

University of Leicester

Molecular analysis of circulating cell-free DNA in breast cancer patients

Thesis submitted for the degree of Doctor of Philosophy at the University of Leicester

by

Kevin Anthony Blighe, BSc Honours (Republic of Ireland),
University of Leicester,
Department of Cancer Studies and Molecular Medicine

2012

Personalis EX2155

Abstract

Identifying disease-related variants is a primary aim of human genetics. In breast cancer, loss of heterozygosity at specific loci was previously demonstrated in paired tumour and circulating plasma cell-free DNA (cfDNA) samples. However, alterations unique to cfDNA were also found in all cases, suggesting disease progression. These results prompted the characterisation of the circulating breast cancer genome in more detail in this thesis, to test the hypothesis that cfDNA acts as a surrogate tumour marker. This was achieved using Affymetrix SNP 6.0 technology and bioinformatics to map SNP and copy number variation (CNV), comparing cfDNA with matched normal lymphocyte and tumour DNA in 65 breast cancer patients and 8 healthy female controls.

Results in this thesis show that comparison of cfDNA SNP genotypes can distinguish between primary breast cancer patients and healthy controls ($p<0.0001$), and between pre-surgical breast cancer patients and those who already had surgery/treatment ($p=0.0016$). A significant difference ($p=0.0006$) was also found between cfDNA samples taken an average of 3 years apart in women on follow-up, again suggesting progression. In addition, CNV amplification was observed in matched tumour and cfDNA at numerous loci on different chromosome arms. Many of these tumour-specific CNVs contributed significantly to disease through logistic regression analysis and remained detectable in cfDNA up to 12 years after diagnosis despite no other evidence of disease. This finding strongly infers breast cancer dormancy in the majority of patients on follow-up.

In addition, candidate CNVs were validated by real-time qPCR and additional bioinformatics revealed key SNP and CNV signatures of breast cancer patients. If validated in other patient series, the results could alter the diagnostic and prognostic landscape of breast cancer.

Future studies will focus on developing high throughput approaches to target common SNPs/CNVs with a view to developing a targeted custom DNA chip for screening and monitoring.

Personalis EX2155

Acknowledgements

This work is dedicated to all of those who have been affected directly or indirectly by breast cancer, including relatives, friends, and acquaintances. I sincerely hope that my work has been beneficial to this sensitive area of research and that it will go toward something of greater potential that can help all afflicted with breast cancer. Also, to my grandmother: I wish that you were still alive today - my aim has always been to make you proud of me.

I also must give thanks to my family - including numerous aunts, uncles, and cousins - and my three sisters-in-law Amanda, Neasa, and Sabrina - for having an on-going interest in my research. Also, lest I forget: Lindsay, Angie, Linda, and Janine, for the smooth running of the laboratory; Dr. Jacqui Shaw for being a great supervisor, friend, and colleague, but also for being a great role model in research; Dr. Karen Page, with whom I worked every day, for also being a great colleague and friend; Youssef, Ali Al-Harth, and their families, Eftychios Papadogeorgakis and Eleni, Drs. J Howard Pringle, Shona Elshaw, and Chetana Ruangpratheep for being great friends and for putting up with my antics and personal life brought into the office!

Also, special thanks to: Karen Kulbicki for always making me smile; Nathalie Zahra for sticking by me through highs and lows; the Maria Tilton 'Hope' Foundation, Cancer Research UK, and Breast Cancer Campaign; Joan in Finance for administrative assistance; Jim Strupish for keeping the corridors alive with the sound of whistle; the Leicester Market for providing the fruit that fuelled the passion for my research; Theresa Visser, Amy Moore, and Ush Taylor for administrative help on Floor 3; Malcolm Rae and Sally Munton for purchasing help; Jon Naylor for our refreshing conversations on computers; all of my former lecturers and friends at the Institute of Technology in Carlow, Ireland; and a 'go rabh míle maith agat' to all of my other Irish friends back home. I would also like to acknowledge Professors R Charles Coombes and Justin Stