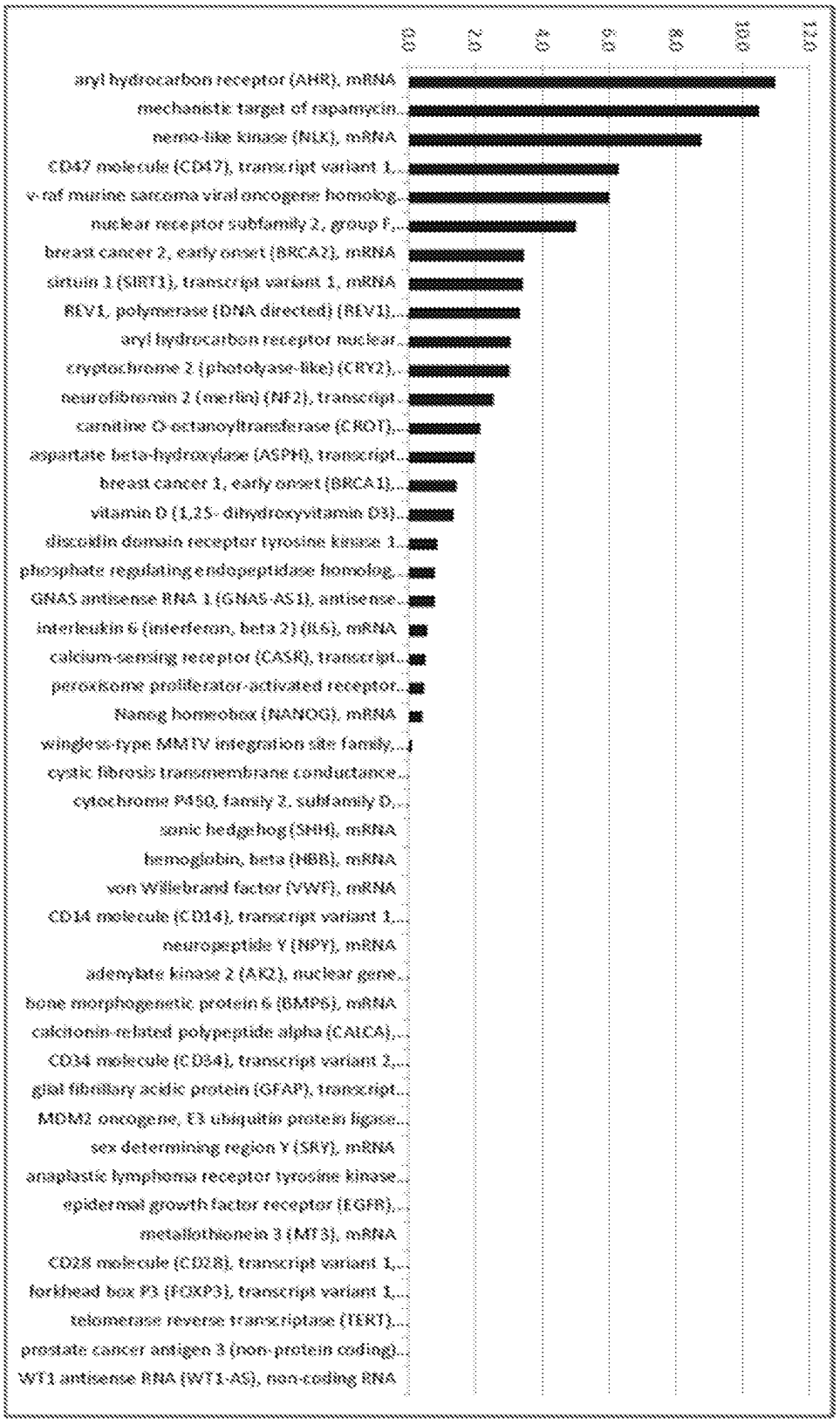

0.0
2.0
4.0
6.0
8.0
10.0
12.0
aryl hydrocarbon receptor (AHR), mRNA
mechanistic target of rapamycin...
nemo-like kinase (NLK), mRNA
CD47 molecule (CD47), transcript variant 1...
v-raf murine sarcoma viral oncogene homolog...
nuclear receptor subfamily 2, group F,...
breast cancer 2, early onset (BRCA2), mRNA
sirtuin 3 (SIRT3), transcript variant 1, mRNA
REV1, polymerase (DNA directed) (REV1),...
aryl hydrocarbon receptor nuclear...
cryptochrome 2 (photolyase-like) (CRY2),...
neurofibromin 2 (merlin) (NF2), transcript...
carnitine O-octanoyltransferase (CROT),...
aspartate beta-hydroxylase (ASPH), transcript...
breast cancer 1, early onset (BRCA1),...
vitamin D (1,25- dihydroxyvitamin D3)...
discoidin domain receptor tyrosine kinase 1...
phosphate regulating endopeptidase homolog,...
GNAS antisense RNA 1 (GNAS-AS1), antisense...
interleukin 6 (interferon, beta 2) (IL6), mRNA
calcium-sensing receptor (CASR), transcript...
peroxisome proliferator-activated receptor...
Nanog homeobox (NANOG), mRNA
wingless-type MMTV integration site family,...
cystic fibrosis transmembrane conductance...
cytochrome P450, family 2, subfamily D,...
sonic hedgehog (SHH), mRNA
hemoglobin, beta (HBB), mRNA
von Willebrand factor (VWF), mRNA
CD14 molecule (CD14), transcript variant 1,...
neuropeptide Y (NPY), mRNA
adenylate kinase 2 (AK2), nuclear gene...
bone morphogenetic protein 6 (BMP6), mRNA
calcitonin-related polypeptide alpha (CALCA),...
CD34 molecule (CD34), transcript variant 2,...
glial fibrillary acidic protein (GFAP), transcript...
MDM2 oncogene, E3 ubiquitin protein ligase...
sex determining region Y (SRY), mRNA
anaplastic lymphoma receptor tyrosine kinase...
epidermal growth factor receptor (EGFR),...
metallothionein 3 (MT3), mRNA
CD28 molecule (CD28), transcript variant 1,...
forkhead box P3 (FOXP3), transcript variant 1,...
telomerase reverse transcriptase (TERT),...
prostate cancer antigen 3 (non-protein coding)...
WT1 antisense RNA (WT1-AS), non-coding RNA

FIG. 64

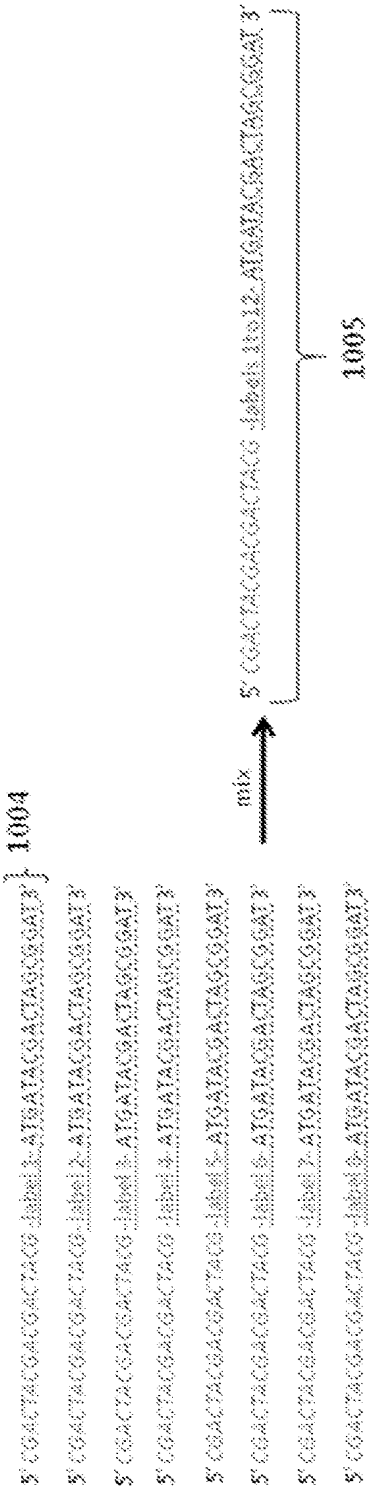

5' CGACTACGACGACTACG label 1 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 2 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 3 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 4 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 5 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 6 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 7 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 8 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 9 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 10 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 11 ATGATACGACTAGCGGAT 3'
5' CGACTACGACGACTACG label 12 ATGATACGACTAGCGGAT 3'
1004
mix
5' CGACTACGACGACTACG label 12 ATGATACGACTAGCGGAT 3'
1005
1001
1002
1003

FIG. 65

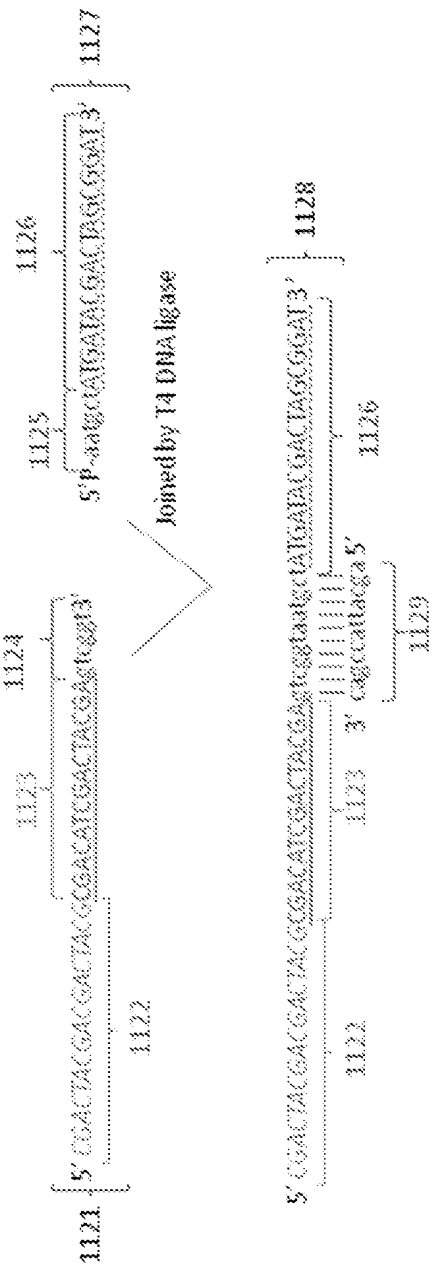

1121
5' CGACTACGACTACGGACATCGACTACGAtcgt3'
1122
1123
1124
1125
5'P-atgcATCGATGACTACTAGCGGAT3'
1126
1127
Joined by T4 DNA ligase
1128
5' CGACTACGACTACGCGACATCGACTACGAtcgtatgcTATCGATGACTAGCGGAT3'
1122
1123
3' gcatacgca 5'
1126
1129

FIG. 66A

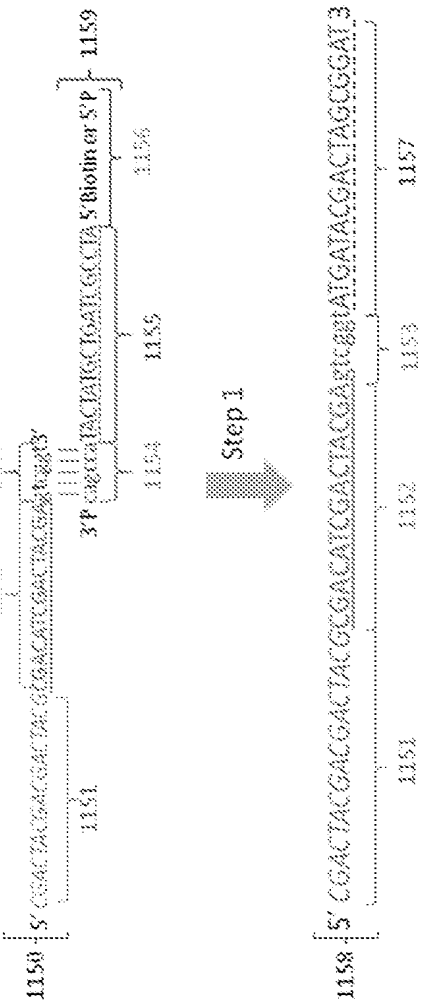


FIG. 66B

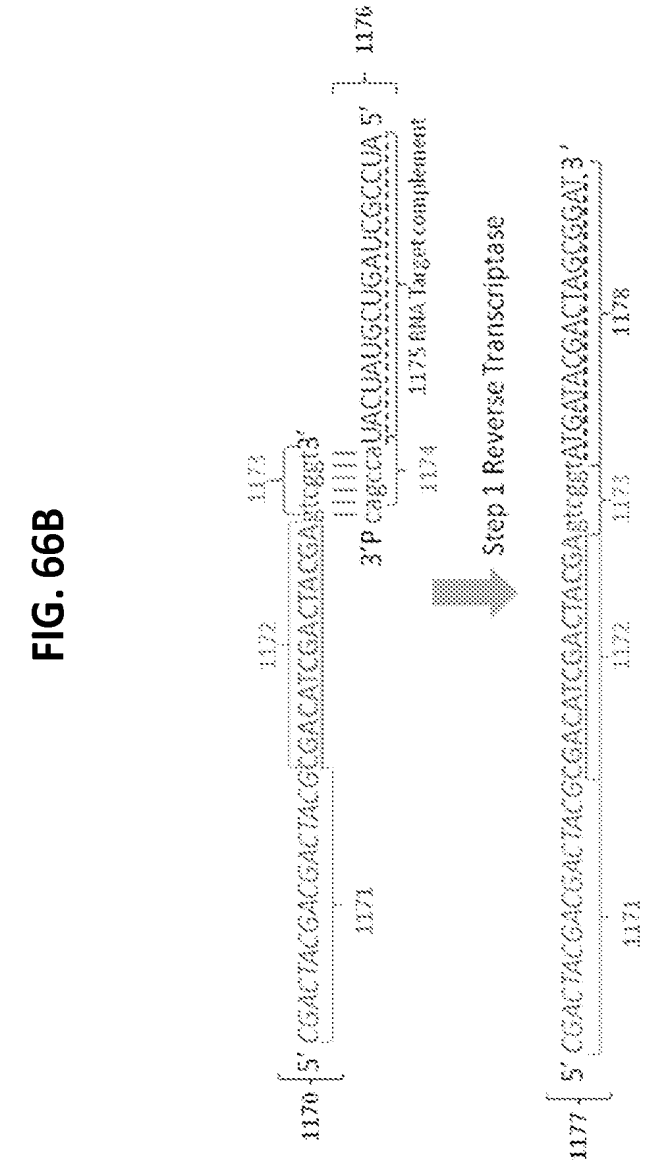


FIG. 66C

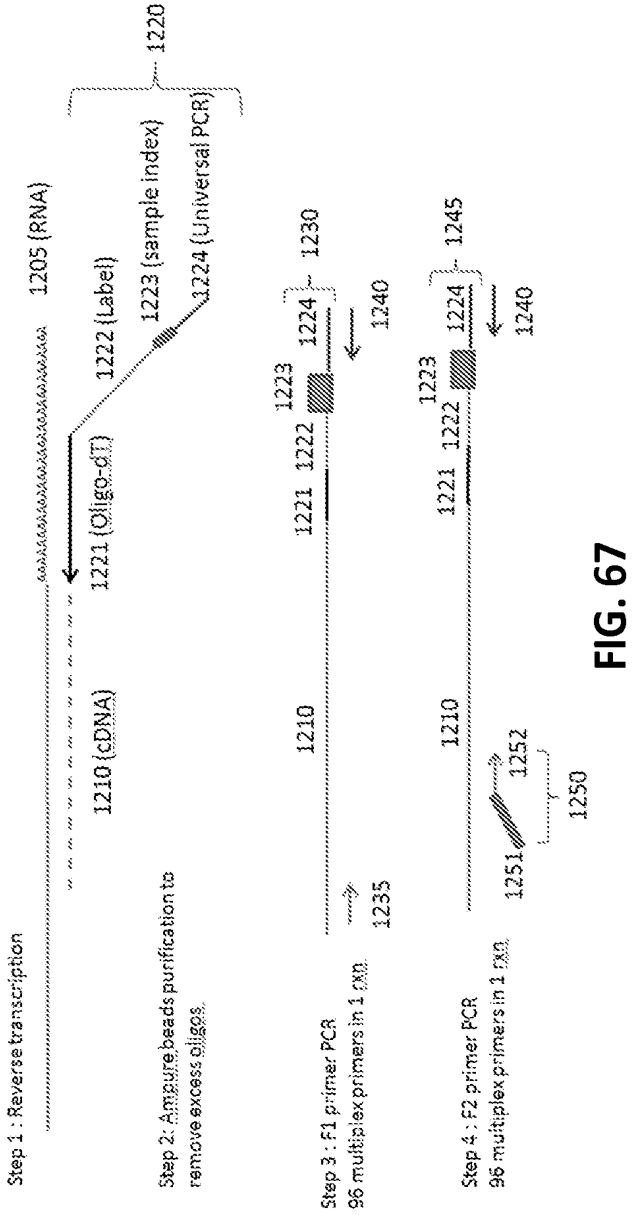

Step 1 : Reverse transcription
1205 (RNA)
1210 (cDNA)
1221 (Oligo-dT)
1222 (Label)
1223 (sample index)
1224 (Universal PCR)
1220
Step 2: Ampure beads purification to remove excess oligos
Step 3 : F1 primer PCR
96 multiplex primers in 1 rxn
1210
1221 1222
1223
1224
1230
1240
1235
Step 4 : F2 primer PCR
96 multiplex primers in 1 rxn
1210
1221 1222
1223
1224
1245
1240
1251
1252
1250

FIG. 67

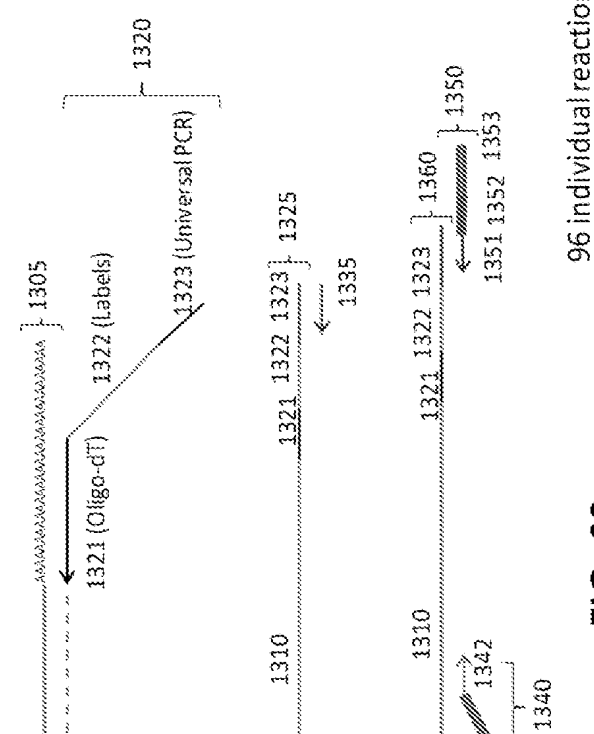

Step 1 : Reverse transcription
1305
1310
1321 (Oligo-dT)
1322 (Labels)
1323 (Universal PCR)
1320
Step 2: Ampure beads purification to remove excess oligos
Step 3 : F1 primer PCR
96 multiplex primers in 1 rxn
1310
1321 1322 1323
1325
1335
1330
Step 4 : F2 primer PCR
96 multiplex primers in 1 rxn
1310
1321 1322 1323
1360
1350
1351 1352 1353
1341
1342
1340
96 individual reactions

FIG. 68

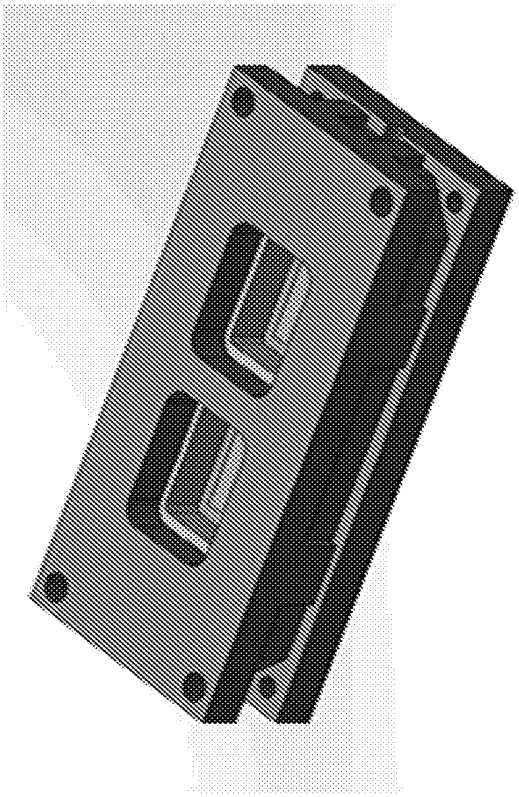


FIG. 70

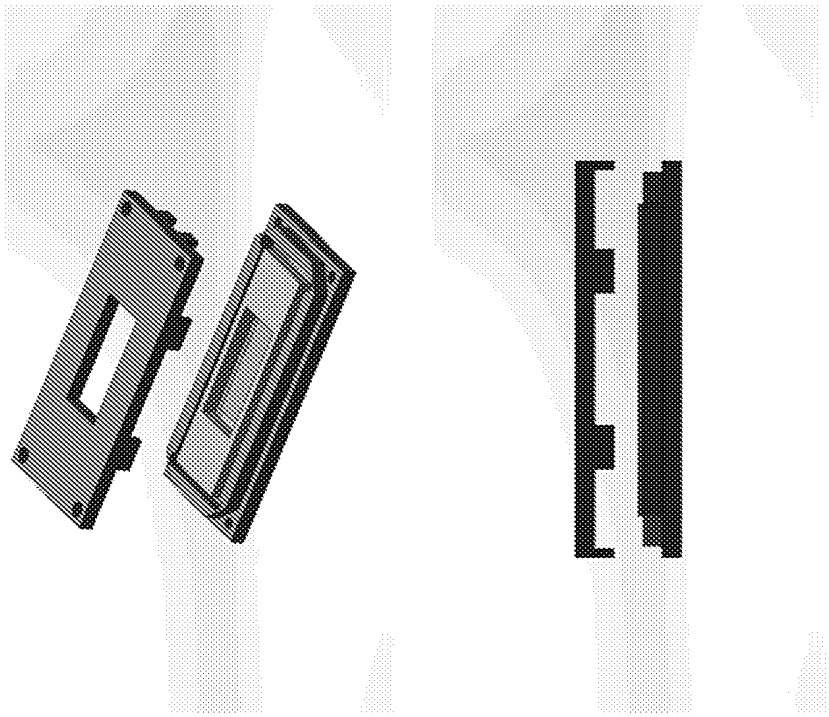


FIG. 69

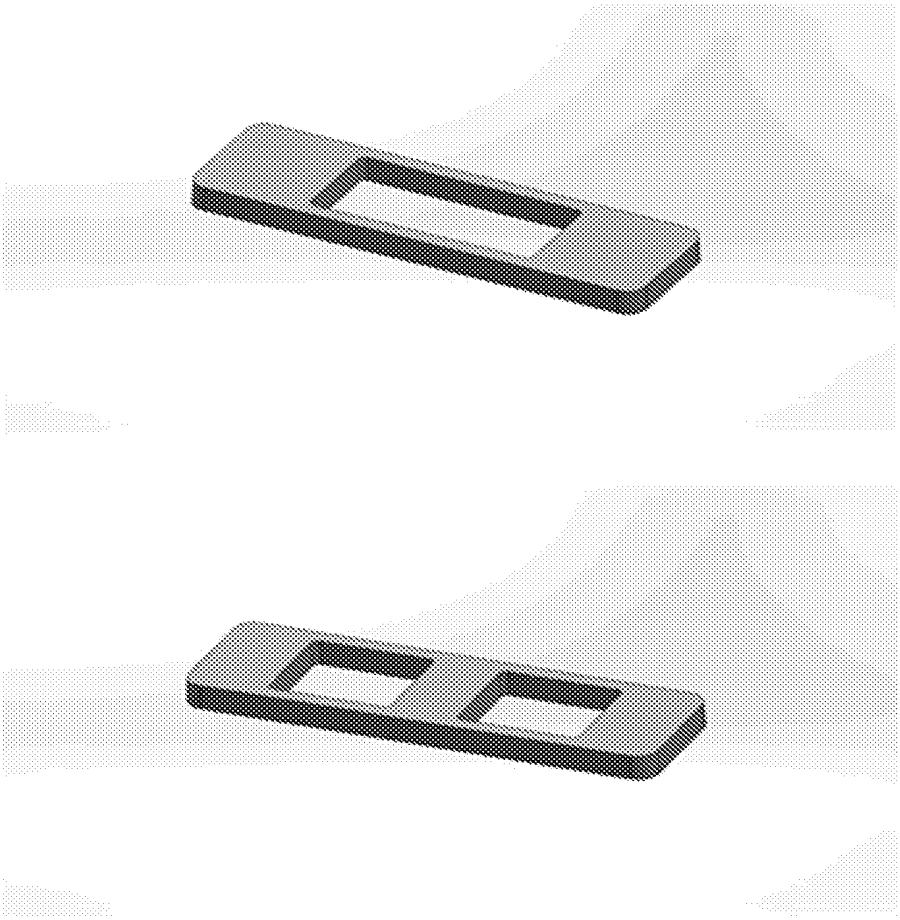


FIG. 71

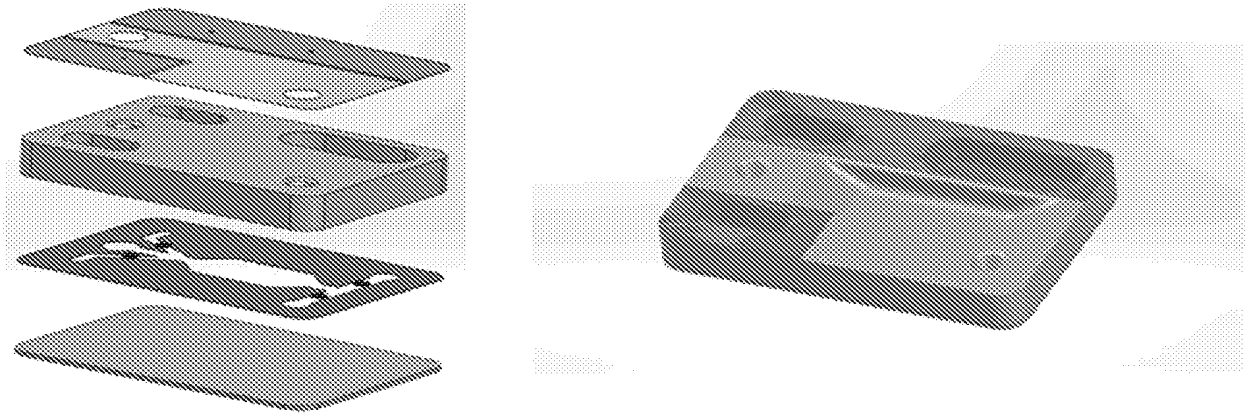


FIG. 72

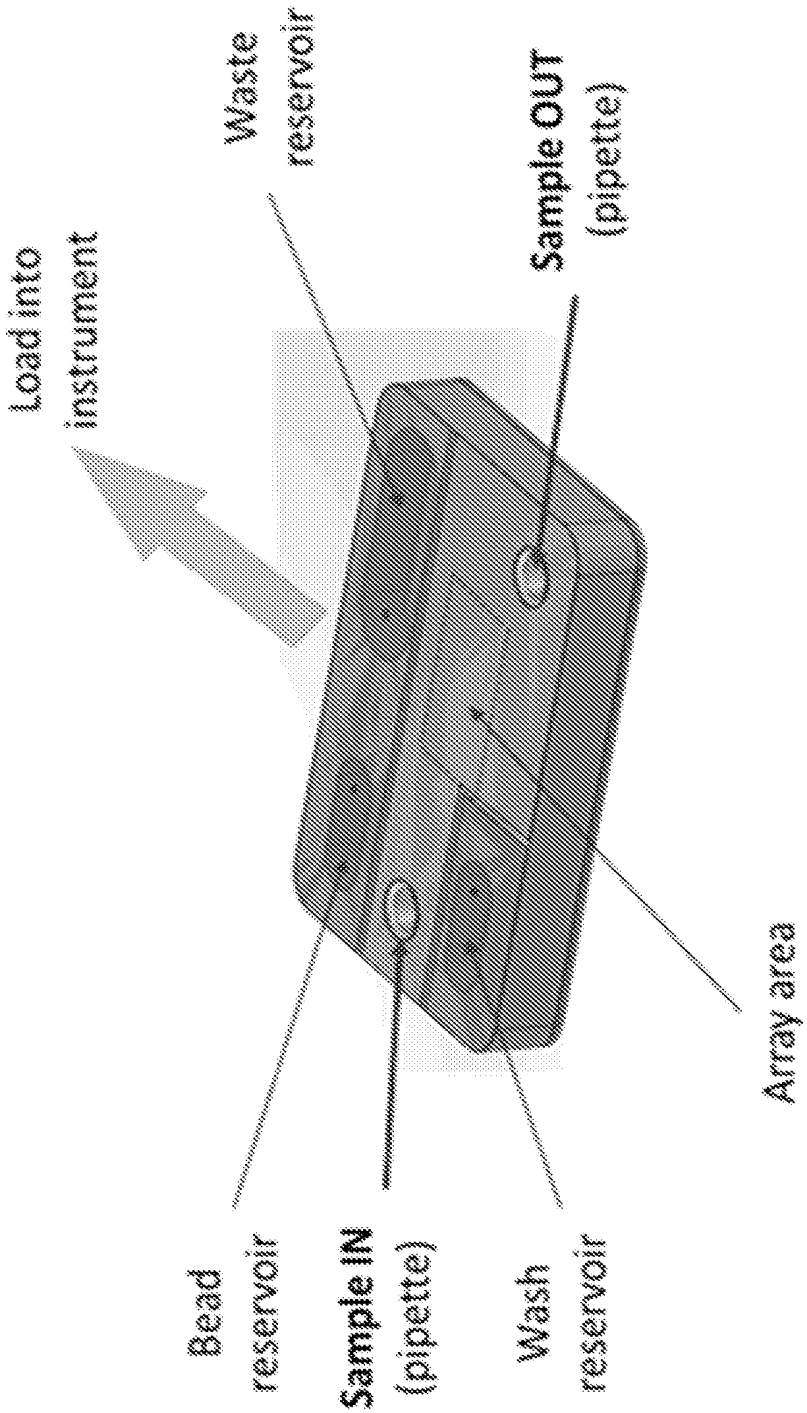

Load into instrument
Waste reservoir
Sample OUT (pipette)
Bead reservoir
Sample IN (pipette)
Wash reservoir
Array area

FIG. 73

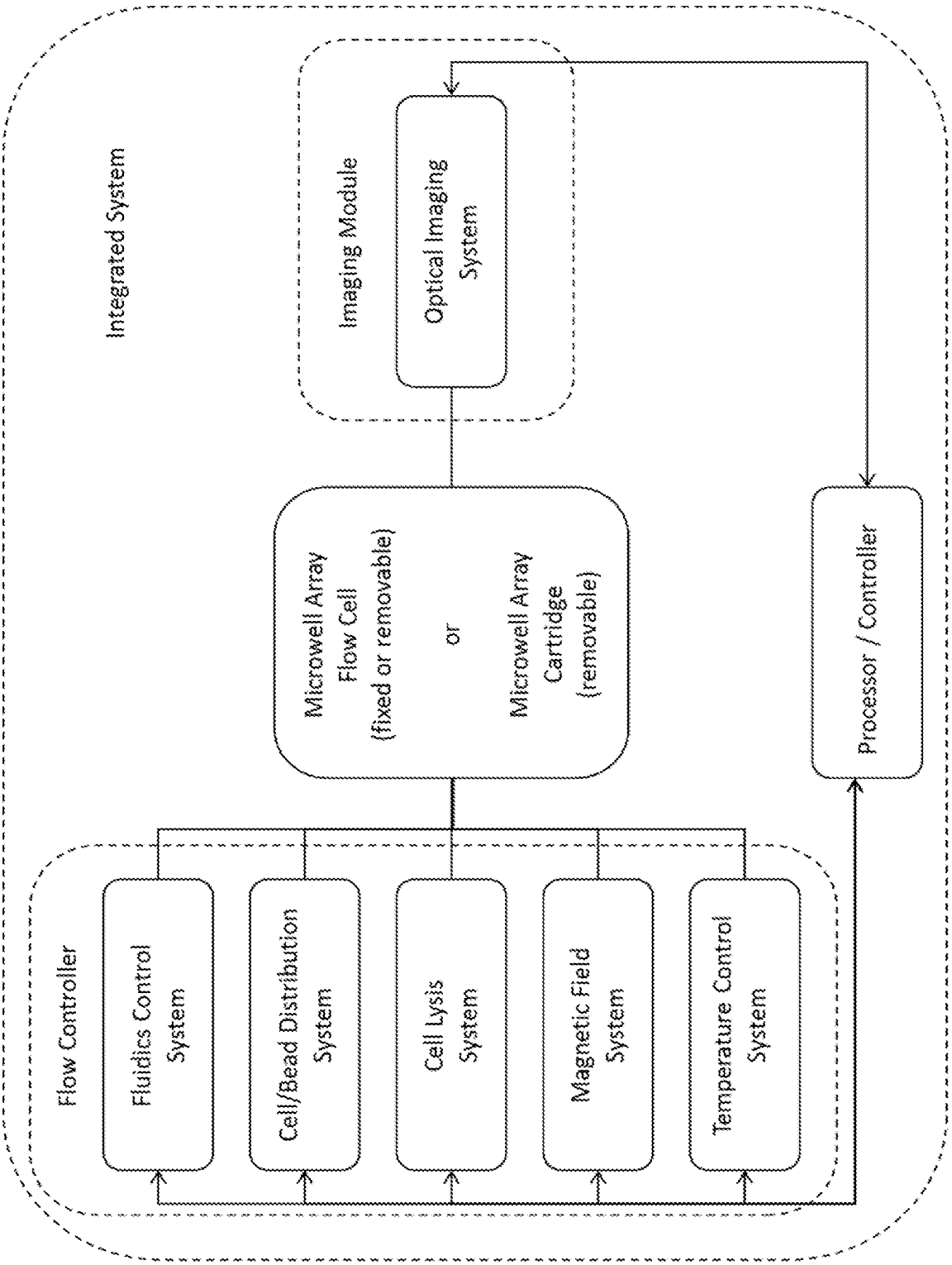

Integrated System
Imaging Module
Optical Imaging System
Microwell Array Flow Cell (fixed or removable)
or
Microwell Array Cartridge (removable)
Processor / Controller
Flow Controller
Fluidics Control System
Cell/Bead Distribution System
Cell Lysis System
Magnetic Field System
Temperature Control System

FIG. 74

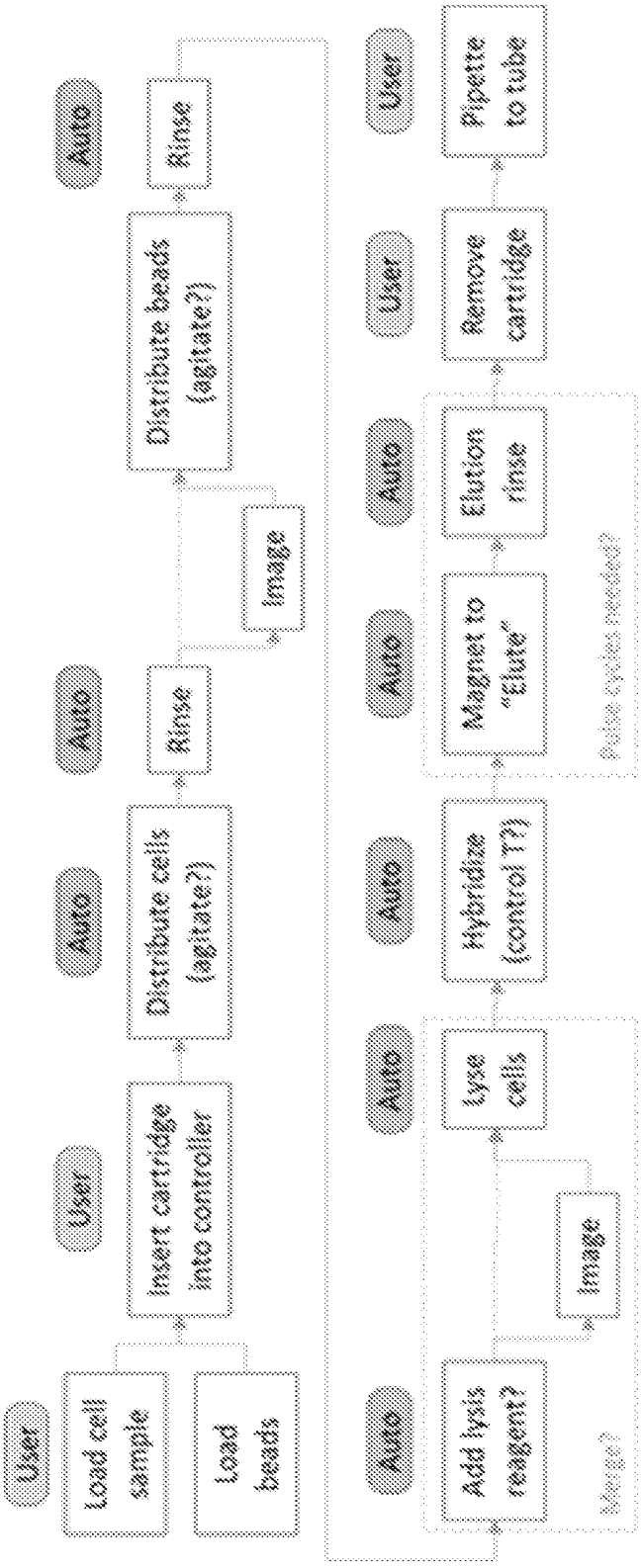

User
Load cell sample
Load beads
User
Insert cartridge into controller
Auto
Distribute cells (agitate?)
Auto
Rinse
Image
Auto
Distribute beads (agitate?)
Rinse
Auto
Add lysis reagent?
Image
Auto
Lyse cells
Auto
Hybridize (control T?)
Auto
Magnet to "Elute"
Auto
Elution rinse
User
Remove cartridge
User
Pipette to tube
Image?
Auto?
Elution cycle needed?

FIG. 75

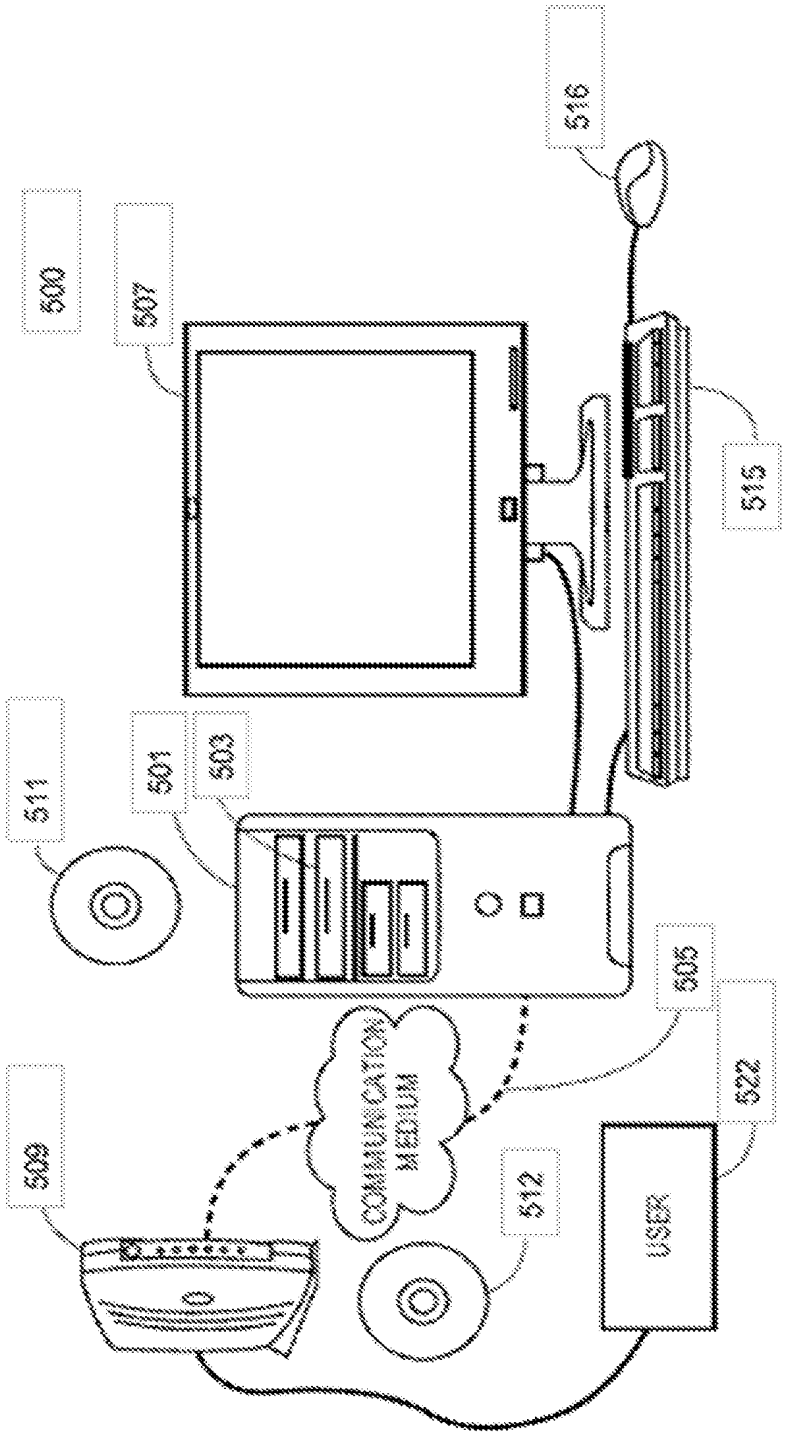

516
500
507
515
511
501
503
505
522
509
COMMUNICATION MEDIUM
512
USER

FIG. 76

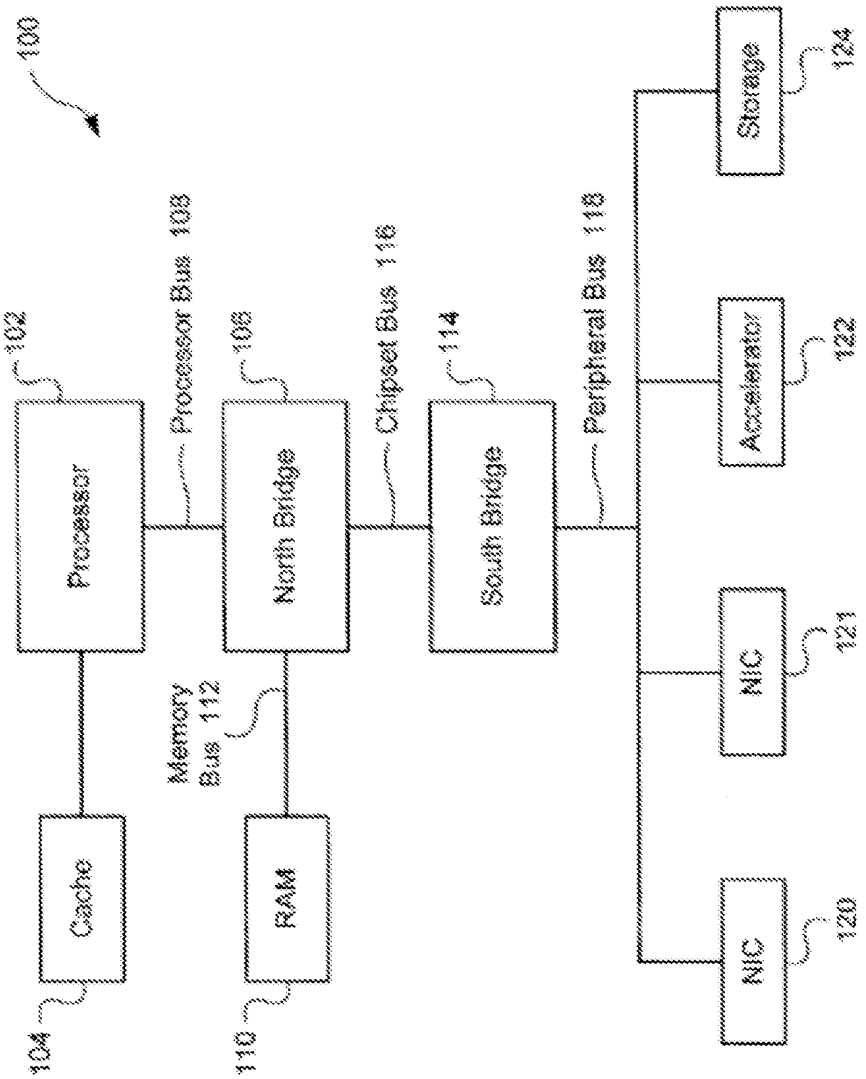

100
Processor
102
Processor Bus 108
North Bridge
106
Chipset Bus 116
South Bridge
114
Peripheral Bus 116
Storage
124
Accelerator
122
NIC
121
NIC
120
Cache
104
Memory Bus 112
RAM
110

FIG. 77

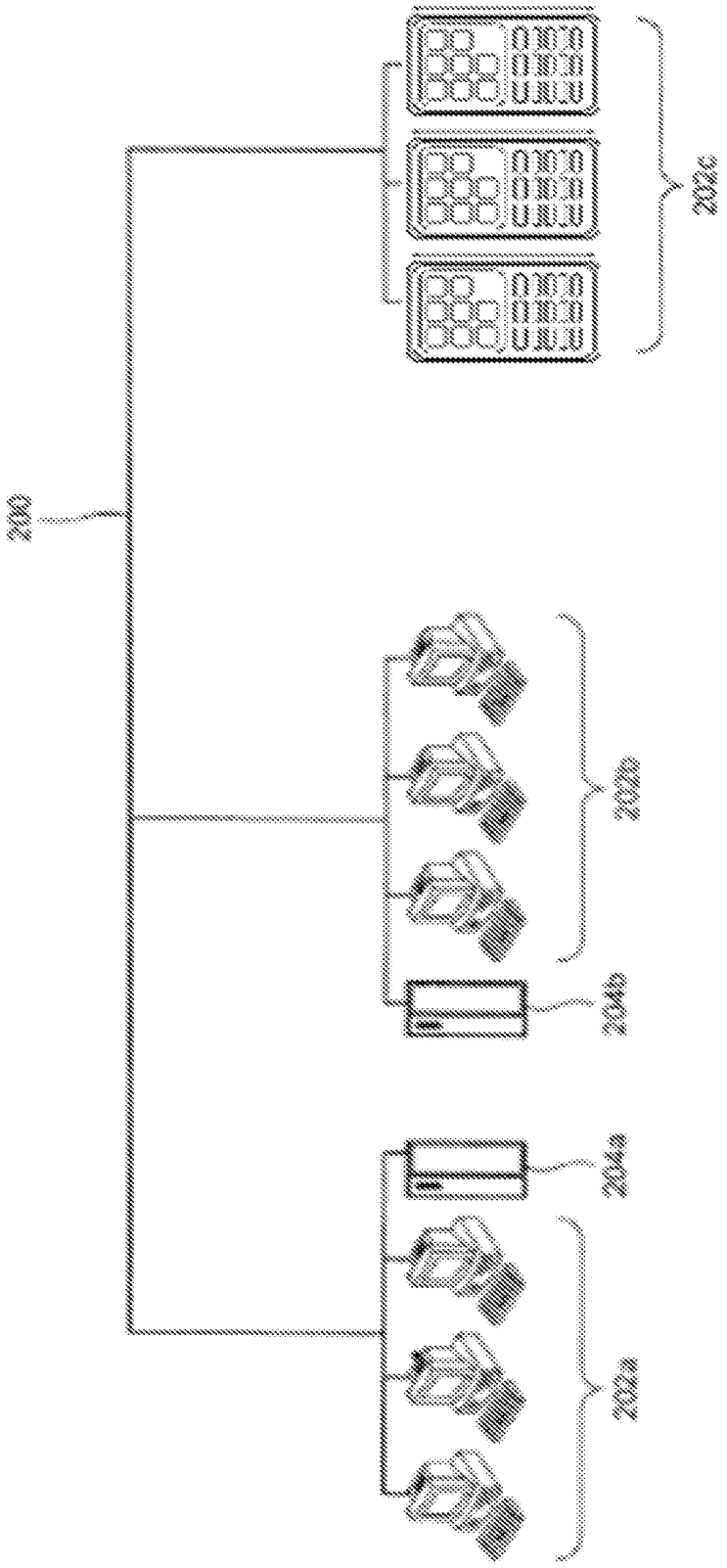


FIG. 78

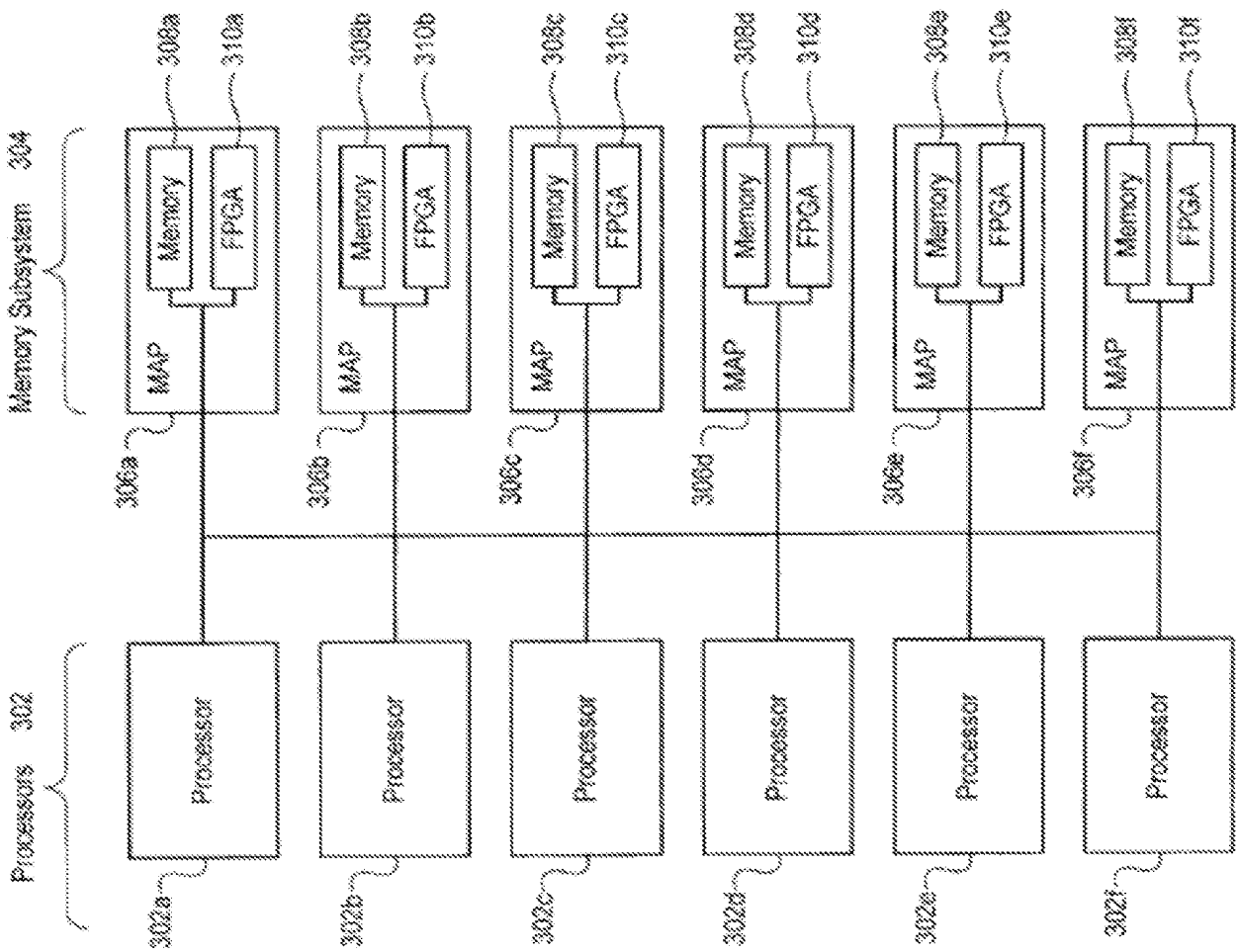

Processors 302
Memory Subsystem 304
Processor
MAP
Memory
FPGA
302a
302b
302c
302d
302e
302f
306a
306b
306c
306d
306e
306f
308a
308b
308c
308d
308e
308f
310a
310b
310c
310d
310e
310f

FIG. 79

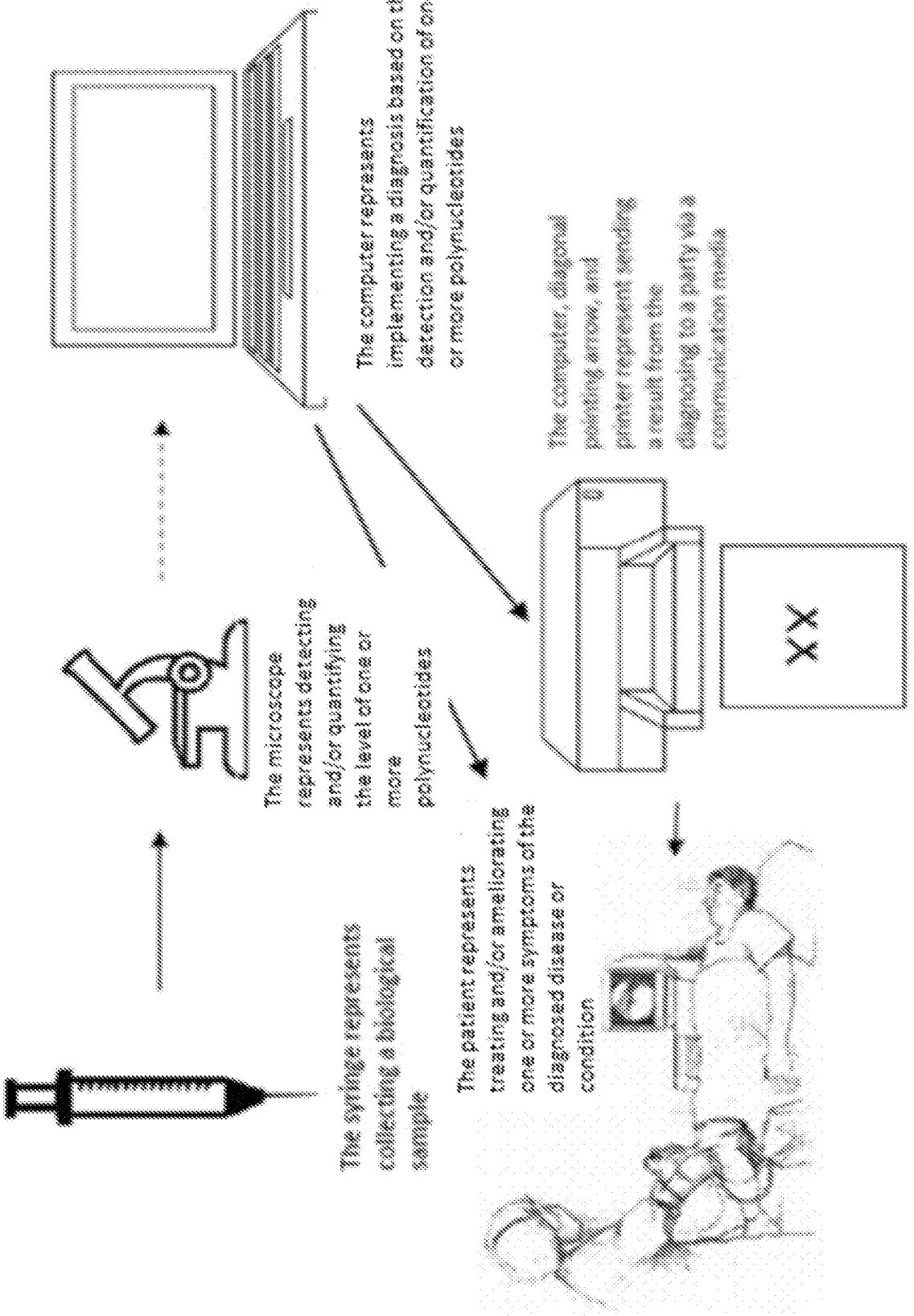

The computer represents implementing a diagnosis based on the detection and/or quantification of one or more polynucleotides
The computer, display, printing device, and printer represent sending a result from the diagnosing to a party via a communication media
The microscope represents detecting and/or quantifying the level of one or more polynucleotides
The syringe represents collecting a biological sample
The patient represents treating and/or ameliorating one or more symptoms of the diagnosed disease or condition
XX

FIG. 80

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2014/053301

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12Q1/68
ADD.

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

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 practicable, search terms used)
EPO-Internal, BIOSIS, MEDLINE, EMBASE, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X	WO 2012/048341 A1 (HARVARD COLLEGE [US]; CHURCH GEORGE M [US]; VIGNEAULT FRANCOIS [US]) 12 April 2012 (2012-04-12) the whole document -----	1-33, 35-50, 52-63, 81-87, 89,90, 92, 94-101
X	WO 2010/117620 A2 (ILLUMINA INC [US]; LINNARSSON STEN [SE]) 14 October 2010 (2010-10-14) the whole document -----	1-33, 35-50, 52-63, 81-87, 89,90, 92, 94-101

-/--

[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 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
23 January 2015	03/02/2015

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	Botz, Jürgen

1

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

IDT Exhibit 1020.4804

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X,P	WO 2014/018460 A1 (SEQUENTA INC [US]; FAHAM MALEK [US]; WILLIS THOMAS [US]; ZHENG JIANBIA) 30 January 2014 (2014-01-30) the whole document	1-33, 35-50, 52-63, 81-87, 89,90, 92, 94-101
X	WO 2012/083225 A2 (GIGAGEN INC [US]; JOHNSON DAVID SCOTT [US]; MEYER EVERETT HURTEAU [US]) 21 June 2012 (2012-06-21) the whole document	1-33, 35-50, 52-63, 81-87, 89,90, 92, 94-101
X	US 2013/116130 A1 (FU GLENN K [US] ET AL) 9 May 2013 (2013-05-09) the whole document	1-33, 35-50, 52-63, 81-87, 89,90, 92, 94-101
X	G. K. FU ET AL: "Counting individual DNA molecules by the stochastic attachment of diverse labels", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, vol. 108, no. 22, 31 May 2011 (2011-05-31) , pages 9026-9031, XP055164127, ISSN: 0027-8424, DOI: 10.1073/pnas.1017621108	1-3
Y	the whole document	4-102
X	C. B. JABARA ET AL: "Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, vol. 108, no. 50, 30 November 2011 (2011-11-30), pages 20166-20171, XP055161674, ISSN: 0027-8424, DOI: 10.1073/pnas.1110064108	1-3,5
Y	the whole document	4,6-102

-/--

1

IDT Exhibit 1020.4805

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
Y	I. KINDE ET AL: "Detection and quantification of rare mutations with massively parallel sequencing", PROCEEDINGS OF THE NATIONAL ACADEMY OF SCIENCES, vol. 108, no. 23, 17 May 2011 (2011-05-17) , pages 9530-9535, XP055164202, ISSN: 0027-8424, DOI: 10.1073/pnas.1105422108 the whole document	1-102
Y	G. MIKE MAKRIGIORGOS ET AL: "A PCR-based amplification method retaining the quantitative difference between two complex genomes", NATURE BIOTECHNOLOGY, vol. 20, no. 9, 1 September 2002 (2002-09-01), pages 936-939, XP055023044, ISSN: 1087-0156, DOI: 10.1038/nbt724 the whole document	1-102
Y	TANG FUCHOU ET AL: "RNA-Seq analysis to capture the transcriptome landscape of a single cell", NATURE PROTOCOLS, NATURE PUBLISHING GROUP, GB, vol. 5, no. 3, 1 March 2010 (2010-03-01), pages 516-535, XP009162232, ISSN: 1750-2799 the whole document	1-102
Y	J. A. CASBON ET AL: "A method for counting PCR template molecules with application to next-generation sequencing", NUCLEIC ACIDS RESEARCH, vol. 39, no. 12, 13 April 2011 (2011-04-13), pages e81-e81, XP055164201, ISSN: 0305-1048, DOI: 10.1093/nar/gkr217 the whole document	1-102

1

IDT Exhibit 1020.4806

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/053301

Box No. I Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application and necessary to the claimed invention, the international search was carried out on the basis of:

   a. (means)

   ☐ on paper

   ☒ in electronic form

   b. (time)

   ☒ in the international application as filed

   ☐ together with the international application in electronic form

   ☐ subsequently to this Authority for the purpose of search

2. ☐ In addition, in the case that more than one version or copy of a sequence listing and/or table relating thereto has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

IDT Exhibit 1020.4807

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2014/053301

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2012048341 A1	12-04-2012	CA 2814049 A1	12-04-2012
		EP 2625320 A1	14-08-2013
		GB 2497912 A	26-06-2013
		GB 2512213 A	24-09-2014
		US 2013274117 A1	17-10-2013
		WO 2012048341 A1	12-04-2012
WO 2010117620 A2	14-10-2010	EP 2414548 A2	08-02-2012
		US 2012010091 A1	12-01-2012
		WO 2010117620 A2	14-10-2010
WO 2014018460 A1	30-01-2014	NONE	
WO 2012083225 A2	21-06-2012	EP 2652155 A2	23-10-2013
		US 2014057799 A1	27-02-2014
		WO 2012083225 A2	21-06-2012
US 2013116130 A1	09-05-2013	NONE	

IDT Exhibit 1020.4808

Espacenet

Bibliographic data: WO2012061832 (A1) — 2012-05-10

LINKING SEQUENCE READS USING PAIRED CODE TAGS

Inventor(s): STEEMERS FRANK J [US]; GUNDERSON KEVIN [US]; ROYCE THOMAS [US]; PIGNATELLI NATASHA [US]; GORYSHIN IGOR YU [US]; CARUCCIO NICHOLAS [US]; MAFFITT MARK [US]; JENDRISAK JEROME [US]; AMINI SASAN [US]; KAPER FIONA [US]; TURK CASEY [US]; KAHLOR REZA [US] ± (STEEMERS, FRANK J, ; GUNDERSON, KEVIN, ; ROYCE, THOMAS, ; PIGNATELLI, NATASHA, ; GORYSHIN, IGOR YU, ; CARUCCIO, NICHOLAS, ; MAFFITT, MARK, ; JENDRISAK, JEROME, ; AMINI, SASAN, ; KAPER, FIONA, ; TURK, CASEY, ; KAHLOR, REZA)

Applicant(s): ILLUMINA INC [US]; STEEMERS FRANK J [US]; GUNDERSON KEVIN [US]; ROYCE THOMAS [US]; PIGNATELLI NATASHA [US]; GORYSHIN IGOR YU [US]; CARUCCIO NICHOLAS [US]; MAFFITT MARK [US]; JENDRISAK JEROME [US]; AMINI SASAN [US]; KAPER FIONA [US]; TURK CASEY [US]; KAHLOR REZA [US] ± (ILLUMINA, INC, ; STEEMERS, FRANK J, ; GUNDERSON, KEVIN, ; ROYCE, THOMAS, ; PIGNATELLI, NATASHA, ; GORYSHIN, IGOR YU, ; CARUCCIO, NICHOLAS, ; MAFFITT, MARK, ; JENDRISAK, JEROME, ; AMINI, SASAN, ; KAPER, FIONA, ; TURK, CASEY, ; KAHLOR, REZA)

Classification: - international: ***C12N15/10; C12N15/90; C40B40/08; C40B50/06***
- cooperative: **C12N15/1065**; **C12N15/1093**; **C40B40/08**; **C40B50/06**

Application number: WO2011US59642 20111107 Global Dossier

Priority number(s): US20100410671P 20101105 ; US201113025022 20110210 ; US201113080345 20110405

Also published as: AU2011323107 (A1) AU2011323107 (B2) CA2821299 (A1) EP2635679 (A1) EP2635679 (B1)

Abstract of WO2012061832 (A1)

IDT Exhibit 1020.4809

Artificial transposon sequences having code tags and target nucleic acids containing such sequences. Methods for making artificial transposons and for using their properties to analyze target nucleic acids.

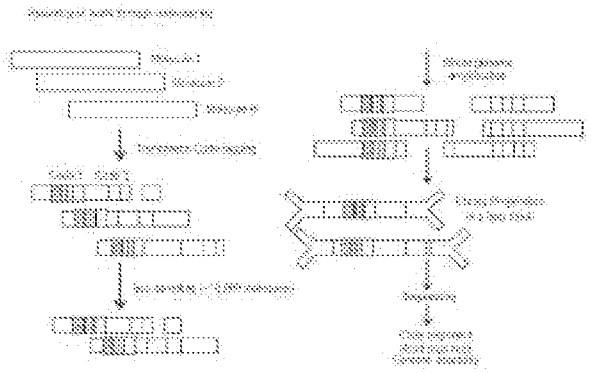


Fig. 31

IDT Exhibit 1020.4810

```
ERROR: ioerror
OFFENDING COMMAND: setfileposition

STACK:

-9
-filestream-
1478
```

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
10 May 2012 (10.05.2012)

PCT

(10) International Publication Number
WO 2012/061832 A1

(51) International Patent Classification:
C12N 15/10 (2006.01) *C40B 40/08* (2006.01)
C12N 15/90 (2006.01) *C40B 50/06* (2006.01)

(21) International Application Number:
PCT/US2011/059642

(22) International Filing Date:
7 November 2011 (07.11.2011)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/410,671 5 November 2010 (05.11.2010) US
13/025,022 10 February 2011 (10.02.2011) US
13/080,345 5 April 2011 (05.04.2011) US

(71) Applicant *(for all designated States except US)*: **ILLUMINA, INC.** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **STEEMERS, Frank J.** [NL/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **GUNDERSON, Kevin** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **ROYCE, Thomas** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **PIGNATELLI, Natasha** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **GORYSHIN, Igor Yu** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **CARUCCIO, Nicholas** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **MAFFITT, Mark** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **JENDRISAK, Jerome** [US/US]; 726 Post Road, Madison, Wisconsin 53713 (US). **AMINI, Sasan** [IR/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **KAPER, Fiona** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **TURK, Casey** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US). **KAHLOR, Reza** [US/US]; 9885 Towne Centre Drive, San Diego, California 92121 (US).

(74) Agents: **FAN, Calvin** et al.; 9885 Towne Centre Drive, San Diego, California 92121 (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, 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.

[Continued on next page]

(54) Title: LINKING SEQUENCE READS USING PAIRED CODE TAGS

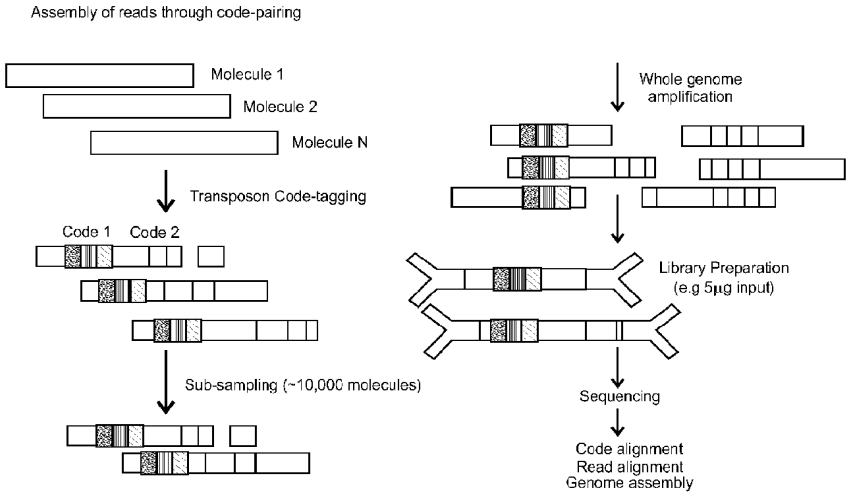


Fig. 31

(57) Abstract: Artificial transposon sequences having code tags and target nucleic acids containing such sequences. Methods for making artificial transposons and for using their properties to analyze target nucleic acids.

WO 2012/061832 A1

IDT Exhibit 1020.4813

(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, 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, RS, 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))*

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 1020.4814

LINKING SEQUENCE READS USING PAIRED CODE TAGS

RELATED APPLICATIONS

This application is a continuation-in-part of U.S. Application No. 13/080,345, filed April 5, 2011, entitled "LINKING SEQUENCE READS USING PAIRED CODE TAGS," which is a continuation-in-part of U.S. Application No. 13/025,022, filed February 10, 2011 entitled "LINKING SEQUENCE READS USING PAIRED CODE TAGS"; and claims the benefit of priority to U.S. provisional Application No. 61/410,671, filed November 5, 2010, entitled "IMPROVED METHOD FOR MAINTAINING MATE-PAIR DNA SEQUENCING LIBRARY DIVERSITY". The contents of these applications are incorporated herein by reference in their entirety.

FIELD OF THE INVENTION

Embodiments of the present invention relate to the fields of biology and genomics. Some embodiments of the present invention relate to methods and compositions that include certain transposon sequences. Some such methods and compositions include analyzing target nucleic acids.

BACKGROUND OF THE INVENTION

The detection of specific nucleic acid sequences present in a biological sample has been used, for example, as a method for identifying and classifying microorganisms, diagnosing infectious diseases, detecting and characterizing genetic abnormalities, identifying genetic changes associated with cancer, studying genetic susceptibility to disease, and measuring response to various types of treatment. A common technique for detecting specific nucleic acid sequences in a biological sample is nucleic acid sequencing.

Nucleic acid sequencing methodology has evolved significantly from the chemical degradation methods used by Maxam and Gilbert and the strand elongation methods used by Sanger. Today several sequencing methodologies are in use which allow for the

IDT Exhibit 1020.4815

parallel processing of nucleic acids all in a single sequencing run. As such, the information generated from a single sequencing run can be enormous.

SUMMARY OF THE INVENTION

Some embodiments of the present invention relate to methods and compositions that include certain transposon sequences. Some such methods and compositions include analyzing target nucleic acids. Some embodiments include methods of preparing a template nucleic acid. Some such methods include: (a) providing a target nucleic acid; (b) providing a plurality of transposon sequences, each transposon sequence comprising a first transposase recognition site, a second transposase recognition site having a barcode disposed therebetween; and (c) contacting the target nucleic acid with the plurality of transposon sequences under conditions such that at least a portion of said plurality of sequences inserts into the target nucleic acid ("tagementation"), thereby preparing a template nucleic acid.

In some embodiments, the barcode of each transposon is different.

In some embodiments, the barcode comprises a first barcode sequence, a second barcode sequence, said first and second barcode sequences being separated by a linker disposed therebetween.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site.

In some embodiments, the linker comprises a fragmentation site. In some embodiments, the fragmentation site comprises a first nickase recognition sequence and a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site. In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site and a fragmentation site. In some embodiments, the linker comprises a

IDT Exhibit 1020.4816

sequencing adapter comprising a first primer site and a second primer site having a fragmentation site disposed therebetween.

Some embodiments include methods of preparing a library of template nucleic acids. Some such methods include: (a) providing a target nucleic acid; (b) providing a plurality of transposon sequences, each transposon sequence comprising a first transposase recognition site and a second transposase recognition site having a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence being separated by a sequencing adapter disposed therebetween, wherein the sequencing adapter comprises a first primer site and a second primer site; and (c) contacting the target nucleic acid with the plurality of transposon sequences under conditions such that at least a portion of said plurality of transposon sequences inserts into the target nucleic acid; and (d) amplifying at least a portion of the target nucleic acid by hybridizing a primer to said first primer site and a primer to said second primer site, such that the amplification product comprises a first barcode sequence and a second barcode sequence, thereby preparing a library of template nucleic acids.

In some embodiments, the barcode of each transposon is different.

Some methods also include a step subsequent to (c) and prior to (d) comprising reducing the number of target nucleic acid molecules comprising inserted transposon sequences.

In some embodiments, the sequencing adapter comprises a first primer site, a second primer site, and a non-amplifiable site therebetween. In some embodiments, the non-amplifiable site comprises a nucleic acid. In some embodiments, the non-amplifiable site comprises at least one nucleotide analogue. In some embodiments, the nucleotide analogue does not significantly base-pair with A, C, G or T.

Some embodiments include methods of preparing a library of template nucleic acids. Some such methods include: (a) providing a target nucleic acid; (b) providing a plurality of transposon sequences, each transposon sequence comprising a first transposase recognition site and a second transposase recognition site having a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence being separated by a linker disposed therebetween, wherein the linker comprises a fragmentation site; (c) contacting the target nucleic acid with the plurality of transposon sequences under conditions such that at least a portion of said plurality of transposon sequences inserts into the target nucleic acid; and (d) fragmenting

IDT Exhibit 1020.4817

said target nucleic acid at said fragmentation sites, such that at least a portion of the fragmented nucleic acids each comprise a first barcode sequence and a second barcode sequence, thereby preparing a library of template nucleic acids.

In some embodiments, the barcode of each transposon is different.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the fragmentation site comprises a first nickase recognition sequence, a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site.

In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some methods, the fragmenting comprises contacting the target nucleic acid with a polymerase. Useful polymerases include those with exonuclease activity (such as 3' to 5' activity, *e.g., E. coli* DNA polymerase III, or 5' to 3' activity, *e.g., E. coli* DNA polymerase I), non-strand-displacing (e.g. T4 DNA polymerase) and strand-displacing activities (e.g. *Bst* DNA polymerase, large fragment).

Some methods also include ligating a first primer site to a first end of at least one fragmented nucleic acid. Some methods also include ligating a second primer site to the second end of the at least one fragmented nucleic acid.

Some methods also include amplifying said at least one fragmented nucleic acid by hybridizing a primer to the first primer site and a primer to the second primer site.

In some embodiments, the linker further comprises a first primer site.

In some embodiments, the linker comprises a sequencing adapter comprises a first primer site, a second primer site having the fragmentation site disposed therebetween. Some methods also include amplifying said at least a portion of the fragmented nucleic acids by hybridizing a primer to said first primer site and a primer to said second primer site.

In some embodiments, the target nucleic acid comprises genomic DNA.

In some embodiments, the first transposase recognition site comprises a mosaic element.

In some embodiments, the sequences of the first barcode and the second barcode comprise the reverse complements of each other.

Some embodiments include libraries of template nucleic acids prepared by any one of the foregoing methods.

IDT Exhibit 1020.4818

Some embodiments include methods of preparing a transposon sequence. Some such methods include: (a) providing a transposon template nucleic acid sequence comprising a first transposase recognition site, a first barcode 3′ thereof, and a linker 3′ thereof; and (b) extending the transposon template sequence by hair-pin extension, such that the extended transposon template sequence further comprises sequences complementary to said barcode and to said transposase recognition site, thereby preparing a transposon comprising a first transposase recognition site, a first barcode, a linker, a second barcode, and a second transposase recognition site.

In some embodiments, the barcode comprises a random sequence.

In some embodiments, the transposase recognition site comprises a mosaic element.

In some embodiments, the barcode comprises at least about five nucleotides.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the linker comprises a fragmentation site. In some embodiments, the fragmentation site comprises a first nickase recognition sequence, a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site. In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some embodiments, the linker comprises a sequencing adapter. In some embodiments, a sequencing adapter comprises a first primer site and second primer site. In some embodiments, the sequencing adapter comprises a first primer site and a second primer site having a fragmentation site disposed therebetween. In some embodiments, the sequencing adapter comprises a first primer site and a second primer site, having a non-amplifiable site disposed therebetween.

In some embodiments, the non-amplifiable site comprises a nucleic acid. In some embodiments, the non-amplifiable site comprises at least one nucleotide analogue. In some embodiments, the nucleotide analogue does not significantly base-pair with A, C, G or T.

Some embodiments include methods of preparing a plurality of transposon sequences. Some such method include repeating steps (a) and (b) of any one of the foregoing methods of preparing a transposon sequence, wherein the barcode of each transposon template nucleic acid is different.

IDT Exhibit 1020.4819

Some embodiments include transposons prepared by the method of any one of the foregoing methods of preparing a transposon sequence.

Some embodiments include methods of analyzing a target nucleic acid. Some such methods include: (a) providing a template nucleic acid, wherein the template nucleic acid comprises the target nucleic acid and a plurality of markers inserted therein, wherein each marker sequence comprises a barcode comprising a first barcode sequence and a second barcode sequence, having a sequencing adapter disposed therebetween, wherein the sequencing adapter comprises a first primer site; (b) obtaining sequence data from said template nucleic acid; and (c) assembling a representation of at least a portion of said target nucleic acid from said sequence data.

In some embodiments, the barcode of each marker is different.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

In some embodiments, obtaining sequence data comprises hybridizing a primer to said first primer site, and extending said primer.

In some embodiments, the sequencing adapter comprises a second primer site, and said obtaining sequence data comprises hybridizing a primer to the second primer site and extending said primer. In some embodiments, the primers hybridize to the sites in opposite orientations.

In some embodiments, the sequences of the first barcode sequence and the second barcode sequence comprise the reverse complements of each other.

In some embodiments, the assembling step comprises identifying more than one sequencing read comprising the first barcode sequence or second barcode sequence of a barcode.

In some embodiments, the presence of a first barcode sequence or second barcode sequence of a barcode in more than one sequencing read is indicative of the more than one sequencing reads representing sequences adjacent to each other in the target nucleic acid.

In some embodiments, each marker sequence comprises a first host tag and second host tag having the barcode disposed therebetween. In some embodiments, the

IDT Exhibit 1020.4820

first and second host tag of a marker comprises the same sequence. In some embodiments, the assembling step further comprises identifying more than one sequencing read comprising the same host tags.

In some embodiments, the target nucleic acid comprises genomic DNA.

Other embodiments include artificial transposon sequences. Some such artificial transposon sequences include: a first transposase recognition site and a second transposase recognition site having a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence, said first and second barcode sequences being separated by a linker.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

In some embodiments, the first transposon recognition site comprises a mosaic element.

In some embodiments, the linker comprises a nucleic acid.

In some embodiments, the linker comprises a fragmentation site. In some embodiments, the fragmentation site comprises a first nickase recognition sequence and a second nickase recognition sequence, wherein the cut site for each recognition sequence is the same site. In some embodiments, the fragmentation site comprises a restriction endonuclease recognition sequence.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site. In some embodiments, the sequencing adapter comprises a second primer site. In some embodiments, the sequencing adapter comprises a first primer site and a second primer site, having a non-amplifiable site disposed therebetween. In some embodiments, the non-amplifiable site comprises a nucleic acid. In some embodiments, the non-amplifiable site comprises at least one nucleotide analogue. In some embodiments, the nucleotide analogue does not significantly base-pair with A, C, G or T.

In some embodiments, the linker comprises a sequencing adapter comprising a first primer site and a second primer site, having a fragmentation site disposed therebetween.

In some embodiments, a sequencing primer is hybridized to said first primer site.

IDT Exhibit 1020.4821

In some embodiments, a sequencing primer is hybridized to said first transposase recognition site.

Some embodiments include populations of artificial transposon sequences comprising a plurality of any one of the foregoing artificial transposon sequences.

Some embodiments include target nucleic acids or copies thereof having a population of the artificial transposon sequences integrated therein, wherein the artificial transposon sequences comprise any one of the artificial transposon sequences described herein.

Some embodiments include populations of genomic DNA fragments or copies thereof having a population of the artificial transposon sequences integrated therein, wherein the artificial transposon sequences comprise any one of the artificial transposon sequences described herein.

Some embodiments include genomes having a population of artificial transposon sequences integrated therein, wherein the artificial transposon sequences comprise any one of the artificial transposon sequences described herein.

Some embodiments include isolated template nucleic acids. Some such isolated template nucleic acids include at least a portion of a target nucleic acid or copy thereof and at least two markers inserted therein, wherein each marker sequence comprises a barcode.

In some embodiments, the barcode of each marker is different.

In some embodiments, the barcode comprises a double-stranded nucleic acid sequence comprising a first strand barcode and a second strand barcode. In some embodiments, the first strand barcode and second strand barcode comprise complementary sequences. In some embodiments, the first strand barcode and second strand barcode comprise non-complementary sequences.

Some embodiments also include a transposase recognition site.

Some embodiments also include a first transposase recognition site and a second transposes recognition site, having the at least a portion of a target nucleic acid or copy thereof disposed therebetween.

Some embodiments also include a host tag. Some embodiments also include a first host tag and a second host tag, having the at least a portion of a target nucleic acid or copy thereof disposed therebetween. Some embodiments also include a first host tag and a second host tag, having a transposase recognition site disposed therebetween, wherein

IDT Exhibit 1020.4822

the transposase recognition site comprises a first transposase recognition site and a second transposase recognition site, having the at least a portion of a target nucleic acid or copy thereof disposed therebetween.

In some embodiments, the host tag comprises nine nucleotides.

In some embodiments, the target nucleic acid comprises genomic DNA.

Some embodiments include a plurality of template nucleic acids comprising the template nucleic acid of any one of the isolated template nucleic acids described herein, wherein a first template nucleic acid of said plurality comprises a first barcode, and a second template nucleic acid comprises a second barcode, wherein the first and second barcodes are indicative of template nucleic acid sequences being adjacent to one another in a sequence representation of the target nucleic acid. In some embodiments, the first barcode comprises the reverse complement sequence of the second barcode.

Some embodiments include substrates comprising nucleotide sequences attached thereto, at least one of said nucleotide sequences comprising any one of the template nucleic acids described herein, or any one of the plurality of template nucleic acids described herein. In some embodiments, the substrate can include a solid support selected from the group consisting of spheres, microparticles, beads, membranes, slides, plates, micromachined chips, tubes, microwells, microfluidic devices, channels, and filters.

BRIEF DESCRIPTION OF THE DRAWINGS

FIG. 1 depicts a schematic a transposon sequence. M indicates a first transposon recognition sequence. C1 indicates a first barcode sequence. The linker sequence is indicated by a rectangle with a double border. C1′ indicates a second barcode sequence. M' indicates a second transposase recognition sequence. Exemplary linker sequences are also shown. For example, linker sequences can have fragmentation sites, such as those with nickase recognition sites (e.g. N1, N2) or a restriction endonuclease recognition sequence (indicated by RE). In other embodiments, linkers can contain one or more primer sites (e.g. A, B), corresponding to complementary primers A′ and B′. A non-amplifiable site is indicated by nA.

FIG. 2 depicts a double-stranded transposon sequence comprising transposon recognition sites (TRS and TRS′), a fragmentation site, and a barcode comprising four barcode sequences (CODE 1, CODE 2, CODE 3, and CODE 4). In the embodiment

IDT Exhibit 1020.4823

shown, CODE 1 and CODE 3, and CODE 2 and CODE 4 are non-complementary sequences.

FIG. 3 illustrates single-stranded transposons having bicodes or unicodes.

FIG. 4 illustrates preparation of transposome having different transposase recognition sequences.

FIG. 5 depicts a transposome complex that has one transposon sequence and two transposases.

FIG. 6 illustrates gap fill-in and ligation reactions.

FIG. 7 depicts "looped" complexes.

FIG. 8 depicts a complex with one transposon and two transposase subunits ensures that a known combination of codes is inserted into target DNA, without fragmenting the target DNA. This configuration allows linking DNA fragments after fragmentation using the code N-Code N' combination.

FIG. 9 illustrates strand-displacement synthesis to copy identical flanking genomic DNA sequences (g-codes).

FIG. 10 illustrates nicking followed by 3'-exo digestion to create gaps in genomic DNA.

FIG. 11 illustrates transposons with engineered bicodes and random bi-codes.

FIG. 12. Transposon-transposase complex. Complexes are formed from two half-transposons, each containing one ME region and a random or specific code region. Upon insertion, the transposase enzyme binds to both transposons preventing DNA fragmentation. The ends of the transposons have "sticky ends" allowing ligation of the two ends. A specific endonuclease fragmentation system nicks each strand. Denaturation creates DNA fragments in which each DNA fragment contains a linked code with the neighboring fragment.

FIG. 13. Transposons with amplifiable and non-amplifiable linkers between the ME sequences.

FIG. 14 depicts use of bi-code junction-tagging.

FIG. 15 depicts a library preparation method that involves ligation of a forked adapter sequence. Steps include: library preparation; extension and A-tailing; adapter ligation; and PCR amplify, cluster, and sequence products.

FIG. 16 depicts an exemplary scheme to prepare a transposon sequence with reverse complements

IDT Exhibit 1020.4824

FIG. 17 depicts a method of preparing a transposon sequence which includes the use of tailed-oligonucleotides comprising mosaic elements (ME). The transposon sequence includes primer sites P3 and P4, a barcode comprising Code X and Code Y, and a fragmentation site.

FIG. 18 depicts the integration of an exemplary Tn5 transposon comprising transposase recognition sequences (indicated by mosaic elements ME) into a target nucleic acid. The insertion results in duplication of the integration site, shown as single-stranded A′B′C′D′E′F′G′H′I′ and single-stranded ABCDEFGHI. The figure also shows an optional repair step to fill in the single-stranded region. In some embodiments, the filled-in region can be used as a host tag.

FIG. 19 shows a tagmentation method using transposomes that have symmetrical transposable end sequences.

FIG. 20 depicts an exemplary embodiment of a contacting a target nucleic acid with a library of transposons, which integrates into the target nucleic acid.

FIG. 21 depicts an optional step of fragmenting a linker into two parts, or an optional step of virtually fragmenting the linker by amplification using a primer A′ and/or a primer B′. Fragmentation of the linker generates a library of template nucleic acids. In this figure, the filled-in region is indicated by a thick line.

FIG. 22 depicts an examplary embodiment using tailed-oligonucleotides to obtain a population of template nucleic acids comprising a first universal primer site (A) and a second universal primer site (B). In step 1, tailed-oligonucleotides comprising universal primer site sequences anneal to a template nucleic acid comprising primer site sequences (P and P′) and are extended. In step 2, the tailed oligonucleotides of step 1 may anneal to the products of step 1 and be extended. Alternatively, step 2 can include an amplification step with the inclusion of oligonucleotides (*e.g.*, oligonucleotides comprising A′ and B′ sequences). The products of step 2 are depicted in step 3.

FIG. 23 illustrates an embodiment for code-tagging individual molecules with unique tags, allowing normalization of the original molecule representation, even after amplification.

FIG. 24 depicts how linked transposon ends result in fragmented DNA, simultaneously linking the top and bottom strand of the target DNA.

FIG. 25. Targeted Tagmentation using ssDNA. (1) Locus-specific oligo probes are annealed to regions of interest in the denatured DNA sample. An optional "gap-fill"

IDT Exhibit 1020.4825

extension (polymerase, ligase, nucleotides, and buffer) can be employed to create dsDNA between sites of probe annealing. (2) Tagmentation mix (i.e. Tn5-mosaic end transposon complexes and appropriate reaction buffer) is added to the annealed DNA. (3) Tagmentation only occurs at regions of dsDNA leading to "targeted" tagmentation (and nicking) at sites where the oligo probes have annealed. (4) Strand-displacement synthesis resolves the nicks into amplifiable library elements flanked by the universal ME primer. This ME primer can be used in a subsequent PCR step to attach P5-Seq1 and P7-Seq2 sequencing tails to the tagments.

FIG. 26. Targeted Tagmentation using dsDNA and Strand Invasion/Triplex formation. D-loop formation via strand invasion using recA coated oligonucleotides or LNA probes can be used to create D-loops and target transposase complexes to loci or regions of interest (ROI) in dsDNA. Targeted transposition can also be accomplished by triplex formation using pyrimidine-rich oligonucleotides. The targeting probe or oligonucleotide is conjugated to the mosaic elements (ME) on the transposase-transposon complex. (1) Strand invasion opens a D-loop in the dsDNA bringing the transposase complex into the vicinity of the targeted region. Ideally two such probes will be used to generate libraries spanning the region between the probes. (2) Localization of the transposon complex leads to insertion in the vicinity of the D-loop or triplex structure. (3) Strand-displacement is use to finish the library element flanked by universal ME primers. PCR can be used append sequencing primers using the ME ends.

FIG. 27 depicts tethering of a circularized intermediate to the surface of a particle.

FIG. 28 depicts a schematic representation of using paired bar codes (indicated by Code 1, Code 2, Code 5) for code-pair sequencing of template contig sequences to assemble the sequence of an original target nucleic acid.

FIG. 29 depicts an exemplary embodiment of a method including fragmentation.

FIG. 30 depicts an exemplary embodiment of a method that includes optional steps of amplifying a target nucleic acid by PCR or whole genome amplification (WGA). In this figure, the filled-in region is indicated by N_1 and N_2.

FIG. 31 depicts an exemplary embodiment of a method for assembling short sequencing reads using a linked read strategy. Optional steps of sub-sampling, code alignment, read assembly, and genome assembly are also illustrated.

FIG. 32 illustrates the preparation of nested sequencing libraries.

IDT Exhibit 1020.4826

FIG. 33 illustrates a two-step tagging method using different transposomes.

FIG. 34 illustrates single cell/DNA tagmentation in microwells.

FIG. 35 illustrates single cell/DNA tagmentation in droplets.

FIG. 36 depicts a graph showing that the proportion of template nucleic acids that are useful can decrease as the average distance between sites of integration increases.

FIG. 37 depicts an embodiment using restriction endonucleases to generate randomer sticky ends.

DETAILED DESCRIPTION

Some embodiments of the present invention relate to methods and compositions that include certain transposon sequences. Some such methods and compositions include analyzing target nucleic acids. Generally, methods of analyzing nucleic acids include preparing a library of template nucleic acids of a target nucleic acid, obtaining sequence data from the library of template nucleic acids, and assembling a sequence representation of the target nucleic acid from such sequence data. However, assembly of a sequence representation using traditional methods is met with several challenges. For example, sequencing data is obtained using short reads and thus there is difficulty assembling contiguous reads through repetitive sequences of a target nucleic acid. In addition, many contigs are required to be constructed to represent a genome such as the human genome. Methods such as paired-end sequencing can be used to mitigate the difficulties of assembling sequence data from many short reads. However, the paired-end sequencing methods require longer template nucleic acids to span repetitive sequences in a target nucleic acid; preparation of template nucleic acids is inefficient and thus requires larger amounts of DNA; the number of different paired-ends is limited; and a reference genome is required to verify any resulting sequence representation.

Some nucleic acids of interest, such as genomic DNAs, comprise long molecules with ordered sequence. Methods to sequence such molecules tend to be highly parallelized and include sequencing short libraries in order to create high throughput. Consequently, such methods require assembly of the shorter reads to obtain ordering information, *i.e.*, a sequence representation of the target nucleic acid. Shotgun sequencing approaches uses DNA fragmentation; however, order information is lost during fragmentation. Advantageously, some methods and compositions provided herein can be used to obtain shorter reads of template nucleic acids in which the order

IDT Exhibit 1020.4827

information is preserved. Thus, assembly of shorter reads can be performed without the requirement of a reference genome.

In an exemplary embodiment, a library of template nucleic acids is prepared from a target nucleic acid. The library is prepared by inserting a plurality of unique barcodes throughout the target nucleic acid.

As will be understood, although a barcode is frequently depicted in the figures as a double-stranded, annealed structure, a useful barcode can be a double-stranded region where the two strands are not complementary or not annealed. For example, two noncomplementary regions can be linked informatically, even though they are not complementary in a conventional Watson-Crick base-pairing.

Each barcode includes a first barcode sequence and a second barcode sequence, having a fragmentation site disposed therebetween. The first barcode sequence and second barcode sequences can be identified or designated to be paired with one another. The pairing can be informatic so that a first barcode is associated with a second barcode. The pairing can also be physically associated to form a junction between the two barcodes. For example, the first barcode sequence and second barcode sequences can be the tandem or reverse complements of each other. The target nucleic acid can be fragmented at the fragmentation sites, and a library of template nucleic acids can be prepared from the fragments. Sequencing information can be obtained from the library of template nucleic acids. Advantageously, the paired barcode sequences can be used to assemble sequencing data from the library of template nucleic acids. For example, identifying a first template nucleic acid comprising a first barcode sequence and a second template nucleic acid comprising a second barcode sequence that is paired with the first indicates that the first and second template nucleic acids represent sequences adjacent to one another in a sequence representation of the target nucleic acid. In essence, two sequences originally adjacent in the target nucleic acid may be separated from each other, introducing one barcode of a barcode pair at their points of separation, so that even when sequenced separately, the detection of the paired barcodes indicates the original proximity of the two sequences. Despite having been separated physically, they can be reunited bioinformatically, much as each divided half sought its other half in Plato's *Symposium*. Thus, a sequence representation of the target nucleic acid can be assembled by identifying further sequencing reads comprising paired barcode sequences. Such methods can be used to assemble a sequence representation of a target nucleic acid *de novo*, without the

IDT Exhibit 1020.4828

requirement of a reference genome. Moreover, such methods are also useful to sequence target nucleic acids comprising highly repetitive sequences. The methods can also be used to reassemble alternative isoforms and splice junctions in cDNAs, as well as reassemble single-molecule haplotypes.

Because the methods herein involve addition of one or more tags (such as an identifier code) at the location that previously existed between two adjacent sequences, these methods can be referred to as "junction tagging". More generally, the addition of an identifier tag to an individual nucleic acid can be termed "code tagging".

Definitions

As used herein the term "nucleic acid" and/or "oligonucleotide" and/or grammatical equivalents thereof can refer to at least two nucleotide monomers linked together. A nucleic acid can generally contain phosphodiester bonds; however, in some embodiments, nucleic acid analogs may have other types of backbones, comprising, for example, phosphoramide (Beaucage, *et al., Tetrahedron*, 49:1925 (1993); 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 (1986)), 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-methylphosphoroamidite linkages (see Eckstein, *Oligonucleotides and Analogues: A Practical Approach*, Oxford University Press), and peptide nucleic acid 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)).

Other analog nucleic acids include those with 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., Nucleosides & Nucleotides*, 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

IDT Exhibit 1020.4829

(1996)) and non-ribose (U.S. Patent No. 5,235,033 and No. 5,034,506, and Chapters 6 and 7, *ASC Symposium* Series 580, "Carbohydrate Modifications in Antisense Research", Ed. Y. S. Sanghui and P. Dan Coo). Nucleic acids may also contain one or more carbocyclic sugars (see Jenkins, *et al.,* Chem. Soc. Rev., (1995) pp. 169 176).

Modifications of the ribose-phosphate backbone may be done to facilitate the addition of additional moieties such as labels, or to increase the stability of such molecules under certain conditions. In addition, mixtures of naturally occurring nucleic acids and analogs can be made. Alternatively, mixtures of different nucleic acid analogs, and mixtures of naturally occurring nucleic acids and analogs may be made. The nucleic acids may be single stranded or double stranded, as specified, or contain portions of both double stranded or single stranded sequence. The nucleic acid may be DNA, for example, genomic or cDNA, RNA or a hybrid, from single cells, multiple cells, or from multiple species, as with metagenomic samples, such as from environmental samples. A nucleic acid can contain any combination of deoxyribo- and ribo-nucleotides, and any combination of bases, including uracil, adenine, thymine, cytosine, guanine, inosine, xanthanine, hypoxanthanine, isocytosine, isoguanine, and base analogs such as nitropyrrole (including 3-nitropyrrole) and nitroindole (including 5-nitroindole), etc.

In some embodiments, a nucleic acid can include at least one promiscuous base. Promiscuous bases can base-pair with more than one different type of base. In some embodiments, a promiscuous base can base-pair with at least two different types of bases and no more than three different types of bases. An example of a promiscuous base includes inosine that may pair with adenine, thymine, or cytosine. Other examples include hypoxanthine, 5-nitroindole, acylic 5-nitroindole, 4-nitropyrazole, 4-nitroimidazole and 3-nitropyrrole (Loakes *et al., Nucleic Acid Res.* 22:4039 (1994); Van Aerschot *et al., Nucleic Acid Res.* 23:4363 (1995); Nichols *et al., Nature* 369:492 (1994); Bergstrom *et al., Nucleic Acid Res.* 25:1935 (1997); Loakes *et al.,* Nucleic Acid Res. 23:2361 (1995); Loakes *et al., J. Mol. Biol.* 270:426 (1997); and Fotin *et al., Nucleic Acid Res.* 26:1515 (1998)). Promiscuous bases that can base-pair with at least three, four or more types of bases can also be used.

As used herein, the term "nucleotide analog" and/or grammatical equivalents thereof can refer to synthetic analogs having modified nucleotide base portions, modified pentose portions, and/or modified phosphate portions, and, in the case of polynucleotides, modified internucleotide linkages, as generally described elsewhere (*e.g.,* Scheit,

IDT Exhibit 1020.4830

Nucleotide Analogs, John Wiley, New York, 1980; Englisch, Angew. *Chem. Int. Ed. Engl.* 30:613-29, 1991; Agarwal, *Protocols for Polynucleotides and Analogs*, Humana Press, 1994; and S. Verma and F. Eckstein, *Ann. Rev. Biochem.* 67:99-134, 1998). Generally, modified phosphate portions comprise analogs of phosphate wherein the phosphorous atom is in the +5 oxidation state and one or more of the oxygen atoms is replaced with a non-oxygen moiety, *e.g.,* sulfur. Exemplary phosphate analogs include but are not limited to phosphorothioate, phosphorodithioate, phosphoroselenoate, phosphorodiselenoate, phosphoroanilothioate, phosphoranilidate, phosphoramidate, boronophosphates, including associated counterions, *e.g.,* H^+, NH_4^+, Na^-, if such counterions are present. Example modified nucleotide base portions include but are not limited to 5-methylcytosine (5mC); C-5-propynyl analogs, including but not limited to, C-5 propynyl-C and C-5 propynyl-U; 2,6-diaminopurine, also known as 2-amino adenine or 2-amino-dA); hypoxanthine, pseudouridine, 2-thiopyrimidine, isocytosine (isoC), 5-methyl isoC, and isoguanine (isoG; *see, e.g.,* U.S. Pat. No. 5,432,272). Exemplary modified pentose portions include but are not limited to, locked nucleic acid (LNA) analogs including without limitation Bz-A-LNA, 5-Me-Bz-C-LNA, dmf-G-LNA, and T-LNA (*see, e.g., The Glen Report*, 16(2):5, 2003; Koshkin *et al., Tetrahedron* 54:3607-30, 1998), and 2'- or 3'-modifications where the 2'- or 3'-position is hydrogen, hydroxy, alkoxy (*e.g.,* methoxy, ethoxy, allyloxy, isopropoxy, butoxy, isobutoxy and phenoxy), azido, amino, alkylamino, fluoro, chloro, or bromo. Modified internucleotide linkages include phosphate analogs, analogs having achiral and uncharged intersubunit linkages (*e.g.,* Sterchak, E. P. *et al.,* Organic Chem., 52:4202, 1987), and uncharged morpholino-based polymers having achiral intersubunit linkages (*see, e.g.,* U.S. Pat. No. 5,034,506). Some internucleotide linkage analogs include morpholidate, acetal, and polyamide-linked heterocycles. In one class of nucleotide analogs, known as peptide nucleic acids, including pseudocomplementary peptide nucleic acids ("PNA"), a conventional sugar and internucleotide linkage has been replaced with a 2-aminoethylglycine amide backbone polymer (*see, e.g.,* Nielsen *et al., Science*, 254:1497-1500, 1991; Egholm *et al., J. Am. Chem. Soc.,* 114: 1895-1897 1992; Demidov *et al., Proc. Natl. Acad. Sci.* 99:5953-58, 2002; *Peptide Nucleic Acids: Protocols and Applications*, Nielsen, ed., Horizon Bioscience, 2004).

As used herein, the term "sequencing read" and/or grammatical equivalents thereof can refer to a repetitive process of physical or chemical steps that is carried out to

IDT Exhibit 1020.4831

obtain signals indicative of the order of monomers in a polymer. The signals can be indicative of an order of monomers at single monomer resolution or lower resolution. In particular embodiments, the steps can be initiated on a nucleic acid target and carried out to obtain signals indicative of the order of bases in the nucleic acid target. The process can be carried out to its typical completion, which is usually defined by the point at which signals from the process can no longer distinguish bases of the target with a reasonable level of certainty. If desired, completion can occur earlier, for example, once a desired amount of sequence information has been obtained. A sequencing read can be carried out on a single target nucleic acid molecule or simultaneously on a population of target nucleic acid molecules having the same sequence, or simultaneously on a population of target nucleic acids having different sequences. In some embodiments, a sequencing read is terminated when signals are no longer obtained from one or more target nucleic acid molecules from which signal acquisition was initiated. For example, a sequencing read can be initiated for one or more target nucleic acid molecules that are present on a solid phase substrate and terminated upon removal of the one or more target nucleic acid molecules from the substrate. Sequencing can be terminated by otherwise ceasing detection of the target nucleic acids that were present on the substrate when the sequencing run was initiated.

As used herein, the term "sequencing representation" and/or grammatical equivalents thereof can refer to information that signifies the order and type of monomeric units in the polymer. For example, the information can indicate the order and type of nucleotides in a nucleic acid. The information can be in any of a variety of formats including, for example, a depiction, image, electronic medium, series of symbols, series of numbers, series of letters, series of colors, etc. The information can be at single monomer resolution or at lower resolution, as set forth in further detail below. An exemplary polymer is a nucleic acid, such as DNA or RNA, having nucleotide units. A series of "A," "T," "G," and "C" letters is a well known sequence representation for DNA that can be correlated, at single nucleotide resolution, with the actual sequence of a DNA molecule. Other exemplary polymers are proteins having amino acid units and polysaccharides having saccharide units.

As used herein the term "at least a portion" and/or grammatical equivalents thereof can refer to any fraction of a whole amount. For example, "at least a portion" can refer to at least about 1%, 2%, 3%, 4%, 5%, 6%, 7%, 8%, 9%, 10%, 15%, 20%, 25%, 30%, 35%,

IDT Exhibit 1020.4832

40%, 45%, 50%, 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 99%, 99.9% or 100% of a whole amount.

Transposomes

A "transposome" is comprised of at least a transposase enzyme and a transposase recognition site. In some such systems, termed "transposomes", the transposase can form a functional complex with a transposon recognition site that is capable of catalyzing a transposition reaction. The transposase or integrase may bind to the transposase recognition site and insert the transposase recognition site into a target nucleic acid in a process sometimes termed "tagmentation". In some such insertion events, one strand of the transposase recognition site may be transferred into the target nucleic acid.

Some embodiments can include the use of a hyperactive Tn5 transposase and a Tn5-type transposase recognition site (Goryshin and Reznikoff, *J. Biol. Chem.*, 273:7367 (1998)), or MuA transposase and a Mu transposase recognition site comprising R1 and R2 end sequences (Mizuuchi, K., *Cell*, 35: 785, 1983; Savilahti, H, *et al., EMBO J.*, 14: 4893, 1995). An exemplary transposase recognition site that forms a complex with a hyperactive Tn5 transposase (*e.g.*, EZ-Tn5™ Transposase, Epicentre Biotechnologies, Madison, Wisconsin) comprises the following 19b transferred strand (sometimes "M" or "ME") and non-transferred strands: 5′ AGATGTGTATAAGAGACAG 3′, (SEQ ID NO:1), 5′ CTGTCT CTTATACACATCT 3′ (SEQ ID NO:2), respectively.

More examples of transposition systems that can be used with certain embodiments provided herein include *Staphylococcus aureus* Tn552 (Colegio *et al., J. Bacteriol.*, 183: 2384-8, 2001; Kirby C *et al., Mol. Microbiol.*, 43: 173-86, 2002), Ty1 (Devine & Boeke, *Nucleic Acids Res.*, 22: 3765-72, 1994 and International Publication WO 95/23875), Transposon Tn7 (Craig, N L, *Science.* 271: 1512, 1996; Craig, N L, Review in: *Curr Top Microbiol Immunol.*, 204:27-48, 1996), Tn/O and IS10 (Kleckner N, *et al., Curr Top Microbiol Immunol.*, 204:49-82, 1996), Mariner transposase (Lampe D J, *et al., EMBO J.*, 15: 5470-9, 1996), Tc1 (Plasterk R H, *Curr. Topics Microbiol. Immunol.*, 204: 125-43, 1996), P Element (Gloor, G B, *Methods Mol. Biol.*, 260: 97-114, 2004), Tn3 (Ichikawa & Ohtsubo, *J Biol. Chem.* 265:18829-32, 1990), bacterial insertion sequences (Ohtsubo & Sekine, *Curr. Top. Microbiol. Immunol.* 204: 1-26, 1996), retroviruses (Brown, *et al., Proc Natl Acad Sci USA*, 86:2525-9, 1989), and retrotransposon of yeast (Boeke & Corces, *Annu Rev Microbiol.* 43:403-34, 1989). More examples include IS5,

IDT Exhibit 1020.4833

Tn10, Tn903, IS911, and engineered versions of transposase family enzymes (Zhang *et al.,* (2009) *PLoS Genet.* 5:e1000689. Epub 2009 Oct 16; Wilson C. et al (2007) *J. Microbiol. Methods* 71:332-5).

Transposon sequences

Some embodiments provided herein include transposon sequences. In some embodiments, a transposon sequence includes at least one transposase recognition site and at least one barcode. In some embodiments, a transposon sequence includes a first transposon recognition site, a second transposon recognition site, and a barcode disposed therebetween. **FIG. 1** depicts a schematic a transposon sequence.

A transposase recognition site can include two complementary nucleic acid sequences, *e.g.,* a double-stranded nucleic acid or a hairpin nucleic acid, that comprise a substrate for a transposase or integrase. An exemplary embodiment of a transposon sequence comprising a double-stranded nucleic acid comprising non-complementary strands is depicted in **FIG. 2**. In some embodiments, a transposon sequence can include at least a portion comprising at least a portion comprising non-complementary sequences.

— single-stranded sequences

In other embodiments, the transposon sequence contains one or more single-stranded portions. Transposons with single-stranded internal sequences can be useful for avoiding the formation of hairpin structures, or when it is undesirable for transposomes to insert into previously transposome-inserted sequences ("self-insertion"), since many transposases will not insert into single-stranded DNA.

Single-stranded transposons with codes can also be prepared by making single-stranded transposons with hair-pinned ends and removing the hair-pin loops by cutting the DNA with endonucleases, such as restriction endonucleases, or uracil DNA deglycosylase (UDG), as shown in **FIG. 3**, upper part. This approach avoids sequence constraints of internal nicking recognition sites. The use of biotin-labeled primers can allow post-amplification return to single-stranded templates.

To avoid hairpins, synthesizing single-stranded transposons can be synthesized with each end forming a hairpin, as shown in **FIG. 3**, lower part. This hairpin prevents opposite-facing M-sites from forming hairpins. These hairpin structures are later

IDT Exhibit 1020.4834

processed to create active transposons with double-stranded M-sites and 5' phosphate and 3' OH groups.

— different first and second transposase recognition sequences

A particularly useful transposon sequence has different transposase recognition sequences (M and M') to reduce the subsequent formation of hairpin structures. For example, instead of having the same M-ends, a transposon can utilize an ME end (CTGTCTCTTATACACATCT or ME1) with an IE (inner end) end (CTGTCTCTTGATCAGATCT or ME2). An OE (outer ends) can also be combined with an IE. Preparation of the transposon is shown in **FIG. 4**. The degree of difference can be determined by screening for how much variation can be present in the ends to allow efficient transposition. A combination of two different ends that yield high transposition efficiency can be chosen from this screening.

The result is a transposome complex that has one transposon sequence and two transposases, as in **FIG. 5**. As shown, the transposase does not insert into single-stranded DNA such as the internal transposon sequence. After tagmentation, the single-stranded gap can be filled and ligated, as shown in **FIG. 6**. In this figure, there is no transposition into the transposon, and no ME-ME' hairpin is formed, since ME1 and ME2 are different.

— looped structures

A useful transposome is has one transposon that binds two transposase subunits forming a "looped" complex (**FIG. 7**). This configuration ensures that transposons are inserted into target DNA while maintaining ordering information of the original target DNA and without fragmenting the target DNA.

The transposon can be "pre-bent" in a configuration that favors the "looped" complex configuration, where transposon DNA ends have to come together and bend to derive the "looped" configuration. Without being tied to a particular mechanism, it is believed that the transposase subunits bind the ME-ends first, followed by dimerization of the subunits to form an active complex. It would be favorable for the two ends of the transposon to already be in a configuration that allows efficient intramolecular dimerization of the two bound transposase units, and can improve the yield of "looped" complex formation and stabilize the final complex. Such a configuration would also minimize undesired intermolecular dimerization products. To facilitate this bending

IDT Exhibit 1020.4835

process, the ends are positioned in close proximity, combining bending of transposon DNA with transposase complex formation.

Methods for bending DNA include surface binding, ion induction, chemical modification of DNA, protein-like bending with proteins such as Cro, histones, integration host factor (IHF), Fos, Yeast MCM1p, Jun, HU, dendrimer binding, and use of specific DNA sequences. For example, an IHF protein-DNA complex can be used to pre-bend because IHF recognizes and bends DNA at a specific sequence. Other proteins or other methods do not require a specific sequence. The introduction of one or more bends allows the creation of a specific 2- or 3-dimensional shape of the transposon that position the ME-ends in a desired position for "looped" complex formation. The agents responsible for bending can also prevent transposition into transposon-transposase complexes, and therefore favor insertion into target DNA.

Another approach favoring "looped" transposon-transposase complex formation is to relieve torsional stress, increasing rotational freedom in the transposon in order to allow the two ME-ends adapt a favorable orientation for "looped" complex formation. DNA molecules of certain sizes can be rigid ordinarily, so that nicks, gaps, single-stranded regions within the transposon can make the transposon flexible and adaptive. For example, formation of transposomes is enhanced when formed with transposons with a ~13bp region that is single-stranded, compared to its double-stranded version.

— pre-cleaved transposomes

In another alternative, pre-cleaved transposomes are prepared with a looped configuration (**FIG. 8**), ensuring that neighboring DNA fragments receive code combinations that can be unambiguously assembled at a later stage of the assay. As shown, the fragmentation maintains the code combination of N and N'. Subsequent fragmentation between the code-pair combination results in transposon-transposase complexes in which the transposon-ends are not linked to each other (*i.e.* a linear configuration). However, the complexes maintain the code-combinations as they were introduced in the looped configuration. Upon insertion of the fragmented transposon into the target, DNA fragments are created with codes on the ends, allowing code-mediated assembly of the sequenced DNA fragments/libraries.

The preparation of complexes with the looped configuration can be favored by various techniques. In one technique, the complexes are prepared in a dilute environment

IDT Exhibit 1020.4836

with excess transposase to favor formation of looped complexes. A subsequent concentration step prepares the complexes at a concentration desired for insertion. In another technique, the slow addition of transposons to transposases creates an environment with low amounts of transposon with excess transposase. This reaction condition ensures that each transposon will bind to two transposase units, forming the intra-molecular complex.

Nick-based methods

The methods herein are not limited to using the transposomes described above. Another embodiment can involve nicking of target DNA, which exploits the genomic DNA to provide the junction tags. Nicking methods can also be used in combination with transposon methods, such as preceding or following tagmentation. Any method that nicks a single strand of the target DNA may be used, preferably those that create random gaps in DNA. When random, the nicking methods can avoid any insertional bias that transposons may have. Alternatively, sequence-specific nicking enzymes can be used, or enzymes or chemical methods to remove a base or nucleotide from a single strand. Sequence-specific nicking enzymes include Nt.CviPII, Nb.BsmI, Nb.BbvCI, Nb.Bsr.DI, Nb.BtsI, Nt.BsmAI, Nt.BbvCI, Nt.BspQI, Nt.AlwI, and Nt.BstNBI (New England Biolabs).

As shown in **FIG. 9**, strand-displacement synthesis is used to copy identical flanking genomic DNA sequences (g-codes) at the end of formerly juxtaposed DNA fragments. The g-code is endogenous and therefore all the sequenced bases belong to the target DNA and not to any exogenously introduced code. After DNA fragments have obtained two identical g-codes, adapters or unique adapters (adapters with unique barcode) can be ligated to allow sequencing. The g-code sequences then serve as junction tags to assemble the individual fragments.

Alternatively, nicking followed by 3'-exo digestion can create gaps in genomic DNA without fragmenting the entire DNA sample, as shown in **FIG. 10**. The gaps can then be filled with code-containing oligonucleotides. For example, the oligonucleotides can be ligated to the two ends that were exposed by the nicking. Depending on the ligation method used (*e.g.* chemical or enzymatic), ligation may also be facilitated by phosphorylation of the oligo or the nicked end.

IDT Exhibit 1020.4837

Barcodes

Generally, a barcode can include one or more nucleotide sequences that can be used to identify one or more particular nucleic acids. The barcode can be an artificial sequence, or can be a naturally occurring sequence, such as a g-code, described herein. A barcode can comprise at least about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more consecutive nucleotides. In some embodiments, a barcode comprises at least about 10, 20, 30, 40, 50, 60, 70 80, 90, 100 or more consecutive nucleotides. In some embodiments, at least a portion of the barcodes in a population of nucleic acids comprising barcodes is different. In some embodiments, at least about 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 99% of the barcodes are different. In more such embodiments, all of the barcodes are different. The diversity of different barcodes in a population of nucleic acids comprising barcodes can be randomly generated or non-randomly generated.

In some embodiments, a transposon sequence comprises at least one barcode. In some embodiments, a transposon sequence comprises a barcode comprising a first barcode sequence and a second barcode sequence. In some such embodiments, the first barcode sequence can be identified or designated to be paired with the second barcode sequence. For example, a known first barcode sequence can be known to be paired with a known second barcode sequence using a reference table comprising a plurality of first and second bar code sequences known to be paired to one another.

In another example, the first barcode sequence can comprise the same sequence as the second barcode sequence. In another example, the first barcode sequence can comprise the reverse complement of the second barcode sequence. In some embodiments, the first barcode sequence and the second barcode sequence are different ("bi-codes").

— bi-codes

As depicted in **FIG. 11**, transposons can be generated with either engineered bi-codes or random bi-codes. In the case of random bi-codes, the code pool may be down-selected (sub-sampled) to create unique code-code pairs. These unique bi-code combinations can then be identified by sequencing the transposons. The particular bi-code combinations (A_i and B_j) present in the final pool are determined and allow creation

IDT Exhibit 1020.4838

of a look-up table, which can be used to reassemble reads as described in the methods below.

An advantage of this approach is that not every code combination needs to be prepared individually. For example, *in situ* array technology can synthesize transposons or precursors with pre-designed code combinations in spatially separated locations, generating greater than 10^{18} different transposons or precursors in a single synthesis. To create unique bi-code maps, the codes are preferably down-selected so that a single first code is paired with only a single second code.

A method for reducing the presence of non-unique code-combinations (*e.g.*, Code1-Code2 and Code1-Code3 both present) is the use of "bottle-necking" or sub-sampling approaches. In these approaches, a large pool of many code combinations is generated. For example, two 15bp random sequences (4^{15}) can result in more than 10^{18} different transposons. Taking a fraction of this high-complexity pool results in unique code-combinations for the sub-sampled fraction.

In another embodiment, a transposon sequence comprising a double-strand nucleic acid, each strand can comprise a different barcode sequence. As will be understood, such a transposon sequence can be useful to inhibit insertion of other sequences into the transposon sequence by transposition. Such transposon sequences can also be used to tag particular strands of a double-stranded target nucleic acid, for example in methods that include haplotypes sequencing. An exemplary transposon sequence comprising a plurality of barcode sequences is depicted in **FIG. 2.**

— **half-transposons**

In another embodiment, code combinations are introduced after formation of the transposon-transposase complex. Complexes with non-linked ME regions can generate linked code information after transposon insertion and subsequent downstream assay steps. Complexes can be prepared as illustrated in **FIG. 12**, in which transposon ends are not linked. For example, complexes can be formed from two "half-transposons", each containing one ME region and a random or specific code region. This generates transposase complexes with randomly inserted transposons ends; no prior knowledge of pairing between the two half-codes is required. Each transposon contains a ME sequence for transposase recognition, a random or specific code, and an endonuclease fragmentation system. Upon insertion, the transposase enzyme can prevent target DNA

IDT Exhibit 1020.4839

fragmentation by binding both transposon ends. The ends of transposons can now be ligated to each other via the sticky ends. Upon nicking each strand (*e.g.*, by a sequence-specific fragmentation system, such as a restriction endonuclease), sticky code-tagged ends can be created that link one DNA fragment to the other. The codes generated at each junction are reverse complements to one another. Denaturation creates DNA fragments in which each DNA fragment contains a linked code with the neighboring fragment. An advantage of this embodiment is that transposon-transposase complexes with specific or random codes can readily be prepared, without the need for making complexes with looped configurations. Furthermore in this embodiment, the code combination does not have to be introduced at the transposase complex level, but can be generated at a later stage in the assay.

As will be described further herein, paired first and second barcode sequences can be used to identify different nucleic acids comprising barcodes linked with one another.

Linkers

Some embodiments include transposon sequences comprising a first barcode sequence and a second barcode sequence having a linker disposed therebetween. In other embodiments, the linker can be absent, or can be the sugar-phosphate backbone that connects one nucleotide to another. The linker can comprise, for example, one or more of a nucleotide, a nucleic acid, a non-nucleotide chemical moiety, a nucleotide analogue, amino acid, peptide, polypeptide, or protein. In preferred embodiments, a linker comprises a nucleic acid. The linker can comprise at least about 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, or more nucleotides. In some embodiments, a linker can comprise at least about 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more nucleotides.

— nonamplifiable linkers

In another embodiment, non-amplifiable linkers can be introduced between ME regions (**FIG. 13**), which can be useful in methods using bi-codes, as in **FIG. 14.** Prior to Round 1 in the figure, it is preferable to perform an amplification with P1 primers to increase the number of copies, and to ensure that all templates in the population are sequenceable. Examples of non-amplifiable linkers include organic chemistry linkers such as alkyl, propyl, PEG; non-natural bases such as IsoC, isoG; or any group that does

IDT Exhibit 1020.4840

not amplify in DNA-based amplification schemes. For example, transposons containing isoC, isoG pairs can be amplified with dNTPs mixtures lacking a complementary isoG and isoC, ensuring that no amplification occurs across the inserted transposons.

In this method, transposons are inserted into target DNA with the following configuration: ME-Code*N*-P1-linker-P1'-Code*N*'-ME so that code*N* and Code*N*' generate the code combination together. P1 is a primer and P1' is its complement. PCR amplification with P1 or P1' (*e.g.,* generates DNA fragments that are code-tagged. Fragmentation at the linker site or a non-amplifiable linker can reduce undesirable amplification across transposons, which may reduce the yield of the libraries or generate undesired products.

Fragmentation sites

In some embodiments, the linker can comprise a fragmentation site. A fragmentation site can be used to cleave the physical, but not the informational association between a first barcode sequence and a second barcode sequence. Cleavage may be by biochemical, chemical or other means. In some embodiments, a fragmentation site can include a nucleotide or nucleotide sequence that may be fragmented by various means. For example, a fragmentation site may be a substrate for an enzyme, such as a nuclease, that will cleave the physical association between a first barcode sequence and a second barcode sequence. For example, the fragmentation site comprises a restriction endonuclease site and may be cleaved with an appropriate restriction endonuclease. In another example, a fragmentation site can comprise at least one ribonucleotide in a nucleic acid that may otherwise comprise deoxyribonucleotides and may be cleaved with an RNAse. Chemical cleavage agents capable of selectively cleaving the phosphodiester bond between a deoxyribonucleotide and a ribonucleotide include metal ions, for example rare-earth metal ions (*e.g.,* La^{3+}, particularly Tm^{3+}, Yb^{3+} or Lu^{3+} (Chen *et al.* Biotechniques. 2002, 32: 518-520; Komiyama *et al.* Chem. Commun. 1999, 1443-1451)), Fe(3) or Cu(3), or exposure to elevated pH, *e.g.,* treatment with a base such as sodium hydroxide. As used herein, selective cleavage of the phosphodiester bond between a deoxyribonucleotide and a ribonucleotide can refer to the chemical cleavage agent is not capable of cleaving the phosphodiester bond between two deoxyribonucleotides under the same conditions.

IDT Exhibit 1020.4841

In another example, the fragmentation site can comprise one or more recognition sequences for a nickase, that is, a nicking endonuclease that breaks one strand of a double-stranded nucleic acid. Thus, the fragmentation site can comprise a first nickase recognition sequence, a second nickase recognition sequence. The cut site for each recognition sequence can be the same site or different site.

In another example, a fragmentation site can include one or more nucleotide analogues that comprise an abasic site and permits cleavage at the fragmentation site in the presence of certain chemical agents, such as polyamine, N,N′-dimethylethylenediamine (DMED) (U.S. Patent Publication No. 2010/0022403). In some embodiments, an abasic site may be created within a fragmentation site by first providing a fragmentation site comprising a deoxyuridine (U) of a double stranded nucleic acid. The enzyme uracil DNA glycosylase (UDG) may then be used to remove the uracil base, generating an abasic site on one strand. The polynucleotide strand including the abasic site may then be cleaved at the abasic site by treatment with endonuclease (e.g. Endo IV endonuclease, AP lyase, FPG glycosylase/AP lyase, Endo VIII glycosylase/AP lyase), heat or alkali. Abasic sites may also be generated at nucleotide analogues other than deoxyuridine and cleaved in an analogous manner by treatment with endonuclease, heat or alkali. For example, 8-oxo-guanine can be converted to an abasic site by exposure to FPG glycosylase. Deoxyinosine can be converted to an abasic site by exposure to AlkA glycosylase. The abasic sites thus generated may then be cleaved, typically by treatment with a suitable endonuclease (e.g. Endo IV, AP lyase). (U.S. Patent Publication No. 2011/0014657).

In another example, a fragmentation site may include a diol linkage which permits cleavage by treatment with periodate (*e.g.,* sodium periodate). In another example, a fragmentation site may include a disulphide group which permits cleavage with a chemical reducing agent, e.g. Tris (2-carboxyethyl)-phosphate hydrochloride (TCEP).

In some embodiments, a fragmentation site may include a cleavable moiety that may be subject to photochemical cleavage. Photochemical cleavage encompasses any method which utilizes light energy in order to achieve cleavage of nucleic acids, for example, one or both strands of a double-stranded nucleic acid molecule. A site for photochemical cleavage can be provided by a non-nucleotide chemical moiety in a nucleic acid, such as phosphoramidite [4-(4,4'-dimethoxytrityloxy)butyramidomethyl)-1-

IDT Exhibit 1020.4842

(2-nitrophenyl)-ethyl]-2-cyanoethyl-(N,N-diisopropyl)-phosphoramidite) (Glen Research, Sterling, Va., USA, Cat No. 10-4913-XX).

In some embodiments, a fragmentation site can include a peptide, for example, conjugate structure in which a peptide molecule is linked to a nucleic acid. The peptide molecule can subsequently be cleaved by a peptidase enzyme of the appropriate specificity, or any other suitable means of non-enzymatic chemical or photochemical cleavage. In some embodiments, a conjugate between peptide and nucleic acid will be formed by covalently linking a peptide to a nucleic acid, *e.g.*, a strand of a double-stranded nucleic acid. Conjugates between a peptide and nucleic acid can be prepared using techniques generally known in the art. In one such technique the peptide and nucleic acid components of the desired amino acid and nucleotide sequence can be synthesized separately, e.g. by standard automated chemical synthesis techniques, and then conjugated in aqueous/organic solution. By way of example, the OPeC™ system commercially available from Glen Research is based on the native ligation of an N-terminal thioester-functionalized peptide to a 5′-cysteinyl oligonucleotide.

Primer sites

In some embodiments, a linker can be a "sequencing adaptor" or "sequencing adaptor site", that is to say a region that comprises one or more sites that can hybridize to a primer. In some embodiments, a linker comprises at least a first primer site. In some embodiments, a linker comprises at least a first primer site and a second primer site. The orientation of the primer sites in such embodiments can be such that a primer hybridizing to the first primer site and a primer hybridizing to the second primer site are in the same orientation, or in different orientations. In one embodiment, the primer sequence in the linker can be complementary to a primer used for amplification. In another embodiment, the primer sequence is complementary to a primer used for sequencing.

In some embodiments, a linker can include a first primer site, a second primer site having a non-amplifiable site disposed therebetween. The non-amplifiable site is useful to block extension of a polynucleotide strand between the first and second primer sites, wherein the polynucleotide strand hybridizes to one of the primer sites. The non-amplifiable site can also be useful to prevent concatamers. Examples of non-amplifiable sites include a nucleotide analogue, non-nucleotide chemical moiety, amino-acid, peptide,

IDT Exhibit 1020.4843

and polypeptide. In some embodiments, a non-amplifiable site comprises a nucleotide analogue that does not significantly base-pair with A, C, G or T.

Some embodiments include a linker comprising a first primer site, a second primer site having a fragmentation site disposed therebetween.

Other embodiments can use a forked or Y-shaped adapter design useful for directional sequencing, as described in U.S. Patent No. 7,741,463. An example is shown in **FIG. 15.**

Affinity tags

In some embodiments, a linker can comprise an affinity tag. Affinity tags can be useful for the bulk separation of target nucleic acids hybridized to hybridization tags. As used herein, the term "affinity tag" and grammatical equivalents can refer to a component of a multi-component complex, wherein the components of the multi-component complex specifically interact with or bind to each other. For example an affinity tag can include biotin or His that can bind streptavidin or nickel, respectively. Other examples of multiple-component affinity tag complexes include, ligands and their receptors, for example, avidin-biotin, streptavidin-biotin, and derivatives of biotin, streptavidin, or avidin, including, but not limited to, 2-iminobiotin, desthiobiotin, NeutrAvidin (Molecular Probes, Eugene, Oreg.), CaptAvidin (Molecular Probes), and the like; binding proteins/peptides, including maltose-maltose binding protein (MBP), calcium-calcium binding protein/peptide (CBP); antigen-antibody, including epitope tags, including c-MYC (*e.g.,* EQKLISEEDL (SEQ ID NO: 3)), HA (*e.g.,* YPYDVPDYA (SEQ ID NO: 4)), VSV-G (*e.g.,* YTDIEMNRLGK (SEQ ID NO: 5)), HSV (*e.g.,* QPELAPEDPED (SEQ ID NO: 6)), V5 (*e.g.,* GKPIPNPLLGLDST (SEQ ID NO: 7)), and FLAG Tag™. (*e.g.,* DYKDDDDKG (SEQ ID NO: 8)), and their corresponding anti-epitope antibodies; haptens, for example, dinitrophenyl and digoxigenin, and their corresponding antibodies; aptamers and their corresponding targets; poly-His tags (*e.g.,* penta-His and hexa-His) and their binding partners including corresponding immobilized metal ion affinity chromatography (IMAC) materials and anti-poly-His antibodies; fluorophores and anti-fluorophore antibodies; and the like.

IDT Exhibit 1020.4844

Reporter moieties

In some embodiments, a linker can comprise a reporter moiety. As used herein, the term "reporter moiety" and grammatical equivalents can refer to any identifiable tag, label, or group. The skilled artisan will appreciate that many different species of reporter moieties can be used with the methods and compositions described herein, either individually or in combination with one or more different reporter moieties. In certain embodiments, a reporter moiety can emit a signal. Examples of signals fluorescent, a chemiluminescent, a bioluminescent, a phosphorescent, a radioactive, a calorimetric, or an electrochemiluminescent signals. Example reporter moieties include fluorophores, radioisotopes, chromogens, enzymes, antigens including epitope tags, semiconductor nanocrystals such as quantum dots, heavy metals, dyes, phosphorescence groups, chemiluminescent groups, electrochemical detection moieties, binding proteins, phosphors, rare earth chelates, transition metal chelates, near-infrared dyes, electrochemiluminescence labels, and mass spectrometer compatible reporter moieties, such as mass tags, charge tags, and isotopes. More reporter moieties that may be used with the methods and compositions described herein include spectral labels such as fluorescent dyes (*e.g.,* fluorescein isothiocyanate, Texas red, rhodamine, and the like), radiolabels (*e.g.,* ^{3}H, ^{125}I, ^{35}S, ^{14}C, ^{32}P, ^{33}P, etc.), enzymes (*e.g.,* horseradish peroxidase, alkaline phosphatase etc.) spectral calorimetric labels such as colloidal gold or colored glass or plastic (e.g. polystyrene, polypropylene, latex, etc.) beads; magnetic, electrical, thermal labels; and mass tags. Reporter moieties can also include enzymes (horseradish peroxidase, etc.) and magnetic particles. More reporter moieties include chromophores, phosphors and fluorescent moieties, for example, Texas red, dixogenin, biotin, 1- and 2-aminonaphthalene, p,p'-diaminostilbenes, pyrenes, quaternary phenanthridine salts, 9-aminoacridines, p,p'-diaminobenzophenone imines, anthracenes, oxacarbocyanine, merocyanine, 3-aminoequilenin, perylene, bis-benzoxazole, bis-p-oxazolyl benzene, 1,2-benzophenazin, retinol, bis-3-aminopyridinium salts, hellebrigenin, tetracycline, sterophenol, benzimidazolylphenylamine, 2-oxo-3-chromen, indole, xanthen, 7-hydroxycoumarin, phenoxazine, calicylate, strophanthidin, porphyrins, triarylmethanes and flavin. Individual fluorescent compounds which have functionalities for linking to an element desirably detected in an apparatus or assay provided herein, or which can be modified to incorporate such functionalities include, *e.g.,* dansyl chloride; fluoresceins such as 3,6-dihydroxy-9-phenylxanthydrol; rhodamineisothiocyanate; N-phenyl 1-amino-

IDT Exhibit 1020.4845

8-sulfonatonaphthalene; N-phenyl 2-amino-6-sulfonatonaphthalene; 4-acetamido-4-isothiocyanato-stilbene-2,2'-disulfonic acid; pyrene-3-sulfonic acid; 2-toluidinonaphthalene-6-sulfonate; N-phenyl-N-methyl-2-aminoaphthalene-6-sulfonate; ethidium bromide; stebrine; auromine-0,2-(9'-anthroyl)palmitate; dansyl phosphatidylethanolamine; N,N'-dioctadecyl oxacarbocyanine: N,N'-dihexyl oxacarbocyanine; merocyanine, 4-(3'-pyrenyl)stearate; d-3-aminodesoxy-equilenin; 12-(9'-anthroyl)stearate; 2-methylanthracene; 9-vinylanthracene; 2,2'(vinylene-p-phenylene)bisbenzoxazole; p-bis(2- -methyl-5-phenyl-oxazolyl))benzene; 6-dimethylamino-1,2-benzophenazin; retinol; bis(3'-aminopyridinium) 1,10-decandiyl diiodide; sulfonaphthylhydrazone of hellibrienin; chlorotetracycline; N-(7-dimethylamino-4-methyl-2-oxo-3-chromenyl)maleimide; N-(p-(2benzimidazolyl)-phenyl)maleimide; N-(4-fluoranthyl)maleimide; bis(homovanillic acid); resazarin; 4-chloro7-nitro-2,1,3-benzooxadiazole; merocyanine 540; resorufin; rose bengal; 2,4-diphenyl-3(2H)-furanone, fluorescent lanthanide complexes, including those of Europium and Terbium, fluorescein, rhodamine, tetramethylrhodamine, eosin, erythrosin, coumarin, methyl-coumarins, quantum dots (also referred to as "nanocrystals": see U.S. Patent No. 6,544,732), pyrene, Malachite green, stilbene, Lucifer Yellow, Cascade Blue™, Texas Red, Cy dyes (Cy3, Cy5, etc.), Alexa Fluor® dyes, phycoerythin, bodipy, and others described in the 6th Edition of the *Molecular Probes Handbook* by Richard P. Haugland,.

Certain methods of making transposon sequences

The transposon sequences provided herein can be prepared by a variety of methods. Example methods include direct synthesis and hairpin extension methods. In some embodiments, transposon sequences may be prepared by direct synthesis. For example, a transposon sequence comprising a nucleic acid may be prepared by methods comprising chemical synthesis. Such methods are well known in the art, *e.g.*, solid phase synthesis using phophoramidite precursors such as those derived from protected 2′-deoxynucleosides, ribonucleosides, or nucleoside analogues.

In some embodiments, a transposon sequence can be prepared by hairpin extension. In some such embodiments, a portion of a transposon sequence may be prepared by chemical synthesis and extended by hairpin extension. In an example embodiment, a precursor transposon sequence comprising a polynucleotide may include a first transposase recognition site, and a first barcode sequence. The precursor transposon

IDT Exhibit 1020.4846

sequence may be extended using an appropriate nucleic acid polymerase by hairpin extension, thereby preparing a hairpin structure comprising a first transposase recognition site, a first barcode sequence, a second barcode sequence, and a second transposase recognition site. In such a transposon sequence, the first transposase recognition site and first barcode sequence can have the reverse complement sequence of, and the second transposase recognition site and the second barcode sequence, respectively.

Some methods of preparing transposons sequences can include preparing barcode sequences. Barcode sequences can be generated randomly and non-randomly. Some barcode sequences may or may not include sequences likely to be found in a target nucleic acid. Some barcode sequences may or may not include restriction sites. In some embodiments, at least a portion of a barcode sequence can be generated randomly. In some embodiments a barcode sequence can be generated using combinatorial methods. In some such methods, barcode sequences can comprise one or more subunits comprising one or more consecutive nucleotides. As will be understood, in some embodiments, a subunit can comprise a nucleotide analogue and/or a nucleotide comprising a label. At least a portion of a barcode sequence can comprise at least one subunit. Barcode sequences comprising at least one subunit can be generated randomly or non-randomly. In some embodiments, at least a portion of a barcode sequences comprising at least one subunit can be generated randomly or non-randomly. As will be understood, in some embodiments, at least a portion of a barcode sequence is known.

Primer sites may be ligated to the ends of the hairpin structure in order to generate a complementary strand to the single-strand of the hairpin structure. An exemplary embodiment of a method of making a transposon sequence is depicted in **FIG. 16**. It will be understood that the primer sites introduced to generate the complementary strand may be removed by a transposase during a transposition reaction.

In some embodiments, methods of making transposon sequences can include incorporating sequences using amplification methods. Templates for amplification that include barcode sequences can be prepared by a variety of systems, for example, using *in situ* oligonucleotide arrays. Some methods of making transposon sequences include the use of tailed-oligonucleotides to incorporate sequences into a transposon sequence. An exemplary embodiment is depicted in **FIG. 17,** where tailed-oligonucleotides comprising mosaic elements (ME) are used to prepare a transposon sequence comprising a barcode comprising Code X and Code Y, and a linker comprising a fragmentation site.

IDT Exhibit 1020.4847

Target nucleic acids

A target nucleic acid can include any nucleic acid of interest. Target nucleic acids can include DNA, RNA, peptide nucleic acid, morpholino nucleic acid, locked nucleic acid, glycol nucleic acid, threose nucleic acid, mixtures thereof, and hybrids thereof. In a preferred embodiment, genomic DNA fragments or amplified copies thereof are used as the target nucleic acid. In another preferred embodiment, mitochondrial or chloroplast DNA is used.

A target nucleic acid can comprise any nucleotide sequence. In some embodiments, the target nucleic acid comprises homopolymer sequences. A target nucleic acid can also include repeat sequences. Repeat sequences can be any of a variety of lengths including, for example, 2, 5, 10, 20, 30, 40, 50, 100, 250, 500, 1000 nucleotides or more. Repeat sequences can be repeated, either contiguously or non-contiguously, any of a variety of times including, for example, 2, 3, 4, 5, 6, 7, 8, 9, 10, 15, 20 times or more.

Some embodiments described herein can utilize a single target nucleic acid. Other embodiments can utilize a plurality of target nucleic acids. In such embodiments, a plurality of target nucleic acids can include a plurality of the same target nucleic acids, a plurality of different target nucleic acids where some target nucleic acids are the same, or a plurality of target nucleic acids where all target nucleic acids are different. Embodiments that utilize a plurality of target nucleic acids can be carried out in multiplex formats so that reagents are delivered simultaneously to the target nucleic acids, for example, in one or more chambers or on an array surface. In some embodiments, the plurality of target nucleic acids can include substantially all of a particular organism's genome. The plurality of target nucleic acids can include at least a portion of a particular organism's genome including, for example, at least about 1%, 5%, 10%, 25%, 50%, 75%, 80%, 85%, 90%, 95%, or 99% of the genome. In particular embodiments the portion can have an upper limit that is at most about 1%, 5%, 10%, 25%, 50%, 75%, 80%, 85%, 90%, 95%, or 99% of the genome

Target nucleic acids can be obtained from any source. For example, target nucleic acids may be prepared from nucleic acid molecules obtained from a single organism or from populations of nucleic acid molecules obtained from natural sources that include one or more organisms. Sources of nucleic acid molecules include, but are not limited to, organelles, cells, tissues, organs, or organisms. Cells that may be used as sources of target

IDT Exhibit 1020.4848

nucleic acid molecules may be prokaryotic (bacterial cells, for example, *Escherichia, Bacillus, Serratia, Salmonella, Staphylococcus, Streptococcus, Clostridium, Chlamydia, Neisseria, Treponema, Mycoplasma, Borrelia, Legionella, Pseudomonas, Mycobacterium, Helicobacter, Erwinia, Agrobacterium, Rhizobium,* and *Streptomyces* genera); archeaon, such as crenarchaeota, nanoarchaeota or euryarchaeotia; or eukaryotic such as fungi, (for example, yeasts), plants, protozoans and other parasites, and animals (including insects (for example, *Drosophila* spp.), nematodes (*e.g., Caenorhabditis elegans*), and mammals (for example, rat, mouse, monkey, non-human primate and human)).

Methods of preparing template nucleic acids

Some embodiments include methods of preparing template nucleic acids. As used herein, the term "template nucleic acid" can refer to a target nucleic acid, a fragment thereof, or any copy thereof comprising at least one transposon sequence, a fragment thereof, or any copy thereof. Accordingly, some methods of preparing template nucleic acids include inserting a transposon sequence into a target nucleic acid, thereby preparing a template nucleic acid. Some methods of insertion include contacting a transposon sequence provided herein with a target nucleic acid in the presence of an enzyme, such as a transposase or integrase, under conditions sufficient for the integration of the transposon sequence into the target nucleic acid.

In some embodiments, insertion of transposon sequences into a target nucleic acid can be non-random. In some embodiments, transposon sequences can be contacted with target nucleic acids comprising proteins that inhibit integration at certain sites. For example, transposon sequences can be inhibited from integrating into genomic DNA comprising proteins, genomic DNA comprising chromatin, genomic DNA comprising nucleosomes, or genomic DNA comprising histones. In some embodiments, transposon sequences can be associated with affinity tags in order to integrate the transposon sequence at a particular sequence in a target nucleic acid. For example, a transposon sequence may be associated with a protein that targets specific nucleic acid sequences, *e.g.,* histones, chromatin-binding proteins, transcription factors, initiation factors, etc., and antibodies or antibody fragments that bind to particular sequence-specific nucleic-acid-binding proteins. In an exemplary embodiment, a transposon sequence is associated with an affinity tag, such as biotin; the affinity tag can be associated with a nucleic-acid-binding protein.

IDT Exhibit 1020.4849

It will be understood that during integration of some transposon sequences into a target nucleic acid, several consecutive nucleotides of the target nucleic acid at the integration site are duplicated in the integrated product. Thus the integrated product can include a duplicated sequence at each end of the integrated sequence in the target nucleic acid. An example of such a duplication event is depicted in **FIG. 18**. As used herein, the term "host tag" can refer to a target nucleic acid sequence that is duplicated at each end of an integrated transposon sequence. Single-stranded portions of nucleic acids that may be generated by the insertion of transposon sequences can be repaired by a variety of methods well known in the art, for example by using ligases, oligonucleotides and/or polymerases.

In some embodiments, a plurality of the transposon sequences provided herein is inserted into a target nucleic acid. Some embodiments include selecting conditions sufficient to achieve integration of a plurality of transposon sequences into a target nucleic acid such that the average distance between each integrated transposon sequence comprises a certain number of consecutive nucleotides in the target nucleic acid.

In some embodiments, conditions for insertion of transposon sequences are sufficient to reduce the likelihood of forming concatameric complexes comprising a transposase associated with more than one transposon sequence. In one example, complexes comprising a transposase and a transposon sequence can be formed under dilute conditions; subsequent steps of inserting the transposon sequences into a target nucleic acid may be carried out at higher concentrations of transposase/transposon sequence complex. In another example, transposase/transposon sequence complexes can be prepared by contacting a circular transposon sequence with a transposase. As will be understood, the transposon sequence may be linearized during formation of the transposase/transposon sequence complex. In another example, a transposase/transposon sequence complex may be prepared by preparing complexes comprising partial transposon sequences comprising a transposon recognition site, and contacting the partial sequences with transposase monomers. Two partial transposon sequences may be ligated to one another to prepare a whole transposon sequence associated with a transposase comprising a dimer.

Some embodiments include selecting conditions sufficient to achieve insertion of a transposon sequence into a target nucleic acid, but not into another transposon sequence. A variety of methods can be used to reduce the likelihood that a transposon

IDT Exhibit 1020.4850

sequence inserts into another transposon sequence. For example, transposon sequences can comprise thiophosphate-modified nucleic acids. In another example, a transposon sequence can comprise a DNA/RNA hybrid, such as an RNA transposon sequence comprising DNA transposase recognition sites. In another example, a transposon sequence comprises a single-stranded sequence, further comprising double-stranded transposase recognition sites. It will be appreciated that more methods can include transposon sequences comprising single-stranded nucleic acids to inhibit insertion into the transposon sequence; transposon sequences comprising RNA to inhibit insertion into the transposon sequence; and transposon sequences associated with nucleic acid binding proteins to inhibit insertion into the transposon sequence.

In some embodiments, conditions may be selected so that the average distance in a target nucleic acid between integrated transposon sequences is at least about 5, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 100, 200, 300, 400, 500, 600, 700, 800, 900, 1000, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 1 kb, 2 kb, 3 kb, 4 kb, 5 kb, 6 kb, 7 kb, 8 kb, 90 kb, 100 kb, or more consecutive nucleotides. In some embodiments, the average distance in a target nucleic acid between integrated transposon sequences is at least about 100 kb, 200 kb, 300 kb, 400 kb, 500 kb, 600 kb, 700 kb, 800 kb, 900 kb, 1000 kb, or more consecutive nucleotides. As will be understood, some conditions that may be selected include contacting a target nucleic acid with a certain number of transposon sequences.

Some embodiments include selecting conditions sufficient to achieve at least a portion of transposon sequences integrated into a target nucleic acid are different. In preferred embodiments, each transposon sequence integrated into a target nucleic acid is different. Some conditions that may be selected to achieve a certain portion of transposon sequences integrated into a target sequences that are different include selecting the degree of diversity of the population of transposon sequences. As will be understood, the diversity of transposon sequences arises in part due to the diversity of the barcodes of such transposon sequences. Accordingly, some embodiments include providing a population of transposon sequences in which at least a portion of the barcodes are different. In some embodiments, at least about 10%, 20%, 30%, 40%, 50%, 60%, 70%,

IDT Exhibit 1020.4851

80%, 90%, 95%, 98%, 99%, or 100% of barcodes in a population of transposon sequences are different.

Some embodiments of preparing a template nucleic acid can include copying the sequences comprising the target nucleic acid. For example, some embodiments include hybridizing a primer to a primer site of a transposon sequence integrated into the target nucleic acid. In some such embodiments, the primer can be hybridized to the primer site and extended. The copied sequences can include at least one barcode sequence and at least a portion of the target nucleic acid. In some embodiments, the copied sequences can include a first barcode sequence, a second barcode sequence, and at least a portion of a target nucleic acid disposed therebetween. In some embodiments, at least one copied nucleic acid can include at least a first barcode sequence of a first copied nucleic acid that can be identified or designated to be paired with a second barcode sequence of a second copied nucleic acid. In some embodiments, the primer can include a sequencing primer. In some embodiments sequencing data is obtained using the sequencing primer.

Some embodiments of preparing a template nucleic acid can include amplifying sequences comprising at least a portion of one or more transposon sequences and at least a portion of a target nucleic acid. In some embodiments, at least a portion of a target nucleic acid can be amplified using primers that hybridize to primer sites of integrated transposon sequences integrated into a target nucleic acid. In some such embodiments, an amplified nucleic acid can include a first barcode sequence, and second barcode sequence having at least a portion of the target nucleic acid disposed therebetween. In some embodiments, at least one amplified nucleic acid can include at least a first barcode sequence of a first amplified nucleic acid that can be identified to be paired with a second barcode sequence of a second amplified sequence.

— symmetrical transposomes, adding sequences by amplification

In an embodiment, transposomes are used that have symmetrical transposable end sequences, as exemplified in **FIG. 19**. Each tagmented fragment therefore contains identical ends, lacking directionality. A single primer PCR, using the transposon end sequences, can then be employed to amplify the template copy number from $2n$ to $2n*2^x$ where x corresponds to the number of PCR cycles. In a subsequent step, PCR with primers can add additional sequences, such as sequencing adapter sequences.

IDT Exhibit 1020.4852

Some embodiments of preparing a template nucleic acid can include fragmenting a target nucleic acid comprising transposon sequences. Methods of fragmenting nucleic acids are well known in the art. In some embodiments, a nucleic acid comprising transposon sequences can be fragmented at random positions along the length of the nucleic acid. In some embodiments, a target nucleic acid comprising transposon sequences can be fragmented at the fragmentation sites of the transposon sequences. In some embodiments, insertion of a transposon sequence can include the duplication of the insertion site so that the inserted transposon sequence is disposed between duplicated single-stranded sequences (*See, e.g.,* **FIG. 18**). In some embodiments, a polymerase may be used to cleave the fragmentation site. Examples of such polymerases include strand-displacing nucleic acid polymerases.

Further embodiments of preparing a template nucleic acid that include fragmenting a target nucleic acid comprising transposon sequences can also include amplifying the fragmented nucleic acids. In some embodiments, the fragmented nucleic acids can be amplified using primers that hybridize to primer sites of transposon sequences. In more embodiments, primer sites can be ligated to the ends of the fragmented nucleic acids. In some such embodiments, the fragmented nucleic acids with ligated primer sites can be amplified from such primer sites.

Some embodiments include reducing the complexity of a library of template nucleic acids. A complexity-reduction step can be performed before or after the fragmentation step in the method. For example, the target nucleic acid comprising the transposon sequences can be diluted so that a small number or a single molecule represents the target diluted before performing subsequent steps.

FIG. 20 depicts an exemplary embodiment of a contacting a target nucleic acid with a library of transposons, which integrates into the target nucleic acid. **FIG. 21** depicts an optional step of fragmenting a linker into two parts, or an optional step of virtually fragmenting the linker by amplification using a primer A′ and/or a primer B′. Fragmentation of the linker generates a library of template nucleic acids. In this figure, the filled-in region is indicated by a thick line.

In some embodiments, it can be advantageous for each template nucleic acid to incorporate at least one universal primer site. For example, a template nucleic acid can include first end sequences that comprise a first universal primer site, and second end sequences that comprise a second universal primer site. Universal primer sites can have

IDT Exhibit 1020.4853

various applications, such as amplifying, sequencing, and/or identifying one or more template nucleic acids. The first and second universal primer sites can be the same, substantially similar, similar, or different. In some embodiments, in order to prepare a template nucleic acid comprising a first universal primer site and a second universal primer site, a transposon sequence is prepared that includes a first transposase recognition site, a second transposase recognition site, a barcode disposed therebetween, wherein the barcode comprises a first barcode sequence and a second barcode sequence, separated by a linker. The linker includes a first primer site and a second primer site with a fragmentation site therebetween. The first primer site can comprise sequences that are the reverse complement of sequences within the second primer. In some embodiments, the first primer site comprises sequences having dyad symmetry to sequences within the second primer. In some embodiments, the first primer site comprises sequences having C_2 symmetry to sequences within the second primer. A plurality of transposons may be inserted into a target nucleic acid by transposition in the presence of a transposase. The incorporated sequences may be cleaved to yield a plurality of target nucleic acids, each comprising the first primer site and second primer site. In some embodiments, a first universal primer site and a second universal primer site can be incorporated into each template nucleic acid by a variety of methods. For example, template nucleic acids can be amplified using the first primer site and second primer site using tailed-oligonucleotides. As is understood in the art, a tailed-oligonucleotide can include sequences complementary to a primer site and additional sequences. In an example embodiment, a first tailed-oligonucleotide comprises sequences complementary to a first primer site and sequences for a first universal primer site, and a second tailed-oligonucleotide comprises sequences complementary to a second primer site and sequences for a second universal primer site. **FIG. 22** depicts an example embodiment series of rounds of extending template using tailed-oligonucleotides to obtain a population of template nucleic acids comprising template nucleic acids that include a first universal primer site and a second universal primer site. With respect to, in step 1, oligonucleotides comprising either a first universal primer sequence (A) or a second universal primer sequence (B) anneal to a primer site (P) of a template nucleic acid, and are extended. In step 2, the extension products provide templates for a further extension step. Alternatively, the extension products can be amplified with the use of additional oligonucleotides. The products of step 2 are depicted in step 3 and include nucleic acid

IDT Exhibit 1020.4854

sequences comprising first universal primer sites only, second universal primer sites only, or first and second universal primer sites. As will be understood, nucleic acid sequences comprising a first universal primer site and second universal primer site may be used in further sequencing methods.

It will be understood that in some embodiments, the vast number of available barcodes permits each template nucleic acid molecule to comprise a unique identification. Unique identification of each molecule in a mixture of template nucleic acids can be used in several applications to identify individual nucleic acid molecules, in samples having multiple chromosomes, genomes, cells, cell types, cell disease states, and species, for example in haplotype sequencing, parental allele discrimination, metagenomic sequencing, and sample sequencing of a genome.

Tagging molecules individually

The present invention also provides methods for tagging molecules uniquely so individual molecules can be tracked and identified, so that bulk data can be deconvoluted back to the individual molecule. The ability to distinguish individual molecules and relate the information back to the molecule of origin is especially important when processes from original molecule to final product change the (stoichiometric) representation of the original population. For example, amplification leads to duplication (*e.g.*, PCR duplicates or biased amplification) that can skew the original representation. This can alter the methylation state call, copy number, allelic ratio due to non-uniform amplification and/or amplification bias. By identifying individual molecules, code-tagging distinguishes between identical molecules after processing. As such, duplications, and amplification bias can be filtered out, allowing accurate determination of the original representation of a molecule or population of molecules.

An advantage to this method is molecules that are otherwise identical in the original pool become uniquely identified by virtue of their tagging. In further downstream analysis these molecules can now be distinguished. This technique can be exploited in assay schemes in which amplification is employed. For example, amplification is known to distort the original representation of a mixed population of molecules. If unique tagging were not employed, the original representation (such as copy number or allelic ratio) would need to account for the biases (known or unknown) for each molecule in the representation. With unique tagging, the representation can

IDT Exhibit 1020.4855

accurately be determined by removing duplicates and counting the original representation of molecules, each having a unique tag. Thus, the cDNAs can be amplified and sequenced, without fear of bias, followed by filtering the data so that only authentic sequences or sequences of interest are selected for further analysis.

It is preferred to tag the original population in the early stages of the assay, although tagging can occur at later stages if the earlier steps do not introduce bias or are not important. In any of these applications, the complexity of the barcode sequences should be larger than the number of individual molecules to be tagged. This ensures that different target molecules receive different and unique tags. As such, a pool of random oligonucleotides of a certain length (5, 10, 20, 30, 40, 50, 100, 200) is desirable. Random pool of tags represents a large complexity of tags with code space 4^n where n is the number of nucleotides. Additional codes (whether designed or random) can be incorporated at different stages to serve as a further check, such as a parity check for error correction.

In one embodiment, individual molecules (such as target DNA) are attached to unique labels, such as unique oligo sequences, barcodes, as shown in **FIG. 23**. Attachment can occur through ligation, coupling chemistry, adsorption, etc. Other means include amplification (such as by PCR), copying (such as addition by a polymerase), and non-covalent interactions.

Specific methods include including barcodes (*e.g.,* designed or random sequences) to PCR primers so that the each template will receive an individual code within the code space, yielding unique amplicons that can be discriminated from other amplicons. This concept can be applied to any method that uses polymerase amplification, such as GoldenGate assays as disclosed in U.S. Patent No. 7,582,420, No. 7,955,794, and No. 8,003,354. Code-tagged target sequences can be circularized and amplified by methods such as rolling-circle amplification to yield code-tagged amplicons. Similarly, the code can also be added to RNA

Methods for making DNA circles

The present invention also provides methods and compositions for making DNA circles by fragmenting and circularizing DNA at the same time by insertion of "hairpin-like" transposons. Previous methods of circularizing DNA are based on ligating "sticky"

IDT Exhibit 1020.4856

or blunt end ends of DNA fragments together or by the use of ligases or related enzymes to ligate 5' and 3' DNA fragments from the same template molecule.

In this embodiment, special transposons are prepared in which DNA recognition sites for the transposase are double-stranded, with the overall transposon forming a hairpin. Two transposons are loaded into a transposon-transposase complex allowing the insertion of the hairpin transposon. Since the transposon ends are linked (as in **FIG. 24**, for example), insertion will fragment the DNA and simultaneously link the top and bottom strand of the target DNA. This process can occur throughout the genome, leading to the creation of many circles of the original starting DNA. A linker that links the top and bottom strand of the transposon together and can be any standard chemical linkers or a nucleic acid linker, such as single stranded DNA or RNA. Single-stranded DNA linkers are particularly useful since any DNA sequence can be designed as a linker, allowing pre-designed DNA to be incorporated into the DNA circles. The single-stranded linker can contain primer sites for amplification purposes.

Amplification of circularized DNA can be useful for sample preparation schemes for sequencing, genotyping, genotyping of FFPE samples. The method can be used for whole genome amplification with specific primers (not randomers) of genomic DNA. For example, primers can be designed into the single-stranded linker of the transposons allowing MDA or whole genome amplification of genomic DNA. One or more primers can be designed to accomplish exponential whole genome amplification. Alternatively, the method can be used to amplify a subset of genomic DNA. The circular genomic DNA can be amplified to prepare DNA balls. Whole genome amplification can be used for sample preparation for whole genome amplification, such as for the Infinium® assay workflow (Illumina, Inc.). In a particular embodiment, the method can be used to amplify FFPE DNA for whole genome genotyping. Short FFPE (Formalin-Fixed Paraffin-Embedded) DNA does not amplify efficiently in whole genome amplification schemes (Klenow/random primers or MD-like approaches. This method can generate circular DNA and these circular templates are efficient substrates for amplification. The amplification can be combined with lesion-proof polymerases to start the amplification.

A further advantage is that pre-designed primers can be used for amplification purposes. As such, whole genome amplification can be performed with one or few primers instead of the standard randomers. This method can lead to higher uniformity of amplification of a DNA sample with less locus dropout.

IDT Exhibit 1020.4857

Targeted insertion

The method is particularly useful when tagmentation is target-specific, so that a desired subset of the genome is made into a sequencing library.

In one implementation, dsDNA is denatured into ssDNA and annealed with oligonucleotide probes (20-200 bases). These probes create sites of dsDNA that can be efficiently tagmented with the transposase-transposon complex. This approach takes advantage of the fact that transposition into dsDNA is much more efficient compared to ssDNA targets. Furthermore, the transposase complex is covalently linked to the targeting oligonucleotides so that annealing brings the transposon complex into the vicinity of the targeted region of interest. At least two targeting oligos can be employed to generate library elements (tagments) that span the region between the targeting oligos (see **FIG. 25**).

The repeat regions may quickly re-hybridize, creating tagmentation sites that are non-specific and not desirable. To prevent this from occurring, a genomic representation using non-complementary bases can be prepared. For example, the DNA circles discussed above can be prepared that are uniformly amplified with specific primers, allowing the incorporation of non-complementary or pseudo-complementary bases. The end result is a genomic DNA representation that does not interact with itself or other amplified genomic DNA molecules. Specific annealing of standard oligonucleotides to these pseudo-complementary, single-stranded, genomic DNA can create specific tagmentation sites.

In another implementation (see **FIG. 26**), dsDNA is targeted directly using oligonucleotides (or LNA) probes coupled to the transposase complex. dsDNA can be targeted using a number of approaches including D-loop formation with recA-coated oligo probes, triplex formation, and other methods described in the literature. As above, two probes should flank the region of interest to generate a final library element spanning the region between the probes. Depending on the length of the linker between the targeting oligo and the transposase complex, the complex will insert a variable distance upstream or downstream of the targeted complex.

In yet another implementation regions of interest in dsDNA can be targeted using sequence-specific DNA binding proteins such as zinc-finger complexes, and other affinity ligands to specific DNA regions. Tagmentation on either end of a targeted region

IDT Exhibit 1020.4858

creates DNA fragments that have primers on the 5' and 3' end. These primers can be used for the amplification of a specific region.

Methods for generating mate-pair libraries

The present invention also provides an improved method for preparing mate-pair libraries. Mate-pair libraries offer powerful means of mapping inserts and deletions in genomes as well as genome finishing and contig assembly. Current mate-pair library preparation methods can involve sequentially: a first fragmentation, size selection, end-repair, adapter ligation, circularization, a secondary fragmentation, junction fragment enrichment, and PCR. Though random fragmentation is desirable in general, the Poisson distribution of fragment sizes in the second fragmentation can yield numerous mate-pair arms that are too small to be of bioinformatic use. Such mate pairs must be discarded and this reduces the size of the mate-pair library.

The invention can involve tethering the circularized intermediate to the surface of a particle to use the steric hindrance of the particle to reduce access of transposomes to the DNA, reducing transposase cleavage immediately adjacent to the site of immobilization.

In the embodiment shown in **FIG. 27**, the process begins by cleaving genomic DNA with transposomes and selecting fragments of the desired size by gel purification or other fragment selection methodology. The 3'-ends of the cleavage products are then melted away and replaced with an excess quantity of a new sequence that has been amended to contain a 3'-palindrome (for subsequent circularization) and a binding residue. An exemplary binding residue is a biotin residue for enrichment on streptavidin-paramagnetic particles (SA-PMPs). In this embodiment, the newly modified fragments retain the 9-bp gaps left by the Tn5 transposase (or other enzyme that is used) and may be repaired in a gap-filling reaction driven by a non-strand displacing DNA polymerase that also lacks 3'-5' exonuclease activity, and a DNA ligase. The DNA ligase used in the gap-filling reaction can also serve to generate DNA circles. Following exonuclease-mediated removal of residual linear molecules, the circular DNAs can be captured onto the surface of a binding surface (*e.g.*, SA-PMPs) and subjected to a second tagmentation reaction. This tagmentation reaction introduces priming sites compatible with the next-generation sequencing platform of choice. In further embodiments, sequencing is performed.

IDT Exhibit 1020.4859

Since the protective effect diminishes as the distance from a tether increases, the steric hindrance reduces the number of less-desirable small mate-pair arms results in mate-pair libraries of greater complexity due to a process that precludes losses from occurring during the second cleavage reaction.

Methods of analyzing template nucleic acids

Some embodiments include methods of analyzing template nucleic acids. Sequencing information can be obtained from a template nucleic acids and a sequence representation of the target nucleic acid can be obtained from such sequencing data.

In some embodiments, a linked read strategy may be used. A linked read strategy can include identifying sequencing data that links at least two sequencing reads. For example, a first sequencing read may contain a first marker, and a second sequencing read may contain a second marker. The first and second markers can identify the sequencing data from each sequencing read to be adjacent in a sequence representation of the target nucleic acid. In some embodiments, markers can comprise a first barcode sequence and a second barcode sequence in which the first barcode sequence can be paired with the second barcode sequence. In more embodiments, markers can comprise a first host tag and a second host tag. In more embodiments, markers can comprise a first barcode sequence with a first host tag, and a second barcode sequence with a second host tag.

An exemplary embodiment of a method for sequencing a template nucleic acid can comprise the following steps. First, sequence the first barcode sequence using a primer hybridizing to the first primer site as the sequencing primer; second, sequence the second barcode sequence using a primer hybridizing to the second primer site as the sequencing primer. The result is two sequence reads that help link the read to its genomic neighbors. Given long enough reads, and short enough library fragments, these two reads can be merged informatically to make one long read that covers the entire fragment. Using the barcode sequence reads and the 9 nucleotide duplicated sequence present from the insertion, reads can now be linked to their genomic neighbors to form much longer "linked reads" *in silico*. As will be understood, a library comprising template nucleic acids can include duplicate nucleic acid fragments. Sequencing duplicate nucleic acid fragments is advantageous in methods that include creating a consensus sequence for

IDT Exhibit 1020.4860

duplicate fragments. Such methods can increase the accuracy for providing a consensus sequence for a template nucleic acid and/or library of template nucleic acids.

In some embodiments, sequence analysis is performed in real time, for example, sequence data can be obtained and simultaneously analyzed. In some embodiments, a sequencing process to obtain sequencing data can be terminated at various points, including after at least a portion of a target nucleic acid sequence data is obtained or before the entire nucleic acid read is sequenced. Exemplary methods, systems, and further embodiments are provided in International Patent Publication WO 2010/062913.

FIG. 28 depicts an exemplary embodiment of a method for assembling short sequencing reads using a linked read strategy. In such an exemplary embodiment, transposon sequences comprising barcodes are inserted into genomic DNA, a library is prepared and sequencing data obtained for the library of template nucleic acids. Blocks of templates are assembled by identifying paired barcodes and larger contigs are then assembled. **FIG. 29**, **FIG. 30** and **FIG. 31** depict exemplary embodiments of methods of assembling sequencing reads using a linked ready strategy.

Some embodiments include error detection and correction. Examples of errors can include errors in base calls during a sequencing process, and errors in assembling fragments into larger contigs. As would be understood, error detection can include detecting the presence or likelihood of errors in a data set, and as such detecting the location of an error or number of errors may not be required. For error correction, information regarding the location of an error and/or the number of errors in a data set is useful. Methods for error correction are well known in the art. Examples include the use of hamming distances, and the use of a checksum algorithm (*See, e.g.*,U.S. Publication 2010/0323348; U.S. Patent No. 7,574,305; U.S. Patent No. 6,654,696).

— nested libraries

An alternative method involves the junction tagging methods above and preparation of nested sequencing libraries, as illustrated in **FIG. 32**. The nested sub-libraries are created from code-tagged DNA fragments. This can allow less frequent transposon tagging across the genome. It can also create a larger diversity of (nested) sequencing reads. These factors can lead to improved coverage and accuracy.

Sub-sampling and whole genome amplification can create many copies of a certain population of starting molecules. DNA fragments are then generated by

IDT Exhibit 1020.4861

transposon-specific fragmentation, where each fragment receives a code that allows one to link the fragment back to the original neighbor having a matching code (whether identical, complementary or otherwise informatically linked). The tagged fragments are fragmented at least a second time by methods random or sequence-specific methods such as enzymatic digestion, random shearing, transposon-based shearing or other methods, thereby creating sub-libraries of the code-tagged DNA fragments. In a useful variation, code-tagged fragments can be preferentially isolated by using transposons that contain a biotin or other affinity functionality for downstream enrichment purposes. Subsequently, library preparation converts the nested DNA fragments into sequencing templates. Paired-end sequencing results in sequencing the of the code-tag of the DNA fragments and of the target DNA. Since nested libraries for the same code-tag are created, long DNA fragments can be sequenced with short reads.

— two-step sub-assembly

In another embodiment, a two-step tagging method is provided using one or more transposition reactions with different transposomes. The transposomes can be of different types (e.g. Mu versus Tn5) or be two sets of Tn5 transposomes with different sequences, for example.

As exemplified in **FIG. 33**, a first transposome can use a Mu-type transposase, such as HyperMu™ (Epicentre). The DNA section of this complex is barcoded at the individual molecule level and presence of a molecular barcode at both ends of each fragment ("end-tags"). Each DNA fragment receives a unique tag (code-tagging) for the purpose of making each DNA fragment unique. The library is then amplified via PCR, so that every fragment is represented multiple times in the amplified reaction. The amplified library is next subjected to a second round of tagmentation with a different transposome, such as Tn5. If desired, the concentration of enzyme can be limited so every molecule is minimally targeted with the second transposome to preserve maximum continuity information. In addition, complexity reduction schemes at various stages throughout the assay can be utilized to ensure that the majority of libraries sequenced cover a significant fraction of the target DNA.

IDT Exhibit 1020.4862

— compartmentalization of single molecules

Some embodiments include the use of the compositions and methods provided herein in single cell applications. In some embodiments, a transposon sequence can be associated with a bead via a cleavable linker. A bead comprising the transposon sequence and transposase is provided in a droplet. The cell is lysed and the transposon sequence cleaved from the bead. The transposition reaction is initiated. Such methods can be used to provide a library of template nucleic acids for each cell in population. Methods particularly useful for single-cell applications include clonal transposon beads, transposase mix, and single cell/single DNA molecule suspensions.

In an embodiment illustrated in **FIG. 34**, droplets or microwells are used to compartmentalize a single transposon bead with a single cell/DNA molecule in the presence of tagmentation reaction mix. The transposon is immobilized to the bead through a cleavable linker, such as a photocleavable linker. Each bead has a unique transposon index sequence for the appropriate barcoding of a particular cell. Additionally, photo-caged Mg^{++} can be used to activate the tagmentation reaction at the same time the oligos are cleaved from the bead. Upon exposure to an appropriate wavelength of light (*e.g.,* 365 nm for many photoreactive compounds), the transposon is cleaved from the bead and assembles with the transposase to create an active complex. In the presence of Mg^{++}, this complex will insert into the available target DNA and fragment the sample. The DNA sample now contains a unique index along with adapter sequences from the transposon. All the compartmentalized reactions are simultaneously lysed/extracted into a single eluant. This eluant can be amplified with adapter-primers to create final sequencing library. For methylation analysis, the eluant can be treated with bisulfite first, and then amplified to enable bisulfite sequence analysis.

In an embodiment exemplified in **FIG. 35,** beads with uniquely indexed transposons (attached or hybridized via identifier sequences) are loaded into a microwell to create a transposon bead array. Accordingly, the present invention provides a transposon bead array. The transposon beads can be created in a number of ways. One way is clonal amplification of single-transposon seed molecules using bridge amplification of emulsion PCR. Another way is split-and-pool oligonucleotide synthesis in which the index portion of the transposon on the bead is generated during a split-and-pool oligonucleotide bead synthesis. A further way is to hybridize indexed transposons with an address sequence to beads containing the cognate address sequence.

IDT Exhibit 1020.4863

Alternatively, the transposon bead array can be preloaded with beads having index sequences and then hybridizing indexed transposons. The transposon bead array is loaded with tagmentation mix and cells/DNA so that each well has less than one cell on average. The transposon bead array can be overlaid with oil and heated to lyse the cells. The transposons are released from the bead via photocleavage. Alternatively, photo-caged Mg^{++} can be released to activate the tagmentation reaction. After tagmentation, the wells are extracted and the eluate is pooled. The eluate can be amplified (such as by PCR) to yield a cell-indexed sequencing library.

Sequencing methods

The methods described herein can be used in conjunction with a variety of sequencing techniques. In some embodiments, the process to determine the nucleotide sequence of a target nucleic acid can be an automated process.

Some embodiments include pyrosequencing techniques. Pyrosequencing detects the release of inorganic pyrophosphate (PP_i) as particular nucleotides are incorporated into the nascent strand (Ronaghi *et al.*, "Real-time DNA sequencing using detection of pyrophosphate release." *Analytical Biochemistry* 242(1): 84-9 (1996); Ronaghi, M. "Pyrosequencing sheds light on DNA sequencing" *Genome Res.* 11(1):3-11 (2001); Ronaghi *et al.*, "A sequencing method based on real-time pyrophosphate." *Science* 281(5375):363 (1998); U.S. Patent No. 6,210,891; U.S. Patent No. 6,258,568 and U.S. Patent No. 6,274,320). In pyrosequencing, released PP_i can be detected by being immediately converted to adenosine triphosphate (ATP) by ATP sulfurylase, and the level of ATP generated is detected via luciferase-produced photons.

In another example type of SBS, cycle sequencing is accomplished by stepwise addition of reversible terminator nucleotides containing, for example, a cleavable or photobleachable dye label as described, for example, in U.S. Patent No. 7,427,67, U.S. Patent No. 7,414,1163 and U.S. Patent No. 7,057,026. This approach is being commercialized by Solexa (now Illumina Inc.), and is also described in WO 91/06678 and WO 07/123744 (filed in the United States patent and trademark Office as US 12/295,337). The availability of fluorescently-labeled terminators in which both the termination can be reversed and the fluorescent label cleaved facilitates efficient cyclic reversible termination (CRT) sequencing. Polymerases can also be co-engineered to efficiently incorporate and extend from these modified nucleotides.

IDT Exhibit 1020.4864

Additional example SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent Application Publication No. 2007/0166705, U.S. Patent Application Publication No. 2006/0188901, U.S. Patent No. 7057026, U.S. Patent Application Publication No. 2006/0240439, U.S. Patent Application Publication No. 2006/0281109, PCT Publication No. WO 05/065814, U.S. Patent Application Publication No. 2005/0100900, PCT Publication No. WO 06/064199 and PCT Publication No. WO 07/010251.

Some embodiments can utilize sequencing by ligation techniques. Such techniques utilize DNA ligase to incorporate nucleotides and identify the incorporation of such nucleotides. Example SBS systems and methods which can be utilized with the methods and systems described herein are described in U.S. Patent No 6,969,488, U.S. Patent No. 6,172,218, and U.S. Patent No. 6,306,597.

Some embodiments can include techniques such as next-next technologies. One example can include nanopore sequencing techniques (Deamer, D.W. & Akeson, M. "Nanopores and nucleic acids: prospects for ultrarapid sequencing." *Trends Biotechnol.* 18, 147–151 (2000); Deamer, D. and D. Branton, "Characterization of nucleic acids by nanopore analysis". *Acc. Chem. Res.* 35:817-825 (2002); Li *et al.*, "DNA molecules and configurations in a solid-state nanopore microscope" *Nat. Mater.* 2:611–615 (2003)). In such embodiments, the target nucleic acid passes through a nanopore. The nanopore can be a synthetic pore or biological membrane protein, such as α-hemolysin. As the target nucleic acid passes through the nanopore, each base-pair can be identified by measuring fluctuations in the electrical conductance of the pore. (U.S. Patent No. 7,001,792; Soni & Meller, "A. Progress toward ultrafast DNA sequencing using solid-state nanopores." *Clin. Chem.* 53, 1996–2001 (2007); Healy, K. "Nanopore-based single-molecule DNA analysis." *Nanomed.* 2:459–481 (2007); Cockroft *et al.*, "A single-molecule nanopore device detects DNA polymerase activity with single-nucleotide resolution." *J. Am. Chem. Soc.* 130:818–820 (2008)). In some such embodiments, nanopore sequencing techniques can be useful to confirm sequence information generated by the methods described herein.

Some embodiments can utilize methods involving the real-time monitoring of DNA polymerase activity. Nucleotide incorporations can be detected through fluorescence resonance energy transfer (FRET) interactions between a fluorophore-bearing polymerase and γ-phosphate-labeled nucleotides as described, for example, in U.S. Patent No. 7,329,492 and U.S. Patent No. 7,211,414 or nucleotide incorporations can

IDT Exhibit 1020.4865

be detected with zero-mode waveguides as described, for example, in U.S. Patent No. 7,315,019 and using fluorescent nucleotide analogs and engineered polymerases as described, for example, in U.S. Patent No. 7,405,281 and U.S. Patent Application Publication No. 2008/0108082 The illumination can be restricted to a zeptoliter-scale volume around a surface-tethered polymerase such that incorporation of fluorescently labeled nucleotides can be observed with low background (Levene, M.J. *et al.* "Zero-mode waveguides for single-molecule analysis at high concentrations." *Science* 299, 682–686 (2003); Lundquist, P.M. *et al.* "Parallel confocal detection of single molecules in real time." *Opt. Lett.* 33, 1026–1028 (2008); Korlach, J. *et al.* "Selective aluminum passivation for targeted immobilization of single DNA polymerase molecules in zero-mode waveguide nanostructures." *Proc. Natl. Acad. Sci. USA* 105, 1176–1181 (2008)). In one example single molecule, real-time (SMRT) DNA sequencing technology provided by Pacific Biosciences Inc can be utilized with the methods described herein. In some embodiments, a SMRT chip or the like may be utilized (U.S. Patent Nos. 7,181,122, 7,302,146, 7,313,308). A SMRT chip comprises a plurality of zero-mode waveguides (ZMW). Each ZMW comprises a cylindrical hole tens of nanometers in diameter perforating a thin metal film supported by a transparent substrate. When the ZMW is illuminated through the transparent substrate, attenuated light may penetrate the lower 20-30 nm of each ZMW creating a detection volume of about 1×10^{-21} L. Smaller detection volumes increase the sensitivity of detecting fluorescent signals by reducing the amount of background that can be observed.

SMRT chips and similar technology can be used in association with nucleotide monomers fluorescently labeled on the terminal phosphate of the nucleotide (Korlach J. *et al.,* "Long, processive enzymatic DNA synthesis using 100% dye-labeled terminal phosphate-linked nucleotides." Nucleosides, Nucleotides and Nucleic Acids, 27:1072–1083, 2008). The label is cleaved from the nucleotide monomer on incorporation of the nucleotide into the polynucleotide. Accordingly, the label is not incorporated into the polynucleotide, increasing the signal: background ratio. Moreover, the need for conditions to cleave a label from labeled nucleotide monomers is reduced.

An additional example of a sequencing platform that may be used in association with some of the embodiments described herein is provided by Helicos Biosciences Corp. In some embodiments, TRUE SINGLE MOLECULE SEQUENCING can be utilized (Harris T.D. *et al.,* "Single Molecule DNA Sequencing of a viral Genome" *Science*

IDT Exhibit 1020.4866

320:106-109 (2008)). In one embodiment, a library of target nucleic acids can be prepared by the addition of a 3′ poly(A) tail to each target nucleic acid. The poly(A) tail hybridizes to poly(T) oligonucleotides anchored on a glass cover slip. The poly(T) oligonucleotide can be used as a primer for the extension of a polynucleotide complementary to the target nucleic acid. In one embodiment, fluorescently-labeled nucleotide monomer, namely, A, C, G, or T, are delivered one at a time to the target nucleic acid in the presence DNA polymerase. Incorporation of a labeled nucleotide into the polynucleotide complementary to the target nucleic acid is detected, and the position of the fluorescent signal on the glass cover slip indicates the molecule that has been extended. The fluorescent label is removed before the next nucleotide is added to continue the sequencing cycle. Tracking nucleotide incorporation in each polynucleotide strand can provide sequence information for each individual target nucleic acid.

An additional example of a sequencing platform that can be used in association with the methods described herein is provided by Complete Genomics Inc. Libraries of target nucleic acids can be prepared where target nucleic acid sequences are interspersed approximately every 20 bp with adaptor sequences. The target nucleic acids can be amplified using rolling circle replication, and the amplified target nucleic acids can be used to prepare an array of target nucleic acids. Methods of sequencing such arrays include sequencing by ligation, in particular, sequencing by combinatorial probe-anchor ligation (cPAL).

In some embodiments using cPAL, about 10 contiguous bases adjacent to an adaptor may be determined. A pool of probes that includes four distinct labels for each base (A, C, T, G) is used to read the positions adjacent to each adaptor. A separate pool is used to read each position. A pool of probes and an anchor specific to a particular adaptor is delivered to the target nucleic acid in the presence of ligase. The anchor hybridizes to the adaptor, and a probe hybridizes to the target nucleic acid adjacent to the adaptor. The anchor and probe are ligated to one another. The hybridization is detected and the anchor-probe complex is removed. A different anchor and pool of probes is delivered to the target nucleic acid in the presence of ligase.

The sequencing methods described herein can be advantageously carried out in multiplex formats such that multiple different target nucleic acids are manipulated simultaneously. In particular embodiments, different target nucleic acids can be treated in a common reaction vessel or on a surface of a particular substrate. This allows convenient

IDT Exhibit 1020.4867

delivery of sequencing reagents, removal of unreacted reagents and detection of incorporation events in a multiplex manner. In embodiments using surface-bound target nucleic acids, the target nucleic acids can be in an array format. In an array format, the target nucleic acids can be typically coupled to a surface in a spatially distinguishable manner. For example, the target nucleic acids can be bound by direct covalent attachment, attachment to a bead or other particle or associated with a polymerase or other molecule that is attached to the surface. The array can include a single copy of a target nucleic acid at each site (also referred to as a feature) or multiple copies having the same sequence can be present at each site or feature. Multiple copies can be produced by amplification methods such as, bridge amplification or emulsion PCR as described in further detail herein.

The methods set forth herein can use arrays having features at any of a variety of densities including, for example, at least about 10 features/cm^2, 100 features/cm^2, 500 features/cm^2, 1,000 features/cm^2, 5,000 features/cm^2, 10,000 features/cm^2, 50,000 features/cm^2, 100,000 features/cm^2, 1,000,000 features/cm^2, 5,000,000 features/cm^2, 10^7 features/cm^2, 5x 10^7 features/cm^2, 10^8 features/cm^2, 5x 10^8 features/cm^2, 10^9 features/cm^2, 5x 10^9 features/cm^2, or higher.

Surfaces

In some embodiments, the nucleic acid template provided herein can be attached to a solid support ("substrate"). Substrates can be two-or three-dimensional and can comprise a planar surface (*e.g.*, a glass slide) or can be shaped. A substrate can include glass (*e.g.*, controlled pore glass (CPG)), quartz, plastic (such as polystyrene (low cross-linked and high cross-linked polystyrene), polycarbonate, polypropylene and poly(methylmethacrylate)), acrylic copolymer, polyamide, silicon, metal (*e.g.*, alkanethiolate-derivatized gold), cellulose, nylon, latex, dextran, gel matrix (*e.g.*, silica gel), polyacrolein, or composites.

Suitable three-dimensional substrates include, for example, spheres, microparticles, beads, membranes, slides, plates, micromachined chips, tubes (*e.g.*, capillary tubes), microwells, microfluidic devices, channels, filters, or any other structure suitable for anchoring a nucleic acid. Substrates can include planar arrays or matrices capable of having regions that include populations of template nucleic acids or primers.

IDT Exhibit 1020.4868

Examples include nucleoside-derivatized CPG and polystyrene slides; derivatized magnetic slides; polystyrene grafted with polyethylene glycol, and the like.

Various methods can be used to attach, anchor or immobilize nucleic acids to the surface of the substrate. The immobilization can be achieved through direct or indirect bonding to the surface. The bonding can be by covalent linkage. *See* Joos *et al.* (1997) Analytical Biochemistry, 247:96-101; Oroskar *et al.* (1996) Clin. Chem., 42:1547-1555; and Khandjian (1986) Mol. Bio. Rep., 11:107-11. A preferred attachment is direct amine bonding of a terminal nucleotide of the template or the primer to an epoxide integrated on the surface. The bonding also can be through non-covalent linkage. For example, biotin-streptavidin (Taylor *et al.* (1991) J. Phys. D: Appl. Phys., 24:1443,) and digoxigenin with anti-digoxigenin (Smith *et al., Science*, 253:1122 (1992)), are common tools for anchoring nucleic acids to surfaces and parallels. Alternatively, the attachment can be achieved by anchoring a hydrophobic chain into a lipid monolayer or bilayer. Other methods known in the art for attaching nucleic acid molecules to substrates can also be used.

The following Examples provide illustrative embodiments and do not in any way limit the inventions provided herein.

EXAMPLES

Example 1—Whole genome amplification using transposon sequences

This example illustrates a method for uniform amplification of genomic DNA with random insertion therein of specific primer sites. Transposon sequences are prepared, each comprising a first transposase recognition site, a second transposase recognition site having a sequencing adaptor disposed therebetween, in which the sequencing adaptor comprises a first primer site and second primer site. The transposon sequences are contacted with genomic DNA in the presence of MuA transposase under conditions sufficient for the transposon sequences to integrate into the genomic DNA. The genomic DNA is amplified using primers that hybridize to the first primer site or second primer site.

IDT Exhibit 1020.4869

Example 2—Landmark sequencing methods using genomes with increased complexity

This example illustrates an embodiment for providing additional markers in a genome. Additional markers can be useful in genomes that include repetitive sequences during subsequent assembly steps to generate a sequence representation of the genome. Transposon sequences are prepared, each comprising a different barcode. The transposon sequences are integrated into genomic DNA in a transposition reaction. The genomic DNA comprising the integrated transposon is amplified by whole genome amplification. A sequencing library is prepared from the amplified template nucleic acids. Sequencing data is obtained from the sequencing library. Sequencing reads can include representations of one or more nucleic acids with the same barcode on each nucleic acid. Such nucleic acids are identified as containing sequences that overlap in a sequence representation of the genomic DNA. The sequencing reads can be assembled by identifying barcodes on overlapping sequences.

Example 3—Predicted average coverage using linked read sequencing strategy

Useable fragment lengths are modeled as a truncated exponential distribution so that the mean useable fragment length can be obtained by setting k = b/d, where d is the mean of the non-truncated exponential (the total fragment distribution) and b is the value for truncation (either 180 or 280 for 100 nucleotides and 150 nucleotide paired-end reads, respectively) and then calculating the mean of the truncated exponential as

$$E(f) = d\,(1-(k+1)e^{-k}) / (1 - e^{-k})$$

The proportion of useable reads is p = C(b) x (1 - D(0,T)) where C is the exponential cumulative distribution function, T is the average repetitions of observing each fragment (num clusters) / complexity, complexity is the genome size times the number of genome copies diluted to divided by d, and D is the Poisson cumulative distribution function

Expected length of linked read is then (E(f) – 9) x 1 / (1 – p) + 9 where p is proportion of useable reads: 9 is subtracted from each read because of the reused 9 nucleotides sequence in neighboring fragments; and 9 nucleotides is added back to the linked read to account for the end read –one of its 9 nucleotides segments is not shared with another read within the linked read.

IDT Exhibit 1020.4870

The distribution of linked read lengths is also exponential with the above expected value. Very long linked reads can be observed, and shown in **FIG. 36**.

Table 1 sets out predicted average coverage for a bacterial genome, human PCR product, and a single copy of a human diploid genome using a linked read strategy.

TABLE 1

Target nucleic acid	Genome size	Read length	Diluted copies	Number of clusters	Average insert distance	Average coverage	Mean linked read length
Bacterial genome	5.0e6	150	10	35,000,000	50 nt	9.96 X	10,815 nt
Human PCR product	20 kb	150	10,000	40,000,000	50 nt	9.96 X	10,811 nt
Single cell, human diploid genome	3e9	150	2	3000,000,000	50 nt	9.66 X	1,191 nt

Example 4—*De novo* sequencing target DNA

This example illustrates an embodiment of assembling sequencing data obtained from a library of template nucleic acids prepared from a target DNA.

A plurality of transposon sequences are integrated into the target DNA in a transposition reaction. Each transposon sequence includes a barcode which comprises a first barcode sequence and a second barcode sequence. The first barcode sequence is the reverse complement of the second barcode sequence. There are more than 10^{18} different barcodes in the plurality of transposon sequences, such that the code space is large and out-competes the complexity of the target DNA. Accordingly, each integrated barcode is likely to be unique. Optionally, the target DNA comprising the integrated transposon sequences is sub-sampled. The sub-sampled target DNA comprising the integrated transposon sequences is amplified by methods of whole genome amplification. A sequencing library is prepared from the amplified nucleic acids. Sequencing data is obtained from the sequencing library. The sequencing data comprises sequencing reads for each amplified nucleic acid.

IDT Exhibit 1020.4871

Sequencing reads that include a barcode sequence that can be paired with a barcode sequence of another sequencing read are aligned. Shorter alignments of sequencing reads are aligned with other short alignments by identifying paired barcode sequences to produce longer alignments. A sequence representation of the target DNA is generated.

Example 5—Preparing a linked library with blunt-end barcodes

This example illustrates an embodiment of preparing a linked library that includes identical barcodes on template nucleic acids that include adjacent sequences in a sequence representation of the target nucleic acid.

Transposon sequences are prepared comprising a first fragmentation site and a second fragmentation site, having a barcode disposed therebetween. Each fragmentation site comprises a site that can be nicked to produce a single-stranded sticky end, *e.g.*, a restriction endonuclease site that produces a single-strand sticky end. The transposon sequences are integrated into the target DNA by a transposition reaction. Optionally, the target DNA comprising the integrated transposon sequences is sub-sampled. The sub-sampled target DNA comprising the integrated transposon sequences is amplified by methods of whole genome amplification. The amplified nucleic acids are fragmented at the first and second fragmentation sites at each integrated transposon sequence to generate nucleic acids comprising sticky ends. The sticky ends are filled-in so that each end comprises a barcode. Adaptors are ligated to the blunt ends of each nucleic acid. The nucleic acids are amplified using primer sites of the adaptors.

A sequencing library is prepared from the amplified nucleic acids. Sequencing data is obtained from the sequencing library. The sequencing data comprises sequencing reads for each amplified nucleic acid. Sequencing reads that include a barcode sequence that can be paired with a barcode sequence of another sequencing read are aligned. Shorter alignments of sequencing reads are aligned with other short alignments by identifying paired barcode sequences to produce longer alignments. A sequence representation of the target DNA is generated.

FIG. 37 depicts the insertion of the transposon sequence into the target nucleic acid (host material), dilution of the template nucleic acids, and subsequent whole genome amplification of the template nucleic acids. The amplified template nucleic acids are fragmented in the presence of a restriction endonuclease that nicks the transposon

IDT Exhibit 1020.4872

sequences at the first restriction endonuclease site and the second restriction endonuclease site to yield two sticky ends.

FIG. 15 depicts a fill-in reaction of the sticky ends, followed by A-tailing the ends and adding adaptors to the tailed ends. The adaptors can be used in subsequent amplification, library preparation, and methods to obtain sequence data. In other embodiments, the fill-in reaction itself can be used as part of a detection assay, such as by incorporating detectably labeled nucleotides in a variety of assay formats. Advantageously, methods that utilize a nicking endonuclease may be carried out with a relatively small number of target nucleic acids, or a relatively dilute concentration of the target nucleic acid.

The above description discloses several methods and systems of the present invention. This invention is susceptible to modifications in the methods and materials, as well as alterations in the fabrication methods and equipment. Such modifications will become apparent to those skilled in the art from a consideration of this disclosure or practice of the invention disclosed herein. For example, the invention has been exemplified using nucleic acids but can be applied to other polymers as well. Consequently, it is not intended that this invention be limited to the specific embodiments disclosed herein, but that it cover all modifications and alternatives coming within the true scope and spirit of the invention.

All references cited herein including, but not limited to, published and unpublished applications, patents, and literature references, are incorporated herein by reference in their entirety and are hereby made a part of this specification. To the extent publications and patents or patent applications incorporated by reference contradict the disclosure contained in the specification, the specification is intended to supersede and/or take precedence over any such contradictory material.

The term "comprising" as used herein is synonymous with "including," "containing," or "characterized by," and is inclusive or open-ended and does not exclude additional, unrecited elements or method steps.

IDT Exhibit 1020.4873

WHAT IS CLAIMED IS:

1. An artificial transposon, comprising:

a first transposase recognition site,

a second transposase recognition site, and

a barcode disposed therebetween, wherein the barcode comprises a double-stranded nucleic acid sequence comprising

a first strand barcode and

a second strand barcode.

2. The transposon of claim 1, wherein the transposase is a hyperactive Tn5 transposase.

3. The transposon of claim 1, wherein the transposase is a Mu transposase.

4. The transposon of claim 1, wherein the transposase is an IS5 or an IS911 transposase.

5. The transposon of claim 1, wherein the transposon is associated with an affinity tag.

6. The transposon of claim 5, wherein the affinity tag is associated with a protein that targets specific nucleic acid sequences.

7. The transposon of claim 1, wherein the first strand barcode and the second strand barcode comprise complementary sequences.

8. The transposon of claim 1, wherein the first strand barcode and the second strand barcode comprise non-complementary sequences.

9. The transposon of claim 1, further comprising at least one universal primer site.

IDT Exhibit 1020.4874

10. An isolated template nucleic acid, comprising at least a portion of a target nucleic acid or copy thereof and at least first and second transposons of claim 1, wherein the barcode of the first transposon is different from the barcode of the second transposon.

11. A method for preparing a library of template nucleic acids comprising:
(a) contacting a target nucleic acid with a plurality of transposon of claim 1 under conditions such that a portion of said of transposons are inserted into the target nucleic acid; and
(b) fragmenting said target nucleic acid.

12. The method of claim 11, wherein the target nucleic acid comprises cDNAs from a single cell.

13. The method of claim 11, wherein the target nucleic acid comprises nucleic acids from multiple species.

14. The method of claim 11, wherein the target nucleic acid comprises nucleic acids from multiple haplotypes.

15. The method of claim 11, wherein the fragmentation comprises contacting the target nucleic acid with a polymerase.

16. The method of claim 15, wherein the polymerase has 3' to 5' exonuclease activity.

17. A method for preparing a library of template nucleic acids, comprising:
(a) contacting a target nucleic acid with a plurality of transposons of claim 1, further comprising a first primer site, under conditions such that a portion of said transposons are inserted into the target nucleic acid; and
(b) hybridizing a primer to the first primer site; and
(c) amplifying a portion of the transposon sequence.

18. A method of analyzing a target nucleic acid comprising:

IDT Exhibit 1020.4875

(a) performing the method of claim 17;

(b) obtaining sequence data from said templates; and

(c) assembling a representation of a portion of said target nucleic acid from said sequence data.

19. The method of claim 18, wherein assembling step (c) comprises identifying the presence of a first barcode in one sequencing read and the presence of a corresponding barcode in another sequencing read, thereby indicating proximity between the two sequencing reads in the target nucleic acid.

20. The method of claim 13, wherein step (b) is terminated after the barcode is identified.

IDT Exhibit 1020.4876

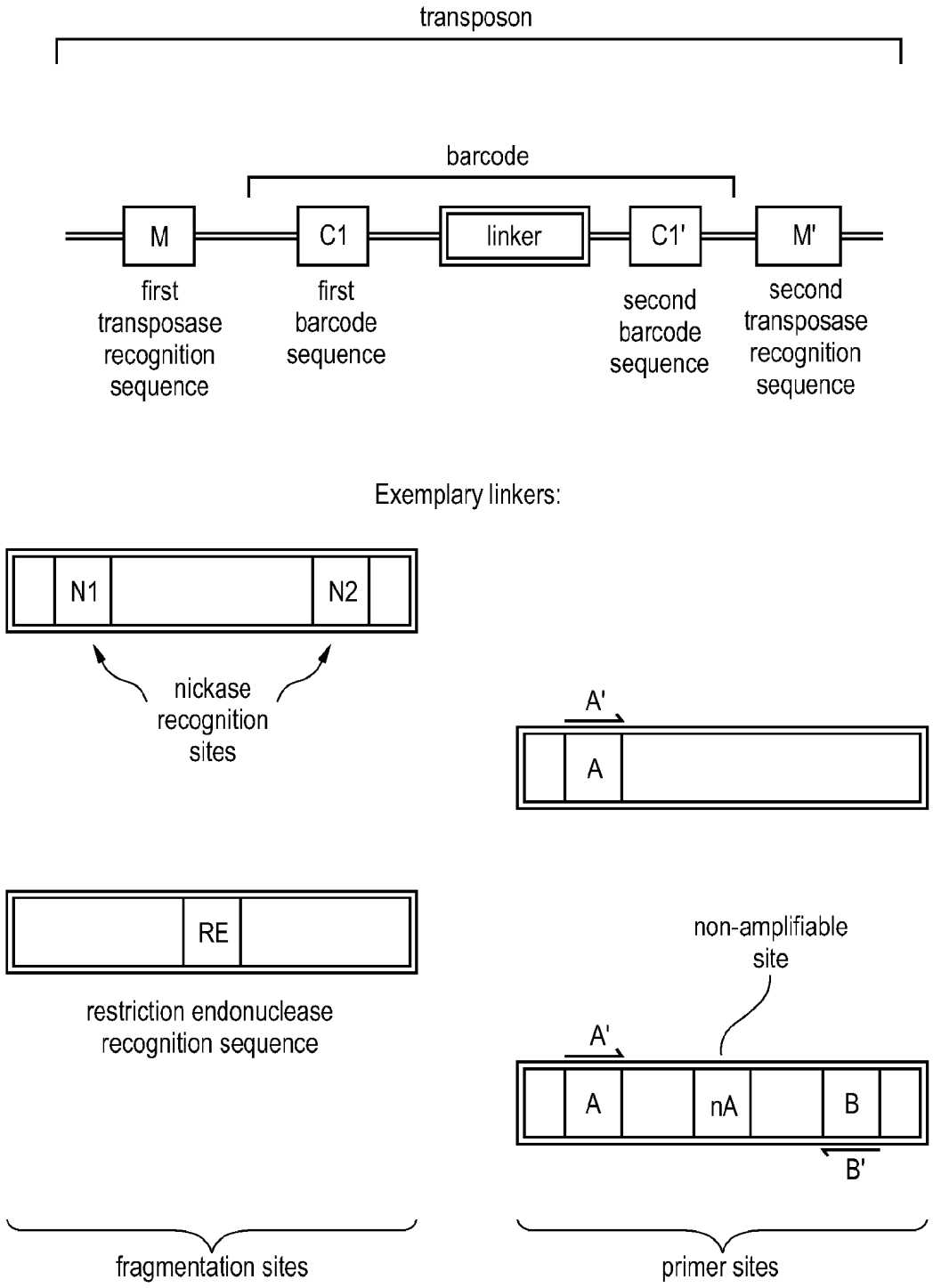


Fig. 1

IDT Exhibit 1020.4877

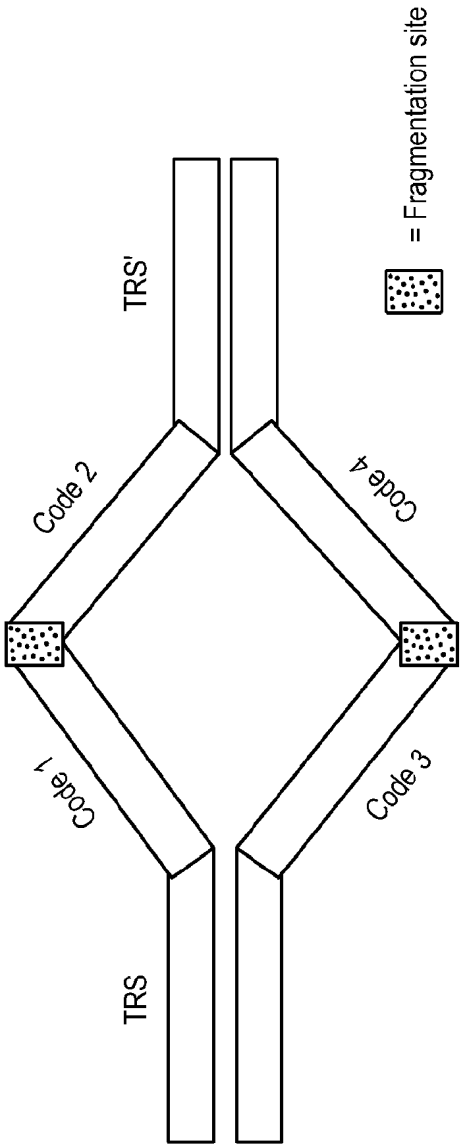


Fig. 2

SUBSTITUTE SHEET (RULE 26)

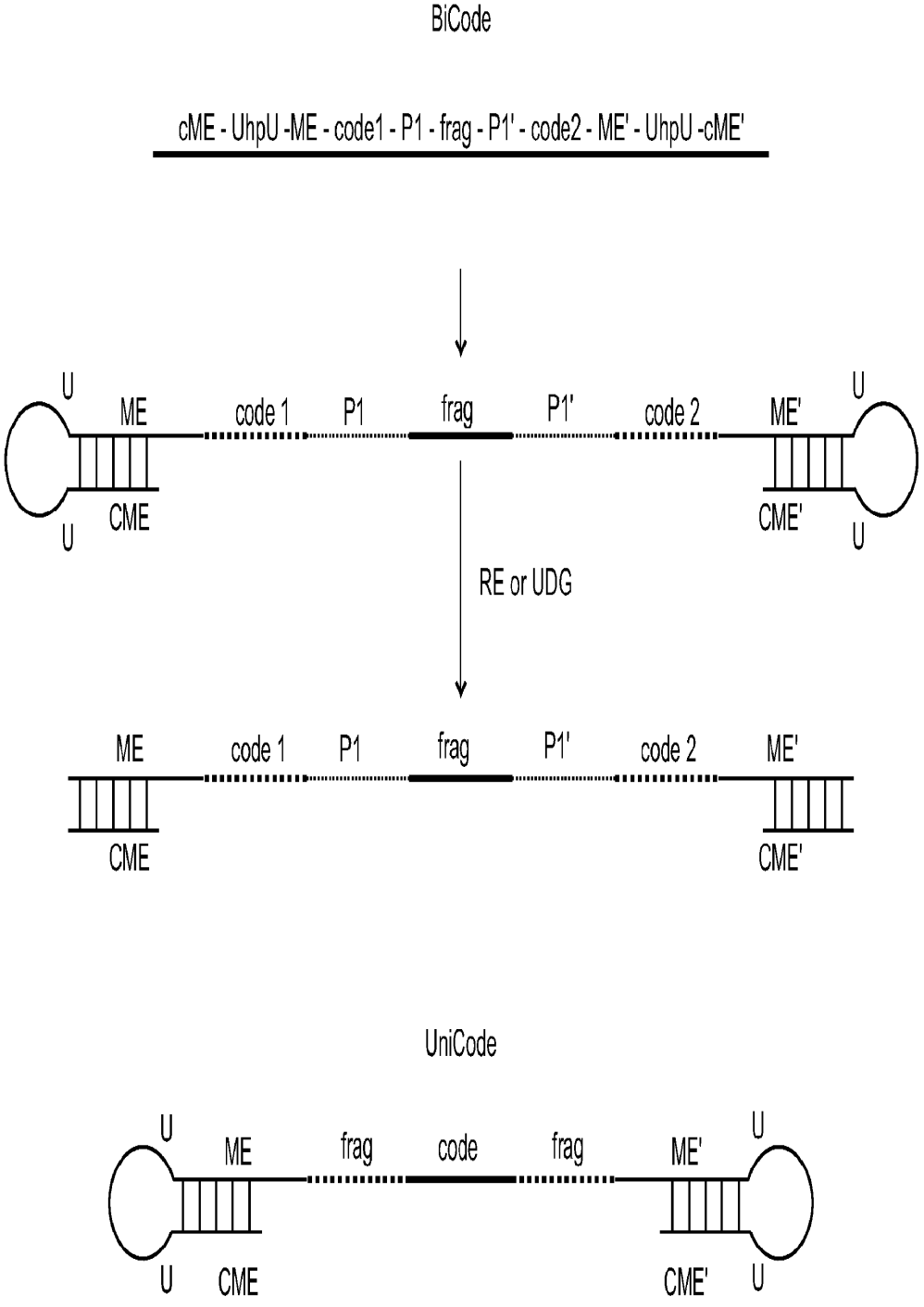


Fig. 3

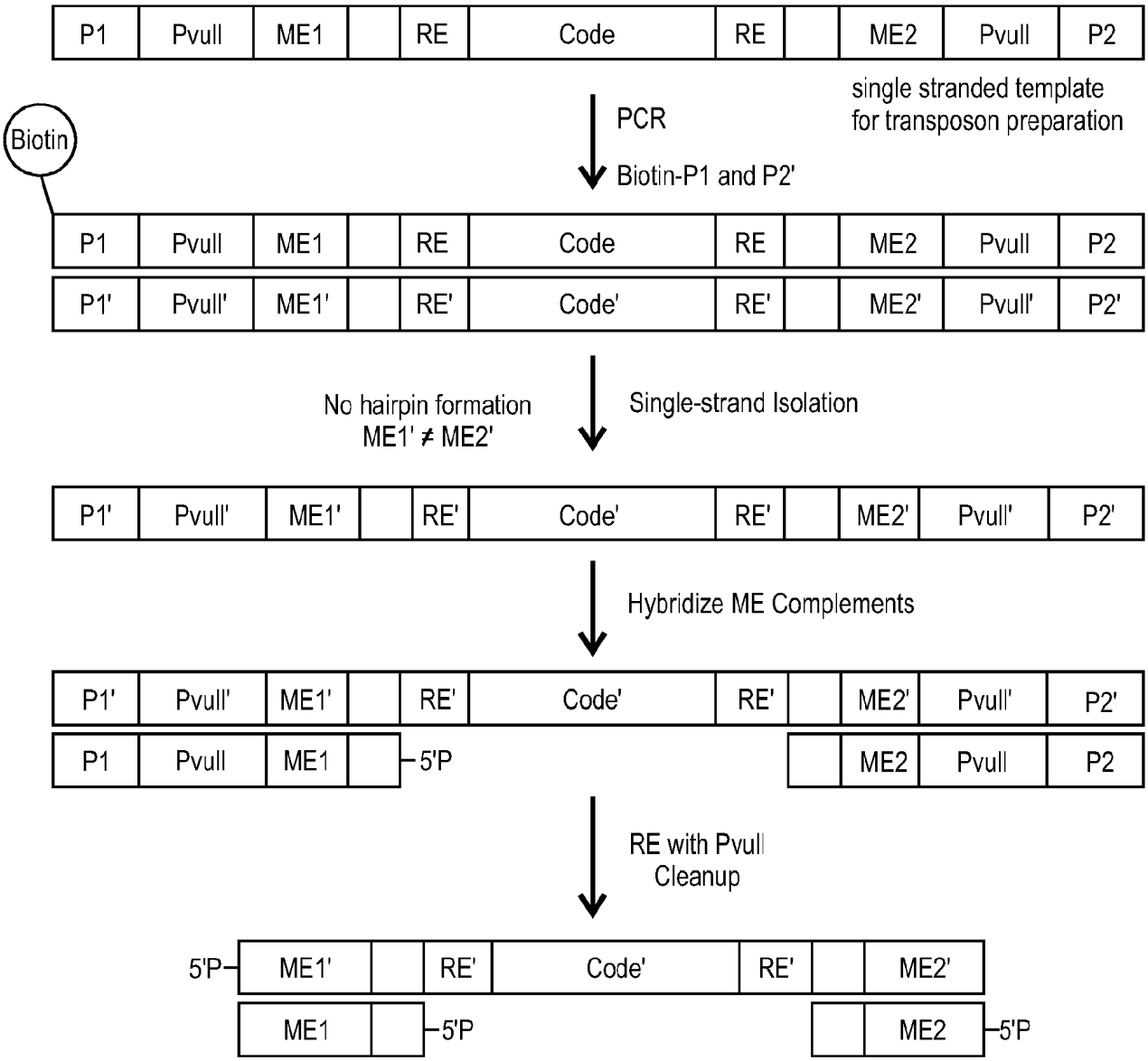


Fig. 4

IDT Exhibit 1020.4880

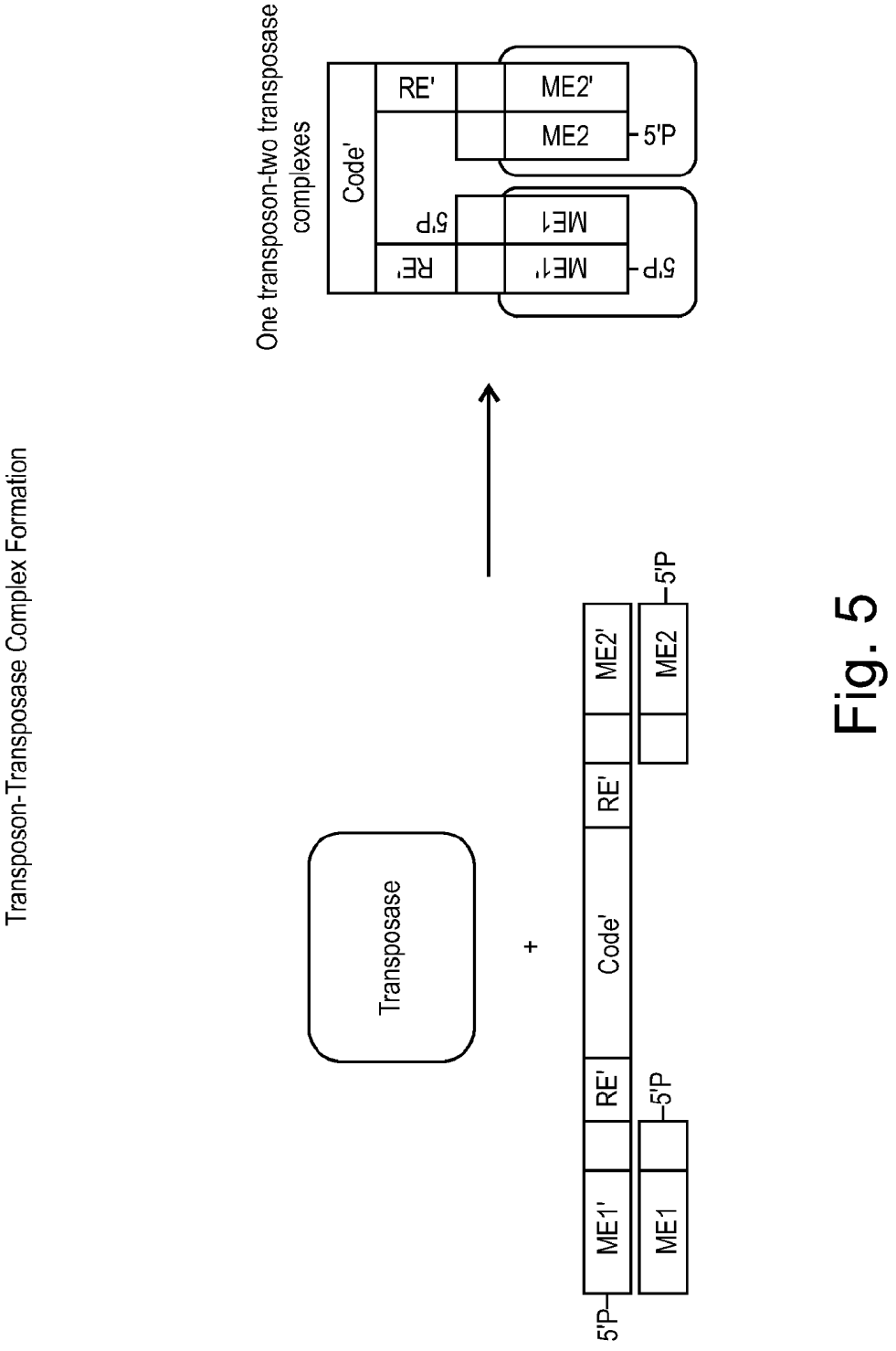

Transposon-Transposase Complex Formation
Transposase
+
5'P
ME1'
RE'
Code'
RE'
ME2'
ME1
5'P
ME2
5'P
One transposon-two transposase complexes
Code'
RE'
ME2'
ME2
5'P
5'P
RE'
ME1'
ME1
5'P

Fig. 5

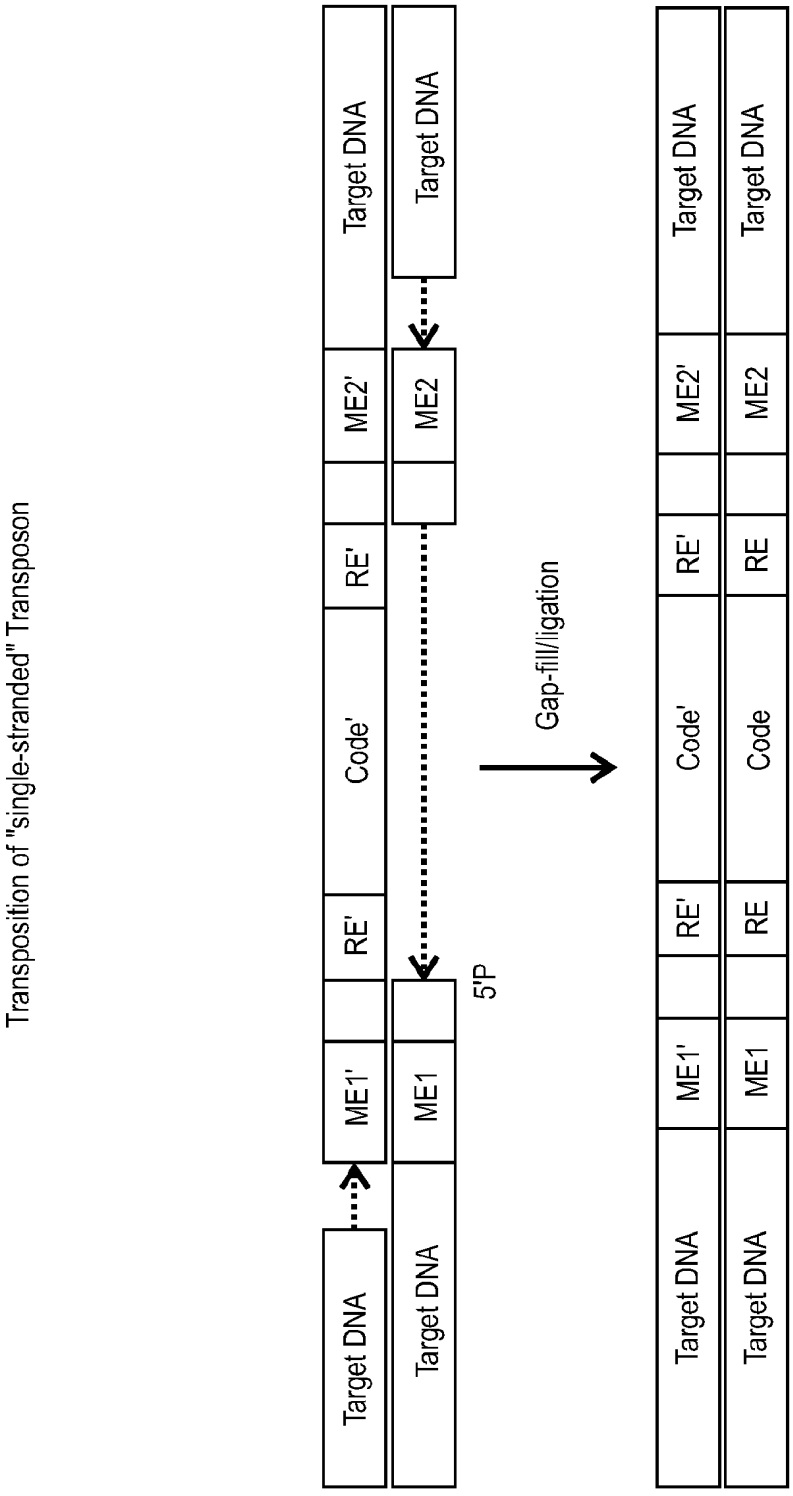

Transposition of "single-stranded" Transposon
Target DNA
ME1'
RE'
Code'
RE'
ME2'
Target DNA
Target DNA
ME1
ME2
5'P
Gap-fill/ligation
Target DNA
ME1'
RE'
Code'
RE'
ME2'
Target DNA
Target DNA
ME1
RE
Code
RE
ME2
Target DNA

Fig. 6

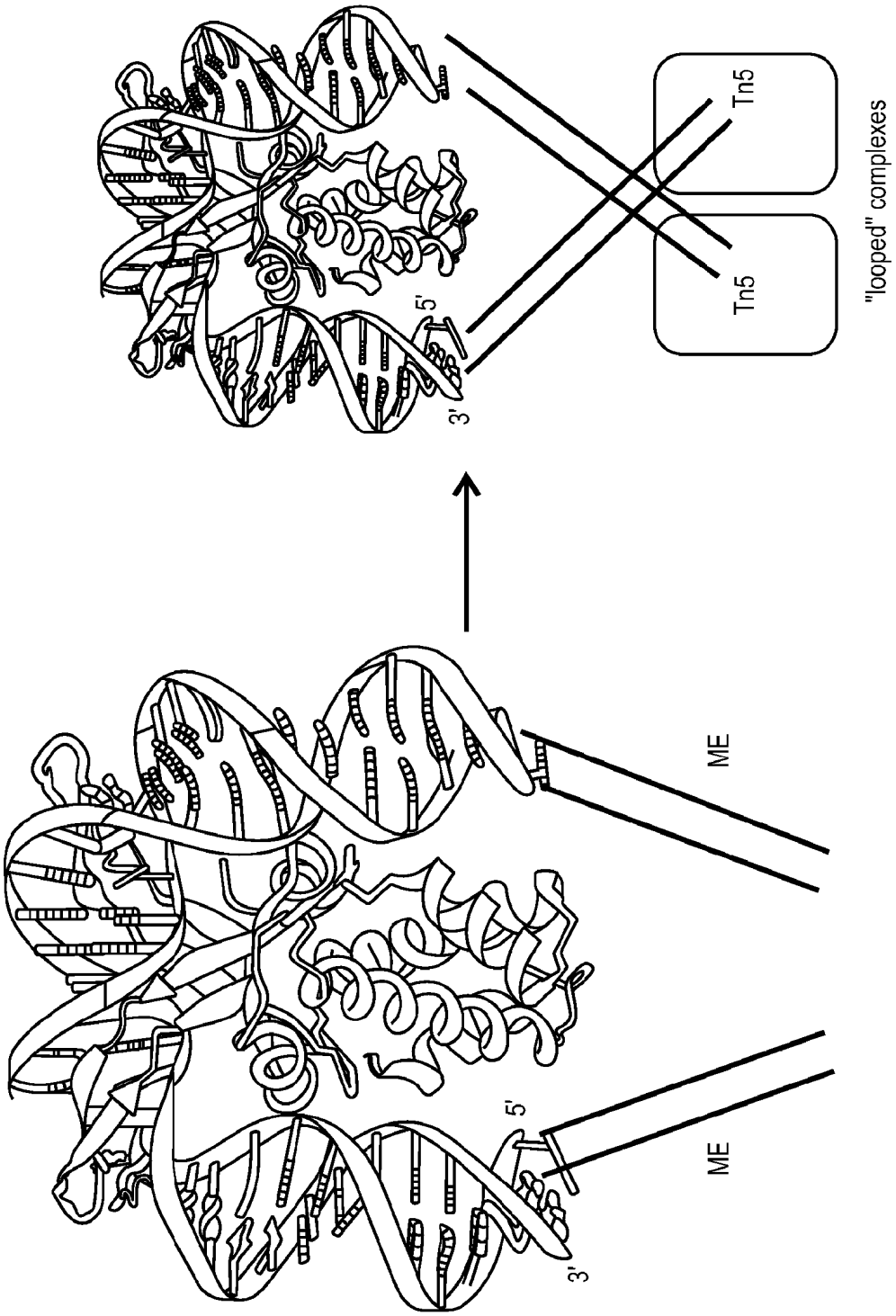


Fig. 7

IDT Exhibit 1020.4883

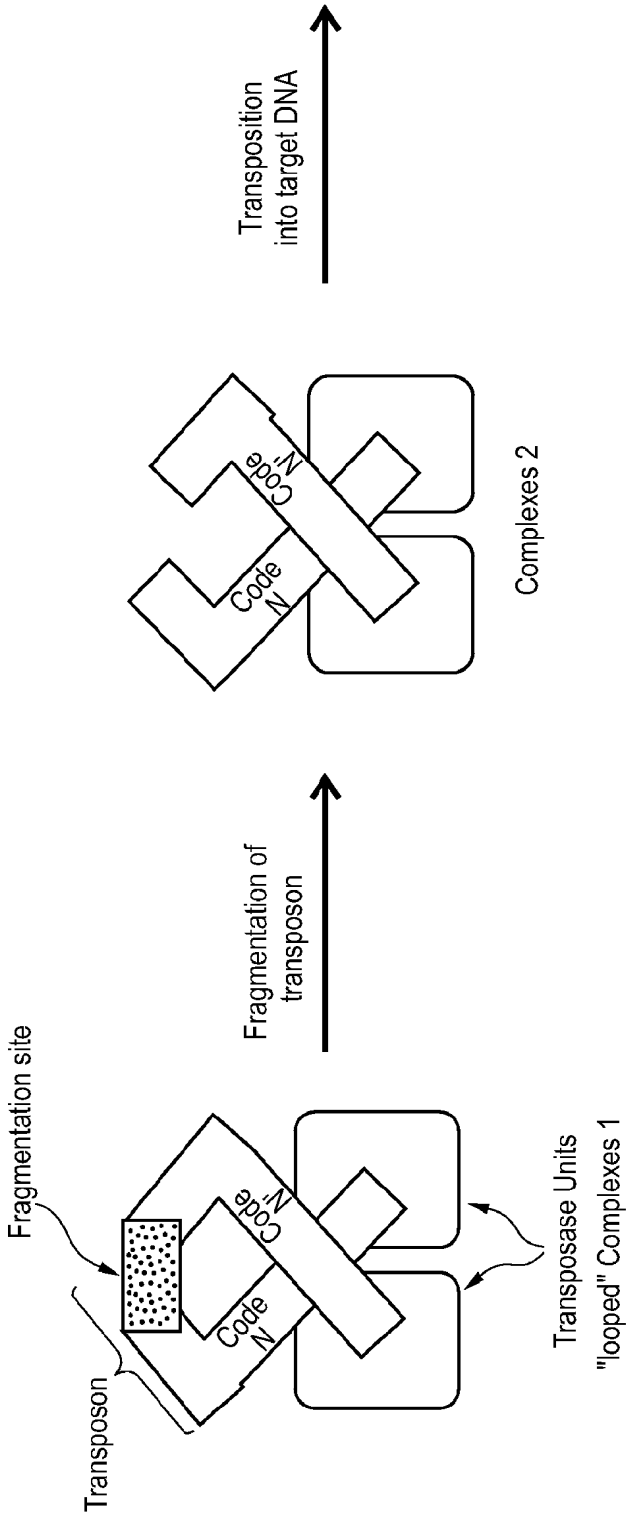

Transposition into target DNA
Complexes 2
Code N
Code N
Fragmentation of transposon
Fragmentation site
Transposase Units
"looped" Complexes 1
Transposon
Code N
Code N

Fig. 8

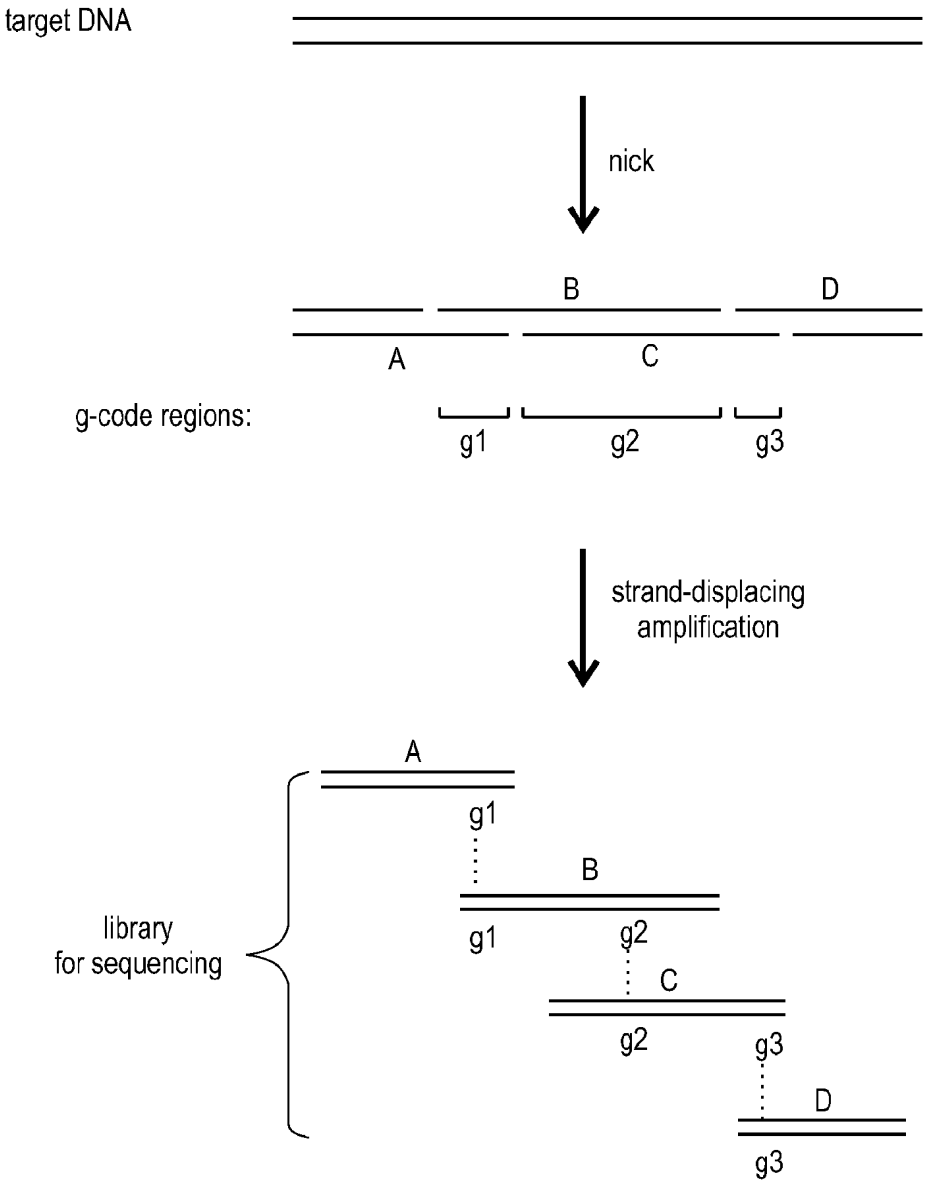


Fig. 9

IDT Exhibit 1020.4885

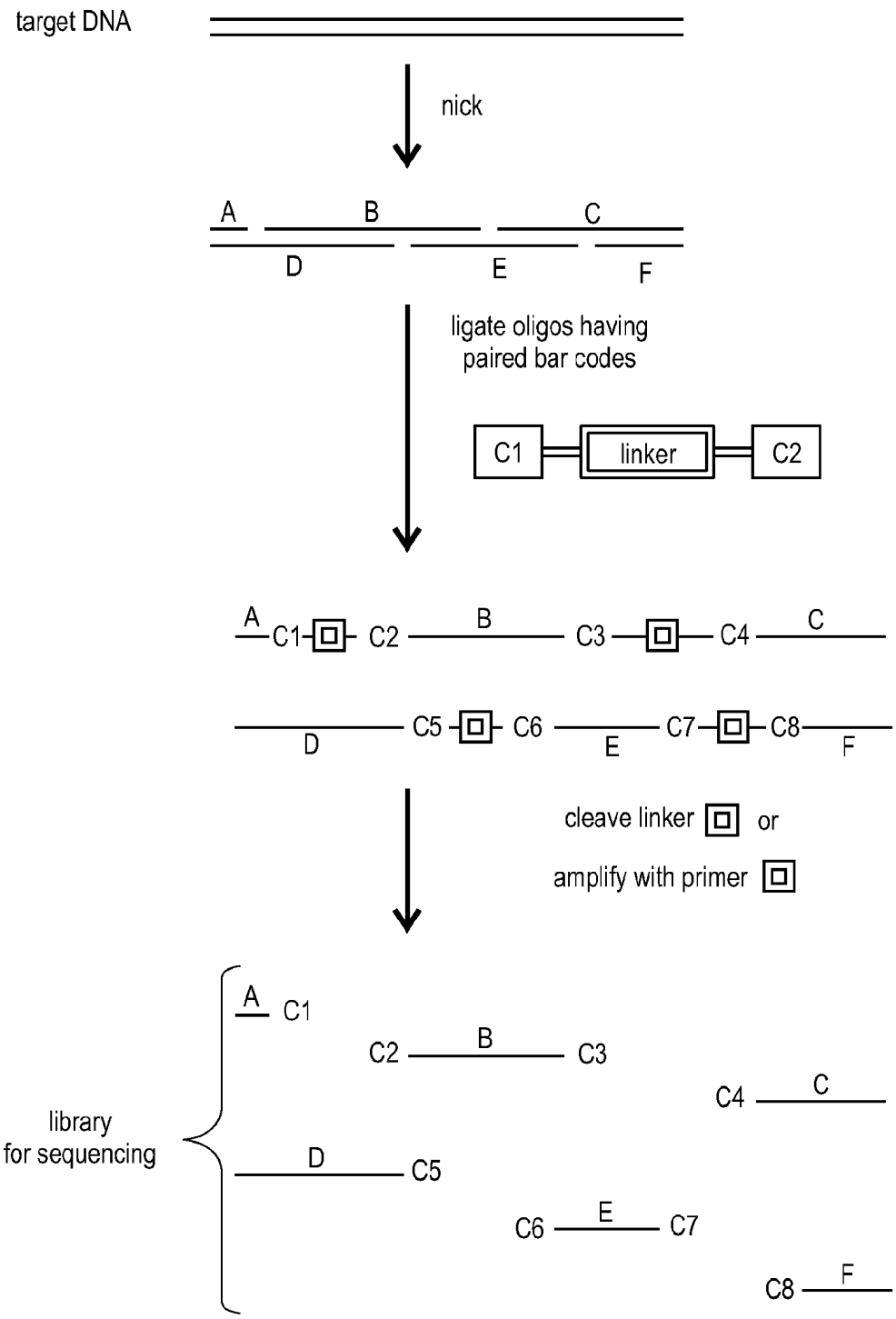
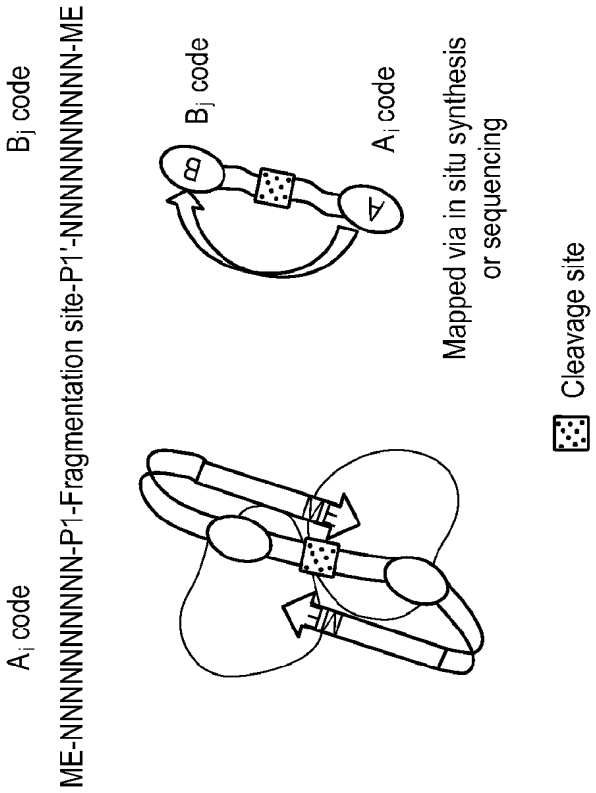


Fig. 10

IDT Exhibit 1020.4886



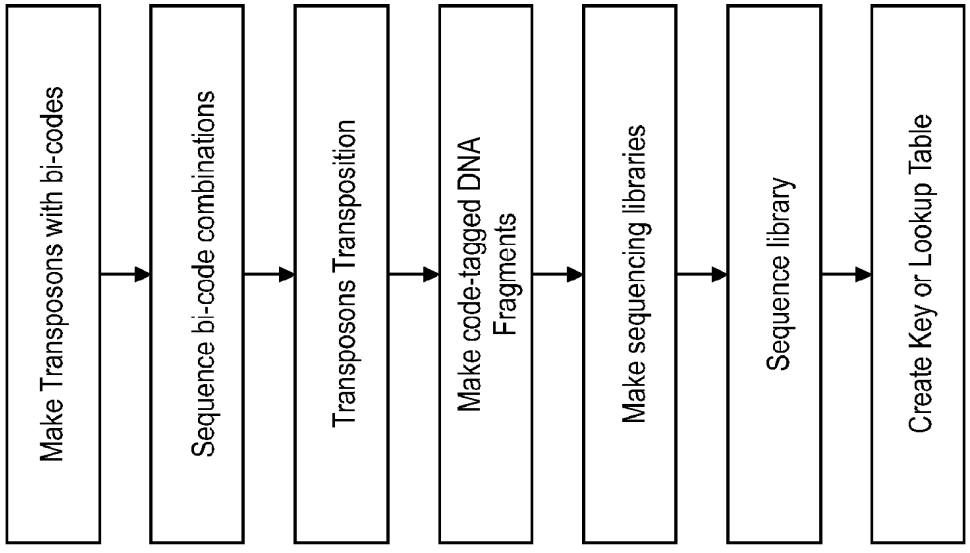


Fig. 11

IDT Exhibit 1020.4887

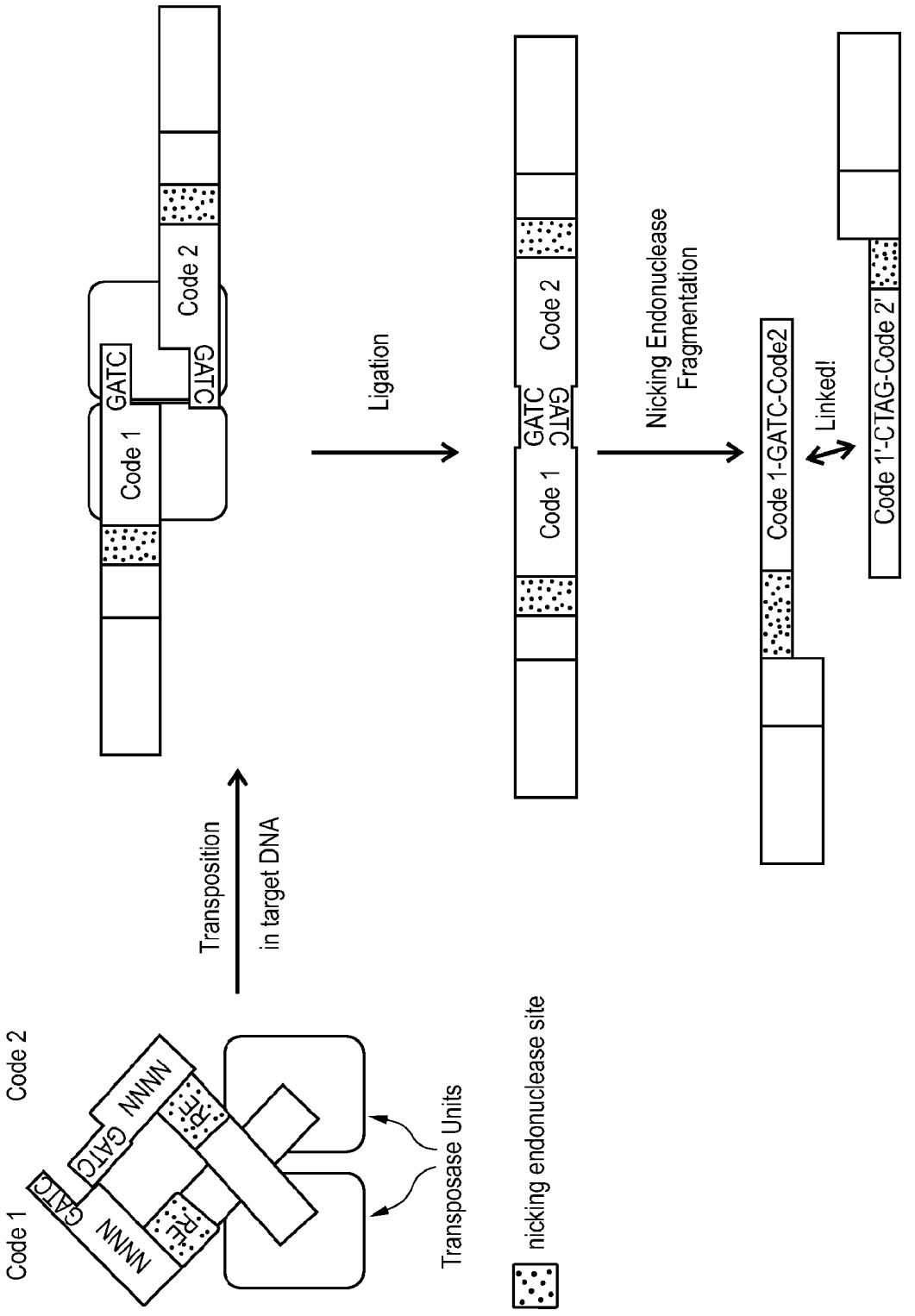

Code 2
NNNN
GATC
RE
Code 1
GATC
NNNN
RE
Transposase Units
nicking endonuclease site
Transposition
in target DNA
Code 2
GATC
Code 1
GATC
Code 2
Code 1
Ligation
Nicking Endonuclease
Fragmentation
Code 1-GATC-Code2
Linked!
Code 1'-CTAG-Code 2'

Fig. 12

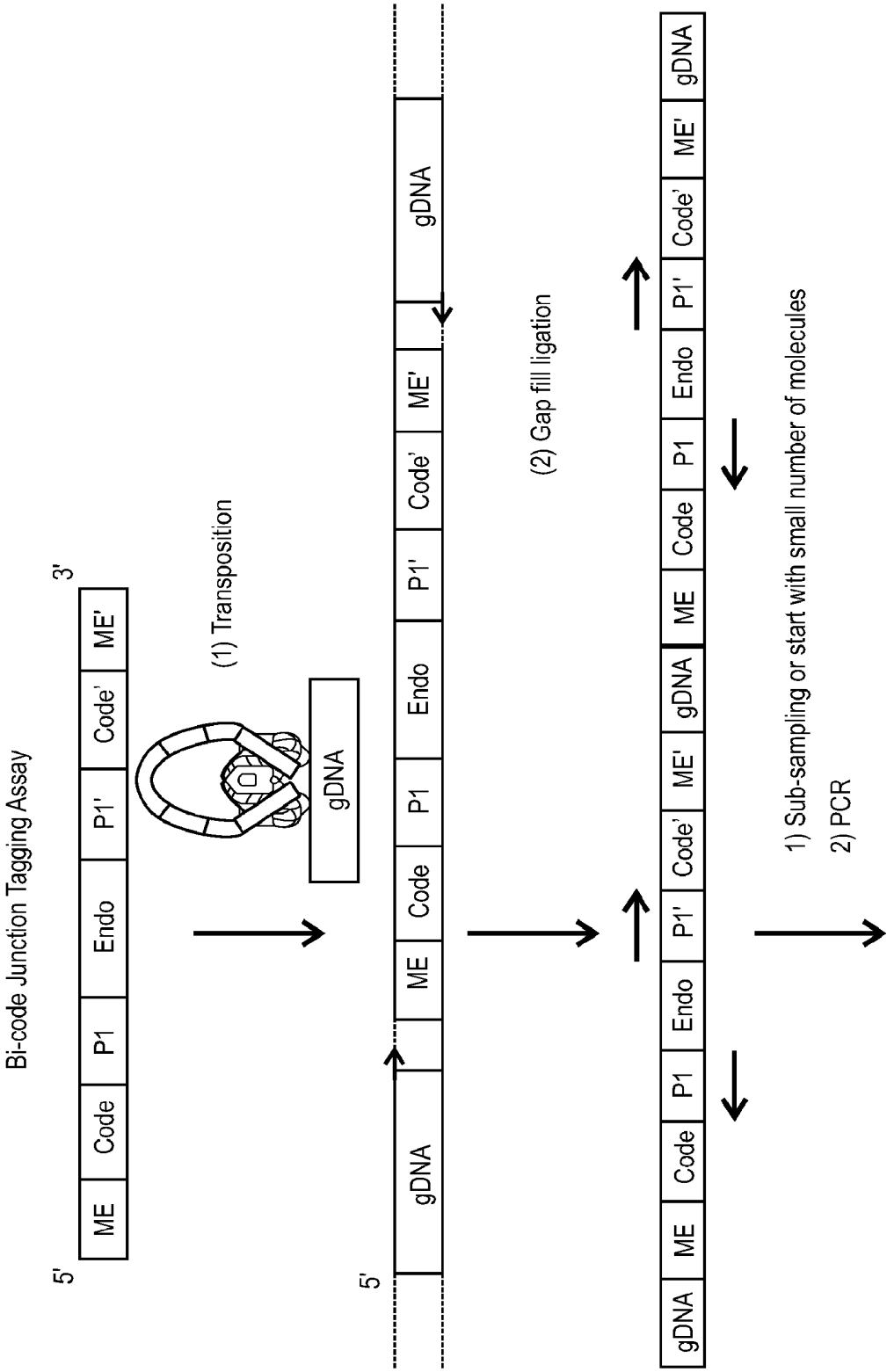

Bi-code Junction Tagging Assay
5'
3'
ME
Code
P1
Endo
P1'
Code'
ME'
(1) Transposition
gDNA
5'
gDNA
ME
Code
P1
Endo
P1'
Code'
ME'
gDNA
(2) Gap fill ligation
gDNA
ME
Code
P1
Endo
P1'
Code'
ME'
gDNA
ME
Code
P1
Endo
P1'
Code'
ME'
gDNA
1) Sub-sampling or start with small number of molecules
2) PCR

Fig. 13

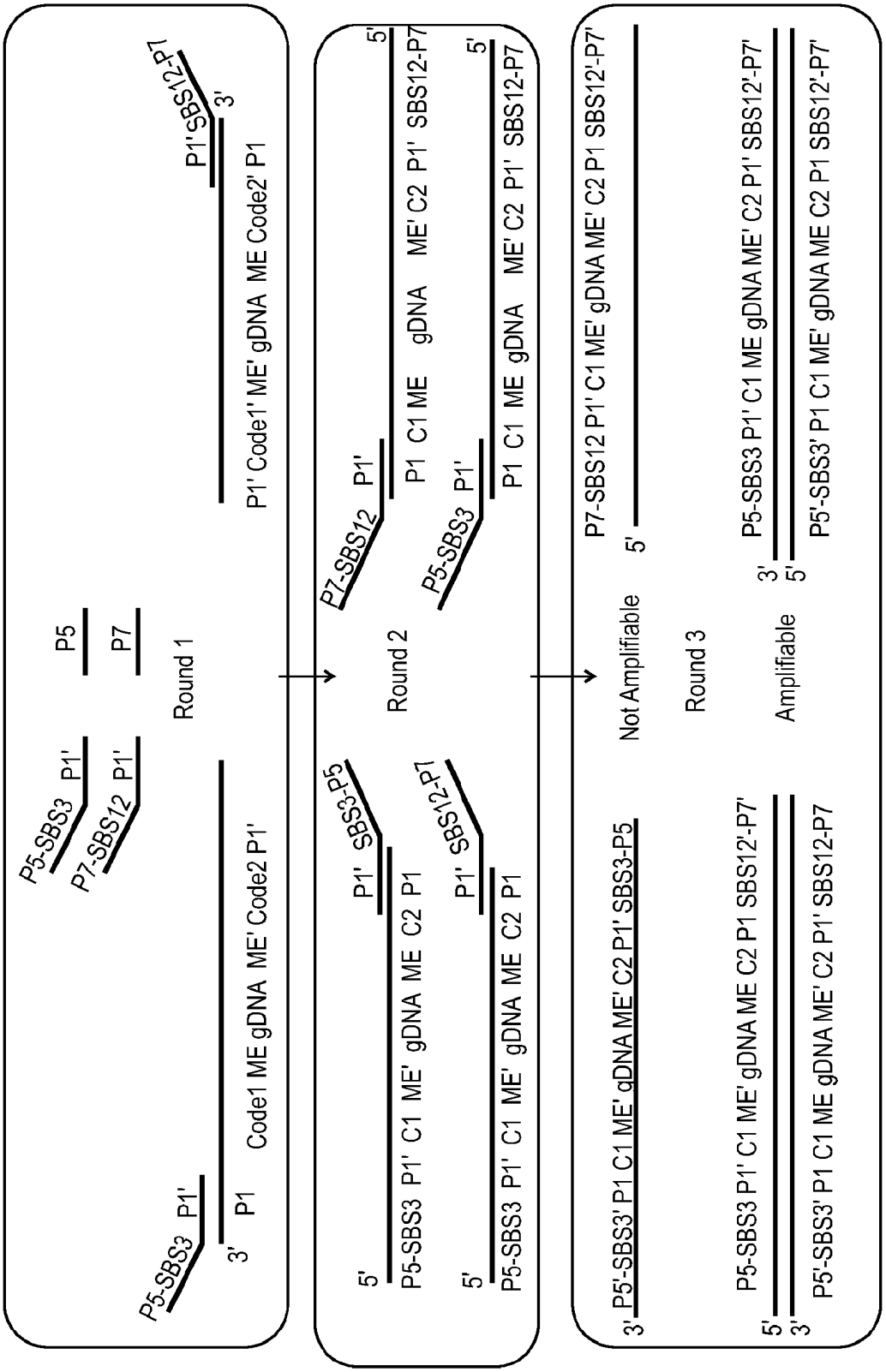

P5-SBS3 P1'
P7-SBS12 P1'
P5
P7
Round 1
P1'SBS12-P7
3'
P1' Code1' ME' gDNA ME Code2' P1
P5-SBS3 P1'
3' P1 Code1 ME gDNA ME' Code2 P1'
P7-SBS12 P1'
5'
P1 C1 ME gDNA ME' C2 P1' SBS12-P7
P5-SBS3 P1'
P1 C1 ME gDNA ME' C2 P1' SBS12-P7
Round 2
P1' SBS3-P5
5'
P5-SBS3 P1' C1 ME' gDNA ME C2 P1
P1' SBS12-P7
5'
P5-SBS3 P1' C1 ME' gDNA ME C2 P1
P7-SBS12 P1' C1 ME' gDNA ME' C2 P1 SBS12'-P7'
5'
Not Amplifiable
3' P5'-SBS3' P1 C1 ME' gDNA ME' C2 P1' SBS3-P5
Round 3
P5-SBS3 P1' C1 ME gDNA ME' C2 P1' SBS12'-P7'
3'
5'
Amplifiable
P5'-SBS3' P1 C1 ME' gDNA ME C2 P1 SBS12'-P7'
5'
3'
P5-SBS3' P1 C1 ME gDNA ME' C2 P1' SBS12-P7

Fig. 14

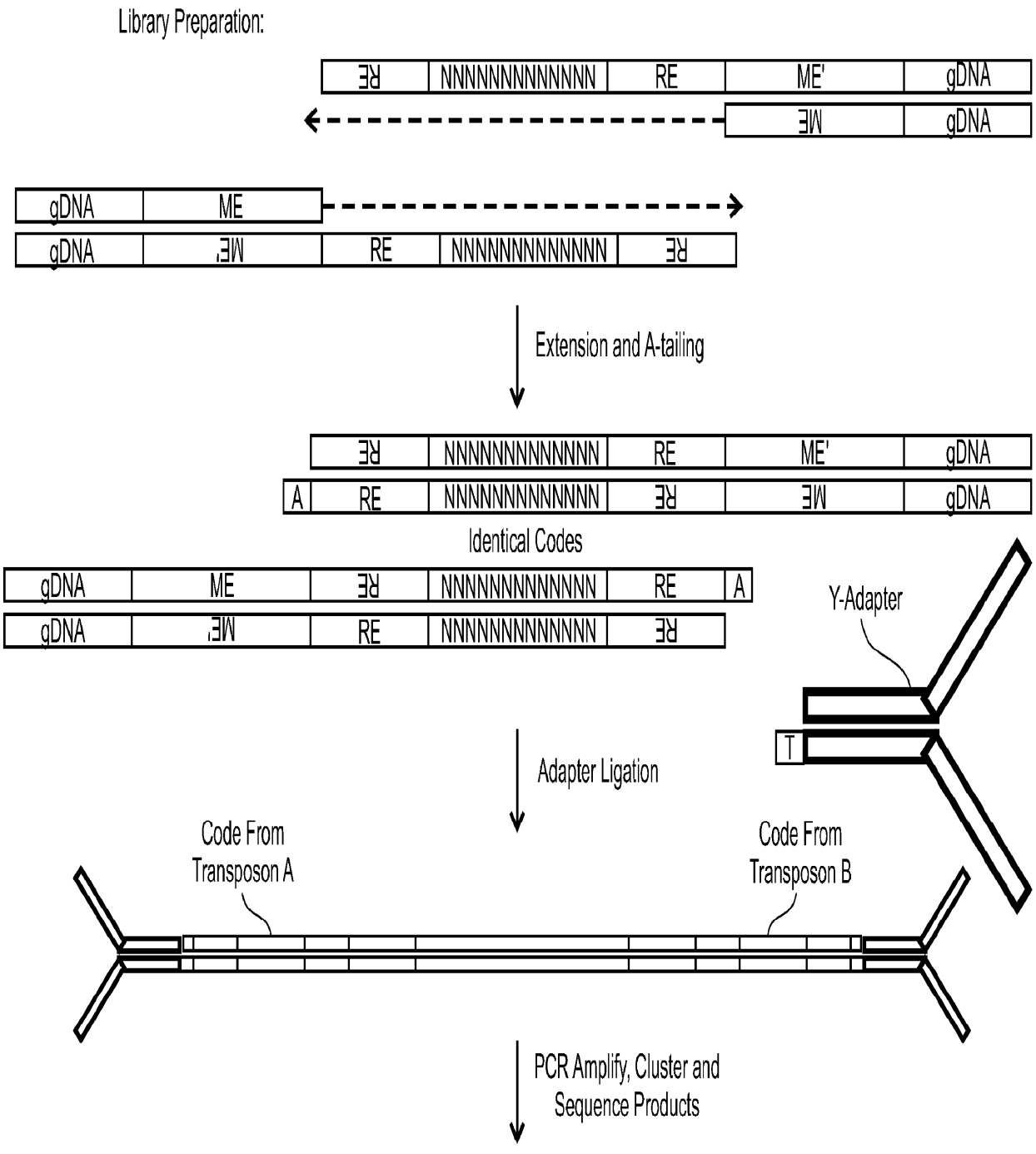
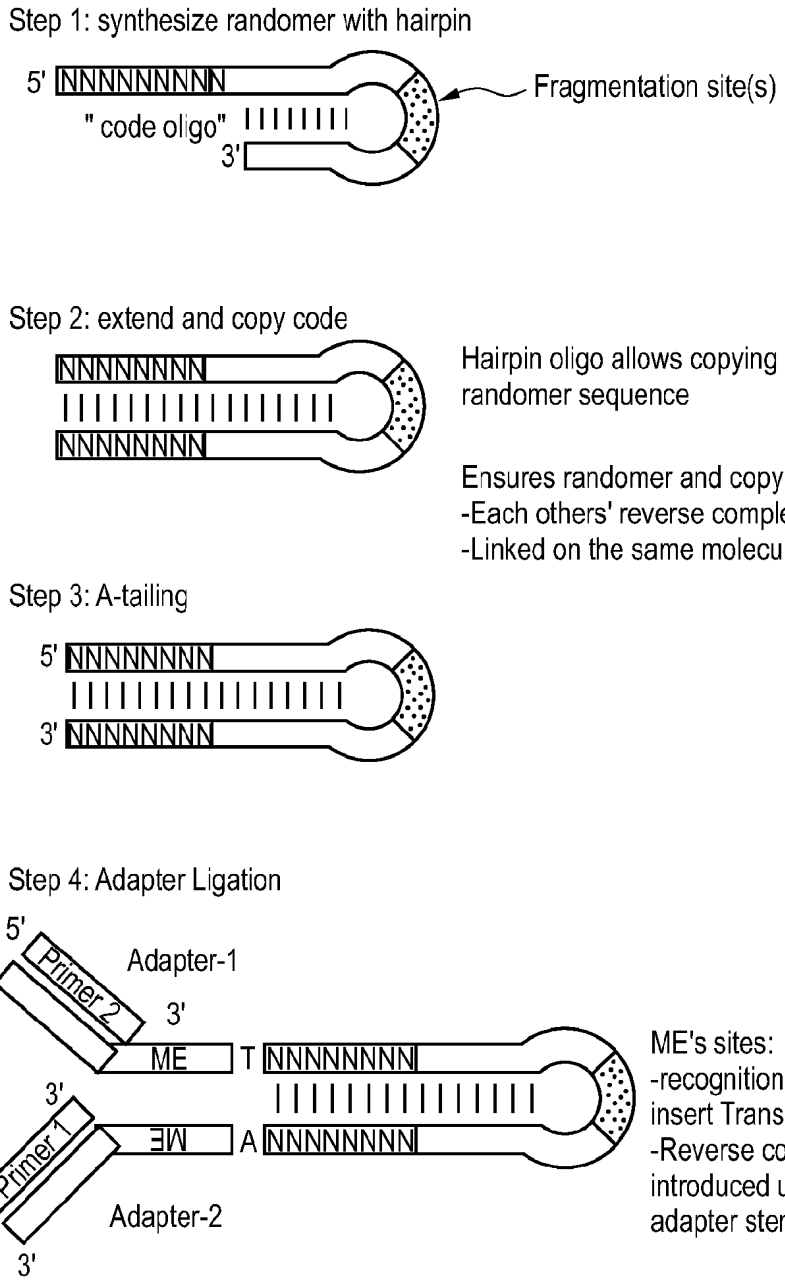


Fig. 15

IDT Exhibit 1020.4891



P2	ME	Code				Code'	ME	P1'
P2'	ME'	Code'				Code	ME'	P1

Fig. 16

IDT Exhibit 1020.4892

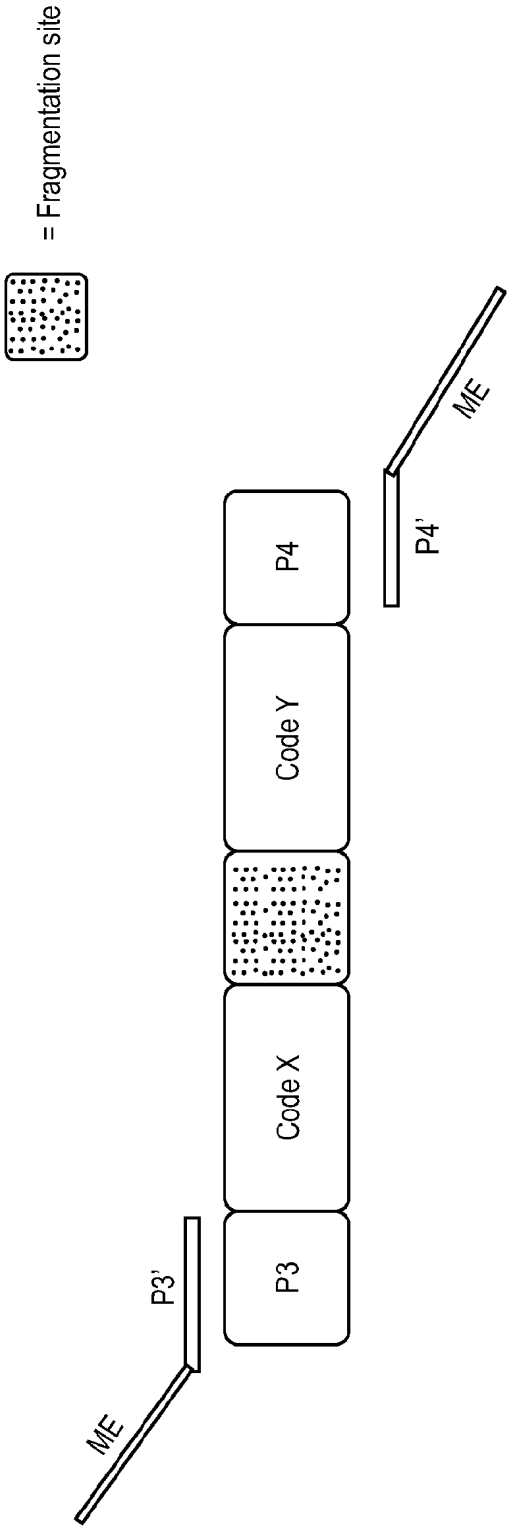


Fig. 17

IDT Exhibit 1020.4893

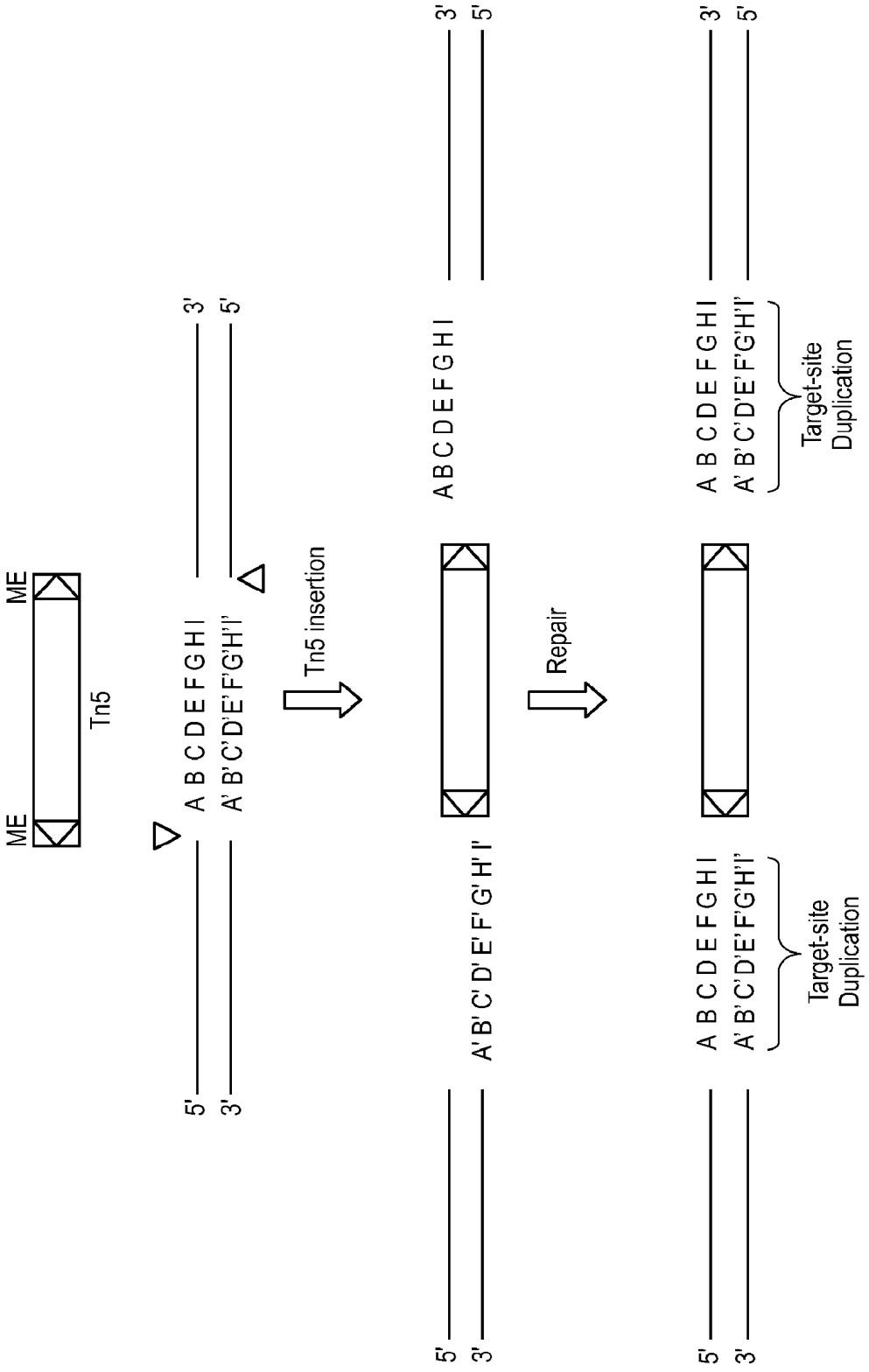


Fig. 18

IDT Exhibit 1020.4894

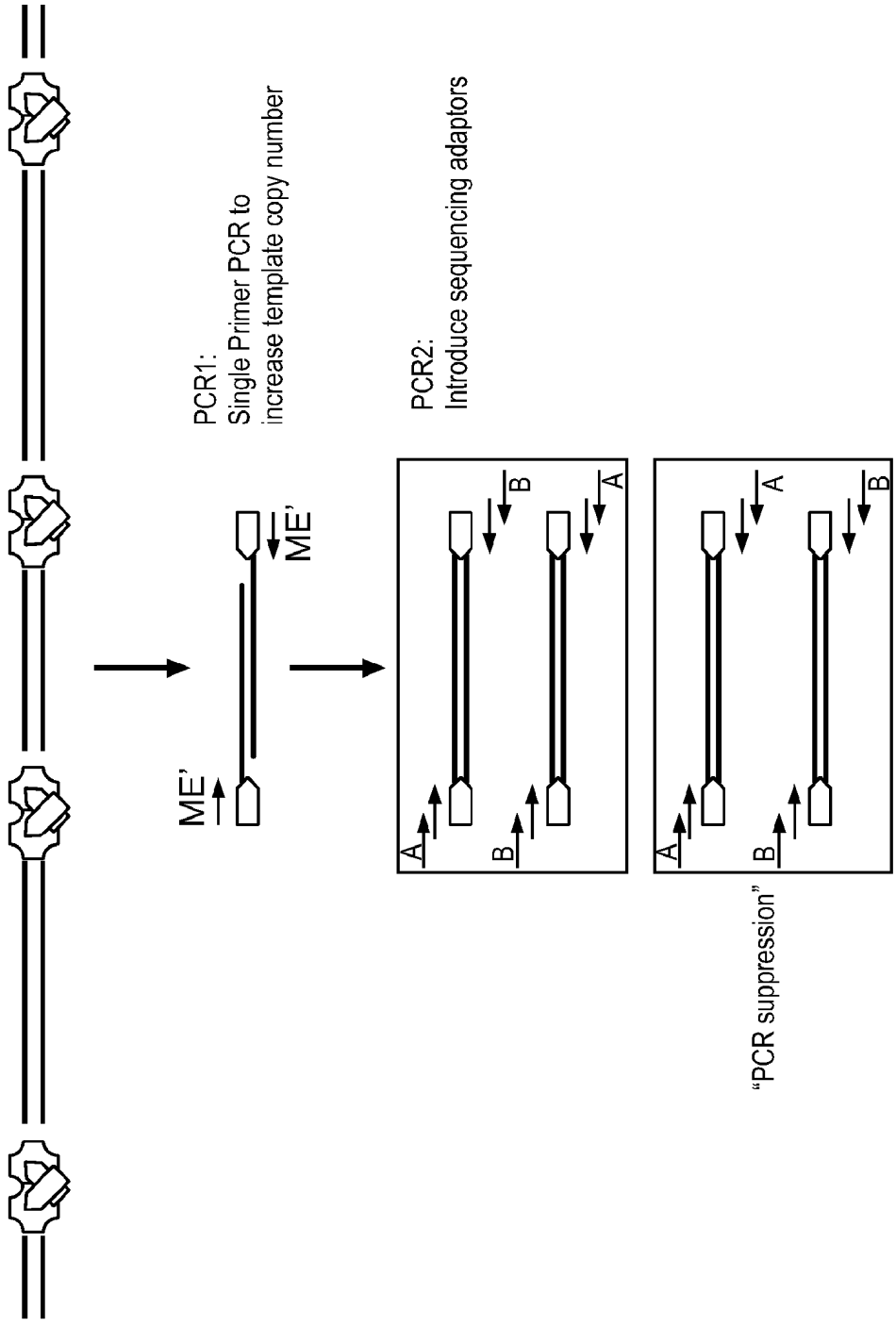


Fig. 19

IDT Exhibit 1020.4895

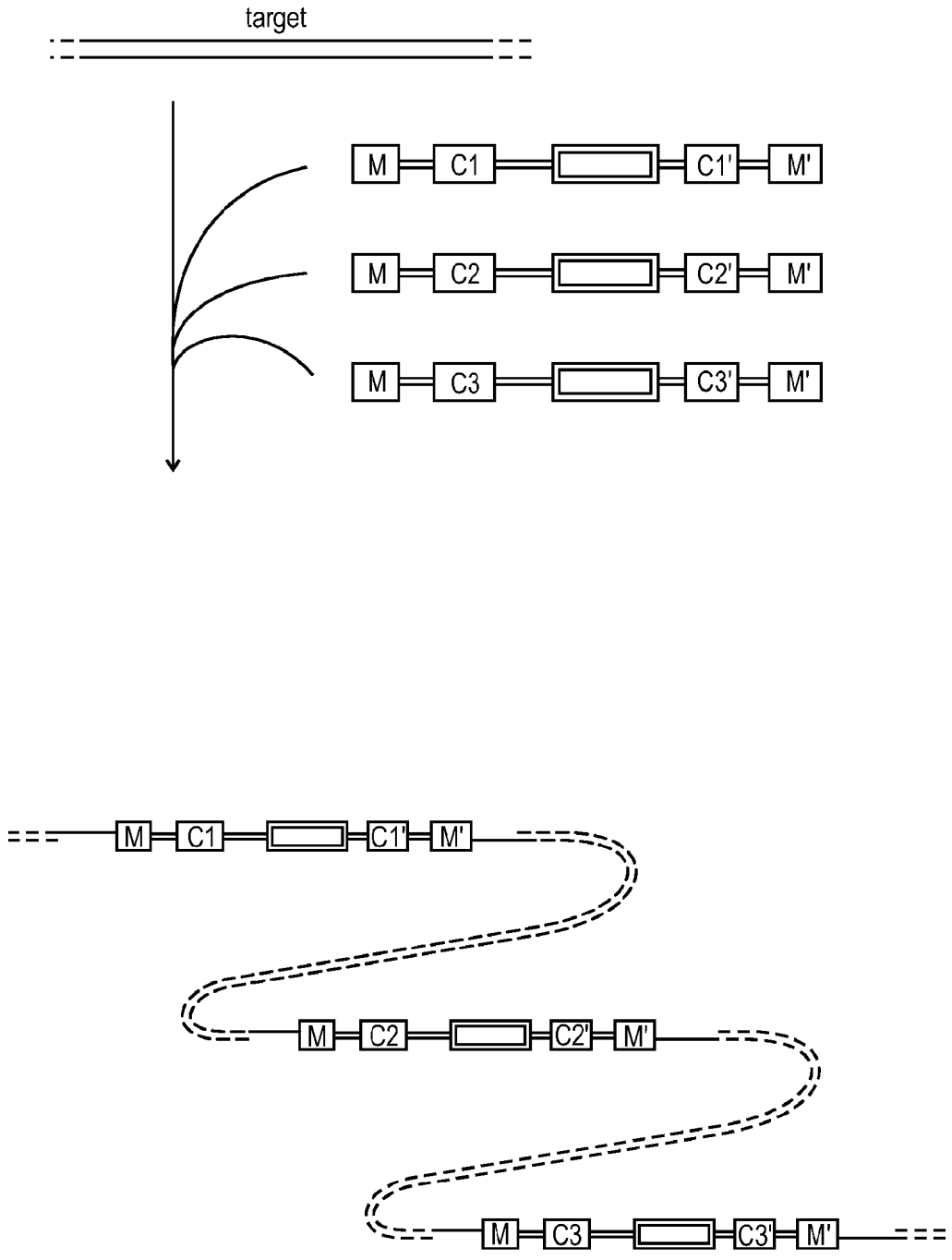


Fig. 20

IDT Exhibit 1020.4896

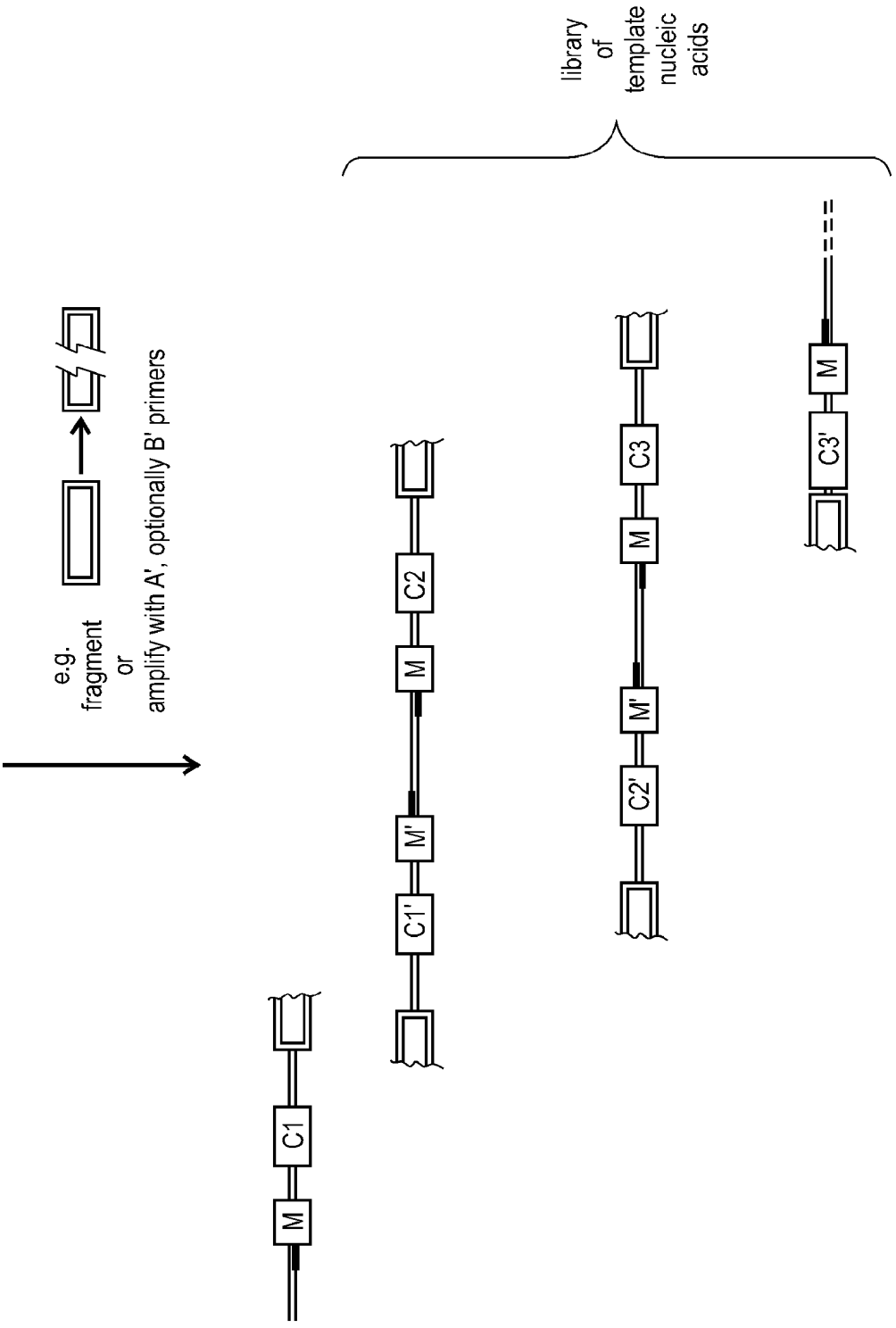


Fig. 21

IDT Exhibit 1020.4897

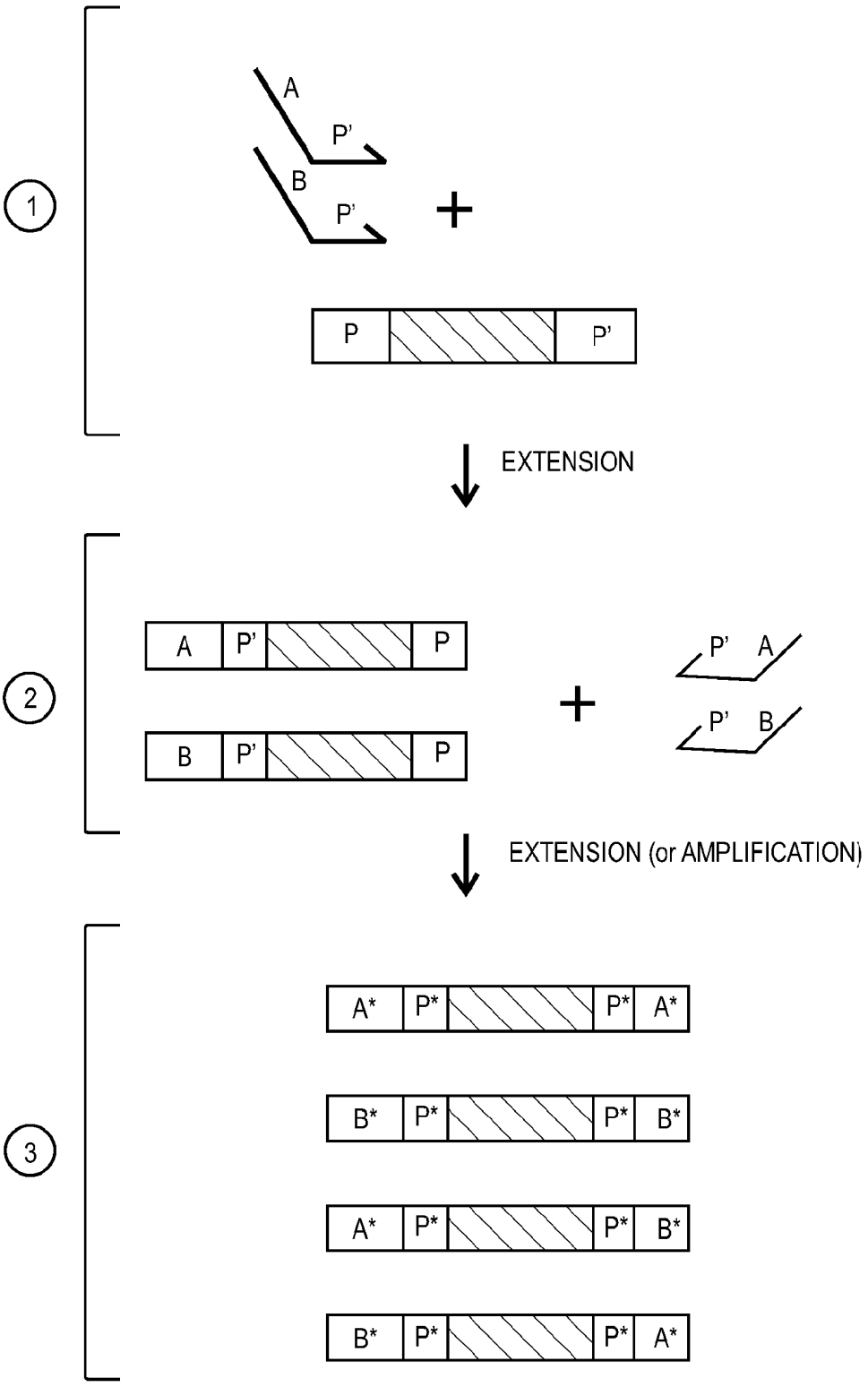


Fig. 22

IDT Exhibit 1020.4898

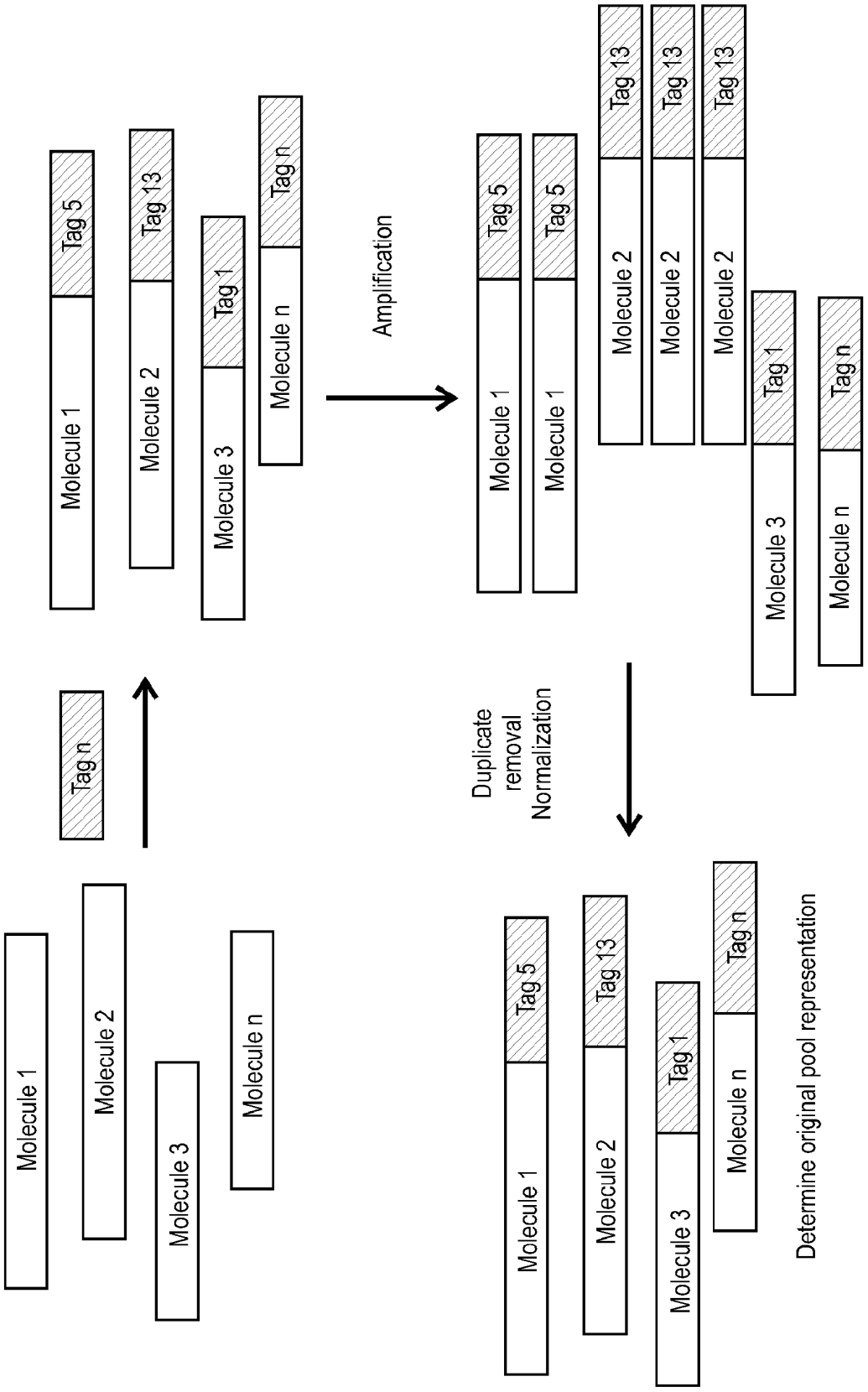

Molecule 1
Molecule 2
Molecule 3
Molecule n
Tag n
Molecule 1
Tag 5
Molecule 2
Tag 13
Molecule 3
Tag 1
Molecule n
Tag n
Amplification
Molecule 1
Tag 5
Molecule 1
Tag 5
Molecule 2
Tag 13
Molecule 2
Tag 13
Molecule 2
Tag 13
Molecule 3
Tag 1
Molecule n
Tag n
Duplicate removal Normalization
Molecule 1
Tag 5
Molecule 2
Tag 13
Molecule 3
Tag 1
Molecule n
Tag n
Determine original pool representation

Fig. 23

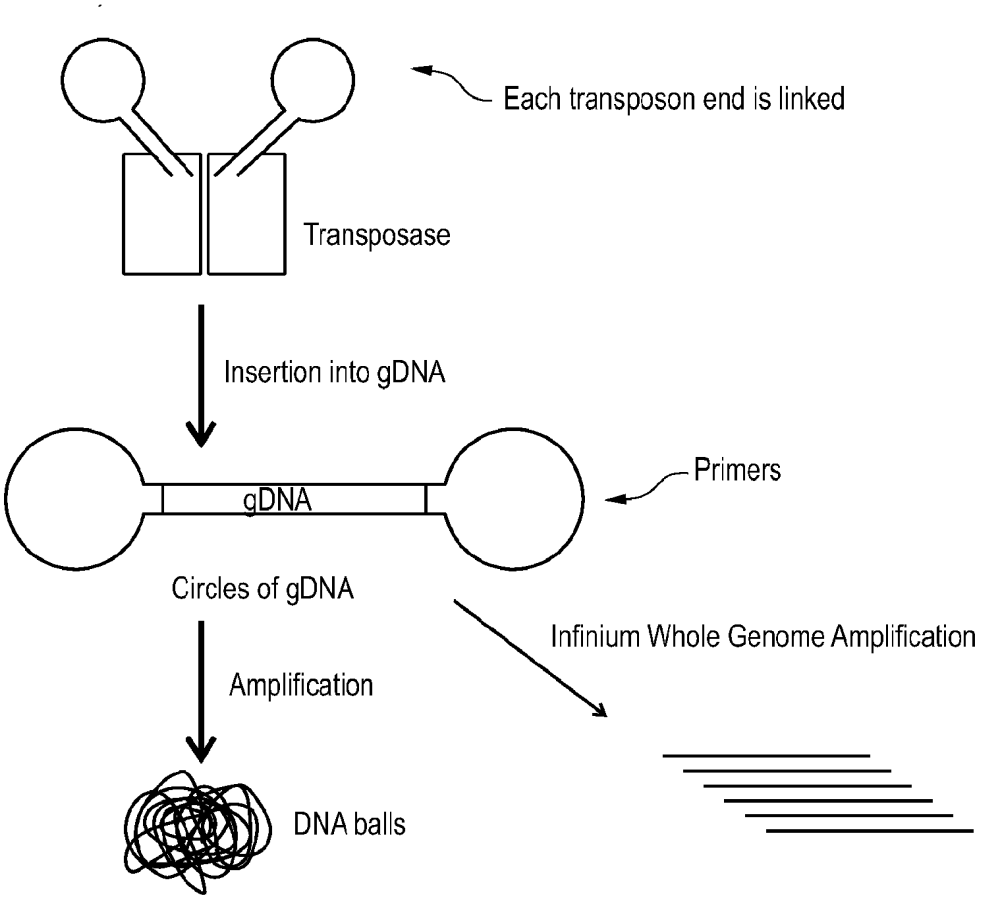


Fig. 24

IDT Exhibit 1020.4900

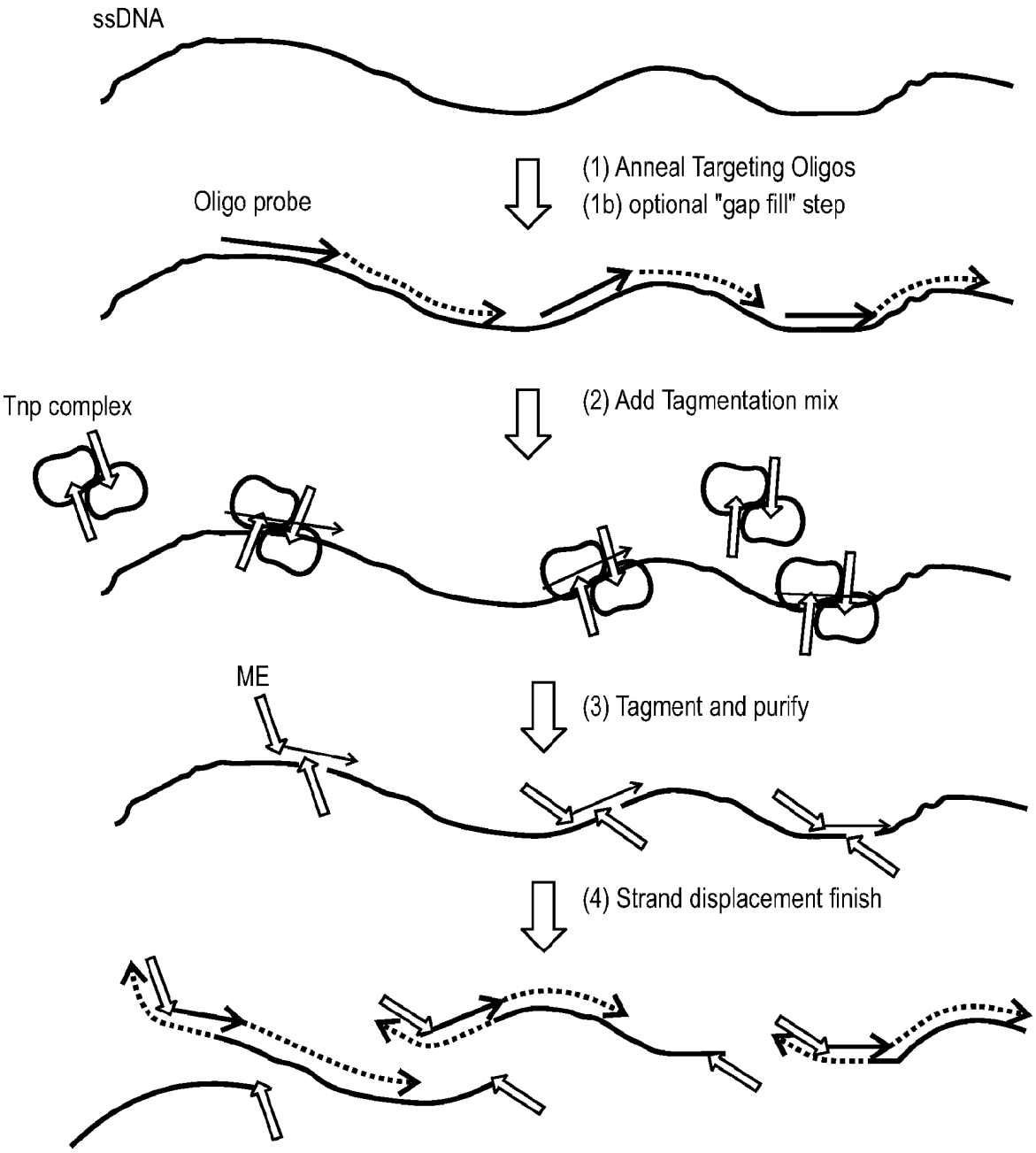


Fig. 25

IDT Exhibit 1020.4901

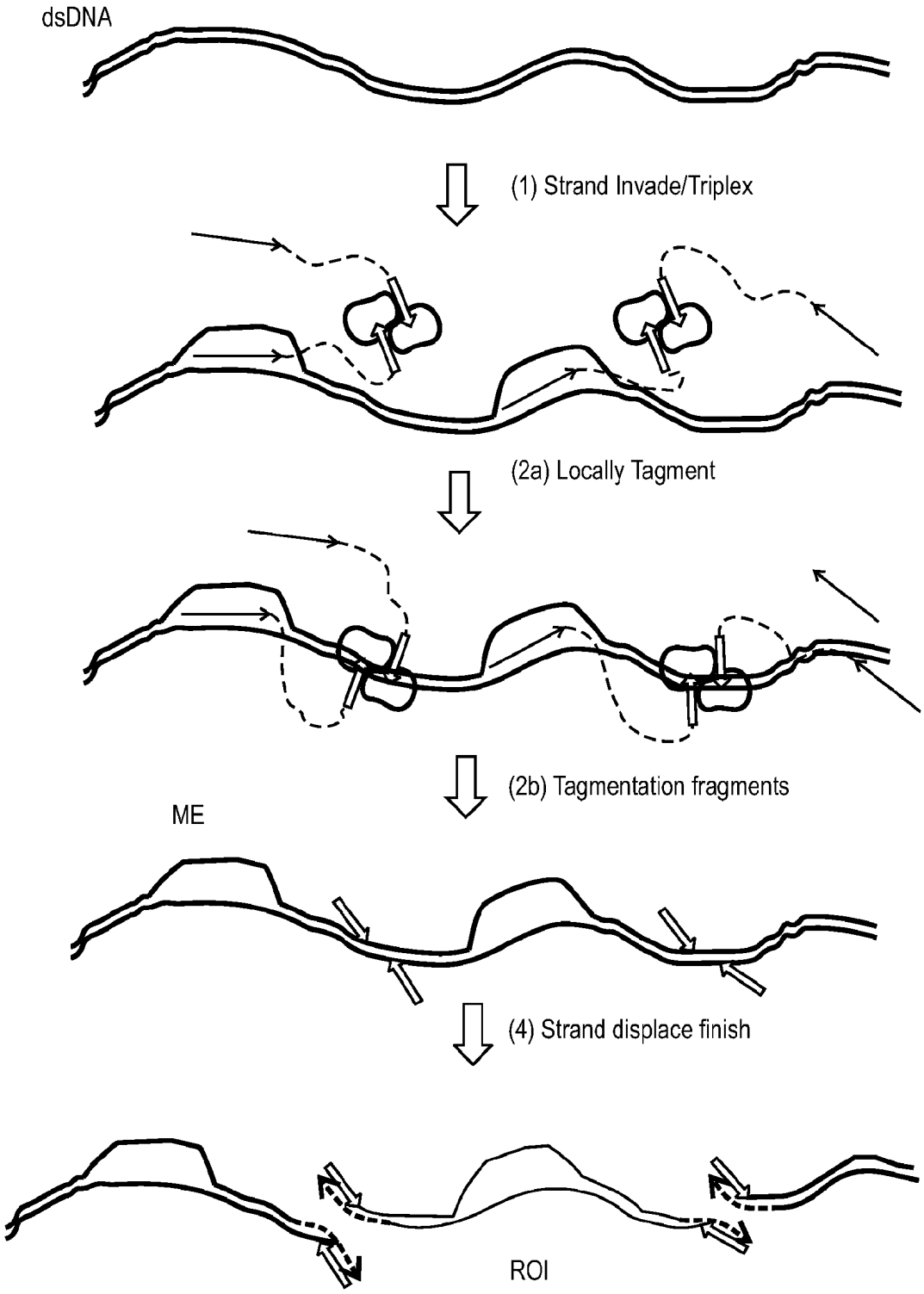


Fig. 26

IDT Exhibit 1020.4902

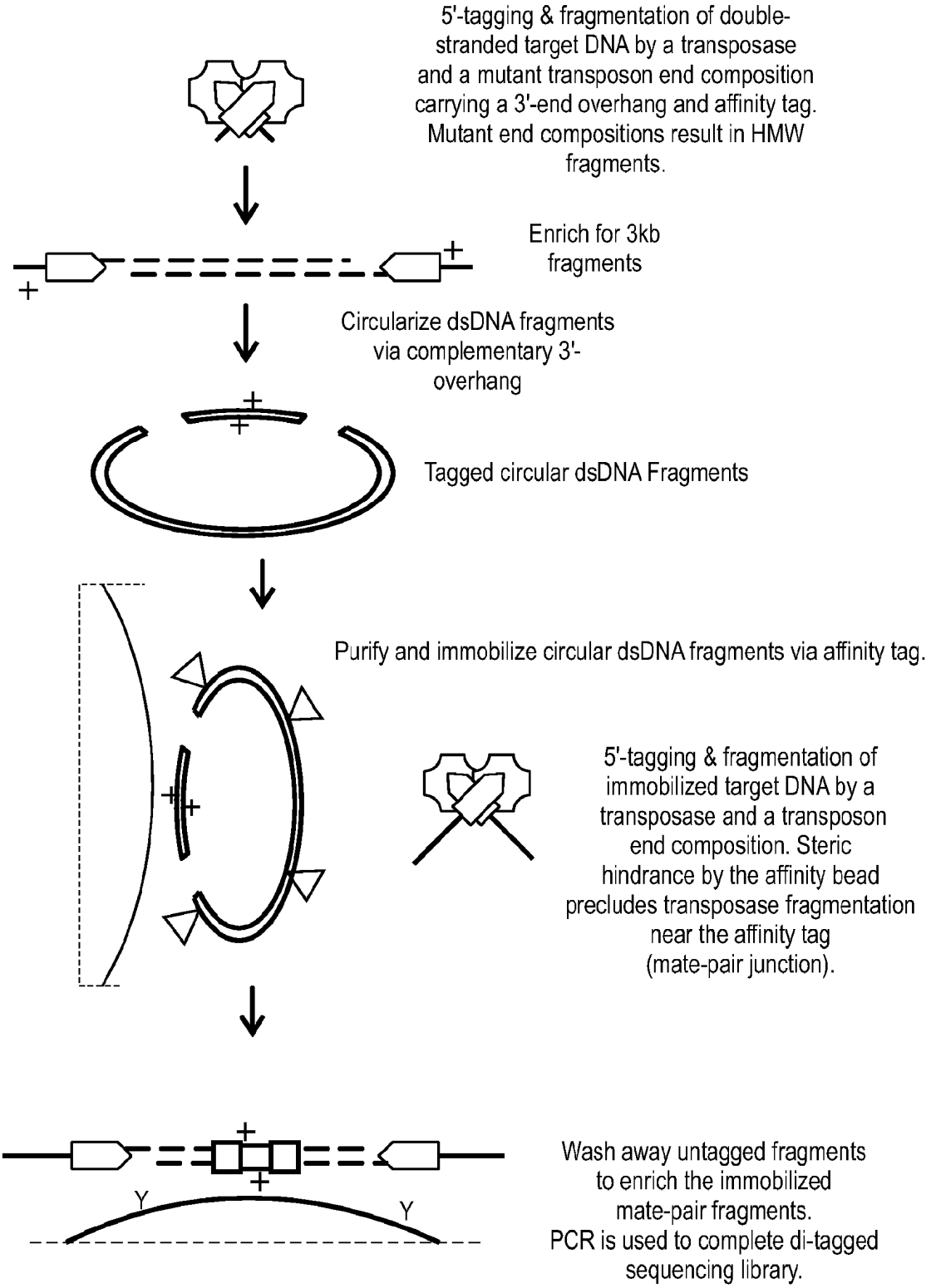


Fig. 27

IDT Exhibit 1020.4903

SUBSTITUTE SHEET (RULE 26)

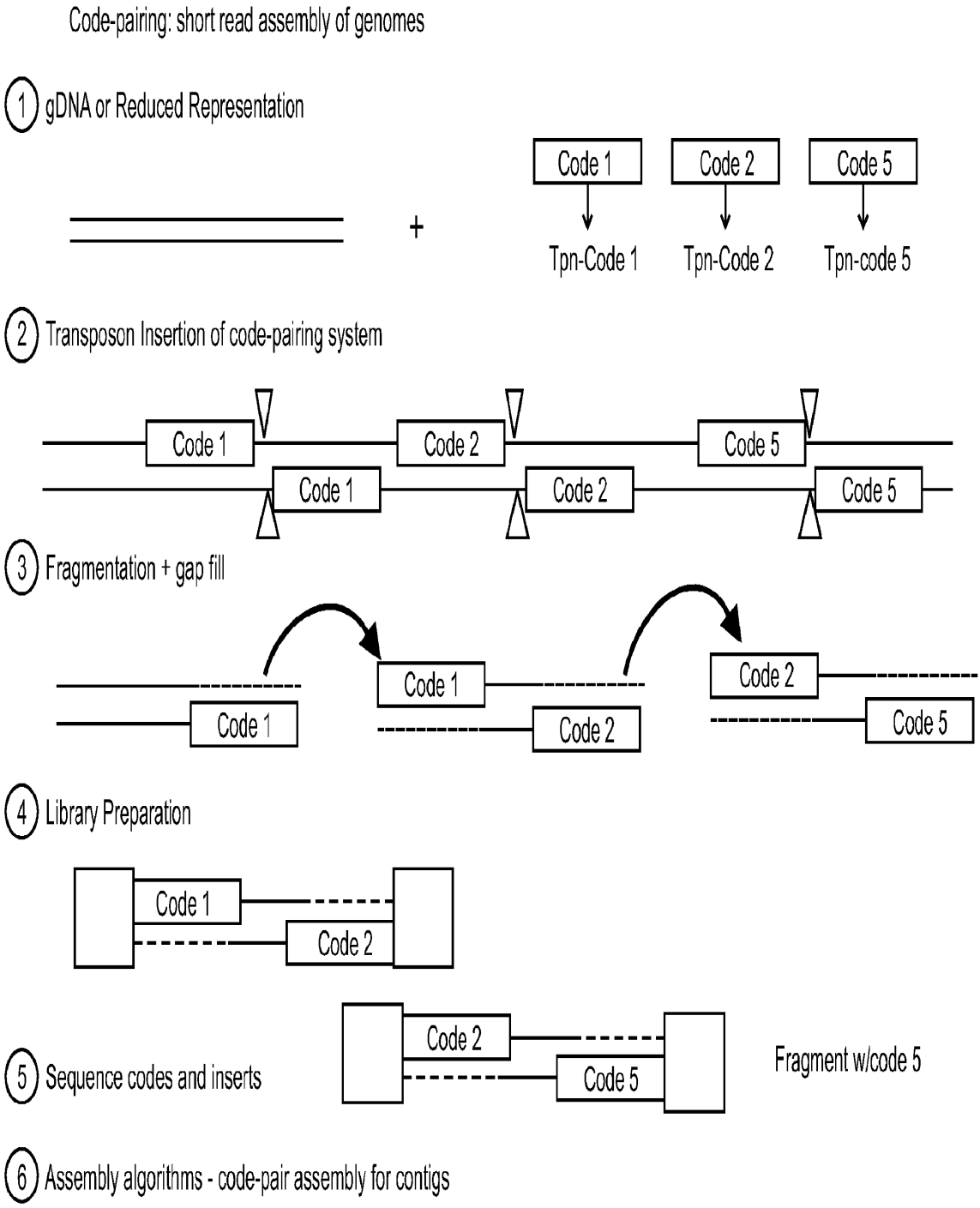


Fig. 28

IDT Exhibit 1020.4904

Fig. 29

SUBSTITUTE SHEET (RULE 26)

IDT Exhibit 1020.4905

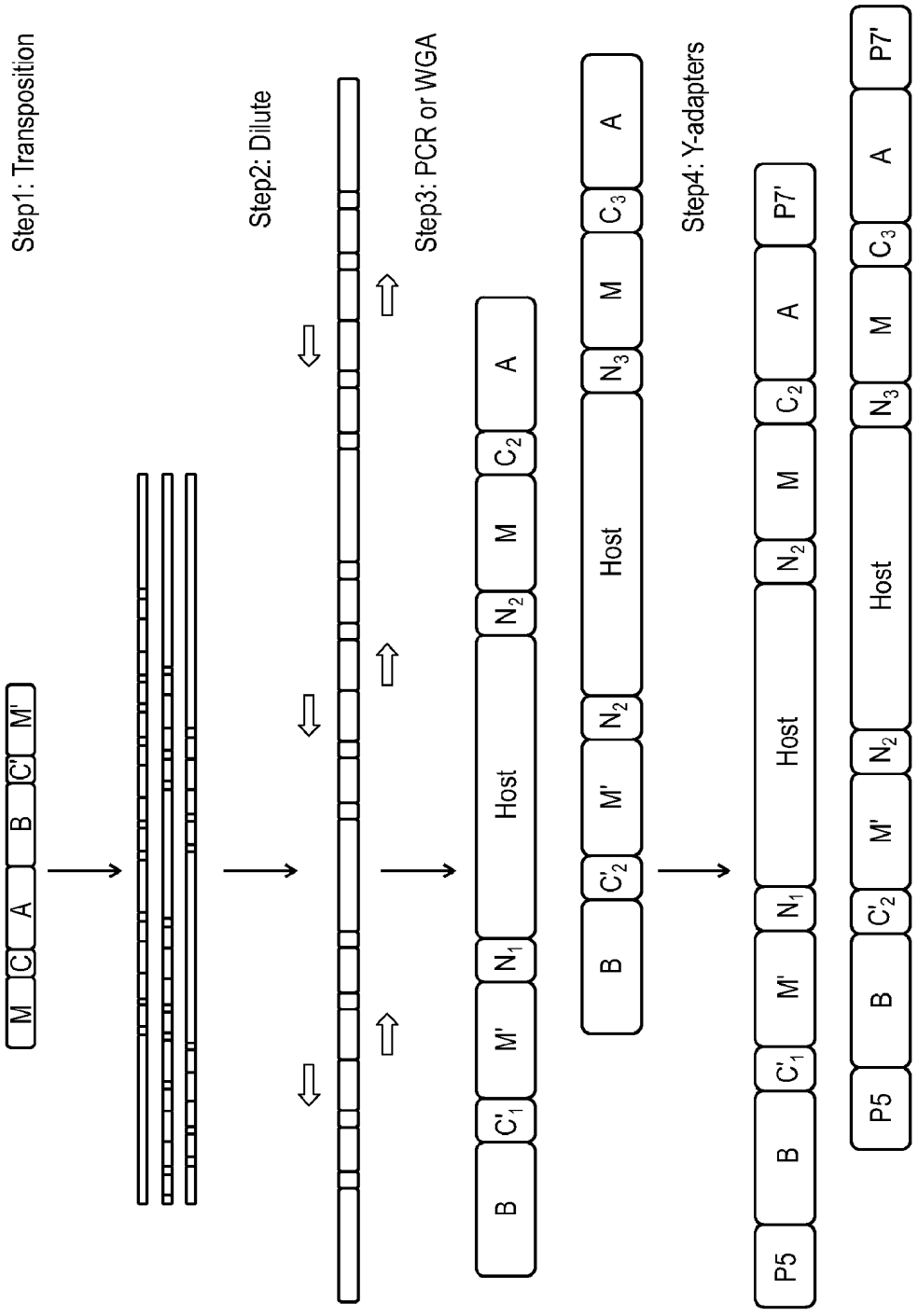

Step1: Transposition
Step2: Dilute
Step3: PCR or WGA
Step4: Y-adapters
M
C
A
B
C'
M'
Host
P5
P7'
C'_1
C'_2
N_1
N_2
N_3
C_2
C_3

Fig. 30

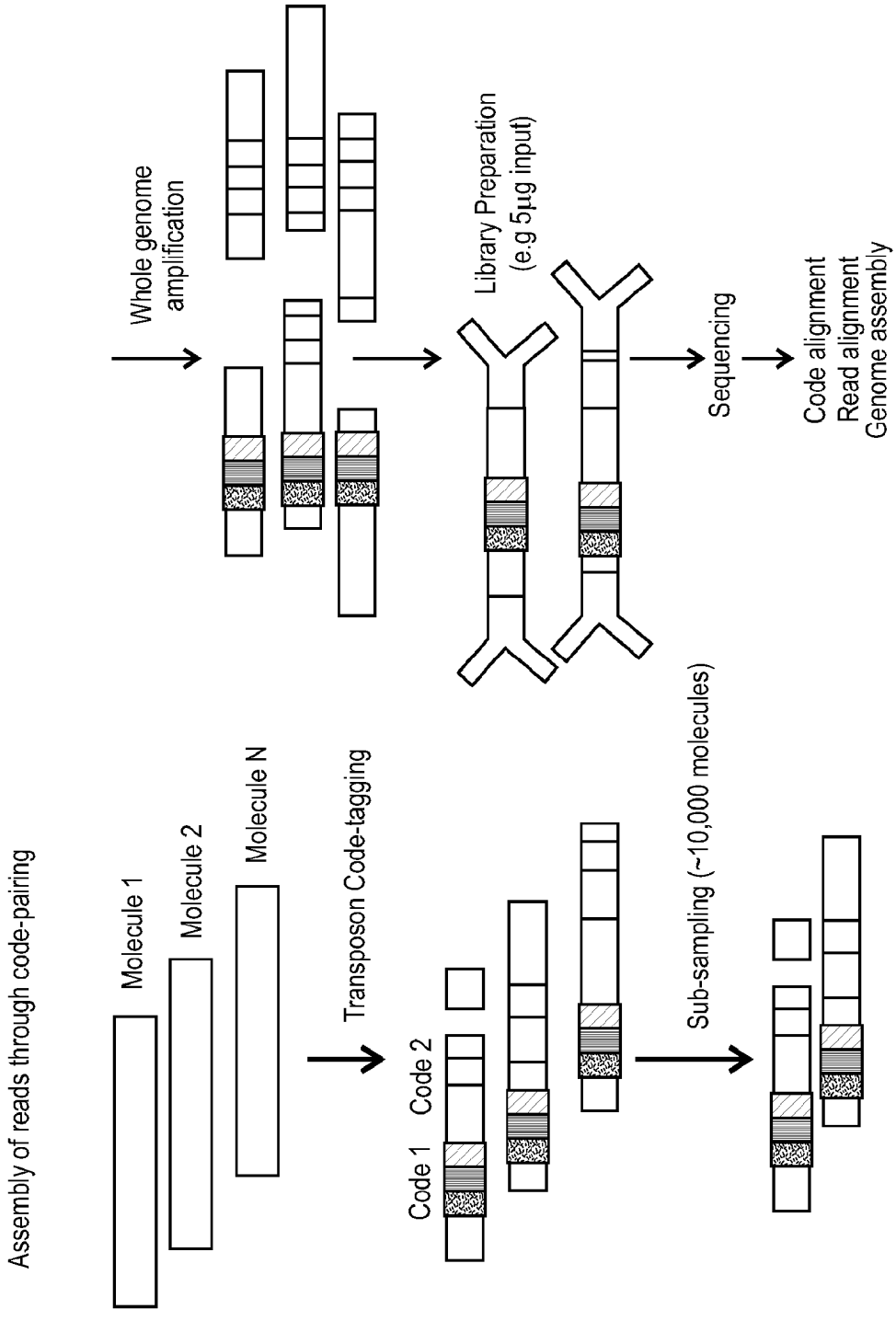

Assembly of reads through code-pairing
Molecule 1
Molecule 2
Molecule N
Transposon Code-tagging
Code 1
Code 2
Sub-sampling (~10,000 molecules)
Whole genome amplification
Library Preparation (e.g 5µg input)
Sequencing
Code alignment
Read alignment
Genome assembly

Fig. 31

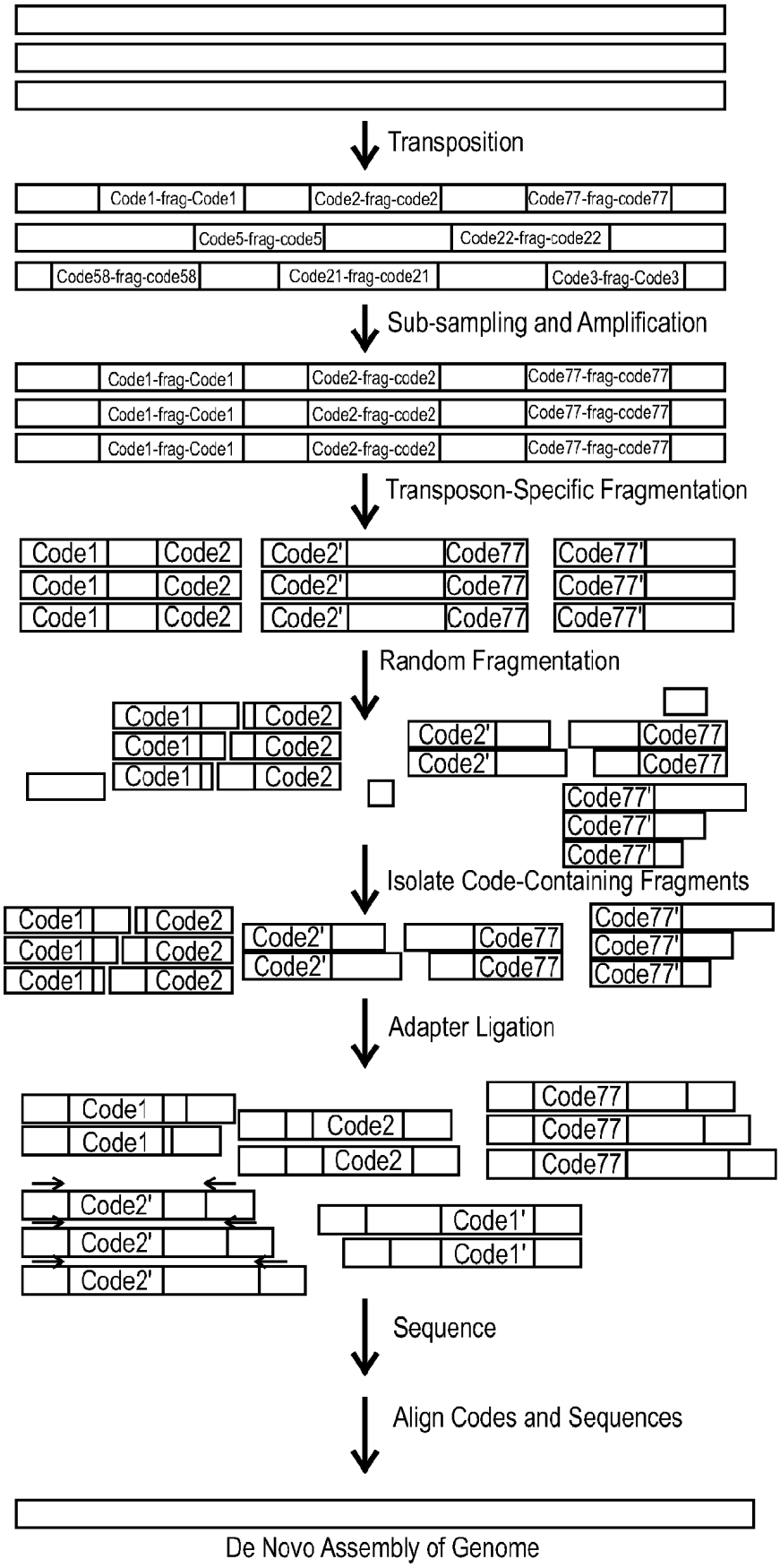

Transposition
Code1-frag-Code1
Code2-frag-code2
Code77-frag-code77
Code5-frag-code5
Code22-frag-code22
Code58-frag-code58
Code21-frag-code21
Code3-frag-Code3
Sub-sampling and Amplification
Transposon-Specific Fragmentation
Code1
Code2
Code2'
Code77
Code77'
Random Fragmentation
Isolate Code-Containing Fragments
Adapter Ligation
Code1'
Sequence
Align Codes and Sequences
De Novo Assembly of Genome

Fig. 32

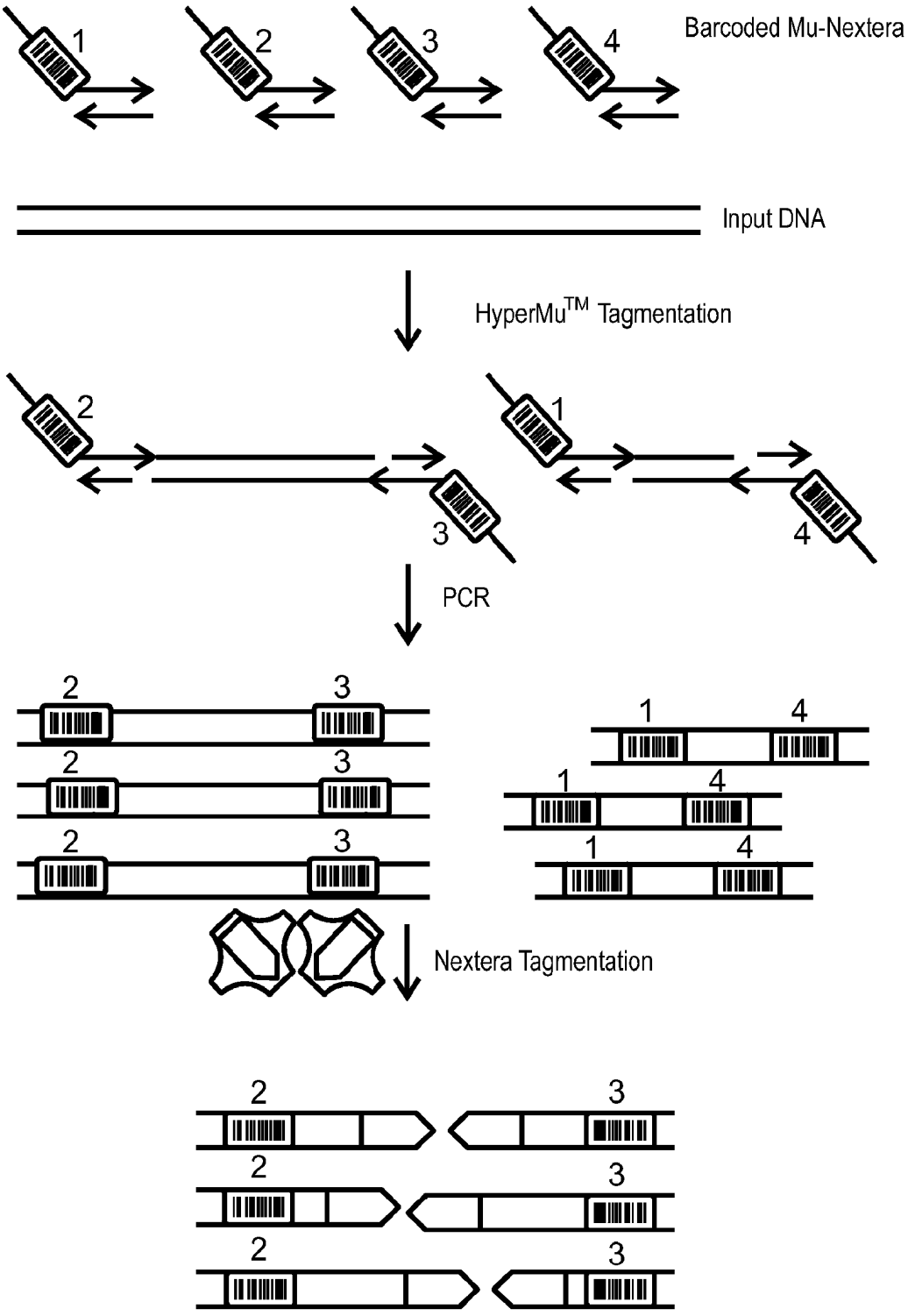

Barcoded Mu-Nextera
1
2
3
4
Input DNA
HyperMu™ Tagmentation
PCR
Nextera Tagmentation

Fig. 33

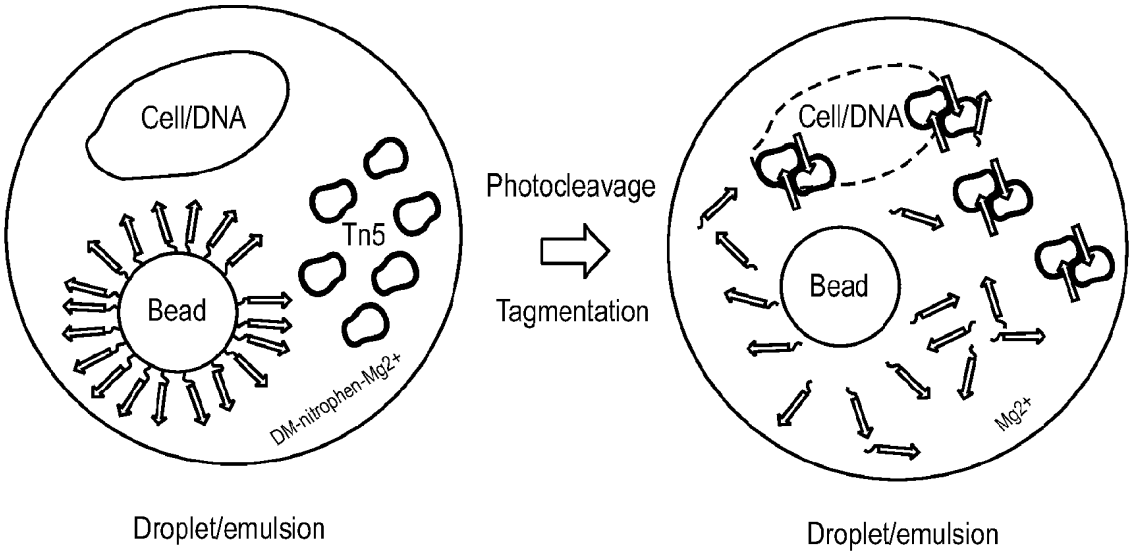

Cell/DNA
Tn5
Bead
DM-nitrophen-Mg2+
Droplet/emulsion
Photocleavage
Tagmentation
Cell/DNA
Bead
Mg2+
Droplet/emulsion

Fig. 34

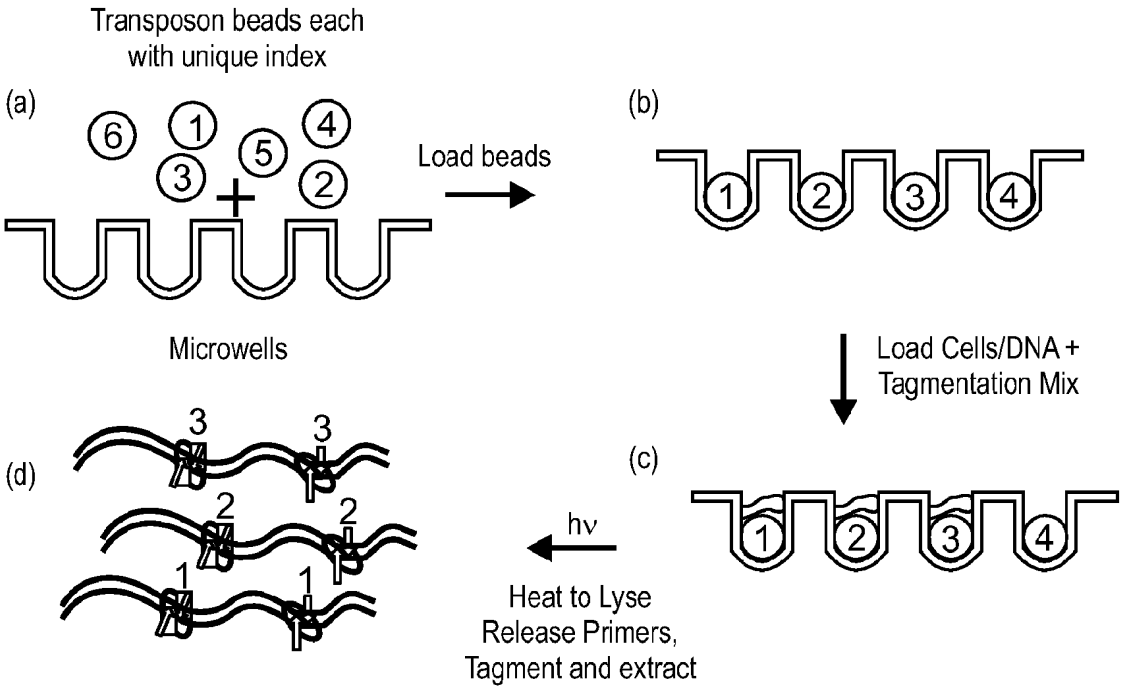

Transposon beads each with unique index
(a)
Load beads
(b)
Microwells
Load Cells/DNA + Tagmentation Mix
(c)
(d)
hν
Heat to Lyse Release Primers, Tagment and extract

Fig. 35

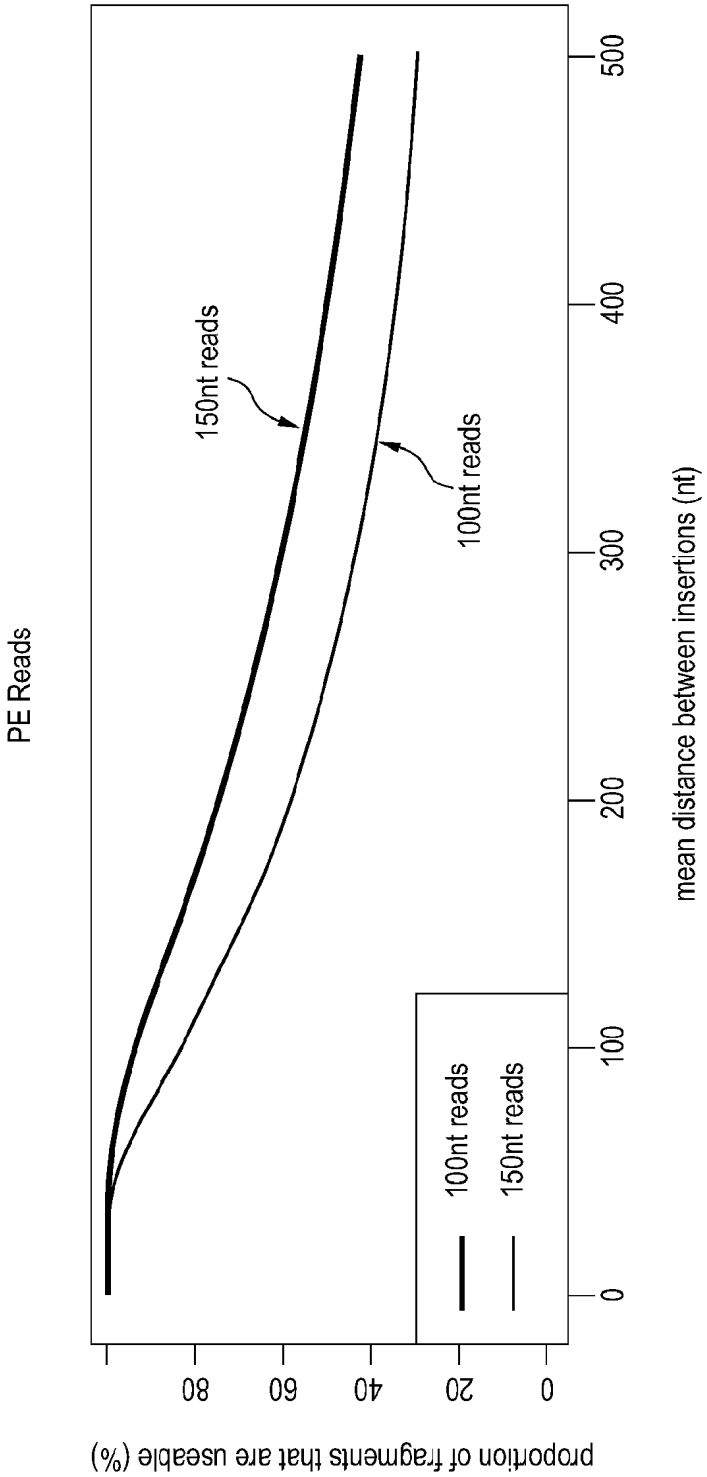


Fig. 36

IDT Exhibit 1020.4911

SUBSTITUTE SHEET (RULE 26)

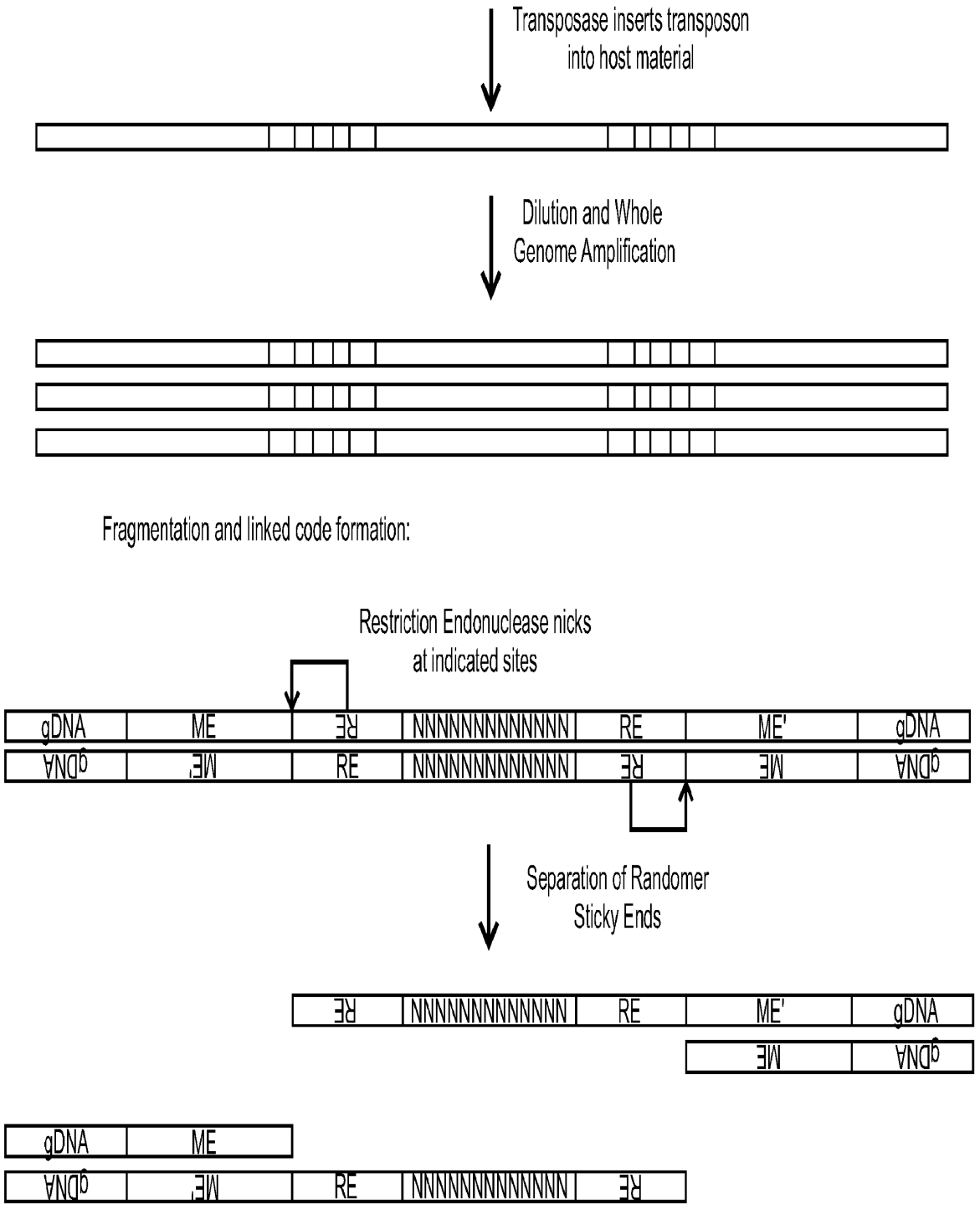


Fig. 37

IDT Exhibit 1020.4912

INTERNATIONAL SEARCH REPORT

International application No
PCT/US2011/059642

A. CLASSIFICATION OF SUBJECT MATTER
INV. C12N15/10 C12N15/90 C40B40/08 C40B50/06
ADD.

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 C40B

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, CAB Data, Sequence Search, WPI Data

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	WO 2010/048605 A1 (EPICT TECHNOLOGIES CORP [US]; JENDRISAK JEROME [US]; DAHL GARY [US]; G) 29 April 2010 (2010-04-29) the whole document -----	1-20
A	SYED FRAZ ET AL: "Optimized library preparation method for next-generation sequencing", NATURE METHODS, vol. 6, no. 10, October 2009 (2009-10), pages I-II, XP002672202, ISSN: 1548-7091 the whole document -----	1-20

-/--

[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
23 March 2012	10/04/2012

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	Hornig, Horst

1

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

IDT Exhibit 1020.4913

C(Continuation). DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
A	BIMBER BENJAMIN N ET AL: "Whole-genome characterization of human and simian immunodeficiency virus intrahost diversity by ultradeep pyrosequencing.", JOURNAL OF VIROLOGY NOV 2010 LNKD-PUBMED:20844037, vol. 84, no. 22, 15 September 2010 (2010-09-15), pages 12087-12092, XP002672203, ISSN: 1098-5514 the whole document	1-20
A	WO 2005/100585 A2 (EPICT [US]; FIANDT MICHAEL J [US]; DAHL GARY A [US]) 27 October 2005 (2005-10-27) the whole document	1-20
A	US 7 138 267 B1 (JENDRISAK JEROME J [US] ET AL) 21 November 2006 (2006-11-21) the whole document	1-20
X,P	WANG HAOYI ET AL: "Calling Cards enable multiplexed identification of the genomic targets of DNA-binding proteins.", GENOME RESEARCH MAY 2011 LNKD-PUBMED:21471402, vol. 21, no. 5, May 2011 (2011-05), pages 748-755, XP002672204, ISSN: 1549-5469 the whole document	1-20
X,P	MARINE RACHEL ET AL: "Evaluation of a transposase protocol for rapid generation of shotgun high-throughput sequencing libraries from nanogram quantities of DNA.", APPLIED AND ENVIRONMENTAL MICROBIOLOGY NOV 2011 LNKD- PUBMED:21948828, vol. 77, no. 22, 23 September 2011 (2011-09-23), pages 8071-8079, XP002672205, ISSN: 1098-5336 the whole document	1-20

IDT Exhibit 1020.4914

INTERNATIONAL SEARCH REPORT

Information on patent family members

International application No
PCT/US2011/059642

Patent document cited in search report	Publication date	Patent family member(s)	Publication date
WO 2010048605 A1	29-04-2010	CA 2750054 A1	29-04-2010
		CN 102264914 A	30-11-2011
		EP 2376517 A1	19-10-2011
		WO 2010048605 A1	29-04-2010
WO 2005100585 A2	27-10-2005	US 2006014169 A1	19-01-2006
		WO 2005100585 A2	27-10-2005
US 7138267 B1	21-11-2006	NONE	

IDT Exhibit 1020.4915

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
21 May 2015 (21.05.2015)

WIPO | PCT

(10) International Publication Number
WO 2015/073711 A1

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

(21) International Application Number:
PCT/US2014/065530

(22) International Filing Date:
13 November 2014 (13.11.2014)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/903,826 13 November 2013 (13.11.2013) US

(71) Applicant: **NUGEN TECHNOLOGIES, INC.** [US/US]; 201 Industrial Road, Suite 310, San Carlos, California 94070 (US).

(72) Inventors: **AMORESE, Douglas**; c/o Nugen Technologies, INC., 201 Industrial Road, Suite 310, San Carlos, California 94070 (US). **SCOLNICK, Jonathan**; c/o Nugen Technologies, INC., 201 Industrial Road, Suite 310, San Carlos, California 94070 (US). **SCHROEDER, Ben**; c/o Nugen Technologies, INC., 201 Industrial Road, Suite 310, San Carlos, California 94070 (US).

(74) Agents: **TUSCAN, Michael S.** et al.; Cooley LLP, 1299 Pennsylvania Avenue, N.W., Suite 700, Washington, District of Columbia 20004-2400 (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, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, 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, SA, 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, ST, 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, KM, ML, MR, NE, SN, TD, TG).

Published:

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

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

WO 2015/073711 A1

(54) **Title**: COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

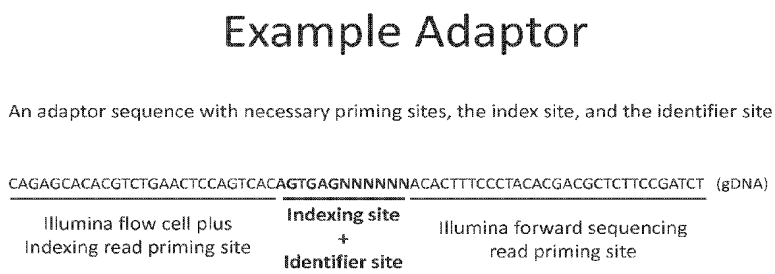


Figure 3

(57) **Abstract**: The present invention provides methods, compositions and kits for detecting duplicate sequencing reads. In some embodiments, the duplicate sequencing reads are removed. The present invention is based, in part, on compositions and methods for discerning duplicate sequencing reads from a population of sequencing reads. The detection and/or removal of duplicate sequencing reads presented herein is a novel approach to increasing the efficacy of evaluating data generated from high throughput sequence reactions, including complex multiplex sequence reactions.

IDT Exhibit 1020.4916

COMPOSITIONS AND METHODS FOR IDENTIFICATION OF A DUPLICATE SEQUENCING READ

CROSS-REFERENCE TO RELATED APPLICATIONS

[0001] This application claims the benefit of U.S. Provisional Application Serial No. 61/903,826, filed November 13, 2013, which is herein incorporated by reference in its entirety.

DESCRIPTION OF THE TEXT FILE SUBMITTED ELECTRONICALLY

[0002] The contents of the text file submitted electronically herewith are incorporated herein by reference in their entirety: A computer readable format copy of the Sequence Listing (filename: NUGN_001_01WO_SeqList_ST25.txt, date recorded: November 12, 2014, file size 3 kilobytes).

FIELD OF THE PRESENT INVENTION

[0003] The present invention relates generally to the field of high throughput sequencing reactions and the ability to discern artifacts arising through sequence duplications from nucleotide molecules that are unique molecules.

BACKGROUND OF THE PRESENT INVENTION

[0004] In RNA sequencing applications, accurate gene expression measurements may be hampered by PCR duplicate artifacts that occur during library amplification. When analyzing RNA sequencing data, when two or more identical sequences are found, it can be difficult to know if these represent unique cDNA molecules derived independently from different RNA molecules, or if they are PCR duplicates derived from a single RNA molecule. In genotyping by

sequencing, duplicate reads can be considered non-informative and may be collapsed down to a single read, thus reducing the number of sequencing reads used in final analysis. Generally, sequencing reads may be determined to be duplicates if both forward and reverse reads have identical starting positions, even though two independently generated molecules can have identical starting positions by random chance. Single primer extension based targeted re-sequencing suffers from an issue in that only one end of a sequencing read is randomly generated, while the other (reverse read) end is generated by a specific probe. This may make it difficult to determine if two reads are duplicates because they have been duplicated by PCR or because by chance they happened to start at the same position.

[0005] In expression analysis studies there may be limited value in doing paired end sequencing since the goal of the experiment is to determine amounts of transcript present as opposed to studying exon usage. In these studies, paired end sequencing adds costs while the only value is in helping distinguish PCR duplicates. The probability of two reads starting in the same position on only one end is higher than the probability of two reads having the same starting position on two ends (forward and reverse read). There is a need for improved methods that allow for low-cost, high throughput sequencing of regions of interest, genotyping or simple detection of RNA transcripts without inherent instrument inefficiencies that drive up sequencing costs due to the generation of unusable or non-desired data reads. The invention described herein fulfills this need. Here, we describe an adaptor approach that allows for the identification of true PCR duplicates and their removal.

[0006] The methods of the present invention provide novel methods for identifying true duplicate reads during sequencing, such as to improve data analysis of sequencing data, and other related advantages.

SUMMARY OF THE PRESENT INVENTION

[0007] The present invention is based, in part, on compositions and methods for discerning duplicate sequencing reads from a population of sequencing reads. The detection and/or removal of duplicate sequencing reads presented herein is a novel approach to increasing the efficacy of evaluating data generated from high throughput sequence reactions, including complex multiplex sequence reactions.

[0008] Accordingly, the present invention provides a method of detecting a duplicate sequencing read from a population of sample sequencing reads, the method comprising ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, wherein the adaptor comprises an indexing primer binding site, an indexing site, an identifier site, and a target sequence primer binding site. The ligated adaptor-nucleic acid fragment products can be amplified, thus generating a population of sequencing reads from the amplified adaptor-nucleic acid ligation products. The sequencing reads with a duplicate identifier site and target sequence can then be detected from the population of sequencing reads. The methods can further include the removal of the sequencing reads with the duplicate identifier site and target sequence from the population of sequence reads.

[0009] In some embodiments, the identifier site is sequenced with the indexing site or the target sequence. In further embodiments, the identifier site is sequenced separately from the indexing site or the target sequence.

[0010] In some embodiments, the adaptor comprises from 5' to 3' the indexing primer binding site; the indexing site; the identifier site; and the target sequence primer binding site. In further embodiments, the adaptor comprises from 5' to 3' the indexing primer binding site; the indexing site; the target sequence primer binding site; and the identifier site.

[0011] In some embodiments, the plurality of nucleic acid fragments is generated from more than one sample. In some embodiments, the nucleic acid fragments from each sample have the same indexing site. In some embodiments, the sequencing reads are separated based on the indexing site. In yet other embodiments, the separation of sequencing reads is performed prior to detecting sequence reads with a duplicate identifier site and target sequence.

[0012] In some embodiments, the nucleic acid fragments are DNA fragments, RNA fragments, or DNA/RNA fragments. In further embodiments, the nucleic acid fragments are genomic DNA fragments or cDNA fragments.

[0013] In some embodiments, the indexing site is between 2 and 8 nucleotides in length. In further embodiments, the indexing site is about 6 nucleotides in length. In some embodiments, the identifier site is between 1 and 8 nucleotides in length. In further embodiments, the identifier site is about 8 nucleotides in length.

[0014] In some embodiments, the indexing primer binding site is a universal indexing primer binding site; and in some embodiments, the target sequence primer binding site is a universal target sequence primer binding site.

[0015] The present invention also encompasses embodiments that include a kit comprising a plurality of adaptors, wherein each adaptor comprise an indexing primer binding site; an indexing site, and identifier site, and a target sequencing primer binding site.

BRIEF DESCRIPTION OF THE DRAWINGS

[0016] A better understanding of the novel features of the invention 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:

[0017] Figure 1 depicts a schematic of generating sequencing reads of a library including where an indexing primer and a target sequence primer anneal.

[0018] Figure 2A depicts a mechanism of the Single Primer Enrichment Technology (SPET) and how an identifier site is carried over into the final library and how sequencing through the indexing site into the identifier site provides data on which nucleic acid molecule is being identified.

[0019] Figure 2B is a continuation of Figure 2A.

[0020] Figure 3 offers a detailed view of the sequence of an example adaptor, among many envisioned embodiments, and the position of the indexing and identifier sites (SEQ ID NO: 1). "N" refers to any nucleic acid.

[0021] Figure 4 depicts a schematic of two separate sequence libraries, indicating where the indexing primers and the target primers anneal in a traditional library as compared to a library using an identifier site.

[0022] Figure 5 depicts a data table demonstrating the accuracy of resolving true duplicates versus apparent or perceived duplicates using an identifier site in the adaptors.

DETAILED DESCRIPTION OF THE PRESENT INVENTION

[0023] The present invention is based, in part, on compositions and methods for discerning duplicate sequencing reads from a population of sequencing reads. The present invention encompasses methods of detecting sequences duplicated in sequencing applications, and further removal of the duplicated sequence reads. The present invention further encompasses kits

comprising components that would allow for customized applications of the method of detecting and removing duplicated sequence reads in high throughput sequencing reactions. The compositions and methods can be used with various applications for genetic sample analysis, such as RNA sequence analysis, copy number variation analysis, methylation sequencing analysis, genotyping and whole genome amplification.

[0024] 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 present invention. On the contrary, the present disclosure is intended to encompass alternatives, modifications and equivalents, which may be included in the spirit and scope of the present invention.

[0025] 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.

[0026] In some embodiments, the methods disclosed herein is for detecting from a population of sequencing reads a sequencing read, such as a duplicate sequencing read with a duplicate identifier site and target sequence. A duplicate sequencing read can be a sequencing read with

the same identifier site and target sequence as another sequencing read in the population of sequencing reads.

Adaptor

[0027] The present invention provides compositions of adaptors and methods comprising use of an adaptor. An adapter refers to an oligonucleotide sequence, the ligation 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. The target polynucleotide molecules may be fragmented or not prior to the addition of adaptors. In some embodiments, a method disclosed herein comprises ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples.

[0028] Various adaptor designs are envisioned which are suitable for generation of amplification-ready products of target sequence regions/strands of interest. For example, the two strands of the adaptor may be self-complementary, non-complementary or partially complementary. In some embodiments, the adaptor can comprise an indexing primer binding site, an indexing site, an identifier site, and a target sequence primer binding site.

[0029] An indexing primer binding site is a nucleotide sequence for binding a primer for an indexing site. An indexing site is a nucleic acid sequence that acts as an index for multiple polynucleotide samples, thus allowing for the samples to be pooled together into a single sequencing run, which is known as multiplexing. In some embodiments, the indexing site is at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides in length. In some embodiments, the indexing site is between 2 and 8 nucleotides in length. In some embodiments, the indexing site is about 6 nucleotides in length.

[0030] An identifier site is a nucleic acid sequence that comprises random bases and is used to identify duplicate sequencing reads. In some embodiments, the identifier site is at least 2, 3, 4, 5, 6, 7, 8, 9, or 10 nucleotides in length. In some embodiments, the identifier site is between 1 and 8 nucleotides in length. In some embodiments, the identifier site is about 8 nucleotides in length. This identifier site can be designed as a set of sequences, or it can be semi-random, or it can be completely random. In addition, this identifier site can be a fixed length, or it can be a variable length. In some embodiments, the identifier sites in a plurality of adaptors are of a fixed length. For example, the identifier sites can all be eight random bases. In another embodiment, the identifier sites in a plurality of adaptors are of a variable length. For example, the identifier sites can range in size from 1 to 8 bases. In yet another embodiment, the identifier sites can be of a defined set of defined sequence. For example, the identifier site of a plurality of adaptors can be one of 96 defined six-base nucleotide sequences.

[0031] A target sequence primer binding site nucleotide sequence for binding a primer for a target sequence. The primer can be used to amplify the target sequence (e.g., a nucleic acid fragment from a sample). Accordingly, in some embodiments, the adaptor comprises an indexing primer binding site and a target sequence primer binding site.

[0032] A primer is a polynucleotide chain, typically less than 200 residues long, most typically between 15 and 100 nucleotides long, but can encompass longer polynucleotide chains. A primer targeting the primer binding sites are typically designed to hybridize to single-stranded nucleic acid strands. In some embodiments, the primers targeting the primer binding sites are designed to hybridize to single-stranded DNA targets. In the case where the sample comprises genomic DNA or other double-stranded DNA, the sample can be first denatured to render the target single stranded and enable hybridization of the primers to the desired sequence regions of

interest. In these embodiments, the methods and compositions described herein can allow for region-specific enrichment and amplification of sequence regions of interest. In some embodiments, the other double-stranded DNA can be double-stranded cDNA generated by first and second strand synthesis of one or more target RNAs.

[0033] In other embodiments, the primers targeting the primer binding sites are designed to hybridize to double-stranded nucleic acid targets, without denaturation of the double stranded nucleic acids. In other embodiments, the primers targeting the primer binding sites are designed to hybridize to a double-stranded DNA target, without denaturation of the dsDNA. In these embodiments, the primers targeting the selected sequence regions of interest are designed to form a triple helix (triplex) at the selected sequence regions of interest. The hybridization of the primers to the double-stranded DNA sequence regions of interest can be carried out without prior denaturation of the double stranded nucleic acid sample. In such embodiments, the methods and compositions described herein can allow for region-specific enrichment as well as strand-specific enrichment and amplification of sequence regions of interest. This method can be useful for generation of copies of strand specific sequence regions of interest from complex nucleic acid without the need to denature the dsDNA input DNA, thus enabling enrichment and analysis of multiplicity of sequence regions of interest in the native complex nucleic acid sample. The method can find use for studies and analyses carried out in situ, enable studies and analysis of complex genomic DNA in single cells or collection of very small well defined cell population, as well as permit the analysis of complex genomic DNA without disruption of chromatin structures.

[0034] The primers of the invention are generally oligonucleotides that are employed in an extension reaction by a polymerase along a polynucleotide template, such as for amplification of a target sequence (e.g., in PCR). The oligonucleotide primer can be 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. In some embodiments, 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 primer binding site.

[0035] In some embodiments, the primer binding site is a binding site for a universal primer. A universal primer is a primer that can be used for amplifying a number of different sequences. In some embodiments, a universal primer is used to amplify different libraries. In some embodiments, an indexing primer binding site is a binding site for a universal indexing primer (i.e., the indexing primer binding site is a universal indexing primer binding site). In some embodiments, the adaptors used for ligating a plurality of nucleic acid fragments have a universal indexing primer binding site. In some embodiments, the universal indexing primer can be used to amplify and/or sequence a number of different indexing sites.

[0036] In some embodiments, a target sequence primer binding site is a binding site for a universal target sequence primer (i.e., the target sequence primer binding site is a universal target sequence primer binding site). In some embodiments, the adaptors used for ligating a plurality of nucleic acid fragments have a universal target sequence primer binding site. In some embodiments, the universal target sequence primer can be used to amplify and/or sequence a number of different target sequences.

[0037] In some embodiments, the adaptor comprises an identifier site 3' to an indexing site. In some embodiments, the adaptor comprises an identifier site 5' to an indexing site. In some embodiments, the adaptor comprises from 5' to 3' an indexing primer binding site, indexing site, identifier site, and target sequence primer binding site. In other embodiments, the adaptor comprises from 5' to 3' an indexing primer binding site, indexing site, identifier site, and target

sequence primer binding site. In yet other embodiments, the adaptor comprises from 5' to 3' an indexing primer binding site, identifier site, indexing site, and target sequence primer binding site.

Samples

[0038] In some embodiments, an adaptor is ligated to a nucleic acid fragment (e.g., the 5' end of the nucleic acid fragment). The nucleic acid fragment can be from a plurality of nucleic acid fragments from one or more samples. The nucleic acid fragment can be RNA, DNA, or complex DNA, for example genomic DNA and PNA, in which case one might use a modified nucleic acid. The nucleic acid fragment may also be cDNA. The cDNA can be generated from RNA, e.g., mRNA.

[0039] The sample can be a biological sample. For example, the sample can be an animal, plant, bacterial, algal, or viral sample. In some embodiments, the sample is a human, rat, or mouse sample. The sample can be from a mixture of genomes of different species such as host-pathogen, bacterial populations and the like. The sample can be cDNA made from a mixture of genomes of different species. In some embodiments, the sample can be from a synthetic source. The sample can be mitochondrial DNA. The sample can be cell-free DNA. The cell-free DNA can be obtained from sources such as a serum or a plasma sample. The sample can comprise one or more chromosomes. For example, if the sample is from a human, the sample can comprise one or more of chromosome 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, X, or Y. In some embodiments, the sample comprises a linear or circular genome. The sample can be plasmid DNA, cosmid DNA, bacterial artificial chromosome (BAC), or yeast artificial chromosome (YAC). The sample can be from more than one individual or organism. The sample

can be double-stranded or single-stranded. The sample can be part of chromatin. The sample can be associated with histones.

[0040] In some embodiments, adaptors are ligated to a plurality of nucleic acid fragments from more than one sample, such as 2, 3, 4, 5, or more samples. In some embodiments, the nucleic acid fragments from each sample have the same indexing site. In some embodiments, a plurality of nucleic acid fragments is generated from a first sample and a second sample and adaptors are ligated to each nucleic acid fragment, in which adaptors ligated to each nucleic acid fragment from the first sample have the same first indexing site and adaptors ligated to nucleic acid fragments from the second sample has the same second indexing site. In some embodiments, the nucleic acid fragments or data associated with the nucleic acid fragments (e.g., sequencing reads) are separated based on the indexing site.

[0041] In some embodiments, a population of nucleic acid fragments generated from a sample is of one or more specific size range(s). In some embodiments, the fragments have an average length from about 10 to about 10,000 nucleotides. In some embodiments, the fragments have an average length from about 50 to about 2,000 nucleotides. In some embodiments, the fragments have an average length from about 100-2,500, 10-1,000, 10-800, 10-500, 50-500, 50-250, or 50-150 nucleotides. In some embodiments, the fragments have an average length less than 10,000 nucleotide, such as less than 5,000 nucleotides, less than 2,500 nucleotides, less than 2,500 nucleotides, less than 1,000 nucleotides, less than 500 nucleotides, such as less than 400 nucleotides, less than 300 nucleotides, less than 200 nucleotides, or less than 150 nucleotides.

[0042] In some embodiments, fragmentation of the nucleic acids can be achieved through methods known in the art. Fragmentation can be achieved through physical fragmentation methods and/or enzymatic fragmentation methods. Physical fragmentation methods can include

nebulization, sonication, and/or hydrodynamic shearing. In some embodiments, the fragmentation can be accomplished mechanically comprising subjecting the nucleic acids in the input sample to acoustic sonication. In some embodiments, the fragmentation comprises treating the nucleic acids in the input sample with one or more enzymes under conditions suitable for the one or more enzymes to generate double-stranded nucleic acid breaks. Examples of enzymes useful in the generation of nucleic acid or polynucleotide fragments include sequence specific and non-sequence specific nucleases. Non-limiting examples of nucleases include DNase I, Fragmentase, restriction endonucleases, variants thereof, and combinations thereof. Reagents for carrying out enzymatic fragmentation reactions are commercially available (e.g., from New England Biolabs). For example, digestion with DNase I can induce random double-stranded breaks in DNA in the absence of Mg^{++} and in the presence of Mn^{++}. In some embodiments, fragmentation comprises treating the nucleic acids in the input sample with one or more restriction endonucleases. Fragmentation can produce fragments having 5′ overhangs, 3′ overhangs, blunt ends, or a combination thereof. In some embodiments, such as when fragmentation comprises the use of one or more restriction endonucleases, cleavage of sample polynucleotides leaves overhangs having a predictable sequence. In some embodiments, the method includes the step of size selecting the fragments via standard methods known in the art such as column purification or isolation from an agarose gel.

[0043] In some embodiments, fragmentation of the nucleic acids is followed by end repair of the nucleic acid fragments. End repair can include the generation of blunt ends, non-blunt ends (i.e sticky or cohesive ends), or single base overhangs such as the addition of a single dA nucleotide to the 3′-end of the nucleic acid fragments, by a polymerase lacking 3′-exonuclease activity. End repair can be performed using any number of enzymes and/or methods known in the art

including, but not limited to, commercially available kits such as the Encore™ Ultra Low Input NGS Library System I. In some embodiments, end repair can be performed on double stranded DNA fragments to produce blunt ends wherein the double stranded DNA fragments contain 5′ phosphates and 3′ hydroxyls. In some embodiments, the double-stranded DNA fragments can be blunt-end polished (or "end-repaired") to produce DNA fragments having blunt ends, prior to being joined to adapters. Generation of the blunt ends on the double stranded fragments can be generated by the use of a single strand specific DNA exonuclease such as for example exonuclease 1, exonuclease 7 or a combination thereof to degrade overhanging single stranded ends of the double stranded products. Alternatively, the double stranded DNA fragments can be blunt ended by the use of a single stranded specific DNA endonuclease, for example, but not limited to, mung bean endonuclease or S1 endonuclease. Alternatively, the double stranded products can be blunt ended by the use of a polymerase that comprises single stranded exonuclease activity such as for example T4 DNA polymerase, or any other polymerase comprising single stranded exonuclease activity or a combination thereof to degrade the overhanging single stranded ends of the double stranded products. In some cases, the polymerase comprising single stranded exonuclease activity can be incubated in a reaction mixture that does or does not comprise one or more dNTPs. In other cases, a combination of single stranded nucleic acid specific exonucleases and one or more polymerases can be used to blunt end the double stranded fragments generated by fragmenting the sample comprising nucleic acids. In still other cases, the nucleic acid fragments can be made blunt ended by filling in the overhanging single stranded ends of the double stranded fragments. For example, the fragments may be incubated with a polymerase such as T4 DNA polymerase or Klenow polymerase or a combination thereof in the presence of one or more dNTPs to fill in the single stranded portions

of the double stranded fragments. Alternatively, the double stranded DNA fragments can be made blunt by a combination of a single stranded overhang degradation reaction using exonucleases and/or polymerases, and a fill-in reaction using one or more polymerases in the presence of one or more dNTPs.

[0044] U.S. Patent Publication Nos. 2013-0231253 A1 and 2014-0274729 A1 further describe methods of generating nucleic acid fragments, methods of modifying the fragments and analysis of the fragments, and are incorporated by reference in their entirety herein.

Ligation of Adaptors

[0045] Ligation of adaptors at the desired end of the sequence regions of interest (e.g., at the 5' or 3' end of a nucleic acid fragment generated form a sample) is suitable for carrying out the methods of the invention. Various ligation modalities are envisioned, dependent on the choice of nucleic acid, nucleic acid modifying enzymes and the resulting ligatable end of the nucleic acid. For example, when a blunt end product comprising the target region/sequence of interest is generated, blunt end ligation can be suitable. Alternatively, where the cleavage is carried out using a restriction enzyme of known sequence specificity, leading to the generation of cleavage sites with known sequence overhangs, suitable ends of the adaptors can be designed to enable hybridization of the adaptor to the cleavage site of the sequence region of interest and subsequent ligation. Ligation also refers to any joining of two nucleic acid molecules that results in a single nucleic acid sequence that can be further modified to obtain the sequence of the nucleic acids in question. Reagents and methods for efficient and rapid ligation of adaptors are commercially available, and are known in the art.

[0046] In some embodiments, the 5′ and/or 3′ end nucleotide sequences of fragmented nucleic acids are not modified or end-repaired prior to ligation with the adapter oligonucleotides of the present invention. For example, fragmentation by a restriction endonuclease can be used to leave

a predictable overhang, followed by ligation with one or more adapter oligonucleotides comprising an overhang complementary to the predictable overhang on a nucleic acid fragment. In another example, cleavage by an enzyme that leaves a predictable blunt end can be followed by ligation of blunt-ended nucleic acid fragments to adapter oligonucleotides comprising a blunt end. In some embodiments, end repair can be followed by an addition of 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20 or more nucleotides, such as one or more adenine, one or more thymine, one or more guanine, or one or more cytosine, to produce an overhang. Nucleic acid fragments having an overhang can be joined to one or more adapter oligonucleotides having a complementary overhang, such as in a ligation reaction. For example, a single adenine can be added to the 3′ ends of end-repaired DNA fragments using a template independent polymerase, followed by ligation to one or more adapters each having a thymine at a 3′ end. In some embodiments, adapter oligonucleotides can be joined to blunt end double-stranded nucleic acid fragments which have been modified by extension of the 3′ end with one or more nucleotides followed by 5′ phosphorylation. In some cases, extension of the 3′ end can be performed with a polymerase such as for example Klenow polymerase or any of the suitable polymerases provided herein, or by use of a terminal deoxynucleotide transferase, in the presence of one or more dNTPs in a suitable buffer containing magnesium. In some embodiments, nucleic acid fragments having blunt ends can be joined to one or more adapters comprising a blunt end. Phosphorylation of 5′ ends of nucleic acid fragments can be performed for example with T4 polynucleotide kinase in a suitable buffer containing ATP and magnesium. The fragmented nucleic acid molecules may optionally be treated to dephosphorylate 5′ ends or 3′ ends, for example, by using enzymes known in the art, such as phosphatases.

[0047] In some embodiments, appending the adaptor to the nucleic acid fragments generated by methods described herein can be achieved using a ligation reaction or a priming reaction. In some embodiments, appendage of an adaptor to the nucleic acid fragments comprises ligation. In some embodiments, ligation of the adaptor to the nucleic acid fragments can be following end repair of the nucleic acid fragments. In another embodiment, the ligation of the adaptor to the nucleic acid fragments can be following generation of the nucleic acid fragments without end repair of the nucleic acid fragments. The adaptor can be any type of adaptor known in the art including, but not limited to, conventional duplex or double stranded adaptors in which the adaptor comprises two complementary strands. In some embodiments, the adaptor can be a double stranded DNA adaptor. In some embodiments, the adaptor can be an oligonucleotide of known sequence and, thus, allow generation and/or use of sequence specific primers for amplification and/or sequencing of any polynucleotides to which the adaptor is appended or attached. In some embodiments, the adaptor can be a conventional duplex adaptor, wherein the adaptor comprises sequence well known in the art. In a some embodiments, the adaptor can be appended to the nucleic acid fragments generated by the methods described herein in multiple orientations. In some embodiment, the methods described herein can involve the use of a duplex adaptor comprising double stranded DNA of known sequence that is blunt ended and can bind to the double stranded nucleic acid fragments generated by the methods described herein in one of two orientations. In some embodiments, the adaptor can be ligated to each of the nucleic acid fragments such that each of the nucleic acid fragments comprises the same adaptor. In other words, each of the nucleic acid fragments comprises a common adaptor. In another embodiment, an adaptor can be appended or ligated to a library of nucleic acid fragments generated by the methods described herein such that each nucleic acid fragment in the library of nucleic acid

fragments comprises the adaptor ligated to one or both ends. In another embodiment, more than one adaptor can be appended or ligated to a library of nucleic acid fragments generated by the methods described herein. The multiple adaptors may occur adjacent to one another, spaced intermittently, or at opposite ends of the nucleic acid fragments. In some embodiments, the adaptor can be ligated or appended to the 5′ and/or 3′ ends of the nucleic acid fragments generated by the methods described herein. The adaptor can comprise two strands wherein each strand comprises a free 3′ hydroxyl group but neither strand comprises a free 5′ phosphate. In some embodiments, the free 3′ hydroxyl group on each strand of the adaptor can be ligated to a free 5′ phosphate present on either end of the nucleic acid fragments of the present invention. In this embodiment, the adaptor comprises a ligation strand and a non-ligation strand whereby the ligation strand can be ligated to the 5′ phosphate on either end of the nucleic acid fragment while a nick or gap can be present between the non-ligation strand of the adaptor and the 3′ hydroxyl on either end of the nucleic acid fragment. In some embodiments, the nick or gap can be filled in by performing a gap repair reaction. In some embodiments, the gap repair can be performed with a DNA dependent DNA polymerase with strand displacement activity. In some embodiments, the gap repair can be performed using a DNA-dependent DNA polymerase with weak or no strand displacement activity. In some embodiments, the ligation strand of the adaptor can serve as the template for the gap repair or fill-in reaction. The gap repair or fill-in reaction may comprise an extension reaction wherein the ligation strand of the adaptor serves as a template and leads to the generation of nucleic acid fragments with complementary termini or ends. In some embodiments, the gap repair can be performed using Taq DNA polymerase. In some embodiments, the ligation of the first adaptor to the nucleic acid fragments generated by the

methods described herein may not be followed by gap repair. The nucleic acid fragments may comprise the adaptor sequence ligated only at the 5′ end of each strand.

[0048] Ligation and, optionally gap repair, of the adaptor to the nucleic acid fragments generates an adaptor-nucleic acid fragment complex. In some embodiments, the adaptor-nucleic acid fragment complex can be denatured. Denaturation can be achieved using any of the methods known in the art including, but not limited to, physical, thermal, and/or chemical denaturation. In some embodiments, denaturation can be achieved using thermal or heat denaturation. In some embodiments, denaturation of the adaptor-nucleic acid fragment complex generates single stranded nucleic acid fragments comprising the adaptor sequence at only the 5′end of the nucleic acid fragments. In another embodiment, denaturation of the first adaptor-nucleic acid fragment complex generates single stranded nucleic acid fragments comprising adaptor sequence at both the 5′end and 3′end of the nucleic acid fragments.

Methods of Amplification

[0049] The methods, compositions and kits described herein can be useful to generate amplification-ready products directly from a nucleic acid source for downstream applications such as next generation sequencing, as well as generation of libraries with enriched population of sequence regions of interest. In some embodiments, the adapter-nucleic fragment ligated products, e.g., from the ligation of an adaptor to a 5’ end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, is amplified.

[0050] Methods of amplification are well known in the art. In some embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR). In other embodiments the amplification method is linear. In other embodiments the amplification method is isothermal. In

some embodiments, the amplification is exponential, e.g. in the enzymatic amplification of specific double stranded sequences of DNA by a polymerase chain reaction (PCR).

[0051] Suitable amplification reactions can be exponential or isothermal and can include any DNA amplification reaction, including but not limited to polymerase chain reaction (PCR), strand displacement amplification (SDA), linear amplification, multiple displacement amplification (MDA), rolling circle amplification (RCA), single primer isothermal amplification (SPIA, see e.g. U.S. Pat. No. 6,251,639), Ribo-SPIA, or a combination thereof. In some cases, the amplification methods for providing the template nucleic acid may be performed under limiting conditions such that only a few rounds of amplification (e.g., 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 etc.), such as for example as is commonly done for cDNA generation. The number of rounds of amplification can be about 1-30, 1-20, 1-15, 1-10, 5-30, 10-30, 15-30, 20-30, 10-30, 15-30, 20-30, or 25-30.

[0052] PCR is an in vitro amplification procedure based on repeated cycles of denaturation, oligonucleotide primer annealing, and primer extension by thermophilic template dependent polynucleotide polymerase, resulting in the exponential increase in copies of the desired sequence of the polynucleotide analyte flanked by the primers. The two different PCR primers, which anneal to opposite strands of the DNA, are positioned so that the polymerase catalyzed extension product of one primer can serve as a template strand for the other, leading to the accumulation of a discrete double stranded fragment whose length is defined by the distance between the 5′ ends of the oligonucleotide primers.

[0053] LCR uses a ligase enzyme to join pairs of preformed nucleic acid probes. The probes hybridize with each complementary strand of the nucleic acid analyte, if present, and ligase is

employed to bind each pair of probes together resulting in two templates that can serve in the next cycle to reiterate the particular nucleic acid sequence.

[0054] SDA (Westin et a 2000, Nature Biotechnology, 18, 199-202; Walker et al 1992, Nucleic Acids Research, 20, 7, 1691-1696), is an isothermal amplification technique based upon the ability of a restriction endonuclease such as HincII or BsoBI to nick the unmodified strand of a hemiphosphorothioate form of its recognition site, and the ability of an exonuclease deficient DNA polymerase such as Klenow exo minus polymerase, or Bst polymerase, to extend the 3′-end at the nick and displace the downstream DNA strand. Exponential amplification results from coupling sense and antisense reactions in which strands displaced from a sense reaction serve as targets for an antisense reaction and vice versa.

[0055] Some aspects of the invention utilize linear amplification of nucleic acids or polynucleotides. Linear amplification generally refers to a method that involves the formation of one or more copies of the complement of only one strand of a nucleic acid or polynucleotide molecule, usually a nucleic acid or polynucleotide analyte. Thus, the primary difference between linear amplification and exponential amplification is that in the latter process, the product serves as substrate for the formation of more product, whereas in the former process the starting sequence is the substrate for the formation of product but the product of the reaction, i.e. the replication of the starting template, is not a substrate for generation of products. In linear amplification the amount of product formed increases as a linear function of time as opposed to exponential amplification where the amount of product formed is an exponential function of time.

Downstream Applications

[0056] One aspect of the present invention is that the methods and compositions disclosed herein can be efficiently and cost-effectively utilized for downstream analyses, such as in next generation sequencing or hybridization platforms, with minimal loss of biological material of interest. The methods of the present invention can also be used in the analysis of genetic information of selective genomic regions of interest (e.g., analysis of SNPs or other disease markers) as well as genomic regions which may interact with the selective region of interest. The methods of the present invention may further be used in the analysis of copy number variation as well as differential expression.

Sequencing

[0057] In some embodiments, a population of sequencing reads is generated from the amplified adapter-nucleic fragment ligated products. In some embodiments, a sequencing read comprises an index read, which comprises the sequence of an indexing site. In some embodiments, an index read comprises the sequence of an indexing site and the sequence of an identifier site. For example, the indexing site is sequenced with the identifier site. In some embodiments, the index read does not include the sequence of an identifier site. For example, the indexing site is not sequenced with the identifier site. In some embodiments, a sequencing read comprises the target sequence. In some embodiments, a sequencing read comprises a target sequence and an identifier sequence. For example, the target sequence is sequenced with the identifier site. In some embodiments, the target sequence is not sequenced with the identifier site.

[0058] The methods of the present invention are useful for sequencing by the method commercialized by Illumina, as described U.S. Pat Nos. 5,750,341; 6,306,597; and 5,969,119.

[0059] In general, double stranded fragment polynucleotides can be prepared by the methods of the present invention to produce amplified nucleic acid sequences tagged at one (e.g., (A)/(A′) or both ends (e.g., (A)/(A′) and (C)/(C′)). In some cases, single stranded nucleic acid tagged at one or both ends is amplified by the methods of the present invention (e.g., by SPIA or linear PCR).The resulting nucleic acid is then denatured and the single-stranded amplified polynucleotides are randomly attached to the inside surface of flow-cell channels. Unlabeled nucleotides are added to initiate solid-phase bridge amplification to produce dense clusters of double-stranded DNA. To initiate the first base sequencing cycle, four labeled reversible terminators, primers, and DNA polymerase are added. After laser excitation, fluorescence from each cluster on the flow cell is imaged. The identity of the first base for each cluster is then recorded. Cycles of sequencing are performed to determine the fragment sequence one base at a time.

[0060] In some embodiments, the methods of the invention are useful for preparing target polynucleotides for sequencing by the sequencing by ligation methods commercialized by Applied Biosystems (e.g., SOLiD sequencing). In other embodiments, the methods are useful for preparing target polynucleotides for sequencing by synthesis using the methods commercialized by 454/Roche Life Sciences, including but not limited to the methods and apparatus described in Margulies et al., *Nature (*2005) 437:376-380 (2005); and U.S. Pat. Nos. 7,244,559; 7,335,762; 7,211,390; 7,244,567; 7,264,929; and 7,323,305. In other embodiments, the methods are useful for preparing target polynucleotide(s) for sequencing by the methods commercialized by Helicos BioSciences Corporation (Cambridge, Mass.) as described in U.S. application Ser. No. 11/167,046, and U.S. Pat. Nos. 7,501,245; 7,491,498; 7,276,720; and in U.S. Patent Application Publication Nos. US20090061439; US20080087826; US20060286566; US20060024711;

US20060024678; US20080213770; and US20080103058. In other embodiments, the methods are useful for preparing target polynucleotide(s) for sequencing by the methods commercialized by Pacific Biosciences as described in U.S. Pat. Nos. 7,462,452; 7,476,504; 7,405,281; 7,170,050; 7,462,468; 7,476,503; 7,315,019; 7,302,146; 7,313,308; and US Application Publication Nos. US20090029385; US20090068655; US20090024331; and US20080206764.

[0061] An example of a sequencing technique that can be used in the methods of the provided invention is semiconductor sequencing provided by Ion Torrent (e.g., using the Ion Personal Genome Machine (PGM)). Ion Torrent technology can use a semiconductor chip with multiple layers, e.g., a layer with micro-machined wells, an ion-sensitive layer, and an ion sensor layer. Nucleic acids can be introduced into the wells, e.g., a clonal population of single nucleic can be attached to a single head, and the bead can be introduced into a well. To initiate sequencing of the nucleic acids on the beads, one type of deoxyribonucleotide (e.g., dATP, dCTP, dGTP, or dTTP) can be introduced into the wells. When one or more nucleotides are incorporated by DNA polymerase, protons (hydrogen ions) are released in the well, which can be detected by the ion sensor. The semiconductor chip can then be washed and the process can be repeated with a different deoxyribonucleotide. A plurality of nucleic acids can be sequenced in the wells of a semiconductor chip. The semiconductor chip can comprise chemical-sensitive field effect transistor (chemFET) arrays to sequence DNA (for example, as described in U.S. Patent Application Publication No. 20090026082). Incorporation of one or more triphosphates into a new nucleic acid strand at the 3' end of the sequencing primer can be detected by a change in current by a chemFET. An array can have multiple chemFET sensors.

[0062] Another example of a sequencing technique that can be used in the methods of the provided invention is nanopore sequencing (see e.g., Soni G V and Meller A. (2007) *Clin Chem*

53: 1996-2001). A nanopore can be a small hole of the order of 1 nanometer in diameter. Immersion of a nanopore in a conducting fluid and application of a potential across it can result in a slight electrical current due to conduction of ions through the nanopore. The amount of current that flows is sensitive to the size of the nanopore. As a DNA molecule passes through a nanopore, each nucleotide on the DNA molecule obstructs the nanopore to a different degree. Thus, the change in the current passing through the nanopore as the DNA molecule passes through the nanopore can represent a reading of the DNA sequence.

Data Analysis

[0063] In some embodiments, the sequence reads are used in the analysis of genetic information of selective genomic regions of interest as well as genomic regions which may interact with the selective region of interest. Amplification methods as disclosed herein can be used in the devices, kits, and methods known to the art for genetic analysis, such as, but not limited to those found in U.S. Pat. Nos. 6,449,562, 6,287,766, 7,361,468, 7,414,117, 6,225,109, and 6,110,709.

[0064] In some embodiments, the sequencing reads are used to detect duplicate sequencing reads. In some embodiments, a sequencing read is identified as a duplicate sequencing read when it contains an identifier site and target sequence that is the same as another sequencing read from the same population of sequencing reads.

[0065] In some embodiments, duplicate sequencing reads are differentiated from one another as being true duplicates versus apparent or perceived duplicates. Apparent or perceived duplicates may be identified from sequencing libraries and using conventional measures of duplicate reads (i.e., reads were mapped using bowtie), where all reads with the same start and end nucleic acid coordinates were counted as duplicates. True duplicates may be identified from sequencing

libraries having had an identifier site introduced through ligation to differentiate between fragments of DNA that randomly have the same start and end mapping coordinates.

[0066] In some embodiments, the sequence of the identifier sites from the index read of two nucleic acids generated from any dsDNA may have the same start site, as determined from the sequencing reads of the target sequence of the generated nucleic acid fragments. If the identifier site from the index read of the two nucleic acid fragments are not identical, then the target sequence reads were not generated from the same original dsDNA molecule and, therefore, are no true duplicate reads. The ligation of random sequences onto dsDNA molecules, accompanied by the methods of the invention, allow for the identification of true duplicate reads versus apparent or perceived duplicate reads.

[0067] In some embodiments, the sequence of the identifier sites from the index read of two nucleic acid fragments generated from a genomic DNA (gDNA) molecule that have the same start site, which can be determined from the sequencing reads of the target sequence of the nucleic acid fragments is determined. If the identifier site from the index read of the two nucleic acid fragments are not identical, then the target sequence reads were not generated from the same original gDNA molecule and, therefore, are not true duplicate reads. In another embodiment, an identifier site is inserted at the adapter insert junction. The sequence of the identifier site is carried through the library amplification step. The identifier site is the first sequence read during the forward read. As the identifier sequence is not logically present adjacent to naturally occurring sequence, this uniquely identifies the DNA fragment. Therefore, by ligating random sequences onto the original gDNA, the methods of the invention identify true duplicate reads.

[0068] In some embodiments, a duplicate sequencing read is detected and analyzed. Duplicate reads can be filtered using 'samtools rmdup', wherein reads with identical external coordinates

are removed, only one read with highest mapping quality is retained. After filtering, a filtered set of deduplicated reads can be used in any downstream analysis. Conversely, this filtering step can be skipped, and downstream analysis can be done using the unfiltered reads, including duplicates.

[0069] In some embodiments, sequencing reads are generated from a number of samples. In some embodiments, adaptors are ligated to a plurality of nucleic acid fragments from the samples, in which the nucleic acid fragments from each sample has the same indexing site. In some embodiments, a plurality of nucleic acid fragments is generated from a first sample and a second sample and adaptors are ligated to each nucleic acid fragment, in which adaptors ligated to each nucleic acid fragment from the first sample have the same first indexing site and adaptors ligated to nucleic acid fragments from the second sample has the same second indexing site. In some embodiments, the data associated with the nucleic acid fragments (e.g., sequencing reads) are separated based on the indexing site before sequencing reads of the target sequence and/or identifier site are analyzed. In some embodiments, the nucleic acid fragments or data associated with the nucleic acid fragments (e.g., sequencing reads) are separated based on the indexing site before duplicate sequencing reads are analyzed and/or removed.

[0070] In some embodiments, a method disclosed herein identifies or detects one or more true duplicate(s) with increased accuracy as compared to other methods. For example, in some embodiments, a method disclosed herein identify true duplicates (in contrast to identifying apparent or perceived duplicates) with increased accuracy as compared to other methods. The increased resolution and/or accuracy in identifying one or more true duplicate(s) can provide a considerable contribution to the state of the art in more accurately identifying true duplicates. In some embodiments, a method disclosed herein identifies or detects a true duplicate with

increased efficiency as compared to other methods, such as paired end sequencing. The increase in accuracy, resolution, and/or efficiency in detecting duplicate reads (e.g., true duplicate(s)) can increase confidence in sequencing results, such as for expression and CNV analysis.

Kits

[0071] Any of the compositions described herein may be comprised in a kit. In a non-limiting example, the kit, in a suitable container, comprises: an adaptor or several adaptors, one or more of oligonucleotide primers and reagents for amplification.

[0072] The containers of the kits will generally include at least one vial, test tube, flask, bottle, syringe or other containers, into which a component may be placed, and preferably, suitably aliquotted. Where there is more than one component in the kit, the kit also will generally contain a second, third or other additional container into which the additional components may be separately placed. However, various combinations of components may be comprised in a container.

[0073] When the components of the kit are provided in one or more liquid solutions, the liquid solution can be an aqueous solution. However, the components of the kit may be provided as dried powder(s). When reagents and/or components are provided as a dry powder, the powder can be reconstituted by the addition of a suitable solvent.

[0074] A kit may include instructions for employing the kit components as well the use of any other reagent not included in the kit. Instructions may include variations that can be implemented.

[0075] In some embodiments, the invention provides kits containing any one or more of the elements disclosed in the above methods and compositions. In some embodiments, a kit comprises a composition of the invention, in one or more containers. In some embodiments, the

invention provides kits comprising adapters, primers, and/or other oligonucleotides described herein. In some embodiments, the kit further comprises one or more of: (a) a DNA ligase, (b) a DNA-dependent DNA polymerase, (c) an RNA-dependent DNA polymerase, (d) a forward adapter (e) one or more oligonucleotides comprising reverse adaptor sequence and (f) one or more buffers suitable for one or more of the elements contained in said kit. The adapters, primers, other oligonucleotides, and reagents can be, without limitation, any of those described above. Elements of the kit can further be provided, without limitation, in any of the amounts and/or combinations (such as in the same kit or same container) described above. The kits may further comprise additional agents, such as those described above, for use according to the methods of the invention. For example, the kit can comprise a first forward adaptor that is a partial duplex adaptor as described herein, a second forward adapter, and a nucleic acid modifying enzyme specific for a restriction and/or cleavage site present in the first forward adaptor. The kit elements can be provided in any suitable container, including but not limited to test tubes, vials, flasks, bottles, ampules, syringes, or the like. The agents can be provided in a form that may be directly used in the methods of the invention, or in a form that requires preparation prior to use, such as in the reconstitution of lyophilized agents. Agents may be provided in aliquots for single-use or as stocks from which multiple uses, such as in a number of reaction, may be obtained.

[0076] In some embodiments, the kit comprises a plurality of adaptor oligonucleotides, wherein each of the adaptor oligonucleotides comprises at least one of a plurality of identifier site sequences, wherein each identifier site sequence of the plurality of identifier site sequences differs from every other identifier site sequence in said plurality of identifier site sequences at at least three nucleotide positions, and instructions for using the same. Adapters comprising

different identifier site sequences can be supplied individually or in combination with one or more additional adapters having a different identifier site sequence. In some embodiments, the kit can comprises a plurality of adapter oligonucleotides.

EXAMPLES

Example 1: Identification of Duplicate Sequencing Reads with NuGEN Ovation Target Enrichment Library System

[0077] Sample Description: l00ng of DNA from a human HapMap sample (NA19238) was fragmented to approximately 500 base pair in length by sonication with a Covaris system (Covaris, Inc., Woburn, MA). The resulting DNA was treated with end repair enzyme mix NuGEN R01280 and R01439 (NuGEN Technologies, Inc., San Carlos, CA) according to supplier's recommendation to produce blunt ended DNA fragments.

[0078] Library Generation, Enrichment, and Identifier Site Incorporation: An oligonucleotide with segments, from 5' to 3' of the top strand, 1) an Illumina indexing read priming site such as AGAGCACACGTCTGAACTCCAGTCAC (SEQ ID NO:2), 2) an indexing site, 3) an identifier site having a random 6 base sequence and, 4) a sequence compatible with an Illumina forward sequencing priming site such as TCTTTCCCTACACGACGCTCTTCCGATCT (SEQ ID NO:3) , was annealed to a second oligonucleotide to form a partially double stranded DNA adapter. Five uM of these adapters were ligated onto the end-repaired DNA using Ligase and Ligase reaction buffer from NuGEN's Ovation Ultralow Library System (NuGEN Technologies, Inc., San Carlos, CA) according to supplier's recommendations. Following 30 minutes of incubation at 25ºC, the reaction mixture was diluted with water, 0.8X volume of Ampure XP magnetic beads (Agencourt Biosciences Corporation, A Beckman Coulter Company, Beverly, MA) was added and the solution thoroughly mixed. The beads were collected, washed and the ligated

DNA fragments eluted according to manufacturer's recommendations. A pool of targeting probes was annealed to the eluted DNA fragments by initially heating the solution to 95ºC then slowly cooling the mixture from 80ºC to 60ºC by 0.6 degrees/minute. Targeting probes that were specifically annealed were extended with Taq DNA polymerase (New England Biolabs, Inc., Ipswich, MA) according to the manufacturer's protocols. Following extension, the DNA fragments were collected on Agencourt magnetic beads, washed and eluted according to manufacturer's recommendations. These libraries were enriched by 30 cycles of PCR using NuGEN library enrichment primers (Ovation Target Enrichment Library System, NuGEN Technologies, Inc., San Carlos, CA) that also contain the Illumina flow cell sequences (Illumina Inc., San Diego, CA) according to supplier's recommendation.

[0079] The resulting libraries were quantitated by qPCR using a kit provided by KAPA, diluted to 2nM and applied to an Illumina MiSeq DNA Sequencer (Illumina Inc., San Diego, CA). The following series was run: 36 base first read, 14 base second read, and 24 base third read.

[0080] *Data Analysis:* The sequencer output was processed in accordance with manufacturer's recommendation. In order to analyze the data, the indexing read was split into two files. The first file contained the first 8 bases of the indexing reads and is utilized as the library index file for standard library parsing. The other file contains only the random bases and is set aside for further sequence parsing.

[0081] Following our data analysis pipeline of sequence alignment with bowtie aligner (Langmead B. et al., Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* 2009, 10:R2.), duplicate reads were identified by their genomic start positions. At this point, sequencing reads that started at the same genomic position were checked against the random bases file to see if they have the same or different set of random

bases that were ligated to them. Where two sequences with the same starting genomic coordinate had the same set of random bases, they were considered to have come from the same initial DNA ligation event, regardless of which Ovation Target Enrichment targeting probe was used to generate the sequence fragment in question. These sequences, therefore, did not provide unique information about the starting genomic DNA and are considered as one sequencing read for the purpose of variant analysis. Two sequence reads with the same starting genomic coordinate that have different random bases were derived from unique ligation events and were considered to both be valid sequencing reads for the purpose of variant identification. **Figure 5** provides the analysis results demonstrating the identification of the duplicate reads. The use of an identifier site allowed for the determination of the number of true duplications versus the number of apparent or perceived duplicates.

[0082] If the sequences of two libraries are identical, their duplication status is unknown since this could occur by chance in any library. If an identifier sequence is used in combination with the library reads, the status can be determined (duplicates if identical, distinct if different). With the SPET system, one end is common so the probability of two libraries having identical ends increases. These would appear to be duplicate sequences and their true status can be determined by looking at an identifier sequence.

[0083] Over the sampling of all randomly selected reads, the use of identifier sites provided an increased resolution on the presence of true duplicates. When evaluating two million random reads, the apparent duplicates were found to comprise 39% of all reads. However, the true duplicates, identified through the use of an identifier site, were found to comprise only 26% of all reads. The methods employing the use of an identifier site were found to considerably increase the resolution of the true number of duplicates within the pool of reads.

Example 2: Removal of Duplicate Sequencing Reads with 8 Base Identifier Site

[0084] In a standard RNA sequencing library, adaptors are ligated to the ends of double stranded cDNA. These adaptors contain universal sequences that allow for PCR amplification and sequencing on high throughput sequencing machines. The adaptors are synthesized with a large population of additional sequences, in which each additional sequence is an identifier site, at the ligation end. The identifier sites are present at the junction between the adaptor and the cDNA. The sequence read starts with the identifier site and follows with the cDNA sequence.

[0085] This pool of identifier sites is used for the detection of PCR duplicates, as PCR duplicates will contain the identical identifier site, whereas two different cDNA molecules will ligate to two different adaptors containing two different identifier sites. This identifier site is designed as eight random bases introduced onto the end of the adaptor. Sequence reads from libraries made with such adaptors contain the 8 bases of the identifier site followed by the cDNA sequence. Standard PCR duplicate removal software such as PICARD, Markduplicates, and/or SAMtools rndup is used to identify and remove the PCR duplicates, leaving behind for analysis any instances of multiple cDNA fragments that happen to have the same sequence.

Example 3: Removal of Duplicate Sequencing Reads with Mixture of Random 1-8 Base Identifier Sites

[0086] In a standard RNA sequencing library, adaptors are ligated to the ends of double stranded cDNA. These adaptors contain universal sequences that allow for PCR amplification and sequencing on high throughput sequencing machines. The adaptors are synthesized with a large population of additional sequences, in which each additional sequence is an identifier site, at the

ligation end. The identifier sites are present at the junction between the adaptor and the cDNA. The sequence read starts with the identifier site and follows with the cDNA sequence.

[0087] This pool of identifier sites is used for the detection of PCR duplicates, as PCR duplicates will contain the identical identifier site, whereas two different cDNA molecules will ligate to two different adaptors containing two different identifier sites.

[0088] One to eight random bases are introduced onto the end of the adaptor. Sequence reads from libraries made with such adaptors contain between 1 and 8 bases of the identifier site followed by the cDNA sequence. Standard PCR duplicate removal software such as PICARD, Markduplicates, and/or SAMtools rndup is used to identify and remove the PCR duplicates, leaving behind for analysis any instances of multiple cDNA fragments that happen to have the same sequence.

Example 4: Removal of Duplicate Sequencing Reads with Mixture of 96 Defined 6-Base Identifier Sites

[0089] In a standard RNA sequencing library, adaptors are ligated to the ends of double stranded cDNA. These adaptors contain universal sequences that allow for PCR amplification and sequencing on high throughput sequencing machines. The adaptors are synthesized with a large population of additional sequences, in which each additional sequence is an identifier site, at the ligation end. The identifier sites are present at the junction between the adaptor and the cDNA. The sequence read starts with the identifier site and follows with the cDNA sequence.

[0090] This pool of identifier sites is used for the detection of PCR duplicates, as PCR duplicates will contain the identical identifier site, whereas two different cDNA molecules will ligate to two different adaptors containing two different identifier sites.

[0091] A mixture of 96 defined six-bases sequences are introduced onto the end of the adaptors. Thus, each six-base sequence is an identifier site. Sequence reads from libraries made with such adaptors contain one of the 96 six-base identifier site followed by the cDNA sequence. Standard PCR duplicate removal software such as PICARD, Markduplicates, and/or SAMtools rndup is used to identify and remove the PCR duplicates, leaving behind for analysis any instances of multiple cDNA fragments that happen to have the same sequence.

Example 5: Identification of Duplicate Sequencing Reads in Determining mRNA Expression Levels

[0092] Sample Description: Total RNA is extracted from tumor and normal adjacent tissue for the purpose of finding differences in expression levels of transcripts between the two sample types. 100ng of each sample is converted into cDNA using the USP primers, reaction buffer, and Reverse Transcriptase provided in NuGEN's Encore Complete Library System (NuGEN Technologies, Inc., San Carlos, CA) according to the supplier's recommendations. This is followed by second strand synthesis, again using materials provided in the kit according to recommendations. Double stranded cDNA was prepared using the SuperScript® Double-Stranded cDNA Synthesis Kit from Life Technologies (Carlsbad, CA) according the manufacturer's instructions. DNA was sheared with a Covaris S-series device (Covaris, Inc., Woburn, MA) using the 200 bp sonication protocol provided with the instrument (10% duty cycle, 200 cycles/burst, 5 intensity, 180 seconds). DNA was treated with 1.5 uL l0X Blunting Buffer, 0.5 uL Blunting Enzyme (New England Biolabs, Inc., Ipswich, MA; p/n E1201) and 1.2 uL of 2.5mM of each dNTP mix in a total volume of 15 uL for 30 minutes at 25°C followed by 10 minutes at 70°C.

[0093] *Library Generation*: The DNA fragments were then subjected to end repair using end repair buffers and enzymes provided in NuGEN's Ovation Ultralow Library System (NuGEN Technologies, Inc., San Carlos, CA).

[0094] Forward adaptor 5'

[0095] AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCC GATCTNNNNNNNN (SEQ ID NO:4),

[0096] Reverse adaptor 1) 5'

[0097] CAAGCAGAAGACGGCATACGAGATTCCCTTGTGACTGGAGTTCAGACGTGTG CTCTTCCGATCTNNNNNNNN (SEQ ID NO:5),

[0098] Reverse adaptor 2) 5'

[0099] CAAGCAGAAGACGGCATACGAGATTGAAGGGTGACTGGAGTTCAGACGTGT GCTCTTCCGATCNNNNNNNN (SEQ ID NO:6), and

[00100] Common partner 5' NNNNNNNNAGATCGGAAGAGC (SEQ ID NO:7) were all ordered from IDT (Integrated DNA Technologies, Coralville, IA). The reverse adaptors each contain a unique identifier (underlined) that enables libraries made with these adapters to be distinguished. In this case N represents an equimolar mixture of A, C, G, and T. A mixture of 5uM forward, 5uM reverse, and 10uM common in 10mM MgCl2, 50mM Tris pH 8 was heated to 95C for 5 minutes, then cooled to 20C. Adaptor ligation was performed by addition of 4.5 uL water, 3 uL Adaptor mix (prepared above), 6 uL 5X NEBNext Quick Ligation Reaction Buffer and 1.5 uL Quick T4 DNA Ligase (New England Biolabs, Inc., Ipswich, MA; p/n E6056), followed by incubation for 30 minutes at 25°C followed by 10 minutes at 70°C. Ligation products were purified by adding 70uL water and 80 uL of Ampure XP beads (Agencourt Genomics), washing twice with 70% ethanol and eluting with 20 uL of 10mM Tris pH 8.0.

Library products were amplified in a 50uL PCR containing 0.5uM each of primer (5' AATGATACGGCGACCACCGA (SEQ ID NO:8) , and 5' CAAGCAGAAGACGGCATACGA (SEQ ID NO:9), 10 mM Tris-HCl, pH 8.3, 50 mM KCl, 2 mM MgC12, 0.2mM each dNTP, and 1 unit Taq polymerase. The reaction was cycled 15 times under the conditions 95C for 15 seconds, 60C for 1 minute. PCR products were purified with 1 volume of Ampure XP beads (Agencourt Biosciences Corporation, A Beckman Coulter Company, Beverly, MA) as described above. The library was analyzed by HS DNA Bioanalyzer (Agilent Technologies, Santa Clara, CA) and quantitated with the KAPA Library Quantification Kit (KAPA Biosystems, Wilmington, MA; p/n KK4835) according to the supplied instructions. The resulting libraries are combined and are compatible with standard TruSeq single end or paired Illumina sequencing protocols for GAIIx, MiSeq, or Hi Seq sequencing instruments (Illumina Inc., San Diego, CA). The following series is run; 50 base first read, 6 base second read. A third read is not required for counting purposes or duplication analysis.

[00101] Data Analysis: The sequencer output was processed in accordance with manufacturer's recommendation. The 6 bases of the index read is used for standard library parsing, separating the data files from the two sample types. The first 50 bases of the target sequence read are compared to each other. Any read that has identical sequence to any other read is identified as a duplicate and removed from the population, thus, only a single copy is retained within the file. Once the duplicate reads have been removed, 8 bases are trimmed from each read. The trimmed reads are aligned to a reference genome. Differential expression is then determined by comparing FPKM (fragments per kilobase per million reads) values between libraries utilizing scripts such as cufflinks or cuffdiff (Trapnell et al. 2010, Nature Biotechnology, 28, 511-515; Trapnell et al. 2013, Nature Biotechnology, 31, 46-53).

Example 6: Sequencing of the Identifier Site With the Target Sequence in Paired End Reads

[00102] The Ovation Library System for Low Complexity Samples (NuGEN Technologies, Inc., San Carlos, CA) was used to generate four libraries, each from a single amplicon, following manufacturer's protocol. Purified libraries were mixed and sequenced as a multiplex on the Illumina MiSeq (Illumina Inc., San Diego, CA) to produce 125nt forward, 8nt index 1, 8nt index 2, and 25nt reverse reads. Because all amplicon reads start and end at the same sequence coordinates, the traditional method of marking library PCR duplicates (marking reads that start and end at the same genomic coordinates as duplicates) cannot be used. Instead, the 0-8nt of random sequence contained in the adaptors ligated to the amplicon were treated as an identifier sequence and used to mark duplicates. Any paired end reads that shared the same length and sequence of these random bases with any other paired end read was called a duplicate. The table below shows the results of this duplicate marking.

Table 1.

Library	Reads	Duplicate Reads	Reads from independent molecules
1	454249	376797	77452
2	439367	364001	75366
3	317760	253057	64703
4	476572	398380	78192

[00103] Table 1 demonstrates the accuracy of the method used to differentiate between duplicate reads and reads from truly independent molecules. The population of reads from independent molecules, depicted in the final column of Table 1, represent sequences from independent amplicon molecules used in generating the libraries.

Example 7: Sequencing of the Identifier Site With the Target Sequence in Reduced Representation Bisulfite (RRBS) Libraries

[00104] Reduced representation bisulfite (RRBS) libraries of the human genome are generated through the complete restriction enzyme digestion of 100ng input sample, followed by selection for short fragments. The resulting pool of fragments are ligated to adaptor sequences comprising an indexing site and an identifier site. The identifier sites comprise either 6 or 8 random nucleotides. The sequences are then sequenced to identify the identifier sites, thus revealing the true number of duplicates in the pool. In the absence of identifier sites, the number of apparent or perceived duplicates are greater than the number of true duplicates. The inclusion of an identifier site results in the identification of the number of true duplicates, as compared to the larger number of apparent or perceived duplicates.

[00105] Unless defined otherwise, all technical and scientific terms 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 be used in the practice or testing of the present invention, the preferred methods and materials are described herein. All publications, patents, and patent publications cited are incorporated by reference herein in their entirety for all purposes.

[00106] The publications discussed herein are provided solely for their disclosure prior to the filing date of the present application. Nothing herein is to be construed as an admission that the present invention is not entitled to antedate such publication by virtue of prior invention.

[00107] All references, articles, publications, patents, patent publications, and patent applications cited herein are incorporated by reference in their entireties for all purposes. However, mention of any reference, article, publication, patent, patent publication, and

patent application cited herein is not, and should not be taken as an acknowledgment or any form of suggestion that they constitute valid prior art or form part of the common general knowledge in any country in the world.

[00108] While the invention has been described in connection with specific embodiments thereof, it will be understood that it is capable of further modifications and this application is intended to cover any variations, uses, or adaptations of the invention following, in general, the principles of the invention and including such departures from the present disclosure as come within known or customary practice within the art to which the invention pertains and as may be applied to the essential features hereinbefore set forth and as follows in the scope of the appended claims.

CLAIMS

1. A method for detecting a duplicate sequencing read from a population of sample sequencing reads comprising:
    a) ligating an adaptor to a 5' end of each nucleic acid fragment of a plurality of nucleic acid fragments from one or more samples, wherein the adaptor comprises:
        (i) an indexing primer binding site;
        (ii) an indexing site;
        (iii) an identifier site; and
        (iv) a target sequence primer binding site;
    b) amplifying the adapter-nucleic acid fragment ligated products;
    c) generating a population of sequencing reads from amplified adapter-nucleic acid fragment ligated products; and
    d) detecting the population of sequencing reads comprising a sequencing read with a duplicate identifier site and target sequence.
2. The method of claim 1, wherein the method further comprises removing from the population of sequence reads the sequencing read with a duplicate identifier site and target sequence.
3. The method of claim 1, wherein the identifier site is sequenced with the indexing site.
4. The method of claim 1, wherein the identifier site is sequenced separately from the indexing site.
5. The method of claim 1, wherein the identifier site is sequenced with the target sequence.
6. The method of claim 1, wherein the identifier site is sequenced separately from the target sequence.
7. The method of claim 1, wherein the adaptor comprises from 5' to 3':

(i) the indexing primer binding site;

(ii) the indexing site;

(iii) the identifier site; and

(iv) the target sequence primer binding site.

8. The method of claim 1, wherein the adaptor comprises from 5' to 3':

(i) the indexing primer binding site;

(ii) the indexing site;

(iii) the target sequence primer binding site; and

(iv) the identifier site.

9. The method of claim 1, wherein the plurality of nucleic acid fragments is generated from more than one sample.

10. The method of claim 9, wherein the nucleic acid fragments from each sample has the same indexing site.

11. The method of claim 10, wherein the sequencing reads are separated based on the indexing site.

12. The method of claim 11, wherein the separation of sequencing reads is performed prior to step d).

13. The method of claim 1, wherein the nucleic acid fragments are DNA fragments, RNA fragments, or DNA/RNA fragments.

14. The method of claim 13, wherein the nucleic acid fragments are genomic DNA fragments or cDNA fragments.

15. The method of claim 1, wherein the indexing site is between 2 and 8 nucleotides in length.

16. The method of claim 1, wherein the indexing site is about 6 nucleotides in length.

17. The method of claim 1, wherein the identifier site is between 1 and 8 nucleotides in length.

18. The method of claim 1, wherein the identifier site is about 8 nucleotides in length.

19. The method of claim 1, wherein the indexing primer binding site is a universal indexing primer binding site.

20. The method of claim 1, wherein the target sequence primer binding site is a universal target sequence primer binding site.

21. A kit comprising a plurality of adaptors, wherein each adaptor comprises:

(i) an indexing primer binding site;

(ii) an indexing site;

(iii) an identifier site; and

(iv) a target sequencing primer binding site.

Standard DNA Sequencing With
Duplication Counter
Sequence Read 1 (forward read)
Read 1 sequence
Indexing read sequence
Forward flow cell sequence plus indexing primer binding site
Indexing site plus identifier site
Forward sequencing primer site
Target sequence
Reverse sequencing primer site
Reverse flow cell sequence
The indexing read results in sequencing of both the index site and the identifier site
Extension of the indexing read beyond the indexing site and into the nucleotides of the identifier site provides data on which nucleic acid molecule is being sequenced.

Figure 1

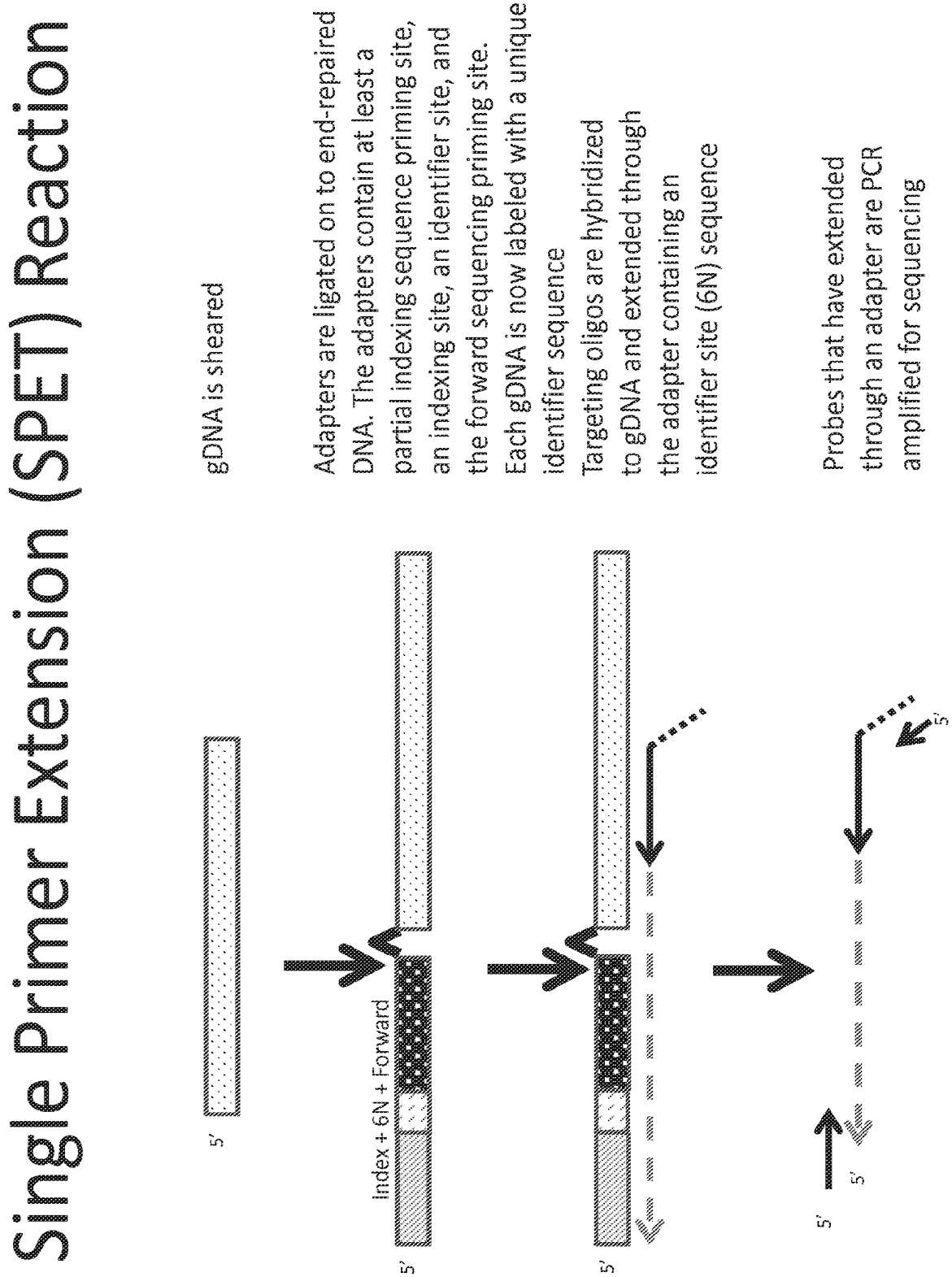

Single Primer Extension (SPET) Reaction
gDNA is sheared
5'
Adapters are ligated on to end-repaired DNA. The adapters contain at least a partial indexing sequence priming site, an indexing site, an identifier site, and the forward sequencing priming site.
Each gDNA is now labeled with a unique identifier sequence
Index + 6N + Forward
5'
Targeting oligos are hybridized to gDNA and extended through the adapter containing an identifier site (6N) sequence
5'
Probes that have extended through an adapter are PCR amplified for sequencing
5'
5'
5'

Figure 2A

SPET Sequencing

- Sequence from random end toward the probe

PCR adds sequence specific primer binding sites and Read 1 of sequencing is performed

Read 1 sequence

Indexing Read performed sequences both the indexing site as well as the identifier site (N6)

Indexing read sequence

- Extending the indexing read past the 6 base indexing site sequence and into the random Ns of the identifier site provides data on which gDNA molecule the extension reaction occurred upon.

Figure 2B

Example Adaptor

An adaptor sequence with necessary priming sites, the index site, and the identifier site

CAGAGCACACGTCTGAACTCCAGTCACAGTGAGNNNNNNACACTTTCCCTACACGACGCTCTTCCGATCT (gDNA)

Illumina flow cell plus Indexing read priming site

Indexing site + identifier site

Illumina forward sequencing read priming site

Figure 3

DNA Sequencing With Duplicate Recognition at the insert junction
Traditional library, sequence read 1 (forward read)
Read 1 Sequence
Duplicate recognition library, sequence read 1 (forward read)
Read 1 Sequence
Read 1 initially records the approximately 8 random bases of the identifier site appended to the library sequences, followed by the target sequence
Forward flow cell sequence
Target sequence primer binding site
Target sequence
Reverse sequencing priming site plus indexing primer binding site
Indexing site
Reverse flow cell sequence
Forward flow cell sequence
Target sequence primer binding site
Identifier site
Target sequence
Reverse sequencing priming site plus indexing primer binding site
Indexing site
Reverse flow cell sequence

Figure 4

Resolution of apparent duplicate reads by identifier sequence ligation

Libraries were generated from 100ng genomic DNA Covaris fragmented to ~500bp in length. These were enriched with the 344 gene cancer panel, ~12k probes covering 969kb of genomic space and sequenced on the ILMN Miseq. Different numbers of reads were randomly selected for this analysis

Total reads randomly selected	Aligned read	Fold Target Coverage	Total duplicates after identifier analysis (true rate)	Percentage duplicates after identifier analysis (true rate)	Total apparent duplicates with no identifier	Percentage apparent duplicates with no identifier
100K	97,919	2X	1854	1.89	3256	3.33
500K	488,423	7X	39731	8.13	68741	14.07
1M	976,662	13X	145589	14.91	237882	24.36
2M	1,954,673	21X	514848	26.34	764825	39.13
10M	9,771,462	43X	6493577	66.45	7515496	76.91

"Apparent" Duplicates – individual targets within libraries consist of one random end (result of fragmentation) and one common end (initiated with the sequence specific probe). Apparent duplicates are derived from sequencing libraries and using conventional measures of duplicate reads ie, reads were mapped using bowtie and all reads with the same start and end genomic coordinates were counted as duplicates.

"True" duplicates - An identifier sequence is introduced through ligation to differentiate between fragments of DNA that randomly have the same start and end mapping coordinates. The analysis method is described in the specification.

Figure 5

IDT Exhibit 1020.4965

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/065530

Box No. I **Nucleotide and/or amino acid sequence(s) (Continuation of item 1.c of the first sheet)**

1. With regard to any nucleotide and/or amino acid sequence disclosed in the international application, the international search was carried out on the basis of a sequence listing filed or furnished:

   a. (means)

   ☐ on paper

   ☒ in electronic form

   b. (time)

   ☒ in the international application as filed

   ☐ together with the international application in electronic form

   ☐ subsequently to this Authority for the purposes of search

2. ☐ In addition, in the case that more than one version or copy of a sequence listing has been filed or furnished, the required statements that the information in the subsequent or additional copies is identical to that in the application as filed or does not go beyond the application as filed, as appropriate, were furnished.

3. Additional comments:

INTERNATIONAL SEARCH REPORT

International application No.
PCT/US2014/065530

A. CLASSIFICATION OF SUBJECT MATTER
IPC(8) - C12N 15/11 (2015.01)
CPC - C12Q 1/6869 (2015.01)
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)
IPC(8) - C12N 15/09, 15/10, 15/11; C12P 19/34, C12Q 1/68; C40B 20/00, 20/04 (2015.01)
USPC - 435/6.1, 6.11; 506/2, 23; 536/23.1, 24.2, 25.3

Documentation searched other than minimum documentation to the extent that such documents are included in the fields searched
CPC - C12Q 1/6869, 1/6874, 2525/16, 2537/143 (2015.01) (keyword delimited)

Electronic data base consulted during the international search (name of data base and, where practicable, search terms used)

PatBase, Google Patents, PubMed, Google
Search terms used: nucleic acid barcode tag* duplicate* identifier* adaptor* sequenc*

C. DOCUMENTS CONSIDERED TO BE RELEVANT

Category*	Citation of document, with indication, where appropriate, of the relevant passages	Relevant to claim No.
X – Y	WO 2008/093098 A2 (SOLEXA LIMITED) 07 August 2008 (07.08.2008) entire document	21 --- 7, 8, 10-12, 20
Y	US 2012/0208705 A1 (STEEMERS et al) 16 August 2012 (16.08.2012) entire document	1-20
Y	MEYERS, R. 'Myers lab microRNA-seq Protocol, Hudson Alpha Institute for Biotechnology,' 02 May 2013 ()02.05.2013), Pgs. 1-15. Retrieved from the Internet:<https://www.encodeproject.org/documents/9fb8e06c-d145-447a-a606-e7f3dc44b549/@@download/attachment/HA_Myers_miRNA-seq_Protocol_2013-05-02.pdf> on 28 February 2015 (28.02.2015). entire document	1-20
Y	WO 2012/061832 A1 (ILLUMINA, INC.) 10 May 2012 (10.05.2012) entire document	2
P, Y	WO 2014/144092 A1 (NUGEN TECHNOLOGIES, INC.) 18 September 2014 (18.09.2014)	1-21

☐ 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
18 February 2015	05 MAR 2015
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: Blaine R. Copenheaver PCT Helpdesk: 571-272-4300 PCT OSP: 571-272-7774

Form PCT/ISA/210 (second sheet) (July 2009)

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

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
11 December 2008 (11.12.2008)

PCT

(10) International Publication Number
WO 2008/150432 A1

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

(21) International Application Number:
PCT/US2008/006822

(22) International Filing Date: 29 May 2008 (29.05.2008)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
60/941,381 1 June 2007 (01.06.2007) US

(63) Related by continuation (CON) or continuation-in-part (CIP) to earlier application:
US 60/941,381 (CIP)
Filed on 1 June 2007 (01.06.2007)

(71) Applicant *(for all designated States except US)*: **454 LIFE SCIENCES CORPORATION** [US/US]; 20 Commercial Street, Branford, CT 06405 (US).

(72) Inventors; and
(75) Inventors/Applicants *(for US only)*: **BRAVERMAN, Michael, S.** [US/US]; 375 Stevenson Road, New Haven, CT 06515 (US). **SIMONS, Jan, Fredrik** [FI/US]; 2515 Poppy Drive, Burlingame, CA 94010 (US). **SRINIVASAN, Maithreyan** [US/US]; 801 Church Street #1398, Mountain View, CA 94041 (US). **TURENCHALK, Gregory, S.** [US/US]; 1448 Quinnipiac Avenue, New Haven, CT 06513 (US).

(74) Agents: **IWAMOTO, Michelle, A.** et al.; Mintz, Levin, Cohn, Ferris, glovsky And Popeo P.C., Chrysler Center, 666 Third Avenue, New York, NY 10017 (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, 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, SV, SY, TJ, TM, TN, TR, TT, TZ, UA, UG, US (patent), 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, 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*
— *with sequence listing part of description published separately in electronic form and available upon request from the International Bureau*

WO 2008/150432 A1

(54) Title: SYSTEM AND METH0D FOR IDENTIFICATION OF INDIVIDUAL SAMPLES FROM A MULTIPLEX MIXTURE

(57) Abstract: An embodiment of an identifier element for identifying an origin of a template nucleic acid molecule is described that comprises 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, where 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.

IDT Exhibit 1020.4968

SYSTEM AND METHOD FOR IDENTIFICATION OF INDIVIDUAL SAMPLES FROM A MULTIPLEX MIXTURE

FIELD OF THE INVENTION

The present invention relates to the fields of molecular biology and bioinformatics. More specifically, the invention relates to associating a unique identifier (UID) element, which is sometimes also referred to as a multiplex identifier (MID), with one or more nucleic acid elements derived from a specific sample, combining the associated elements from the sample with associated elements from one or more other samples into a multiplex mixture of said samples, and identifying each identifier and its associated sample from data generated by what are generally referred to as "Sequencing" techniques.

BACKGROUND OF THE INVENTION

There are a number of "sequencing" techniques known in the art amenable for use with the presently described invention such as, for instance, techniques based upon what are referred to as Sanger sequencing methods commonly known to those of ordinary skill in the art that employ termination and size separation techniques. Other classes of powerful high throughput sequencing techniques for determining the identity or sequence composition of one or more nucleotides in a nucleic acid sample include what are referred to as "Sequencing-by-synthesis" techniques (SBS), "Sequencing-by-Hybridization" (SBH), or "Sequencing-by-Ligation" (SBL) techniques. Of these, SBS methods provide many desirable advantages over previously employed sequencing methods that include, but are not limited to the massively parallel generation of a large volume of high quality sequence information at a low cost relative to previous techniques. The term "massively parallel" as used herein generally refers to the simultaneous generation of sequence information from many different template molecules in parallel where the individual template molecule or population of substantially identical template molecules are separated or compartmentalized and simultaneously exposed to sequencing processes which may include a iterative series of reactions thereby producing an independent sequence read representing the nucleic acid composition of each template molecule. In other words, the advantage includes the ability to simultaneously sequence multiple nucleic acid elements associated with many different samples or different nucleic acid elements existing within a sample.

Typical embodiments of SBS methods comprise the stepwise synthesis of a single strand of polynucleotide molecule complementary to a template nucleic acid molecule whose

IDT Exhibit 1020.4969

nucleotide sequence composition is to be determined. For example, SBS techniques typically operate by adding a single nucleic acid (also referred to as a nucleotide) species to a nascent polynucleotide molecule complementary to a nucleic acid species of a template molecule at a corresponding sequence position. The addition of the nucleic acid species to the nascent molecule is generally detected using a variety of methods known in the art that include, but are not limited to what are referred to as pyrosequencing or fluorescent detection methods such as those that employ reversible terminators or energy transfer labels including fluorescent resonant energy transfer dyes (FRET). Typically, the process is iterative until a complete (i.e. all sequence positions are represented) or desired sequence length complementary to the template is synthesized.

Further, as described above many embodiments of SBS are enabled to perform sequencing operations in a massively parallel manner. For example, some embodiments of SBS methods are performed using instrumentation that automates one or more steps or operation associated with the preparation and/or sequencing methods. Some instruments employ elements such as plates with wells or other type of microreactor configuration that provide the ability to perform reactions in each of the wells or microreactors simultaneously. Additional examples of SBS techniques as well as systems and methods for massively parallel sequencing are described in US Patent Nos. 6,274,320; 6,258,568; 6,210,891, 7,211,390; 7,244,559; 7,264,929; 7,335,762; and 7,323,305 each of which is hereby incorporated by reference herein in its entirety for all purposes; and US Patent Application Serial No. 11/195,254, which is hereby incorporated by reference herein in its entirety for all purposes.

It may also be desirable in some embodiments of SBS, to generate many substantially identical copies of each template nucleic acid element that for instance, provides a stronger signal when one or more nucleotide species is incorporated in each nascent molecule in a population comprising the copies of a template nucleic acid molecule. There are many techniques known in the art for generating copies of nucleic acid molecules such as, for instance, amplification using what are referred to as bacterial vectors, "Rolling Circle" amplification (described in US Patent Nos. 6,274,320 and 7,211,390, incorporated by reference above), isothermal amplification techniques, and Polymerase Chain Reaction (PCR) methods, each of the techniques are applicable for use with the presently described invention. One PCR technique that is particularly amenable to high throughput applications include what are referred to as emulsion PCR methods.

IDT Exhibit 1020.4970

Typical embodiments of emulsion PCR methods include creating stable emulsion of two immiscible substances and are resistant to blending together where one substance is dispersed within a second substance. The emulsions may include droplets suspended within another fluid and are sometimes also referred to as compartments, microcapsules, microreactors, microenvironments, or other name commonly used in the related art. The droplets may range in size depending on the composition of the emulsion components and formation technique employed. The described emulsions create the microenvironments within which chemical reactions, such as PCR, may be performed. For example, template nucleic acids and all reagents necessary to perform a desired PCR reaction may be encapsulated and chemically isolated in the droplets of an emulsion. Thermo cycling operations typical of PCR methods may be executed using the droplets to amplify an encapsulated nucleic acid template resulting in the generation of a population comprising many substantially identical copies of the template nucleic acid. Also in the present example, some or all of the described droplets may further encapsulate a solid substrate such as a bead for attachment of nucleic acids, reagents, labels, or other molecules of interest.

Embodiments of an emulsion useful with the presently described invention may include a very high density of droplets or microcapsules enabling the described chemical reactions to be performed in a massively parallel way. Additional examples of emulsions and their uses for sequencing applications are described in US Patent Application Serial Nos. 10/861,930; 10/866,392; 10/767,899; 11/045,678 each of which are hereby incorporated by reference herein in its entirety for all purposes.

Those of ordinary skill in the related art will appreciate that advantages provided by the massively parallel nature of the amplification and sequencing methods described herein may be particularly to amenable for processing what may be referred to as a "Multiplex" sample. For example, a multiplex composition may include representatives from multiple samples such as samples from multiple individuals. It may be desirable in many applications to combine multiple samples into a single multiplexed sample that may be processed in one operation as opposed to processing each sample separately. Thus the result may typically include a substantial savings in reagent, labor, and instrument usage and cost as well as a significant savings in processing time invested. The described advantages of multiplex processing become more pronounced as the numbers of individual samples increase. Further, multiplex processing has application in research as well as diagnostic contexts. For example, it may be desirable in many applications to employ a single multiplexed sample in an

IDT Exhibit 1020.4971

amplification reaction and subsequently processing the amplified multiplex composition in a single sequencing run.

One problem associated with processing a multiplex composition then becomes identifying the association between each sample of origin and the sequence data generated from a template molecule derived from said sample. A solution to this problem includes associating an identifier such as a nucleic acid sequence that specifically identifies the association of each template molecule with its sample of origin. An advantage of this solution is that the sequence information of the associated nucleic acid sequence is embedded in the sequence data generated from the template molecule and may be bioinformatically analyzed to associate the sequence data with its sample of origin.

Previous studies have described associating nucleic acid sequence identifiers with 5' primers coupled with target sequences for multiplex processing. One such study is that of Binladen et al. (Binladen J, Gilbert MTP, Bollback JP, Panitz F, Bendixen C (2007) The use of coded PCR Primers Enables High-Throughput Sequencing of Multiple Homolog Amplification Products by Parallel 454 Sequencing. PLoS ONE 2(2): e197.doi:10.1371/journal.pone.0000197 (published online February 14, 2007, which is hereby incorporated by reference herein in its entirety for all purposes). As mentioned above, Binladen et al. describe associating short sequence identifiers with target sequences to be processed in a multiplex sample producing sequence data that is subsequently bioinformatically analyzed to associate the short identifiers with their sample of origin. However, there are limitations to simply attaching a nucleic acid identifier of generic sequence composition to a template molecule and identifying the sequence of said identifier in the generated sequence data. Of primary concern is the introduction of error into the sequence data from various mechanisms. Such mechanisms typically work in combination with each other and are generally not individually identifiable from the sequence data. Thus because of introduced error, an end user may not be able to identify the association between the sequence data with its sample of origin, or possibly worse fail to identify that an error has occurred and mis-assign sequence data to a sample of origin that is incorrect.

There are two important sources of error introduction to consider, although other sources may also exist. First is error introduced by the sequencing operation that may in some cases be referred to a "flow error". For example, flow error may include polymerase errors that include incorporation of an incorrect nucleotide species by a polymerase enzyme. A sequencing operation may also introduce what may be referred to as phasic synchrony error that include what are referred to as "carry forward" and "incomplete extension" (the

IDT Exhibit 1020.4972

combination of phasic synchrony error is sometimes referred to as CAFIE error). Phasic synchrony error and methods of correction are further described in PCT Application Serial No. US2007/004187, titled "System and Method for Correcting Primer Extension Errors in Nucleic Acid Sequence Data", filed February 15, 2007 which is hereby incorporated by reference herein in its entirety for all purposes.

Second is error introduced from processes that are independent of the sequencing operations such as primer synthesis or amplification error. For example, oligonucleotide primers synthesized for PCR may include one or more UID elements of the presently described invention, where error may be introduced in the synthesis of the primer/UID element that is then employed as a sequencing template. High fidelity sequencing of the UID element faithfully reproduces the synthesized error in sequence data. Also in the present example, polymerase enzymes commonly employed in PCR methods are known for having a measure of replication error, where for instance an error in replication may be introduced by the polymerase in 1 of every 10,000; 100,000; or 1,000,000 bases amplified.

Therefore, it is significantly advantageous to employ unique identifiers that are 1) resistant to error introduction; 2) enable detection of introduced error; and 3) enable correction of introduced error. The presently described invention addresses these problems and provides systems and methods for associating unique identifiers that provide better recognition and identification characteristics resulting in improved data quality and experimental efficiency.

SUMMARY OF THE INVENTION

Embodiments of the invention relate to the determination of the sequence of nucleic acids. More particularly, embodiments of the invention relate to methods and systems for correcting errors in data obtained during the sequencing of nucleic acids and associating the nucleic acids with their origin.

An embodiment of an identifier element for identifying an origin of a template nucleic acid molecule is described that comprises 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, where 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.

Also, an embodiment of a method for identifying an origin of a template nucleic acid molecule is described that comprises the steps of identifying a first identifier sequence from

IDT Exhibit 1020.4973

sequence data generated from a template nucleic acid molecule; 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.

In some implementations, the method further comprises 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.

Further, an embodiment of a kit for identifying an origin of a template nucleic acid molecule is described that comprises 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.

In addition, an embodiment of a computer comprising executable code stored in system memory is described where 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.

The above embodiments and implementations are not necessarily inclusive or exclusive of each other and may be combined in any manner that is non-conflicting and otherwise possible, whether they be presented in association with a same, or a different, embodiment or implementation. The description of one embodiment or implementation is not intended to be limiting with respect to other embodiments and/or implementations. Also, any one or more function, step, operation, or technique described elsewhere in this specification may, in alternative implementations, be combined with any one or more

IDT Exhibit 1020.4974

function, step, operation, or technique described in the summary. Thus, the above embodiment and implementations are illustrative rather than limiting.

BRIEF DESCRIPTION OF THE DRAWINGS

The above and further features will be more clearly appreciated from the following detailed description when taken in conjunction with the accompanying drawings. In the drawings, like reference numerals indicate like structures, elements, or method steps and the leftmost digit of a reference numeral indicates the number of the figure in which the references element first appears (for example, element 160 appears first in Figure 1). All of these conventions, however, are intended to be typical or illustrative, rather than limiting.

Figure 1 is a functional block diagram of one embodiment of a sequencing instrument and computer system amenable for use with the presently described invention;

Figure 2A is a simplified graphical representation of one embodiment of an adaptor element amenable for use with genomic libraries comprising a UID component;

Figure 2B is a simplified graphical representation of one embodiment of an adaptor element amenable for use with amplicons comprising a UID component; and

Figure 3 is a simplified graphical representation of one embodiment of computed error balls representing compatibility of UID elements of different sequence composition.

DETAILED DESCRIPTION OF THE INVENTION

As will be described in greater detail below, embodiments of the presently described invention include systems and methods for associating a unique identifier hereafter referred to as a UID element with one or more nucleic acid molecules from a sample. The UID elements are resistant to introduced error in sequence data, and enable detection and correction of error. Further, the invention includes combining or pooling those UID associated nucleic acid molecules with similarly UID associated (sometimes also referred to as "labeled") nucleic acid molecules from one or more other samples, and sequencing each nucleic acid molecule in the pooled sample to generate sequence data for each nucleic acid. The presently described invention further includes systems and methods for designing the sequence composition for each UID element and analyzing the sequence data of each nucleic acid to identify an embedded UID sequence code and associating said code with the sample identity.

a. General

The terms "flowgram" and "pyrogram" may be used interchangeably herein and generally refer to a graphical representation of sequence data generated by SBS methods.

Further, the term "read" or "sequence read" as used herein generally refers to the entire sequence data obtained from a single nucleic acid template molecule or a population of a plurality of substantially identical copies of the template nucleic acid molecule.

The terms "run" or "sequencing run" as used herein generally refer to a series of sequencing reactions performed in a sequencing operation of one or more template nucleic acid molecule.

The term "flow" as used herein generally refers to a serial or iterative cycle of addition of solution to an environment comprising a template nucleic acid molecule, where the solution may include a nucleotide species for addition to a nascent molecule or other reagent such as buffers or enzymes that may be employed to reduce carryover or noise effects from previous flow cycles of nucleotide species.

The term "flow cycle" as used herein generally refers to a sequential series of flows where a nucleotide species is flowed once during the cycle (i.e. a flow cycle may include a sequential addition in the order of T, A, C, G nucleotide species, although other sequence combinations are also considered part of the definition). Typically the flow cycle is a repeating cycle having the same sequence of flows from cycle to cycle.

The term "read length" as used herein generally refers to an upper limit of the length of a template molecule that may be reliably sequenced. There are numerous factors that contribute to the read length of a system and/or process including, but not limited to the degree of GC content in a template nucleic acid molecule.

A "nascent molecule" generally refers to a DNA strand which is being extended by the template-dependent DNA polymerase by incorporation of nucleotide species which are complementary to the corresponding nucleotide species in the template molecule.

The terms "template nucleic acid", "template molecule", "target nucleic acid", or "target molecule" generally refer to a nucleic acid molecule that is the subject of a sequencing reaction from which sequence data or information is generated.

The term "nucleotide species" as used herein generally refers to the identity of a nucleic acid monomer including purines (Adenine, Guanine) and pyrimidines (Cytosine, Uracil, Thymine) typically incorporated into a nascent nucleic acid molecule.

The term "monomer repeat" or "homopolymers" as used herein generally refers to two or more sequence positions comprising the same nucleotide species (i.e. a repeated nucleotide species).

IDT Exhibit 1020.4976

The term "homogeneous extension", as used herein, generally refers to the relationship or phase of an extension reaction where each member of a population of substantially identical template molecules is homogenously performing the same extension step in the reaction.

The term "completion efficiency" as used herein generally refers to the percentage of nascent molecules that are properly extended during a given flow.

The term "incomplete extension rate" as used herein generally refers to the ratio of the number of nascent molecules that fail to be properly extended over the number of all nascent molecules.

The term "genomic library" or "shotgun library" as used herein generally refers to a collection of molecules derived from and/or representing an entire genome (i.e. all regions of a genome) of an organism or individual.

The term "amplicon" as used herein generally refers to selected amplification products such as those produced from Polymerase Chain Reaction or Ligase Chain Reaction techniques.

The term "keypass" or "keypass mapping" as used herein generally refers to a nucleic acid "key element" associated with a template nucleic acid molecule in a known location (i.e. typically included in a ligated adaptor element) comprising known sequence composition that is employed as a quality control reference for sequence data generated from template molecules. The sequence data passes the quality control if it includes the known sequence composition associated with a Key element in the correct location.

The term "blunt end" or "blunt ended" as used herein generally refers to a linear double stranded nucleic acid molecule having an end that terminates with a pair of complementary nucleotide base species, where a pair of blunt ends are always compatible for ligation to each other.

Some exemplary embodiments of systems and methods associated with sample preparation and processing, generation of sequence data, and analysis of sequence data are generally described below, some or all of which are amenable for use with embodiments of the presently described invention. In particular the exemplary embodiments of systems and methods for preparation of template nucleic acid molecules, amplification of template molecules, generating target specific amplicons and/or genomic libraries, sequencing methods and instrumentation, and computer systems are described.

In typical embodiments, the nucleic acid molecules derived from an experimental or diagnostic sample must be prepared and processed from its raw form into template molecules

IDT Exhibit 1020.4977

amenable for high throughput sequencing. The processing methods may vary from application to application resulting in template molecules comprising various characteristics. For example, in some embodiments of high throughput sequencing it is preferable to generate template molecules with a sequence or read length that is at least the length a particular sequencing method can accurately produce sequence data for. In the present example, the length may include a range of about 25-30 base pairs, about 30-50 base pairs, about 50-100 base pairs, about 100-200 base pairs, about 200-300 base pairs, or about 350-500 base pairs, or other length amenable for a particular sequencing application. In some embodiments, nucleic acids from a sample, such as a genomic sample, are fragmented using a number of methods known to those of ordinary skill in the art. In preferred embodiments, methods that randomly fragment (i.e. do not select for specific sequences or regions) nucleic acids are employed that include what is referred to as nebulization or sonication. It will however, be appreciated that other methods of fragmentation such as digestion using restriction endonucleases may be employed for fragmentation purposes. Also in the present example, some processing methods may employ size selection methods known in the art to selectively isolate nucleic acid fragments of the desired length.

Also, it is preferable in some embodiments to associate additional functional elements with each template nucleic acid molecule. The elements may be employed for a variety of functions including, but not limited to, primer sequences for amplification and/or sequencing methods, quality control elements, unique identifiers that encode various associations such as with a sample of origin or patient, or other functional element. For example, some embodiments may associate priming sequence elements or regions comprising complementary sequence composition to primer sequences employed for amplification and/or sequencing. Further, the same elements may be employed for what may be referred to as "strand selection" and immobilization of nucleic acid molecules to a solid phase substrate. In the present example, two sets of priming sequence regions (hereafter referred to as priming sequence A, and priming sequence B) may be employed for strand selection where only single strands having one copy of priming sequence A and one copy of priming sequence B is selected and included as the prepared sample. The same priming sequence regions may be employed in methods for amplification and immobilization where, for instance priming sequence B may be immobilized upon a solid substrate and amplified products are extended therefrom.

Additional examples of sample processing for fragmentation, strand selection, and addition of functional elements and adaptors are described in U.S. Patent Application Serial

IDT Exhibit 1020.4978

No. 10/767,894, titled "Method for preparing single-stranded DNA libraries", filed January 28, 2004; and U.S. Provisional Application Serial No. 60/941,381, titled "System and Method for Identification of Individual Samples from a Multiplex Mixture", filed June 1, 2007, each of which is hereby incorporated by reference herein in its entirety for all purposes.

Various examples of systems and methods for performing amplification of template nucleic acid molecules to generate populations of substantially identical copies are described. It will be apparent to those of ordinary skill that it is desirable in some embodiments of SBS to generate many copies of each nucleic acid element to generate a stronger signal when one or more nucleotide species is incorporated into each nascent molecule associated with a copy of the template molecule. There are many techniques known in the art for generating copies of nucleic acid molecules such as, for instance, amplification using what are referred to as bacterial vectors, "Rolling Circle" amplification (described in US Patent Nos. 6,274,320 and 7,211,390, incorporated by reference above) and Polymerase Chain Reaction (PCR) methods, each of the techniques are applicable for use with the presently described invention. One PCR technique that is particularly amenable to high throughput applications include what are referred to as emulsion PCR methods (also referred to as emPCR™ methods).

Typical embodiments of emulsion PCR methods include creating a stable emulsion of two immiscible substances creating aqueous droplets within which reactions may occur. In particular, the aqueous droplets of an emulsion amenable for use in PCR methods may include a first fluid such as a water based fluid suspended or dispersed in what may be referred to as a discontinuous phase within another fluid such as an oil based fluid. Further, some emulsion embodiments may employ surfactants that act to stabilize the emulsion that may be particularly useful for specific processing methods such as PCR. Some embodiments of surfactant may include non-ionic surfactants such as sorbitan monooleate (also referred to as Span™ 80), polyoxyethylenesorbitsan monooleate (also referred to as Tween™ 80), or in some preferred embodiments dimethicone copolyol (also referred to as Abil® EM90), polysiloxane, polyalkyl polyether copolymer, polyglycerol esters, poloxamers, and PVP/hexadecane copolymers (also referred to as Unimer U-151), or in more preferred embodiments a high molecular weight silicone polyether in cyclopentasiloxane (also referred to as DC 5225C available from Dow Corning).

The droplets of an emulsion may also be referred to as compartments, microcapsules, microreactors, microenvironments, or other name commonly used in the related art. The aqueous droplets may range in size depending on the composition of the emulsion components or composition, contents contained therein, and formation technique employed.

IDT Exhibit 1020.4979

The described emulsions create the microenvironments within which chemical reactions, such as PCR, may be performed. For example, template nucleic acids and all reagents necessary to perform a desired PCR reaction may be encapsulated and chemically isolated in the droplets of an emulsion. Additional surfactants or other stabilizing agent may be employed in some embodiments to promote additional stability of the droplets as described above. Thermocycling operations typical of PCR methods may be executed using the droplets to amplify an encapsulated nucleic acid template resulting in the generation of a population comprising many substantially identical copies of the template nucleic acid. In some embodiments, the population within the droplet may be referred to as a "clonally isolated", "compartmentalized", "sequestered", "encapsulated", or "localized" population. Also in the present example, some or all of the described droplets may further encapsulate a solid substrate such as a bead for attachment of template or other type of nucleic acids, reagents, labels, or other molecules of interest.

Embodiments of an emulsion useful with the presently described invention may include a very high density of droplets or microcapsules enabling the described chemical reactions to be performed in a massively parallel way. Additional examples of emulsions employed for amplification and their uses for sequencing applications are described in US Patent Application Serial Nos. 10/861,930; 10/866,392; 10/767,899; 11/045,678 each of which are hereby incorporated by reference herein in its entirety for all purposes.

Also, an exemplary embodiment for generating target specific amplicons for sequencing is described that includes using sets of nucleic acid primers to amplify a selected target region or regions from a sample comprising the target nucleic acid. Further, the sample may include a population of nucleic acid molecules that are known or suspected to contain sequence variants and the primers may be employed to amplify and provide insight into the distribution of sequence variants in the sample.

For example a method for identifying a sequence variant by specific amplification and sequencing of multiple alleles in a nucleic acid sample may be performed. The nucleic acid is first subjected to amplification by a pair of PCR primers designed to amplify a region surrounding the region of interest or segment common to the nucleic acid population. Each of the products of the PCR reaction (amplicons) is subsequently further amplified individually in separate reaction vessels such as an emulsion based vessel described above. The resulting amplicons (referred to herein as second amplicons), each derived from one member of the first population of amplicons, are sequenced and the collection of sequences, from different emulsion PCR amplicons, are used to determine an allelic frequency.

IDT Exhibit 1020.4980

Some advantages of the described target specific amplification and sequencing methods include a higher level of sensitivity than previously achieved. Further, embodiments that employ high throughput sequencing instrumentation such as for instance embodiments that employ what is referred to as a PicoTiterPlate® array of wells provided by 454 Life Sciences Corporation, the described methods can be employed to sequence over 100,000 or over 300,000 different copies of an allele per run or experiment. Also, the described methods provide a sensitivity of detection of low abundance alleles which may represent 1% or less of the allelic variants. Another advantage of the methods includes generating data comprising the sequence of the analyzed region. Importantly, it is not necessary to have prior knowledge of the sequence of the locus being analyzed.

Additional examples of target specific amplicons for sequencing are described in U.S. Patent Application Serial No. 11/104,781, titled "Methods for determining sequence variants using ultra-deep sequencing", filed April 12, 2005, which is hereby incorporated by reference herein in its entirety for all purposes.

Further, embodiments of sequencing may include Sanger type techniques, what is referred to as polony sequencing techniques, nanopore and other single molecule detection techniques, or reversible terminator techniques. As described above a preferred technique may include Sequencing by Synthesis methods. For example, some SBS embodiments sequence populations of substantially identical copies of a nucleic acid template and typically employ one or more oligonucleotide primers designed to anneal to a predetermined, complementary position of the sample template molecule or one or more adaptors attached to the template molecule. The primer/template complex is presented with a nucleotide species in the presence of a nucleic acid polymerase enzyme. If the nucleotide species is complementary to the nucleic acid species corresponding to a sequence position on the sample template molecule that is directly adjacent to the 3' end of the oligonucleotide primer, then the polymerase will extend the primer with the nucleotide species. Alternatively, in some embodiments the primer/template complex is presented with a plurality of nucleotide species of interest (typically A, G, C, and T) at once, and the nucleotide species that is complementary at the corresponding sequence position on the sample template molecule directly adjacent to the 3' end of the oligonucleotide primer is incorporated. In either of the described embodiments, the nucleotide species may be chemically blocked (such as at the 3'-O position) to prevent further extension, and need to be deblocked prior to the next round of synthesis. It will also be appreciated that the process of adding a nucleotide species to the

IDT Exhibit 1020.4981

end of a nascent molecule is substantially the same as that described above for addition to the end of a primer.

As described above, incorporation of the nucleotide species can be detected by a variety of methods known in the art, e.g. by detecting the release of pyrophosphate (PPi) (examples described in US Patent Nos. 6,210,891; 6,258,568; and 6,828,100, each of which is hereby incorporated by reference herein in its entirety for all purposes), or via detectable labels bound to the nucleotides. Some examples of detectable labels include but are not limited to mass tags and fluorescent or chemiluminescent labels. In typical embodiments, unincorporated nucleotides are removed, for example by washing. Further, in some embodiments the unincorporated nucleotides may be subjected to enzymatic degradation such as, for instance, degradation using the apyrase enzyme as described in U.S. Provisional Patent Application Serial No. 60/946,743, titled System and Method For Adaptive Reagent Control in Nucleic Acid Sequencing, filed June 28, 2007, which is hereby incorporated by reference herein in its entirety for all purposes. In the embodiments where detectable labels are used, they will typically have to be inactivated (e.g. by chemical cleavage or photobleaching) prior to the following cycle of synthesis. The next sequence position in the template/polymerase complex can then be queried with another nucleotide species, or a plurality of nucleotide species of interest, as described above. Repeated cycles of nucleotide addition, extension, signal acquisition, and washing result in a determination of the nucleotide sequence of the template strand. Continuing with the present example, a large number or population of substantially identical template molecules (e.g. 10^3, 10^4, 10^5, 10^6 or 10^7 molecules) are typically analyzed simultaneously in any one sequencing reaction, in order to achieve a signal which is strong enough for reliable detection.

In addition, it may be advantageous in some embodiments to improve the read length capabilities and qualities of a sequencing process by employing what may be referred to as a "paired-end" sequencing strategy. For example, some embodiments of sequencing method have limitations on the total length of molecule from which a high quality and reliable read may be generated. In other words, the total number of sequence positions for a reliable read length may not exceed 25, 50, 100, or 150 bases depending on the sequencing embodiment employed. A paired-end sequencing strategy extends reliable read length by separately sequencing each end of a molecule (sometimes referred to as a "tag" end) that comprise a fragment of an original template nucleic acid molecule at each end joined in the center by a linker sequence. The original positional relationship of the template fragments is known and thus the data from the sequence reads may be re-combined into a single read having a longer

IDT Exhibit 1020.4982

high quality read length. Further examples of paired-end sequencing embodiments are described in US Patent Application Serial No. 11/448,462, titled "Paired end sequencing", filed June 6, 2006, and in US Provisional Patent Application Serial No. 60/026,319, titled "Paired end sequencing", filed February 5, 2008, each of which is hereby incorporated by reference herein in its entirety for all purposes.

Some examples of SBS apparatus may implement some or all of the methods described above may include one or more of a detection device such as a charge coupled device (i.e. CCD camera), a microfluidics chamber or flow cell, a reaction substrate, and/or a pump and flow valves. Taking the example of pyrophosphate based sequencing, embodiments of an apparatus may employ a chemiluminescent detection strategy that produces an inherently low level of background noise.

In some embodiments, the reaction substrate for sequencing may include what is referred to as a PicoTiterPlate® array (also referred to as a PTP® plate) formed from a fiber optics faceplate that is acid-etched to yield hundreds of thousands of very small wells each enabled to hold a population of substantially identical template molecules. In some embodiments, each population of substantially identical template molecule may be disposed upon a solid substrate such as a bead, each of which may be disposed in one of said wells. For example, an apparatus may include a reagent delivery element for providing fluid reagents to the PTP plate holders, as well as a CCD type detection device enabled to collect photons of light emitted from each well on the PTP plate. Further examples of apparatus and methods for performing SBS type sequencing and pyrophosphate sequencing are described in US Patent No 7,323,305 and US Patent Application Serial No. 11/195,254 both of which are incorporated by reference above.

In addition, systems and methods may be employed that automate one or more sample preparation processes, such as the emPCR™ process described above. For example, microfluidic technologies may be employed to provide a low cost, disposable solution for generating an emulsion for emPCR processing, performing PCR Thermocycling operations, and enriching for successfully prepared populations of nucleic acid molecules for sequencing. Examples of microfluidic systems for sample preparation are described in U.S. Provisional Patent Application Serial No. 60/915,968, titled "System and Method for Microfluidic Control of Nucleic Acid amplification and Segregation", filed May 4, 2007, which is hereby incorporated by reference herein in its entirety for all purposes.

Also, the systems and methods of the presently described embodiments of the invention may include implementation of some design, analysis, or other operation using a

IDT Exhibit 1020.4983

computer readable medium stored for execution on a computer system. For example, several embodiments are described in detail below to process detected signals and/or analyze data generated using SBS systems and methods where the processing and analysis embodiments are implementable on computer systems.

An exemplary embodiment of a computer system for use with the presently described invention may include any type of computer platform such as a workstation, a personal computer, a server, or any other present or future computer. Computers typically include known components such as a processor, an operating system, system memory, memory storage devices, input-output controllers, input-output devices, and display devices. It will be understood by those of ordinary skill in the relevant art that there are many possible configurations and components of a computer and may also include cache memory, a data backup unit, and many other devices.

Display devices may include display devices that provide visual information, this information typically may be logically and/or physically organized as an array of pixels. An interface controller may also be included that may comprise any of a variety of known or future software programs for providing input and output interfaces. For example, interfaces may include what are generally referred to as "Graphical User Interfaces" (often referred to as GUI's) that provide one or more graphical representations to a user. Interfaces are typically enabled to accept user inputs using means of selection or input known to those of ordinary skill in the related art.

In the same or alternative embodiments, applications on a computer may employ an interface that includes what are referred to as "command line interfaces" (often referred to as CLI's). CLI's typically provide a text based interaction between an application and a user. Typically, command line interfaces present output and receive input as lines of text through display devices. For example, some implementations may include what are referred to as a "shell" such as Unix Shells known to those of ordinary skill in the related art, or Microsoft Windows Powershell that employs object-oriented type programming architectures such as the Microsoft .NET framework.

Those of ordinary skill in the related art will appreciate that interfaces may include one or more GUI's, CLI's or a combination thereof.

A processor may include a commercially available processor such as a Centrino®, Core™ 2, Itanium® or Pentium® processor made by Intel Corporation, a SPARC® processor made by Sun Microsystems, an Athalon™ or Opteron™ processor made by AMD corporation, or it may be one of other processors that are or will become available. Some

embodiments of a processor may include what is referred to as Multi-core processor and/or be enabled to employ parallel processing technology in a single or multi-core configuration. For example, a multi-core architecture typically comprises two or more processor "execution cores". In the present example each execution core may perform as an independent processor that enables parallel execution of multiple threads. In addition, those of ordinary skill in the related will appreciate that a processor may be configured in what is generally referred to as 32 or 64 bit architectures, or other architectural configurations now known or that may be developed in the future.

A processor typically executes an operating system, which may be, for example, a Windows®-type operating system (such as Windows® XP or Windows Vista®) from the Microsoft Corporation; the Mac OS X operating system from Apple Computer Corp. (such as 7.5 Mac OS X v10.4 "Tiger" or 7.6 Mac OS X v10.5 "Leopard" operating systems); a Unix® or Linux-type operating system available from many vendors or what is referred to as an open source; another or a future operating system; or some combination thereof. An operating system interfaces with firmware and hardware in a well-known manner, and facilitates the processor in coordinating and executing the functions of various computer programs that may be written in a variety of programming languages. An operating system, typically in cooperation with a processor, coordinates and executes functions of the other components of a computer. An operating system also provides scheduling, input-output control, file and data management, memory management, and communication control and related services, all in accordance with known techniques.

System memory may include any of a variety of known or future memory storage devices. Examples include any commonly available random access memory (RAM), magnetic medium such as a resident hard disk or tape, an optical medium such as a read and write compact disc, or other memory storage device. Memory storage devices may include any of a variety of known or future devices, including a compact disk drive, a tape drive, a removable hard disk drive, USB or flash drive, or a diskette drive. Such types of memory storage devices typically read from, and/or write to, a program storage medium (not shown) such as, respectively, a compact disk, magnetic tape, removable hard disk, USB or flash drive, or floppy diskette. Any of these program storage media, or others now in use or that may later be developed, may be considered a computer program product. As will be appreciated, these program storage media typically store a computer software program and/or data. Computer software programs, also called computer control logic, typically are stored in

IDT Exhibit 1020.4985

system memory and/or the program storage device used in conjunction with memory storage device.

In some embodiments, a computer program product is described comprising a computer usable medium having control logic (computer software program, including program code) stored therein. The control logic, when executed by a processor, causes the processor to perform functions described herein. In other embodiments, some functions are implemented primarily in hardware using, for example, a hardware state machine. Implementation of the hardware state machine so as to perform the functions described herein will be apparent to those skilled in the relevant arts.

Input-output controllers could include any of a variety of known devices for accepting and processing information from a user, whether a human or a machine, whether local or remote. Such devices include, for example, modem cards, wireless cards, network interface cards, sound cards, or other types of controllers for any of a variety of known input devices. Output controllers could include controllers for any of a variety of known display devices for presenting information to a user, whether a human or a machine, whether local or remote. In the presently described embodiment, the functional elements of a computer communicate with each other via a system bus. Some embodiments of a computer may communicate with some functional elements using network or other types of remote communications.

As will be evident to those skilled in the relevant art, an instrument control and/or a data processing application, if implemented in software, may be loaded into and executed from system memory and/or a memory storage device. All or portions of the instrument control and/or data processing applications may also reside in a read-only memory or similar device of the memory storage device, such devices not requiring that the instrument control and/or data processing applications first be loaded through input-output controllers. It will be understood by those skilled in the relevant art that the instrument control and/or data processing applications, or portions of it, may be loaded by a processor in a known manner into system memory, or cache memory, or both, as advantageous for execution.

Also a computer may include one or more library files, experiment data files, and an internet client stored in system memory. For example, experiment data could include data related to one or more experiments or assays such as detected signal values, or other values associated with one or more SBS experiments or processes. Additionally, an internet client may include an application enabled to accesses a remote service on another computer using a network and may for instance comprise what are generally referred to as "Web Browsers". In the present example some commonly employed web browsers include Microsoft® Internet

IDT Exhibit 1020.4986

Explorer 7 available from Microsoft Corporation, Mozilla Firefox® 2 from the Mozilla Corporation, Safari 1.2 from Apple Computer Corp., or other type of web browser currently known in the art or to be developed in the future. Also, in the same or other embodiments an internet client may include, or could be an element of, specialized software applications enabled to access remote information via a network such as a data processing application for SBS applications.

A network may include one or more of the many various types of networks well known to those of ordinary skill in the art. For example, a network may include a local or wide area network that employs what is commonly referred to as a TCP/IP protocol suite to communicate. A network may include a network comprising a worldwide system of interconnected computer networks that is commonly referred to as the internet, or could also include various intranet architectures. Those of ordinary skill in the related arts will also appreciate that some users in networked environments may prefer to employ what are generally referred to as "firewalls" (also sometimes referred to as Packet Filters, or Border Protection Devices) to control information traffic to and from hardware and/or software systems. For example, firewalls may comprise hardware or software elements or some combination thereof and are typically designed to enforce security policies put in place by users, such as for instance network administrators, etc.

b. Embodiments of the presently described invention

As described above, the presently described invention comprises associating one or more embodiments of a UID element having a known and identifiable sequence composition with a sample, and coupling the embodiments of UID element with template nucleic acid molecules from the associated samples. The UID coupled template nucleic acid molecules from a number of different samples are pooled into a single "Multiplexed" sample or composition that can then be efficiently processed to produce sequence data for each UID coupled template nucleic acid molecule. The sequence data for each template nucleic acid is de-convoluted to identify the sequence composition of coupled UID elements and association with sample of origin identified. For example, a multiplexed composition may include representatives from about 384 samples, about 96 samples, about 50 samples, about 20 samples, about 16 samples, about 10 samples, or other number of samples. Each sample may be associated with a different experimental condition, treatment, species, or individual in a research context. Similarly, each sample may be associated with a different tissue, cell, individual, condition, or treatment in a diagnostic context. Those of ordinary skill in the

IDT Exhibit 1020.4987

related art will appreciate that the numbers of samples listed above are for the purposes of example and thus should not be considered limiting.

Typically, systems and methods are employed for processing samples to generate sequence data as well as for interpretation of the sequence data. Figure 1 provides an illustrative example of sequencing instrument 100 employed to execute sequencing processes using reaction substrate 105 that for instance may include the PTP® plate substrate described above. Also illustrated in Figure 1 is computer 130 that may for instance execute system software or firmware for processing as well as perform analysis functions. In the example of Figure 1, computer 130 may also store application 135 in system memory for execution, where application 135 may perform some or all of the data processing functions described herein. It will also be understood that application 135 may be stored on other computer or server type structures for execution and perform some or all of its functions remotely communicating over networks or transferring information via standard media. For instance, processed target molecules in a multiplex sample may be loaded onto reaction substrate 105 by user 101 or some automated embodiment then sequenced in a massively parallel manner using sequencing instrument 100 to produce sequence data representing the sequence composition of each target molecule. Importantly, user 101 may include any user such as independent researcher, university, or corporate entity. In the present example, sequencing instrument 100, reaction substrate 105, and/or computer 130 may include some or all of the components and characteristics of the embodiments generally described above.

In preferred embodiments, the sequence composition of each UID element is easily identifiable and resistant to introduced error from sequencing processes. Some embodiments of UID element comprise a unique sequence composition of nucleic acid species that has minimal sequence similarity to a naturally occurring sequence. Alternatively, embodiments of a UID element may include some degree of sequence similarity to naturally occurring sequence.

Also, in preferred embodiments the position of each UID element is known relative to some feature of the template nucleic acid molecule and/or adaptor elements coupled to the template molecule. Having a known position of each UID is useful for finding the UID element in sequence data and interpretation of the UID sequence composition for possible errors and subsequent association with the sample of origin.

For example, some features useful as anchors for positional relationship to UID elements may include, but are not limited to the length of the template molecule (i.e. the UID element is known to be so many sequence positions from the 5' or 3' end), recognizable sequence

IDT Exhibit 1020.4988

markers such as a Key element (described in greater detail below) and/or one or more primer elements positioned adjacent to a UID element. In the present example, The Key and primer elements generally comprise a known sequence composition that typically does not vary from sample to sample in the multiplex composition and may be employed as positional references for searching for the UID element. An analysis algorithm implemented by application 135 may be executed on computer 130 to analyze generated sequence data for each UID coupled template to identify the more easily recognizable Key and/or primer elements, and extrapolate from those positions to identify a sequence region presumed to include the sequence of the UID element. Application 135 may then process the sequence composition of the presumed region and possibly some distance away in the flanking regions to positively identify the UID element and its sequence composition.

Also, as will be described in greater detail below in some embodiments the sequence data generated from each Key and/or one or more primer elements may be analyzed to determine a measure of the relative error rate for the sequencing run. The measure of error rate may then be employed in the analysis of the sequence data generated for the UID element. For example, if the error rate is excessive and is above a predetermined threshold it may also be assumed that a similar rate of error exists in the sequence data generated for the UID element, and thus the sequence data for the entire template may be filtered out as suspect. Further, in embodiments where a UID element is coupled to each end of a linear template molecule an error rate may be established for each end and asymmetrically analyzed. Importantly, it will be appreciated that in some embodiments, particularly sequencing technology capable of producing "long" read lengths (i.e. of about 100 base pairs or greater) the error rate in the sequence data may differ between the 5' end and the 3' end.

In preferred embodiments, a UID element is associated with an adaptor enabled to operatively couple with the end of a template nucleic acid molecule. In typical high throughput sequencing applications it is desirable that the template nucleic acid molecules are linear where an adaptor may be coupled to each end. Figures 2A and 2B provide illustrative examples of embodiments of adaptor composition for various applications comprising one or more UID elements. It will, however, be appreciated that various adaptor configurations may be employed for different amplification and sequencing strategies. Figure 2A provides an illustrative example of adaptor element 200 that comprises an embodiment of an adaptor amenable for use with amplification and sequencing of Genomic Libraries. It will also be appreciated that adaptor element 200 may also be amenable for libraries of template molecules independently amplified with target specific sequences independently of the

IDT Exhibit 1020.4989

adaptor element described herein. Adaptor element 200 comprises several components that include primer 205, key 207, and UID 210. Also, Figure 2B provides an illustrative example of one embodiment of adaptor 220 amenable for use with amplification and sequencing of Amplicons. Adaptor element 220 comprises several similar components to adaptor 200 that include primer 205, key 207, UID 210, with the addition of target specific element 225. It will be appreciated that the relative arrangement of components provided in Figures 2A and 2B are for illustrative purposes and should not be considered limiting.

In some alternative embodiments, the UID 210 elements are not associated with adaptor elements as described above. Rather, the UID 210 elements may be considered separate elements that may be independently coupled to an already adapted template molecule, or non-adapted template molecule. This strategy may be useful in some circumstances to avoid negative effects associated with a particular step or assay. For example, it may be advantageous in some embodiments to ligate the UID 210 elements to each population of substantially identical template molecules after copies have been produced from an amplification step. By coupling the UID elements to the adapted template molecules post-amplification, errors introduced by the amplification method are avoided. In the present example, PCR amplification methods that employ polymerases are known to have a certain rates of introduced error based, at least in part, upon the type of polymerase or polymerase blends (i.e. a blend may include a mixture of what may be referred to as a "high fidelity" polymerase and a polymerase with "proof reading" capability) employed and the number of cycles of amplification.

It will also be appreciated that multiple embodiments of adaptor 200 or 220 may be employed with each template molecule, such as one embodiment of adaptor 200 or 220 at each end of a linear template molecule prepared for sequencing. However, in some embodiments the positional arrangement of elements within adaptor 200 or 220 may be reversed (i.e. the elements of adaptor 200 or 220 are in a palindromic arrangement from the example illustrated in Figure 2A or 2B) at the 3' end relative the arrangement of elements in adaptor 200 or 220 at the 5' end. For example, an embodiment of element 220 may be positioned on each end of substantially every template molecule from a library of amplicons in a multiplex composition, thus 2 embodiments of UID 210 may be employed in a combinatorial manner for identification which will be discussed in greater detail below.

Primer 205 may include a primer species (or a primer of a primer pair) such as is described above with respect to emulsion PCR embodiments (i.e. Primer A and Primer B). Also, primer 205 may include a primer species employed for an SBS sequencing reaction

IDT Exhibit 1020.4990

also as described above. Further, primer 205 may include what is referred to as a bipartite PCR/sequencing primer useable for both the emulsion PCR and SBS sequencing processes. Key 207 may include what may be referred to as a "discriminating key sequence" that refers to a short sequence of nucleotide species such as a combination of the four nucleotide species (i.e., A, C, G, T). Typically, key 207 may employed for quality control of sequence data, where for example key 207 may be located immediately adjacent primer 205 or within close proximity and include one of each of the four nucleotide species in a known sequence arrangement (i.e. TCAG). Therefore, the fidelity of the sequencing method should be represented in the sequence data for each of the 4 nucleotide species in key 207 and may pass quality control metrics if each of the 4 nucleotide species is faithfully represented. For example, an error for one of the nucleotide species represented in the sequence data generated from key 207 could indicate a problem in the sequencing process associated with that nucleotide species. Such error may be from mechanical failure of one or more components of sequencing instrument 100, low quality or supply of reagent, operating script error, or other source of systematic type error that may occur. Thus, if such systematic type error is detected in key 207 that sequence data generated for the run of that template molecule may not pass quality metrics and will typically be rejected.

The same discriminating sequence for key 207 can be used for an entire library of DNA fragments, or alternatively different sequence compositions may be associated with portions of the library for different purposes. Further examples of primer and key elements associated with primer 205 and key 207 are described in U.S. Patent Application Serial No. 10/767,894, incorporated by reference above.

Target specific element 225 includes a sequence composition that specifically recognizes a region of a genome. For example, Target specific element 225 may be employed as a primer sequence to amplify and produce amplicon libraries of specific targeted regions for sequencing such as those found within genomes, tissue samples, heterogeneous cell populations or environmental samples. These can include, for example, PCR products, candidate genes, mutational hot spots, evolutionary or medically important variable regions. It could also be used for applications such as whole genome amplification with subsequent whole genome sequencing by using variable or degenerate amplification primers. Further examples describing the use of target specific sequences with bipartite primers are described in U.S. Patent Application Serial No. 11/104,781, titled "Methods for determining sequence variants using ultra-deep sequencing", filed April 12, 2005, which is hereby incorporated by reference herein in its entirety for all purposes.

IDT Exhibit 1020.4991

Some embodiments of UID 210 may be particularly amenable for use with relatively small numbers of sample associations in a multiplex sample. In particular, when there are only a small number of associations to identify in a multiplex sample, each sample is associated with a distinct implementation of UID 210 comprising a sequence composition that is sufficiently unique from each other as to enable easy detection and correction of introduced error. In some embodiments, groups of compatible UID 210 sequence elements are clustered into "sets" as will be described in greater detail below. For example, a set of UID 210 elements may include 14 members that may be employed to uniquely identify up to 14 associations with samples, where each member is associated with a single sample.

It will be appreciated that as the number of associations to identify grows, it becomes increasingly difficult to design distinct embodiments of UID 210 for each association that meet the design criteria and desired characteristics. In such cases, it may be advantageous to employ multiple UID 210 elements combinatorially to uniquely associate the template molecules with their sample of origin, where one embodiment of UID 210 may be positioned at each end of a linear template molecule. For example, the number of associations to identify between the sequence data generated from template molecules and the sample of origin may become too large to accommodate given the necessary design parameters and characteristics of UID 210. In particular, it is undesirable in many embodiments to employ a distinct UID element for each association when the number a samples would require a sequence length for UID 210 that is undesirably long for the design criteria that includes a specific number of flow cycle iterations and number of sequence positions taken up by the UID element. In the present example, in embodiments of sequencing technology that generate "long" read lengths UID 210 may comprise up to 10 sequence positions. Alternatively, other embodiments of sequencing technology may generate relatively short read lengths of about 25-50 sequence positions, and thus it is desirable that UID 210 is short in order to optimize the read length for the template molecule. In the present example, UID 210 may be designed for short read lengths comprising up to 4 sequence positions, up to 6 sequence positions, or up to 8 sequence positions, depending, at least in part, upon the application.

As described above, embodiments for design and implementation of UID 210 amenable for both small and large numbers of associations is to employ a "set" of UID 210 elements each meeting the preferred design criteria and characteristics. In some applications, such as the design of UID 210 elements with sequence composition that enable accurate error detection and correction features it is desirable to use the "set" strategy presently described.

IDT Exhibit 1020.4992

For example, as will be described in greater detail below the sequence composition for the UID elements in a set must be sufficiently distinct from each other in order to enable error detection and correction thereby limiting the compatible members available for a particular set. However, UID 210 members from multiple sets may be combinatorially employed with a template molecule where the members of each set are located at different relative positions and are thus easily interpretable.

In order to overcome the problems of a large number of associations to identify described above, two or more members from a set of UID 210 elements may be employed in a combinatorial manner. For example, a set of UID 210 elements may include 10, 12, 14, or other number of members comprising a 10-mer sequence length. In some embodiments, two UID 210 elements may be associated with each template molecule and used combinatorially to identify up to 144 different associations (i.e. 12 UID members for use with element 1 multiplied by 12 UID members for use with element 2 results in 144 possible combinations of UID elements 1 and 2 that may be employed to uniquely identify an association).

Those of ordinary skill in the related art will appreciate that alternative embodiments may be employed where each UID 210 element associated with a template molecule may include a subset of the total number of UID members from the set (i.e. use a portion of the members of the set). In other words, of the 12 members of a complete set, only 8 may be employed at one element position. There are a number of reasons why it may be desirable to use a subset of UID members that includes having a need for a smaller number of associations to identify (i.e. smaller number of combinations), physical or practical experimental conditions such as equipment or software limitations, or preferred combinations of UID members of a set in element positions. For instance, a first element may employ all 12 UID members from a set and a second element may employ a subset of 8 UID members from the same or different set yielding 96 possible combinations.

UID 210 elements used in combinatorial strategies may be configured in a variety of positional arrangements relative to the position of the template molecule. For example, a strategy that utilizes 2 UID 210 elements combinatorially to identify the association of each template molecule with its sample of origin may include a UID element positioned at each end of a linear template molecule (i.e. one UID 210 element at the 5' end and another at the 3' end). In the present example, each UID 210 element may be associated with an adaptor element, such as adaptor 200 or 220, employed in a target specific amplicon or genomic library sequencing strategy as discussed above. Thus, the sequence data associated with a template molecule would include the sequence composition of a UID element at each end of

IDT Exhibit 1020.4993

the amplicon. The combination of the UID elements may then be used to associate the sequence data with the sample of origin of the template molecule.

In some alternative embodiments, a UID 210 element may be incorporated in an adaptor element at each end of a linear template molecule as described above. However, the read length of the template molecule may be greater than the ability of the sequencing technology to handle. In such a case, the template molecule may be sequenced from each end independently (i.e. a separate sequencing run for each end), where the UID 210 element associated with the end may be employed as a single UID 210 identifier.

In addition it may be desirable in some embodiments to assign more that one UID 210 element per sample, or more than one combinations of UID 210 elements. Such a strategy may provide redundancy to protect against possible unintended biases introduced by various source, which could include the UID 210 element itself. For example, a sample with a population of template molecules may be sub-divided in sub-samples each using a distinctive UID 210 element for the association. In such a case, the redundancy of the different UID 210 elements for the same population of template molecules from a sample provides for greater confidence that the correct associations will be identified or if the error is too great to make a correct identification of the association with confidence.

As generally described above, embodiments of the presently described invention include one or more UID 210 elements operatively coupled to each template molecule for the purpose of identifying the association between the template molecule and the sequence data generated therefrom with a sample of origin. One or more embodiments of a UID element may be operatively coupled to one or more components of an adaptor and a template molecule using a variety of methods known in the art that include but are not limited to ligation techniques. Methods for ligating nucleic acid molecules to one another are generally known in the art and include employing a ligase enzyme for what is referred to as sticky end or blunt end ligation. Further examples of coupling adaptor elements to template molecules using ligation as described in U.S. Patent Application Serial No. 10/767,894, titled "Method for preparing single-stranded DNA libraries", filed January 28, 2004; and U.S. Provisional Patent Application Serial No. 60/031,779, titled "System and Method for Improved Processing of Nucleic Acids for Production of Sequencable Libraries" filed February 27, 2008, each of which is hereby incorporated by reference herein in its entirety for all purposes). For example, a large template nucleic acid or whole genomic DNA sample may be fragmented by mechanical (i.e. nebulization, sonication) or enzymatic means (i.e. DNase I), the resulting ends of each fragment may be polished for compatibility with adaptor

IDT Exhibit 1020.4994

elements (i.e. polishing using what is referred to as an exonuclease, such as BAL32 nuclease or Mung Bean nuclease), and each fragment may be ligated to one or more adaptor elements (i.e. using T4 DNA ligase). In the present example, each adaptor element is directionally ligated to the fragment such as for instance by selective binding between the 3' end of the adaptor and the 5' end of the fragment.

In some embodiments, UID 210 elements may be provided to user 101 in the form of a kit, where the kit could include adaptors comprising incorporated UID 210 elements as illustrated in Figures 2A and 2B. Or, the kit could include UID 210 as independent elements that enable user 101 to incorporate as they desire.

As described above, embodiments of UID 210 should comprise a number of preferred characteristics or design criteria that include but are not limited to a) each UID element comprises a minimal sequence length requiring a minimal number of synthesis or flow cycles, b) each UID element comprises sequence distinctiveness, c) each UID element comprises resistance to introduced error, and d) each UID element does not interfere with amplification methods (such as PCR, or cloning into vectors).

Also, some embodiments of UID element design may also consider physical characteristics or design criteria of nucleic acids that include some or all of i) UID sequence composition selected to resist formation of what are referred to as "hairpins" (also referred to as a "hairpin loop" or "stem loop") and "primer dimers"; ii) UID elements comprise preferred melting temperature (i.e. 40°C) and/or Gibbs free energy (i.e. ΔG cutoff of -1.5) characteristics. Aspects of some of the desirable characteristics and their impact on UID design are described in greater detail below.

One important characteristic of a UID element is that it should include a minimal number of bases or sequence positions required to satisfy the needs of other characteristic requirements. For example, each UID element should comprise the minimum sequence length required to uniquely identify a desired number of associations between the template molecule/sequence data and their samples of origin. A desired number of associations may include identification of template molecules/sequence data associated with at least 12 different samples, at least 96 different samples, at least 384 different samples, or a greater number of samples that may be contemplated in the future. In other words the sequence length of the UID should be no longer than necessary in order to conserve the number of positions (i.e. what may be referred to as "sequence real estate") of the read length for the template molecule. Further, the minimum sequence length should consume or require a minimum number of flow cycles of the set of nucleotide species to generate the sequence

IDT Exhibit 1020.4995

data for each UID element. Minimizing the number of nucleotide species flow cycles required to generate sequence data for the UID elements provides advantages in reagent cost, instrument usage (i.e. processing time), data quality, and read length. For instance, each additional flow cycle increases the probability of introducing CAFIE error, and reagent usage. In the present example, it is preferable that each 10-mer UID element require only 5 nucleotide species flow cycles to generate sequence data for each UID element.

Another important characteristic includes sequence distinctiveness of each UID element. The term "sequence distinctiveness" as used herein generally refers to a distinguishable difference between a plurality of UID sequences such that each sequence is easily recognizable from every other UID sequence that is the subject of comparison. In particular each UID element needs to comprise a measure of sequence distinctiveness that enables easy detection of introduced error and correction of some or all of the error. Further, it is generally preferable that each UID element be free of repetitive sequence composition and should not include a sequence composition recognized by restriction enzymes. In other words it is undesirable for UID elements to include consecutive monomers having the same composition of nucleotide species. For example, preferred embodiments of the sequence distinctiveness of each UID element enable detection of up to 3 sequence positions with introduced errors and correction of up to 2 sequence positions with introduced errors in a 10-mer element (i.e. 10 total sequence positions). Those of ordinary skill will appreciate that the introduced error may include what are referred to as "insertions", "deletions", "substitutions", or some combination thereof (i.e. a combination of an insertion and deletion at the same sequence position will appear to be a substitution and would be counted as a single error event). Also, the level of error detection and correction may depend, at least in part, upon the sequence length of the UID element. Further, introduced errors outside (i.e. upstream or downstream) of UID 210 may have effects on the interpretation of sequence composition for UID 210. This will be discussed further below in the context of decoding or analysis of sequence data for UID identification.

A further characteristic that is also desirable comprises resistance to introduced error. For example, monomer repeats in nucleic acid sequence such as that of the template molecule or other sequence elements may cause errors in a sequence read. The error may include an over or under representation or call of the number of repeated monomers. It is therefore desirable that the UID elements do not begin or end with the same nucleotide species as the adjacent monomer of a neighboring sequence element (i.e. creating monomer repeats between sequence elements or components). In the present example, a neighboring sequence

IDT Exhibit 1020.4996

element, such as key 207 illustrated in Figures 2A and 2B, may end with a "G" nucleotide species. Therefore, a UID element such as UID 210, should not begin with the same "G" nucleotide species to avoid the increased possibility introduced error from the repeated "G" species.

Another source of error that is particularly relevant in SBS contexts, include what are referred to as "carry forward" or "incomplete extension" effects (sometimes referred to as CAFIE effects). For example, a small fraction of template nucleic acid molecules in each amplified population of a nucleic acid molecule from a sample (i.e. a population of substantially identical copies amplified from a nucleic acid molecule template) loses or falls out of phasic synchronism with the rest of the template nucleic acid molecules in the population (that is, the reactions associated with the fraction of template molecules either get ahead of, or fall behind, the other template molecules in the sequencing reaction run on the population) . Additional description of CAFIE mechanisms and methods of correcting CAFIE error are further described in PCT Application Serial No US2007/004187, titled "System and Method For Correcting Primer Extension Errors in Nucleic Acid Sequence Data", filed February 15, 2007, which is hereby incorporated by reference herein in its entirety for all purposes.

Also, it will be appreciated that some types of error may occur at higher frequency than other types and/or have greater consequences than other types of error. For example, deletion error may have more significant impact than substitution error. It is therefore advantageous to design each UID element so that it is weighted more heavily to deal with the more frequent or more deleterious types of error.

As stated previously, it is not typically desirable to randomly or non-selectively design the sequence composition of UID elements. An illustrative example of two improperly designed UID elements and the potential for problems with error detection/correction using such UID elements is presented in Table 1.

Table 1:

UID Element 1	Generated UID Sequence	UID Element 2
A*C*TGA (SEQ ID NO: 1)	A*G*T*G*A (SEQ ID NO: 2)	AG*C*GA (SEQ ID NO: 3)

In the example of table 1, it is apparent that the UID sequence represented as generated UID sequence contains an error (i.e. the presence of at least one error is detected) if either UID element 1 or 2 is the original sequence element. However, it is not clear from the

IDT Exhibit 1020.4997

sequence composition of the Generated UID sequence whether UID element 1 or UID element 2 was the actual UID element because a single error in either could result in the generated sequence. In other words, it is possible that one error was introduced in UID element 1 transforming the "C" nucleotide species at the second position to a "G" species. It is also possible that one error was introduced in UID element 2 transforming the "C" nucleotide species at the third position to a "T" species. Given the sequence information, the error is detected but it is not possible to infer which UID element was the original element and thus cannot be corrected. Therefore, the association of the generated UID sequence with either UID element 1 or 2 cannot be positively made, and thus the sample of origin for the template molecule coupled to one of the UID elements cannot be identified and the generated sequence information may need to be thrown out. In other words, the design of UID elements 1 and 2 are not sufficiently distinct from each other to recover from the described type of introduced error.

The potential result of poor UID design is further exemplified in Table 2.

Table 2:

UID Element 1	UID Element 2
CTACC (SEQ ID NO: 4)	CTGCC (SEQ ID NO: 5)

The example of Table 2 provides an even clearer picture of the potential consequences where a substitution event in UID element 1 of an A nucleotide species at the third position to a G nucleotide species, which is one of the most common types of error introduced by PCR processes, results in an exact match with the sequence composition of UID 210 element. Thus the poor UID 210 design results in an undetectable error that would likely result in the mis-assignment of the sequence data to a sample of origin.

Various methods may be employed to design UID elements comprising sequence composition that meets the necessary design criteria. Also, application 135 illustrated in Figure 1 may be employed for designing UID 210 using some or all of the methods described herein. For example, "Brute Force" methods may be employed that compute every possible sequence composition for a given length and the possible conflicts with other sequence composition given a set of parameters associated with the design criteria. In the present example, the sequence composition of 10 mer UID elements may be computed for detection of up to 3 sequence positions with introduced errors and correction of up to 2 sequence positions with introduced errors.

IDT Exhibit 1020.4998

Design of a preferred sequence composition for members of a set of UID 210 elements meeting the most stringent design criteria given the characteristics described above presents a computational challenge. Mathematical methods known to those of skill in the art may be applied to compute the possible sequence composition for members of a set given the design constraints. For example, mathematical transformations of all possible combinations of sequence composition may be computed given the design constraints to generate what may be referred to as "Error Balls" or "Error Clouds" to determine the potential compatibility of each UID element with the other members in a set. Compatibility of sequence composition for potential UID elements may be visually illustrated as non-overlapping error balls. For example, Figure 3 provides an illustrative representation of what may be referred to as "space potential" for computed error balls for UID 310, UID 320, UID 330, UID 340, and UID 350 comprising some or all of the design criteria described above such as number of flow cycles, and sequence length requirements. As illustrated in Figure 3 the error balls for UID 310, UID 320, and UID 330 do not overlap and thus represent sequence composition of compatible UID 210 elements. Further, UID 340 overlaps with UID 320 and UID 350 representing a sequence composition for a UID element that is not compatible. However UID 340 does not overlap with UID 310 and UID 330 and thus represents compatible sequence composition for each non-overlapping UID element.

Alternatively, a more computationally efficient approach may be employed that uses what is referred to in the art as "Dynamic Programming" techniques. The term "Dynamic Programming" as used herein generally refers to methods for solving problems that comprise overlapping sub-problems and optimal structure. Dynamic programming techniques are typically substantially more computationally efficient than methods with no *a priori* knowledge.

Some embodiments of dynamic programming technique include computing what may be referred to as the "minimum edit distance" for strings of characters such as strings of nucleic acid species. In other words, each UID member element in a set may be considered a string of characters representing the nucleic acid species composition. The term "minimum edit distance" as used herein generally refers to the minimum number of point mutations required to change a first string into a second string. Further, the term "point mutation" as used herein generally refers to and includes a change of character composition at a location in a string referred to as a substitution of a character for another in a string; an insertion of a character into a string; or a deletion of a character from a string. For example, the minimum edit distance may be computed for each potential member of a set of UID 210 elements

IDT Exhibit 1020.4999

against all other members of the set. Subsequently the minimum edit distances may be compared and members of the set of UID 210 elements selected based, at least in part, upon each member of the set having a sufficiently high minimum edit distance from all other members to meet the specified criteria. Systems and methods for computing minimum edit distance are well known to those of ordinary skill in the related art and may be implemented in a number of ways.

Another important aspect of the presently described invention is directed to the analysis of sequence data to "decode" or identify the UID 210 sequence elements within the data. In some embodiments an algorithm may be implemented in computer code as application 135 that processes the sequence data from each run and identify UID 210 as well as perform any error detection or corrections functions. It is important to recognize that methods of error detection and correction in strings of information have been employed in the computer arts particularly in the area of electronically stored and transmitted data. For example, the problem of "inversion" of bits of data from one form into another occurs when data is transmitted over networks or stored in electronic media. The inversion of bits presents a problem with respect to the integrity of stored or transmitted data and is analogous to the presently described substitution type of error. Methods of detection and correction of inversion error is described in J. F. Wakerly, "Detection of unidirectional multiple errors using low cost arithmetic codes," IEEE Trans. Comput., vol. C-24, pp. 210-212, Feb. 1975.; and J. F. Wakerly, Error Detecting Codes, Self-Checking Circuits and Applications. Amsterdam, The Netherlands: North-Holland, 1978, both of which are hereby incorporated by reference herein in their entireties for all purposes.

However, the methods of detecting and correcting inversion error described above are not applicable to the problem of error detection and correction in sequence data and more specifically errors in UID elements. Importantly, the problem in sequence data is substantially more complex because it deals with the problems of substitutions and deletions as well as substitutions that create phasing problems and complicate the interpretation of information at each sequence position.

As described above, UID 210 may be located at a known position relative to other easily identifiable elements such as primer 205, key 207, the 5' or 3' end of the sequence, etc. However, just as introduced error within UID 210 has deleterious effects, error outside of the region of the UID 210 element may also affect the efficiency of identifying each UID 210 element. Further, some types of error outside of the region defined by UID 210 may contribute to and count as errors within UID 210 sequence. For example, insertion events

may occur and be represented in the sequence data preceding (i.e. upstream of) UID 210 element that may be difficult to interpret. In the present example, an insertion event could include the insertion of one or more G nucleotide species bases at the end of key 207 comprising a TCAG sequence composition as may occur when a nucleotide species at a sequence position is "overcalled". However, an application that interprets the data will not know that it is an insertion event and cannot rule out the possibility of a substitution event that provided a G nucleotide in place of a different nucleotide species at the first sequence position of UID 210. In other words, the error outside of UID 210 will force the algorithm to decide if the error is an insertion that shifts where it should look for the first sequence position of UID 210 or whether it is a substitution event.

Continuing the example from above, an algorithm or user may look for the UID 210 element immediately adjacent to another known element such as key 207 as illustrated in Figures 2A and 2B, but the insertion of one base between key 207 and UID 210 may typically be assigned as belonging to UID 210 (counts as a first insertion error). Additionally, the algorithm or user expects UID 210 to be a certain length (i.e. 10 sequence positions) and thus truncates the last sequence position of the actual UID element because of the first insertion (counts as a second deletion error). Thus, it is clear that errors outside of the UID region can have substantial effect on finding and interpreting the sequence composition of UID 210.

In some embodiments, errors outside of the region defined by UID 210 may be particularly troublesome at the 3' end of a nascent molecule. For example, some embodiments of SBS sequence from 5' to 3' ends (i.e. adding nucleotide species to 3' end of nascent molecule) where cumulative errors (such as CAFIE type error described above) and the rate of introduced error may be increasingly higher as the sequence run gets longer at the 3' end. Thus, it may be more practical and effective to use certain assumptions rather than stringent criteria to identify UID 210. Also as described above, assumptions used for the 5' may be different than assumptions employed for the 3' end and may be referred to as "Asymmetric". For example, it may be assumed that there will never be more than 3 sequence position errors present at the 5' end which would be consistent with empirical evidence. However, in the present example at the 3' end it may be assumed that there will never be more than 4 sequence position errors due to the increased possibility of error at the 3' end. Because of the asymmetric difference in detectable error at each end, it may also be inferred that the amount of that error that is correctable may also be different. In the present example, the correctable error at the 5' end may be 2 sequence positions as described above, however the correctable error at the 3' end may only be 1 sequence position. Also, further

IDT Exhibit 1020.5001

assumptions may be employed at the 3' end that may not be employed for the 5' end. Such an assumption could include the existence of one or more "no called" positions in close proximity to UID 210.

In the present example, an embodiment of adaptor element 200 or 220 is present at the 3' end of a template nucleic acid in a palindromic arrangement to that illustrated in Figure 2A or 2B (as described above). It will be appreciated however, that the present example refers to a difference in the arrangement of elements and that the elements associated with each adaptor do not need to have the same composition (i.e. the 3' end may include the sequence composition of a first UID element and the 5' end may include a UID elements with different sequence composition). It will further be appreciated that some embodiments will not necessarily include the same composition of elements in each adaptor (i.e. an adaptor at the 5' end may include a UID 210 element and the adaptor on the 3' may not, or vice versa). Also, there may be inherent internal controls of the sequence quality of primer element 205 with respect to resistance to introduced error. For instance, error introduced into the sequence composition of primer 205 would negatively affect its hybridization qualities to its respective target and thus not be amplified in a PCR process and therefore not represented in populations of template molecule for sequencing. This inherent quality control of primer 205 is useful for finding UID 210, because the sequence composition of primer 205 is known and can be assumed to be substantially free of error with the exception of some sequencing related error. Also as described above, key element 207 is employed for quality control purposes and it also useful as a positional reference in the same context. Thus, in the present example primer 205 and/or key 207 may serve as easily identifiable anchor points of reference for identifying UID 210 using the known positional relationships between elements. For instance, a user or algorithm, such as an algorithm implemented by application 135, may look for UID 210 located immediately adjacent to key 207, or some known distance away, based, at least in part, upon the assumptions.

Furthermore, once a user or algorithm has identified the sequence composition of a putative UID 210 element, the step of error identification and correction occurs. Embodiments of the presently described invention compare the sequence composition of the putative UID 210 element against the sequence compositions of the UID 210 members in the set. A perfect match is associated with its sample of origin. If no perfect match is found, then the closest UID 210 elements having a sequence composition to the putative sequence are analyzed to determine possible insertion, deletion, or substitution errors that could have occurred. For example, the closest UID 210 element to the putative UID 210 element is

IDT Exhibit 1020.5002

identified or the putative UID 210 element is deemed to have too many errors. In the present example, the minimum edit distance may be computed between sequence composition of the putative UID 210 element against the sequence composition of all members of the UID 210 set or select members. The minimum edit distance may be computed using the parameters of detecting up to 3 sequence position errors with the possibility of correcting up to 2 sequence position errors. In the present example, the UID 210 member with the closest or shortest minimum edit distance to the putative UID 210 element given the parameter constraints (i.e. detection/correction) may be assigned as the sequence composition of the putative UID 210 element. Also, if the minimum edit distance calculation determines that 3 sequence position errors have occurred then, the putative UID 210 element may be assigned as unusable and not associated with a sample of origin.

Those of ordinary skill in the art will appreciate that when the UID 210 elements are employed in a combinatorial manner, each UID 210 element is typically independently analyzed. Then the combination of identified UID 210 elements may be compared against the known combinations assigned to samples of origin to identify the association of the sequence data and its specific sample of origin.

In preferred embodiments, a UID 210 finding algorithm is implemented using application 135 stored for execution on computer 130 as described above. Further, the same or other application may perform the step of associating the identified UID 210 from sequence data with the sample of origin and providing the results to a user via an interface and/or storing the results in electronic media for subsequent analysis or use.

Example 1- Design of UID elements considering a limited number of design constraints

The design of sequence composition for potential UID elements were computed considering detection, correction, and hairpin design constraints.

First a sequence length of 10 base pairs for each UID element were computed yielding 1,048,576 possible elements.

Next, of those possible elements UID elements were selected that have no monomer repeats, require only 5 flow cycles (20 flows) or less, do not begin with the "G" nucleotide species were computed yielding 34,001 possible elements.

A further step of filtering to exclude hairpins at a temperature of 40°C with a ΔG=-1.5 yielded 26,278 possible elements.

IDT Exhibit 1020.5003

Finally, 5,000 of those possible elements were selected randomly to search for compatible sets or clusters that could correct 2 sequence position errors and detect 3 sequence position errors, yielding:

32,999 sets of 12 members

3,625 sets of 13 members

24 sets of 14 members

Example 2- Exemplary computer code for creating UID sequence elements

UIDCreate.java class file that runs a search using 1 of 3 techniques, comprising (1) based on error clouds, (2) based on edit distance, and (3) based on edit distance, with an additional efficiency strategy of using a "safety map" to precompute the edit distance which gives the software the ability to effectively look ahead in the search in advance of trying candidate selections.

```
package com.fourfivefour.amplicons;


import java.util.ArrayList;
import java.util.Arrays;
import java.util.Comparator;
import java.util.HashMap;
import java.util.HashSet;
import java.util.Iterator;
import java.util.List;
import java.util.Map;
import java.util.Set;

public class UIDCreate {
        static int maxReportSize = 0;

        public enum SearchType { ErrorCloud, EditDistance, SafetyMapEditDistance; }

        /**
         * Driver to search for a compatible set of UID sequences of a given set size
         * using one of three different search methods.
         *
         * @param candidateSequences  A set of sequences from which to select.
         * @param setSize             The desired number of sequences in the
         *                                                      solution.
         * @param errsToCorrect               The number of errors that must be able to be
         *                    corrected to obtain the original sequence.
         * @param errsToDetect        The number of errors that must be able to be
         *                    detected, but not necessarily corrected.
         * @param searchType                  The search strategy to use.
         * @return                                    A set of sequences of the requested set
size
```


IDT Exhibit 1020.5004

```
         *                  that is a subset of the candidate sequences
         *                  and which can be distinguished according to
         *                  the error correction criteria.
         */
        public static Set<Sequence> searchCompatibleSet(Set<Sequence>
candidateSequences,
                        int setSize, int errsToCorrect, int errsToDetect, SearchType
searchType) {

                if (errsToCorrect > errsToDetect) {
                        throw new RuntimeException("The numbers of errors to correct
("+errsToCorrect+
                                                        ") must be <= the
number to detect ("+errsToDetect+")");
                }

                List<Sequence> candidates = new
ArrayList<Sequence>(candidateSequences);

                switch (searchType) {
                        case ErrorCloud:
                                return searchCompatibleSetUsingErrorClouds(candidates,
                                                setSize, errsToCorrect, errsToDetect, new
HashMap<Sequence,List<Set<Sequence>>>());
                        case EditDistance:
                                return searchCompatibleSetUsingEditDistance(candidates,
                                                setSize, errsToCorrect, errsToDetect, new
HashSet<Sequence>()
                                                );
                        case SafetyMapEditDistance:
                                return
searchCompatibleSetUsingSafetyMap(candidateSequences,

                 setSize, errsToCorrect, errsToDetect);
                        default:
                                return null;
                }
        }


        /**
         * Return a string representing the current search state.
         *
         * The resulting string encodes the current subset of sequences currently
         * under consideration, independent of the order in which they are
         * being considered.  By keeping track of this string one may ensure
         * that equivalent regions of the search space that just happen to
         * encounter sequences in a different order are not redundantly searched
         *
         * @param currentState  A set of sequences defining the current
```

```
         *                search state.
         * @return          A string that is unique for a set of sequences
         *                and independent of their order.
         */
        public static String searchStateString (Set<Sequence> currentState) {
                Sequence [] currentStateArray = currentState.toArray(new Sequence[0]);
                Arrays.sort(currentStateArray);
                StringBuffer sb = new StringBuffer();
                for (Sequence state : currentStateArray) {
                        sb.append(state.getSequence());
                        sb.append(';');
                }
                return sb.toString();
        }


        /**
         * From the set of candidates, select a set of compatible sequences that
         * can correct and detect the desired number of errors.
         *
         * This strategy uses a pre-calculated edit distance between pairs of
         * sequences to know, in advance, which sequences are compatible with
         * each other.  This pair-wise information forms a "safety map" of
         * compatible pairs that allow more efficient forward-looking knowledge
         * of what sequences will still be compatible after other sequences
         * are added to the solution set.
         *
         * @param candidateSequences Sequence set from which to select.
         * @param setSize          The desired set size to generate.
         * @param errsToCorrect     The numbers of errors that must be able
         *                        to be corrected.
         * @param errsToDetect      The number of errors that must merely be
         *                        detected.
         *
         * @return A set of sequences that meet the desired criteria.
         */
        public static Set<Sequence> searchCompatibleSetUsingSafetyMap (
                        Set<Sequence> candidateSequences,
                        int setSize, int errsToCorrect, int errsToDetect)
        {
                Map<Sequence,Set<Sequence>> safetyMap = new
HashMap<Sequence,Set<Sequence>>();

                /* Safety map is computed on a subset of the candidates in
                 * case there are too many candidates.
                 */
                final int maxSubsetSize = 3000;

                Sequence[] fullSequenceList = candidateSequences.toArray(new
Sequence[0]);
```

```
            final int numSequencesToUse =
Math.min(maxSubsetSize,candidateSequences.size());

            /* Compute the safety map, possibly using a subset of the sequences.
             * An alternative/preferential implementation might choose a random subset
             * of the sequences rather than just the fist numSequencesToUse that are
             * encountered.
             */

            for (int i=0; i < numSequencesToUse; i++) {
                Set<Sequence> iSafety = new HashSet<Sequence>();
                iSafety.add(fullSequenceList[i]);
                safetyMap.put(fullSequenceList[i], iSafety);
            }

            /* Create the safety map, ensuring that there is enough distance
             * between every pair of sequences such that sequences as far away
             * as the distance of error detection from one sequence and error
             * correction from the other sequence could not possibly overlap
             */
            int safeDistance = errsToCorrect + errsToDetect + 1;
            for (int i=0; i < numSequencesToUse; i++) {
                Set<Sequence> iSafety = safetyMap.get(fullSequenceList[i]);
                for (int j=i+1; j < numSequencesToUse; j++) {
                    Set<Sequence> jSafety = safetyMap.get(fullSequenceList[j]);

                    if (Library.minEditDistance(fullSequenceList[i],

        fullSequenceList[j])
                                                                    >=
safeDistance) {
                        iSafety.add(fullSequenceList[j]);
                        jSafety.add(fullSequenceList[i]);
                    }
                }
            }

            // Perform that actual search with the safety-map in hand
            return searchCompatibleSetUsingSafetyMap(safetyMap.keySet(),setSize,
                    errsToCorrect,errsToDetect,
                    safetyMap,new HashSet<Sequence>(), new
HashSet<String>());
        }

        /**
         * From the set of candidates, select a set of compatible sequences that
         * can correct and detect the desired number of errors.
         *
         * This strategy uses a pre-calculated edit distance between pairs of
         * sequences to know, in advance, which sequences are compatible with
```

```
     * each other.  This pair-wise information forms a "safety map" of
     * compatible pairs that allow more efficient forward-looking knowledge
     * of what sequences will still be compatible after other sequences
     * are added to the solution set.
     *
     * @param candidateSequences Sequence set from which to select.
     * @param setSize          The desired set size to generate.
     * @param errsToCorrect    The numbers of errors that must be able
     *                         to be corrected.
     * @param errsToDetect     The number of errors that must merely be
     *                         detected.
     * @param safetyMap        The map indicating the set of sequences
     *                         that are compatible with any given other
     *                         sequence.
     * @param selectedSequences  The current set of sequences selected in
     *                         the partial solution so far.
     * @param searchStateHistory A history of what subsets of sequences have
     *                         already been considered.
     *
     * @return A set of sequences that meet the desired criteria.
     */
    public static Set<Sequence> searchCompatibleSetUsingSafetyMap (
                    Set<Sequence> candidateSequences, int setSize,
                    int errsToCorrect, int errsToDetect,
                    Map<Sequence,Set<Sequence>> safetyMap,
                    Set<Sequence> selectedSequences,
                    Set<String> searchStateHistory) {


        String searchStateString = searchStateString(selectedSequences);
        if (searchStateHistory.contains(searchStateString)) {
                /* some other branch of the search has already explored this area
                 * so no need to search twice
                 */
                return null;
        } else {
                /*
                 * Record that we are initiating a search in the given state
                 */
                searchStateHistory.add(searchStateString);
        }

        if (selectedSequences.size() > maxReportSize) {
                // Provide intermediate reporting to show progress
                maxReportSize = selectedSequences.size();
                reportSet(selectedSequences);
        }
        if (selectedSequences.size() >= setSize) {
                return selectedSequences;
        }
```

```
                if (selectedSequences.size() + candidateSequences.size() < setSize) {
                        // If there aren't enough candidates to fill out the required set size,
                        // return immediately with failure.
                        return null;
                }

                /* find the candidate Sequence with the greatest number of compatible choices
of the
                 * Sequences still available:
                 */
                final Map<Sequence,Integer> remainingSafetyMapSize = new
HashMap<Sequence,Integer>();
                Sequence candidates[] = candidateSequences.toArray(new Sequence[0]);
                for (Sequence sequence : candidates) {
                        Set<Sequence> currentSafety = new
HashSet<Sequence>(safetyMap.get(sequence));
                        currentSafety.retainAll(candidateSequences);
                        remainingSafetyMapSize.put(sequence,currentSafety.size());
                }

                Comparator<Sequence> candidateSorter = new Comparator<Sequence>() {
                        public int compare(Sequence seq1, Sequence seq2) {
                                int seq1Val = remainingSafetyMapSize.get(seq1);
                                int seq2Val = remainingSafetyMapSize.get(seq2);
                                if (seq1Val > seq2Val) {
                                        return -1;
                                } else if (seq1Val < seq2Val) {
                                        return 1;
                                } else {
                                        return (seq1).compareTo(seq2);
                                }
                        }
                };

                Arrays.sort(candidates,candidateSorter);

                Set<Sequence> result = null;
                final int searchBreadth = 10;
                for (int i=0; i < candidates.length && i < searchBreadth && result == null;
i++) {
                        Sequence nextCandidate = candidates[i];
                        int numNextCompatible =
remainingSafetyMapSize.get(nextCandidate);
                        if (numNextCompatible == 1) {
                                /* the next candidate is only compatible with itself, so
                                 * the search is done
                                 */
                                break;
                        } else {
```

```
                                Set<Sequence> nextCandidateSequences =
                                        new HashSet<Sequence>(candidateSequences);

        nextCandidateSequences.retainAll(safetyMap.get(nextCandidate));
                                nextCandidateSequences.remove(nextCandidate);
                                Set<Sequence> nextSelectedSeqs = new
HashSet<Sequence>(selectedSequences);
                                nextSelectedSeqs.add(nextCandidate);
                                result =

        searchCompatibleSetUsingSafetyMap(nextCandidateSequences,

         setSize,

         errsToCorrect,

         errsToDetect,

         safetyMap,

         nextSelectedSeqs,

         searchStateHistory);
                        }
                }
                return result;
        }


        /**
         * Generate the "error cloud" sequence sets out to the indicated number
         * of mutations of the given input sequence.
         *
         * This routine implements a breadth-first enumeration of all possible
         * mutations of a given sequence, while avoiding redundant, equivalent
         * subtrees of the search space.
         *
         * @param inputSequence
         *          The sequence to mutate.
         * @param numMutations
         *          The number of plys of mutations to generate
         * @return A list of sets of mutated sequences such that the i'th list
         *       element represents the set of mutated sequences that may
         *       derived via i mutations from the initial sequence and, further,
         *       cannot be derived by fewer than i mutations of the original
         *       sequence.
         */
        public static List<Set<Sequence>> generateErrorClouds(
                        Sequence inputSequence, int numMutations) {
```

```
        List<Set<Sequence>> mutationPlys = new ArrayList<Set<Sequence>>();

        // The 0'th mutation ply is simply the initial sequence itself:
        Set<Sequence> initialPly = new HashSet<Sequence>();
        initialPly.add(inputSequence);
        mutationPlys.add(initialPly);

        for (int ply = 1; ply <= numMutations; ply++) {
            Set<Sequence> lastPlySequences =
                mutationPlys.get(ply - 1);
            Set<Sequence> nextPlyCandidates =
                Sequence.generateSingleMutations(lastPlySequences);
            Iterator<Sequence> candidateIter =
                nextPlyCandidates.iterator();
            /* Remove any candidates that appeared in a previous ply
             * (meaning that they are derivable via a shorter,
             *  more direct sequence of mutations).
             */
            while (candidateIter.hasNext()) {
                Sequence candidate = candidateIter.next();
                for (Set<Sequence> previousPly : mutationPlys) {
                    if (previousPly.contains(candidate)) {
                        candidateIter.remove();
                        break;
                    }
                }
            }

            // Candidates that remain at this point represent the next ply
            mutationPlys.add(nextPlyCandidates);
        }
        return mutationPlys;
    }


    /**
     * Validate the given collection of sequences to ensure that they
     * are compatible with respect to the desired number of errors
     * to correct and detect.
     *
     * Two sequences are compatible if there is no intersection of the
     * error cloud out to the correction distance of either sequence with
     * the error cloud out to the detection distance of the other sequence.
     * If all pairs of sequences in a collection of sequences are compatible,
     * then the collection itself is compatible.
     *
     * This definition of compatibility logically follows from the fact
     * that incompatibilities exist when a potentially observable sequence
     * could be derived, via mutation, from more than one sequence
     * in the collection (thereby preventing a definitive correction to be
```

```
         * performed) and if that observable sequence is within the desired
         * error correction distance of any of those sequences.  Note that
         * mutated sequences that are within the error detection distance of
         * two sequences are not a problem as there is no expectation of being
         * able to correct such mutated sequences (only to detect that a
         * mutation has occurred).
         *
         * @param sequences          The collection of sequences to test.
         * @param errsToCorrect The desired number of errors to correct.
         * @param errsToDetect  The desired number of errors to detect.
         * @return true, if the collection of sequences are compatible,
         *       false otherwise.
         */
        public static boolean validateCollection(Set<Sequence> sequences,
                                                                    int
errsToCorrect,
                                                                    int
errsToDetect)
        {
                /* First generate the error clouds for each sequence and
                 * collapse the separate cloud levels into two sets: those
                 * sequences within the "errsToCorrect" distance of the
                 * sequence center, and those within the the "errsToDetect"
                 * distance from the center.
                 */
                Map<Sequence,Set<Sequence>> correctLevelMap =
                        new HashMap<Sequence,Set<Sequence>>();
                Map<Sequence,Set<Sequence>> detectLevelMap =
                        new HashMap<Sequence,Set<Sequence>>();
                for (Sequence sequence: sequences) {
                        List<Set<Sequence>> errorClouds =
                                generateErrorClouds(sequence, errsToDetect);
                        Set<Sequence> errsToCorrectSequences = new HashSet<Sequence>();
                        for (int errLevel=0; errLevel <= errsToCorrect; errLevel++) {
                                errsToCorrectSequences.addAll(errorClouds.get(errLevel));
                        }
                        Set<Sequence> errsToDetectSequences =
                                new HashSet<Sequence>(errsToCorrectSequences);
                        for (int errLevel=errsToCorrect+1; errLevel <= errsToDetect;
                                 errLevel++) {
                                errsToDetectSequences.addAll(errorClouds.get(errLevel));
                        }
                        correctLevelMap.put(sequence,errsToCorrectSequences);
                        detectLevelMap.put(sequence,errsToDetectSequences);
                }

                /* Validate that none of the mutations out to the "errsToDetect"
                 * error distance of one sequence are in common with any of the
                 * mutations out to the "errsToCorrect" error distance another
                 * sequence
```

```
                */
                Sequence seqList[] = sequences.toArray(new Sequence[0]);
                boolean foundOverlap = false;

                for (int i=0; !foundOverlap && i < seqList.length; i++) {
                        for (int j=0; !foundOverlap && j < seqList.length; j++) {
                                if (i == j) {
                                        continue;
                                }
                                Set<Sequence> overlapSet =
                                        new
HashSet<Sequence>(detectLevelMap.get(seqList[i]));
                                overlapSet.retainAll(correctLevelMap.get(seqList[j]));
                                foundOverlap = !overlapSet.isEmpty();
                        }
                }

                return !foundOverlap;
        }

        /**
         * Validate the given collection of sequences to ensure that they
         * are compatible with respect to the desired number of errors
         * to correct and detect.
         *
         * Two sequences are compatible if there is no intersection of the
         * error cloud out to the correction distance of either sequence with
         * the error cloud out to the detection distance of the other sequence.
         * If all pairs of sequences in a collection of sequences are compatible,
         * then the collection itself is compatible.
         *
         * This definition of compatibility logically follows from the fact
         * that incompatibilities exist when a potentially observable sequence
         * could be derived, via mutation, from more than one sequence
         * in the collection (thereby preventing a definitive correction to be
         * performed) and if that observable sequence is within the desired
         * error correction distance of any of those sequences.  Note that
         * mutated sequences that are within the error detection distance of
         * two sequences are not a problem as there is no expectation of being
         * able to correct such mutated sequences (only to detect that a
         * mutation has occurred).
         *
         * This routine performs the calculation without creating the
         * actual error clouds, but instead uses the minimum edit distance
         * between sequences to ensure compatibility.
         *
         * @param sequences          The collection of sequences to test.
         * @param errsToCorrect The desired number of errors to correct.
         * @param errsToDetect  The desired number of errors to detect.
         * @return true, if the collection of sequences are compatible,
```

```
         *       false otherwise.
         */
        public static boolean validateCollectionUsingDistance(
                                                    Set<Sequence> sequences,
                                                    int errsToCorrect, int errsToDetect) {
                Sequence seqList[] = sequences.toArray(new Sequence[0]);
                /* The error detection cloud and error correction cloud of two
                 * sequences cannot overlap if the minimum distance between the
                 * sequences is one greater than the sum of the errors.
                 */
                int safeDistance = errsToCorrect + errsToDetect + 1;
                for (int i=0; i < seqList.length; i++) {
                        for (int j=i+1; j < seqList.length; j++) {
                                int ijDist= Library.minEditDistance(seqList[i],seqList[j]);
                                if (ijDist < safeDistance) {
                                        System.out.println("Validation failed, dist = "+ijDist+"
for "+seqList[i]+" vs "+seqList[j]);
                                        return false;
                                }
                        }
                }
                return true;
        }


        /**
         * From the set of candidates, select a set of compatible sequences that
         * can correct and detect the desired number of errors.
         *
         * This strategy uses the edit distance between pairs of sequences
         * to ensure that they do not overlap in a manner that would prevent
         * error detection/correction. This technique is much more efficient
         * than explicitly enumerating and comparing the error clouds between
         * sequences.
         *
         * @param candidateSequences Sequence set from which to select.
         * @param setSize          The desired set size to generate.
         * @param errsToCorrect     The numbers of errors that must be able
         *                  to be corrected.
         * @param errsToDetect      The number of errors that must merely be
         *                  detected.
         * @param currentSequences   State of search indicating the currently
         *                  selected candidates.
         *
         * @return A set of sequences that meet the desired criteria.
         */
        public static Set<Sequence> searchCompatibleSetUsingEditDistance (
                        List<Sequence> candidateSequences, int setSize,
                        int errsToCorrect, int errsToDetect,
                        Set<Sequence> currentSequences) {
```

```
            if (currentSequences.size() > maxReportSize) {
                /**
                 * Provide intermediate level reporting to ensure progress
                 * is being made.
                 */
                maxReportSize = currentSequences.size();
                reportSet(currentSequences);
            }
            if (currentSequences.size() >= setSize) {
                return currentSequences;
            }

            /* The error detection cloud and error correction cloud of two
             * sequences cannot overlap if the minimum distance between the
             * sequences is one greater than the sum of the errors.
             */
            int safeDistance = errsToCorrect + errsToDetect + 1;

            for (int nextIdx = 0; nextIdx < candidateSequences.size(); nextIdx++) {
                /* Check to see that nextSequence is far enough away from all the
                 * other sequences in the prefix solution.
                 */
                Sequence nextSequence = candidateSequences.get(nextIdx);
                boolean overlappingClouds = false;
                for (Sequence currentSequence : currentSequences) {
                    if (Library.minEditDistance(currentSequence,nextSequence) <
safeDistance) {
                        overlappingClouds = true;
                        break;
                    }
                }

                if (!overlappingClouds) {
                    List<Sequence> remainingCandidates =
                        candidateSequences.subList(nextIdx+1,
candidateSequences.size());
                    currentSequences.add(nextSequence);

                    Set<Sequence> solution =
                        searchCompatibleSetUsingEditDistance(
                            remainingCandidates,
                            setSize, errsToCorrect, errsToDetect,
                            currentSequences);
                    if (solution == null) {
                        currentSequences.remove(nextSequence);
                    } else {
                        return solution;
                    }
                }
```

```
            }
            return null;
        }

        /**
         * From the set of candidates, select a set of compatible sequences that
         * can correct and detect the desired number of errors.
         *
         * This strategy uses an explicit enumeration of the "error clouds"
         * surrounding the candidate sequences to ensure that they do not
         * overlap in a manner that would prevent error detection/correction.
         * This technique may not be appropriate, efficient, or even practically
         * usable for large error clouds.
         *
         * @param candidateSequences Sequence set from which to select.
         * @param setSize          The desired set size to generate.
         * @param errsToCorrect    The numbers of errors that must be able
         *                         to be corrected.
         * @param errsToDetect     The number of errors that must merely be
         *                         detected.
         * @param currentErrorClouds State of search mapping from the current
         *                         selected candidates and their corresponding
         *                         error clouds.
         * @return A set of sequences that meet the desired criteria.
         */
        public static Set<Sequence> searchCompatibleSetUsingErrorClouds(List<Sequence>
candidateSequences,
                                                                        int
setSize,
                                                                        int
errsToCorrect, int errsToDetect,

Map<Sequence, List<Set<Sequence>>> currentErrorClouds) {

            if (currentErrorClouds.size() > maxReportSize) {
                /**
                 * Provide intermediate level reporting to ensure progress
                 * is being made.
                 */
                maxReportSize = currentErrorClouds.size();
                reportSet(currentErrorClouds.keySet());
            }
            if (currentErrorClouds.size() >= setSize) {
                return currentErrorClouds.keySet();
            }

            for (int nextIdx = 0; nextIdx < candidateSequences.size(); nextIdx++) {
                Sequence nextSequence = candidateSequences.get(nextIdx);
                List<Set<Sequence>> nextErrorCloud =
                        generateErrorClouds(nextSequence, errsToDetect);
```

```
                    /*
                     * Check to see that nextErrorCloud doesn't overlap with any of the
                     * currentErrorClouds except possibly in the errsToDetect category,
                     * where it can only overlap with other errsToDetect distant clouds
                     */
                    boolean overlappingClouds = false;
                    for (List<Set<Sequence>> currentClouds :
currentErrorClouds.values()) {
                        for (int currentErrorLevel = 0; !overlappingClouds
                                && currentErrorLevel <= errsToCorrect;
currentErrorLevel++) {
                            for (int nextErrorLevel = 0; !overlappingClouds
                                    && nextErrorLevel <= errsToDetect;
nextErrorLevel++) {
                                Set<Sequence> currentLevelErrorCloud =
currentClouds.get(currentErrorLevel);
                                for (Sequence nextErrorSequence :
nextErrorCloud.get(nextErrorLevel)) {
                                    if
(currentLevelErrorCloud.contains(nextErrorSequence)) {
                                        overlappingClouds = true;
                                        break;
                                    }
                                }
                            }
                        }
                    }
                    /*
                     * Check to see that none of the error clouds surrounding the current
                     * codes in the range of errors to merely detect overlap with the
                     * nextErrorCloud clouds in the range of errors to correct
                     */
                    if (!overlappingClouds) {
                        for (List<Set<Sequence>> currentClouds :
currentErrorClouds.values()) {
                            for (int currentErrorLevel = errsToCorrect+1;
!overlappingClouds
                                    && currentErrorLevel <= errsToDetect;
currentErrorLevel++) {
                                for (int nextErrorLevel = 0; !overlappingClouds
                                        && nextErrorLevel <=
errsToCorrect; nextErrorLevel++) {
                                    Set<Sequence> currentLevelErrorCloud
= currentClouds.get(currentErrorLevel);
                                    for (Sequence nextErrorSequence :
nextErrorCloud.get(nextErrorLevel)) {
                                        if
(currentLevelErrorCloud.contains(nextErrorSequence)) {
                                            overlappingClouds = true;
```


IDT Exhibit 1020.5017

```
                                                    break;
                                                }
                                            }
                                        }
                                    }
                                }

                                if (!overlappingClouds) {
                                    List<Sequence> remainingCandidates =
                                        candidateSequences.subList(nextIdx+1,
candidateSequences.size());
                                    currentErrorClouds.put(nextSequence, nextErrorCloud);

                                    Set<Sequence> solution =
searchCompatibleSetUsingErrorClouds(
                                                    remainingCandidates,
                                                    setSize, errsToCorrect, errsToDetect,
currentErrorClouds);
                                    if (solution == null) {
                                        currentErrorClouds.remove(nextSequence);
                                    } else {
                                        return solution;
                                    }
                                }
                            }
                            return null;
                        }

                        /**
                         * Generate the set of all possible flow patterns that would produce
                         * sequences of a given length, with no monomer repeats, using at most
                         * a given number of flows, and avoiding a given starting flow as the
                         * first positive flow of the pattern.
                         *
                         * @param numBases      The number of called bases that should
                         *                      be produced if the given flowgram were
                         *                      observed for a sequence.
                         * @param startFlow     The conceptual flow index of the first
                         *                      flow in the recursive call to this routine.
                         * @param maxFlow       The maximum number of flows (positive or
                         *                      negative) that may be consumed by the pattern.
                         * @param avoidStartFlow  A flow that may not be the first positive flow.
                         * @return Set of flow patterns (Strings of 0's and 1's representing
                         *      negative and positive flows that meet the desired
                         *      criteria).
                         */
                        public static Set<String> generateCandidateFlowPatterns(int numBases,
                                int startFlow, int maxFlow, int avoidStartFlow) {
                            Set<String> candidates = new HashSet<String>();
```

```
                if (numBases > 0) {
                        if (startFlow == maxFlow) {
                                return candidates;
                        } else {
                                String zeroPrefix = "";
                                for (int flowSkip = 0; flowSkip < (startFlow == 0 ? 4 : 3);
flowSkip++) {
                                        int nextPositiveFlow = startFlow + flowSkip;
                                        if (nextPositiveFlow < maxFlow
                                                        && (startFlow > 0 || nextPositiveFlow !=
avoidStartFlow)) {
                                                String currentPrefix = zeroPrefix + "1";
                                                for (String candidateSuffix :
generateCandidateFlowPatterns(
                                                                numBases - 1, nextPositiveFlow
+ 1, maxFlow,
                                                                avoidStartFlow)) {
                                                        candidates.add(currentPrefix +
candidateSuffix);
                                                }
                                        }
                                        zeroPrefix += "0";
                                }
                        }
                } else {
                        candidates.add("");
                }
                return candidates;
        }

        /**
         * Generate a set of candidate sequences of a given base content length, using
         * a maximum number of flow cycles, based on the given flow order such that
         * none of the sequences begin with a particular base.
         *
         * @param numBases   Number of bases each candidate sequence should be.
         * @param maxCycles  Maximum number of flow cycles allowed.
         * @param flowOrder  Flow order for sequencing.
         * @param avoidStartBase Base to avoid at start of sequence.
         * @return Candidate set of sequences that meet the given criteria.
         */
        public static Set<Sequence> generateCandidateSequences(int numBases, int
maxCycles,
                        String flowOrder, char avoidStartBase)
        {
                Set<String> candidateFlows =
                        generateCandidateFlowPatterns(numBases,
                                                        0,
                                                        maxCycles*4,
```

```
flowOrder.indexOf(avoidStartBase));

                Set<Sequence> candidateSequences = new HashSet<Sequence>();
                for (String flows : candidateFlows) {
                        Flowgram f = new
Flowgram(flowOrder,Flowgram.flowValueStringToFlowValues(flows));
                        candidateSequences.add(new Sequence(f.baseCall()));
                }

                /*
                 * At this point, one might further filter the candidate sequences by some
                 * molecular biological criteria, such as propensity to form hairpin turns
                 */
                return candidateSequences;
        }


        /**
         * Produce a printed representation of a set of sequences
         * assuming a default flow oder of TACG.
         *
         * @param sequences  The set of sequences to report.
         */
        public static void reportSet(Set<Sequence> sequences) {
                reportSet(sequences, "TACG");
        }

        /**
         * Produce a printed representation of a set of sequences
         * along with their conceptual flowgrams, assuming the
         * given flow order.
         *
         * @param sequenceSet The set of sequences to report.
         * @param flowOrder   The flow order that would be used
         *                    in sequencing.
         */

        public static void reportSet(Set<Sequence> sequenceSet,String flowOrder) {
                Sequence sequences[] = sequenceSet.toArray(new Sequence[0]);
                Arrays.sort(sequences);

                for (int i = 0; i < sequences.length; i++) {
                        System.out.println("Sequence #" + (i + 1) + "\t"
                                        + new
Flowgram(flowOrder,sequences[i].getSequence()).getFlowValueString() + "\t"
                                        + sequences[i].getSequence());
                }
                System.out.println("---------");
                System.out.flush();
```


IDT Exhibit 1020.5020

```
        }

        /**
         * Example use of search code to generate a set of compatible
         * sequences that meet given design criteria of base composition,
         * maximum flow usage, and ability to correct and detect errors.
         *
         * @param args
         */
        public static void main(String[] args) {
                char fivePrimeAvoidBase    = 'G';
                final String flowOrder     = "TACG";
                final int   uidBaseLength  = 10;
                final int   uidMaxCycles   = 5;
                final int   errorsToCorrect = 2;
                final int   errorsToDetect  = 3;
                final int   desiredSetSize  = 12;
                System.out.println("Computation Begun");

                Set<Sequence> sequenceSet =
                        generateCandidateSequences(uidBaseLength,
                                                   uidMaxCycles,
                                                   flowOrder,
                                                   fivePrimeAvoidBase);
                System.out.println("Total candidate sequencess=" + sequenceSet.size());

                Set<Sequence> compatibleSequences =
                        searchCompatibleSet(sequenceSet,desiredSetSize,
                                            errorsToCorrect,errorsToDetect,
                                            SearchType.EditDistance);
                if (compatibleSequences == null) {
                        System.out.println("No solution");
                } else {
                        reportSet(compatibleSequences,flowOrder);
                }
        }
}
```

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.

Example 3- Table of computed UID sequences, cluster ID, and Flowgram script

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO

IDT Exhibit 1020.5021

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127176	14	01100101010110011010	ACAGAGTGTC	10	7
C1127176	14	01111010100101010100	ACGTCTGAGA	10	8
C1127176	14	01010111001001101010	AGACGCACTC	10	9
C1127176	14	01001010110010101011	ATCTATCTCG	10	10
C1127176	14	00110100111100111000	CGATACGCGT	10	11
C1127176	14	00110011001110010011	CGCGCGTGCG	10	12
C1127176	14	00111101010011010010	CGTAGATAGC	10	13
C1127176	14	00111001101010101100	CGTGTCTCTA	10	14
C1127176	14	00101010011001110110	CTCACACGAC	10	15
C1127176	14	11101010010010111000	TACTCATCGT	10	16
C1127176	14	11010011010011100100	TAGCGATACA	10	17
C1127176	14	11001001110111001000	TATGTAGTAT	10	18
C1127176	14	10101001001101101001	TCTGCGACTG	10	19
C1127176	14	10010110010110100101	TGACAGTCAG	10	20
C1127177	14	01101101001101010100	ACTAGCGAGA	10	21
C1127177	14	01010111010011001100	AGACGATATA	10	22
C1127177	14	01001010100101111010	ATCTGACGTC	10	23
C1127177	14	01001001101011010011	ATGTCTAGCG	10	24
C1127177	14	00110100111100111000	CGATACGCGT	10	25
C1127177	14	00110011001110010011	CGCGCGTGCG	10	26
C1127177	14	00111010011001010110	CGTCACAGAC	10	27
C1127177	14	00111001101010101100	CGTGTCTCTA	10	28
C1127177	14	11101010010101001001	TACTCAGATG	10	29
C1127177	14	11010010011010101010	TAGCACTCTC	10	30
C1127177	14	11001100111001100100	TATATACACA	10	31
C1127177	14	10100100101110100101	TCATCGTCAG	10	32
C1127177	14	10010101100100110110	TGAGTGCGAC	10	33
C1127177	14	10011001010111011000	TGTGAGTAGT	10	34
C1127178	14	01100110101010010110	ACACTCTGAC	10	35
C1127178	14	01010101010101101001	AGAGAGACTG	10	36
C1127178	14	01001111110010101000	ATACGTATCT	10	37
C1127178	14	01001011101101010100	ATCGTCGAGA	10	38
C1127178	14	00100110010111011100	CACAGTAGTA	10	39
C1127178	14	00110100111100111000	CGATACGCGT	10	40
C1127178	14	00110011001110010011	CGCGCGTGCG	10	41
C1127178	14	00111001101010101100	CGTGTCTCTA	10	42
C1127178	14	00101001110101001011	CTGTAGATCG	10	43
C1127178	14	11101001010100110010	TACTGAGCGC	10	44
C1127178	14	11010010101111001000	TAGCTCGTAT	10	45
C1127178	14	11001100111001100100	TATATACACA	10	46
C1127178	14	10110010011001101010	TCGCACACTC	10	47
C1127178	14	10101100100110011001	TCTATGTGTG	10	48
C1127179	14	01101011011111000000	ACTCGACGTA	10	49
C1127179	14	01010110100111010100	AGACTGTAGA	10	50
C1127179	14	01010101010101101001	AGAGAGACTG	10	51

IDT Exhibit 1020.5022

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127179	14	01001001101011010011	ATGTCTAGCG	10	52
C1127179	14	00100110111011001001	CACTACTATG	10	53
C1127179	14	00110100111100111000	CGATACGCGT	10	54
C1127179	14	00110011001110010011	CGCGCGTGCG	10	55
C1127179	14	00111010011001010110	CGTCACAGAC	10	56
C1127179	14	00111001101010101100	CGTGTCTCTA	10	57
C1127179	14	11110101001001010010	TACGAGCAGC	10	58
C1127179	14	11010010010010111001	TAGCATCGTG	10	59
C1127179	14	11001110011010100100	TATACACTCA	10	60
C1127179	14	10101001100110010110	TCTGTGTGAC	10	61
C1127179	14	10011101111001001000	TGTAGTACAT	10	62
C1127180	14	01101011010010101010	ACTCGATCTC	10	63
C1127180	14	01010110100111010100	AGACTGTAGA	10	64
C1127180	14	01010101010101101001	AGAGAGACTG	10	65
C1127180	14	01001001101011010011	ATGTCTAGCG	10	66
C1127180	14	00100110111011001001	CACTACTATG	10	67
C1127180	14	00110100111100111000	CGATACGCGT	10	68
C1127180	14	00110011001110010011	CGCGCGTGCG	10	69
C1127180	14	00111010011001010110	CGTCACAGAC	10	70
C1127180	14	00111001101010101100	CGTGTCTCTA	10	71
C1127180	14	11110101001001010010	TACGAGCAGC	10	72
C1127180	14	11010010010010111001	TAGCATCGTG	10	73
C1127180	14	11001110011010100100	TATACACTCA	10	74
C1127180	14	10101001100110010110	TCTGTGTGAC	10	75
C1127180	14	10011101111001001000	TGTAGTACAT	10	76
C1127181	14	01100110011100101001	ACACACGCTG	10	77
C1127181	14	01110100101001001101	ACGATCATAG	10	78
C1127181	14	01010101010101100110	AGAGAGACAC	10	79
C1127181	14	01001110110010010110	ATACTATGAC	10	80
C1127181	14	00110011001110010011	CGCGCGTGCG	10	81
C1127181	14	00111001101010101100	CGTGTCTCTA	10	82
C1127181	14	00101111011001011000	CTACGACAGT	10	83
C1127181	14	00101001110101001011	CTGTAGATCG	10	84
C1127181	14	11010010010110101100	TAGCAGTCTA	10	85
C1127181	14	11011001001100111000	TAGTGCGCGT	10	86
C1127181	14	10101100100110011001	TCTATGTGTG	10	87
C1127181	14	10101011001010100110	TCTCGCTCAC	10	88
C1127181	14	10010100111011101000	TGATACTACT	10	89
C1127181	14	10011010110101010100	TGTCTAGAGA	10	90
C1127182	14	01100101101011110000	ACAGTCTACG	10	91
C1127182	14	01010111001001101010	AGACGCACTC	10	92
C1127182	14	01010010111001001101	AGCTACATAG	10	93
C1127182	14	01011010100110010110	AGTCTGTGAC	10	94
C1127182	14	01001101010110011100	ATAGAGTGTA	10	95
C1127182	14	00110011001110010011	CGCGCGTGCG	10	96

IDT Exhibit 1020.5023

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127182	14	00111001101010101100	CGTGTCTCTA	10	97
C1127182	14	00101110110100101001	CTACTAGCTG	10	98
C1127182	14	00101001010101110101	CTGAGACGAG	10	99
C1127182	14	11011001001100111000	TAGTGCGCGT	10	100
C1127182	14	10100111110010010100	TCACGTATGA	10	101
C1127182	14	10111010010101001010	TCGTCAGATC	10	102
C1127182	14	10101100111001100100	TCTATACACA	10	103
C1127182	14	10010100110110110010	TGATAGTCGC	10	104
C1127183	14	01110100101100111000	ACGATCGCGT	10	105
C1127183	14	01101010110010011001	ACTCTATGTG	10	106
C1127183	14	01010010011001101101	AGCACACTAG	10	107
C1127183	14	01001110010101011010	ATACAGAGTC	10	108
C1127183	14	01001100101010100111	ATATCTCACG	10	109
C1127183	14	00100101110011110010	CAGTATACGC	10	110
C1127183	14	00110011001110010011	CGCGCGTGCG	10	111
C1127183	14	00111001101010101100	CGTGTCTCTA	10	112
C1127183	14	00101111111001001000	CTACGTACAT	10	113
C1127183	14	11001111001010010100	TATACGCTGA	10	114
C1127183	14	10110110010010101010	TCGACATCTC	10	115
C1127183	14	10110010110101100100	TCGCTAGACA	10	116
C1127183	14	10010101100100110110	TGAGTGCGAC	10	117
C1127183	14	10011001010111011000	TGTGAGTAGT	10	118
C1127184	14	01100111001010100110	ACACGCTCAC	10	119
C1127184	14	01110100101100111000	ACGATCGCGT	10	120
C1127184	14	01010111010101010100	AGACGAGAGA	10	121
C1127184	14	01010010100111001110	AGCTGTATAC	10	122
C1127184	14	01001101100101001011	ATAGTGATCG	10	123
C1127184	14	00100110111001101001	CACTACACTG	10	124
C1127184	14	00110011001110010011	CGCGCGTGCG	10	125
C1127184	14	00111101011101100000	CGTAGACGAC	10	126
C1127184	14	00111001101010101100	CGTGTCTCTA	10	127
C1127184	14	11100100110010010101	TACATATGAG	10	128
C1127184	14	10101010101101100100	TCTCTCGACA	10	129
C1127184	14	10101001010100101101	TCTGAGCTAG	10	130
C1127184	14	10010101010011101010	TGAGATACTC	10	131
C1127184	14	10011110100110011000	TGTACTGTGT	10	132
C1127185	14	01100100101110101001	ACATCGTCTG	10	133
C1127185	14	01110010100111011000	ACGCTGTAGT	10	134
C1127185	14	01010101010101100110	AGAGAGACAC	10	135
C1127185	14	01011010010100111100	AGTCAGCGTA	10	136
C1127185	14	01001111001001110100	ATACGCACGA	10	137
C1127185	14	00100100100111010111	CATGTAGACG	10	138
C1127185	14	00110011001110010011	CGCGCGTGCG	10	139
C1127185	14	00111001101010101100	CGTGTCTCTA	10	140
C1127185	14	00101110010010011110	CTACATGTAC	10	141

IDT Exhibit 1020.5024

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127185	14	11101110100100101000	TACTACTGCT	10	142
C1127185	14	11010101010010011001	TAGAGATGTG	10	143
C1127185	14	10100101011011010100	TCAGACTAGA	10	144
C1127185	14	10011100101101010010	TGTATCGAGC	10	145
C1127185	14	10011011111001001000	TGTCGTACAT	10	146
C1127186	14	01100100101110101001	ACATCGTCTG	10	147
C1127186	14	01110010100111011000	ACGCTGTAGT	10	148
C1127186	14	01010101010101100110	AGAGAGACAC	10	149
C1127186	14	01011010010100111100	AGTCAGCGTA	10	150
C1127186	14	01001111001001110100	ATACGCACGA	10	151
C1127186	14	00100100100111010111	CATGTAGACG	10	152
C1127186	14	00110011001110010011	CGCGCGTGCG	10	153
C1127186	14	00111001101010101100	CGTGTCTCTA	10	154
C1127186	14	00101110010010011110	CTACATGTAC	10	155
C1127186	14	11101110100100101000	TACTACTGCT	10	156
C1127186	14	11010101010010011001	TAGAGATGTG	10	157
C1127186	14	10100101011011010100	TCAGACTAGA	10	158
C1127186	14	10110010011001101010	TCGCACACTC	10	159
C1127186	14	10011100101101010010	TGTATCGAGC	10	160
C1127187	14	01100111001010100110	ACACGCTCAC	10	161
C1127187	14	01110010100111011000	ACGCTGTAGT	10	162
C1127187	14	01011010010010111010	AGTCATCGTC	10	163
C1127187	14	01011001010101100101	AGTGAGACAG	10	164
C1127187	14	01001101010110011100	ATAGAGTGTA	10	165
C1127187	14	00100110010011110101	CACATACGAG	10	166
C1127187	14	00110011001110010011	CGCGCGTGCG	10	167
C1127187	14	00111001101010101100	CGTGTCTCTA	10	168
C1127187	14	00101010110101101010	CTCTAGACTC	10	169
C1127187	14	11001110101001010100	TATACTCAGA	10	170
C1127187	14	11001011110010110000	TATCGTATCG	10	171
C1127187	14	10111110010011001000	TCGTACATAT	10	172
C1127187	14	10101001100110010110	TCTGTGTGAC	10	173
C1127187	14	10010111011100101000	TGACGACGCT	10	174
C1127188	14	01100100101110101001	ACATCGTCTG	10	175
C1127188	14	01110010011101001100	ACGCACGATA	10	176
C1127188	14	01010101110100111000	AGAGTAGCGT	10	177
C1127188	14	01011110011001101000	AGTACACACT	10	178
C1127188	14	01011010100110010110	AGTCTGTGAC	10	179
C1127188	14	00100110111011010010	CACTACTAGC	10	180
C1127188	14	00110011001110010011	CGCGCGTGCG	10	181
C1127188	14	00111001101010101100	CGTGTCTCTA	10	182
C1127188	14	00101110010110011100	CTACAGTGTA	10	183
C1127188	14	00101001110101001011	CTGTAGATCG	10	184
C1127188	14	11001011001101011000	TATCGCGAGT	10	185
C1127188	14	10110110010100100101	TCGACAGCAG	10	186

IDT Exhibit 1020.5025

Cluster Id	Member Count	Flowgram TACGTACGTACGTACGTACG (SEQ ID NO: 6)	UID	UID Length	SEQ ID NO
C1127188	14	10101010010011110100	TCTCATACGA	10	187
C1127188	14	10010101010011001110	TGAGATATAC	10	188
C1127189	14	01100101001010110110	ACAGCTCGAC	10	189
C1127189	14	01101011010011100100	ACTCGATACA	10	190
C1127189	14	01010100110101101100	AGATAGACTA	10	191
C1127189	14	01010011001110011001	AGCGCGTGTG	10	192
C1127189	14	01001001101011010011	ATGTCTAGCG	10	193
C1127189	14	00100111110111001000	CACGTAGTAT	10	194
C1127189	14	00110110011100100101	CGACACGCAG	10	195
C1127189	14	00111001010101010110	CGTGAGAGAC	10	196
C1127189	14	00111001101010101100	CGTGTCTCTA	10	197
C1127189	14	11101100101100101000	TACTATCGCT	10	198
C1127189	14	11011001001001100101	TAGTGCACAG	10	199
C1127189	14	10110010010101111000	TCGCAGACGT	10	200
C1127189	14	10101110100110010100	TCTACTGTGA	10	201
C1127189	14	10010111101001001010	TGACGTCATC	10	202
C1127190	14	01100101011001001101	ACAGACATAG	10	203
C1127190	14	01110011001110011000	ACGCGCGTGT	10	204
C1127190	14	01010111001001101010	AGACGCACTC	10	205
C1127190	14	01010010110010110101	AGCTATCGAG	10	206
C1127190	14	01001100100110011110	ATATGTGTAC	10	207
C1127190	14	01001001110101111000	ATGTAGACGT	10	208
C1127190	14	00100110101110100110	CACTCGTCAC	10	209
C1127190	14	00111001010101010110	CGTGAGAGAC	10	210
C1127190	14	00111001101010101100	CGTGTCTCTA	10	211
C1127190	14	00101010011011011001	CTCACTAGTG	10	212
C1127190	14	11101100110010101000	TACTATATCT	10	213
C1127190	14	11001110101001010100	TATACTCAGA	10	214
C1127190	14	10110010010111100100	TCGCAGTACA	10	215
C1127190	14	10010101100100110011	TGAGTGCGCG	10	216
C1127191	14	01100101011010011010	ACAGACTGTC	10	217
C1127191	14	01010011001111100100	AGCGCGTACA	10	218
C1127191	14	01011011010100101001	AGTCGAGCTG	10	219
C1127191	14	01001110110101010010	ATACTAGAGC	10	220
C1127191	14	01001011101001111000	ATCGTCACGT	10	221
C1127191	14	00110101001001110011	CGAGCACGCG	10	222
C1127191	14	00111001010101010110	CGTGAGAGAC	10	223
C1127191	14	00111001101010101100	CGTGTCTCTA	10	224
C1127191	14	00101010010010111101	CTCATCGTAG	10	225
C1127191	14	11100110101100101000	TACACTCGCT	10	226
C1127191	14	11001001110111001000	TATGTAGTAT	10	227
C1127191	14	10110010100110011001	TCGCTGTGTG	10	228
C1127191	14	10101100111001100100	TCTATACACA	10	229
C1127191	14	10010111101001001010	TGACGTCATC	10	230
C1127192	14	01101001100101001011	ACTGTGATCG	10	231

IDT Exhibit 1020.5026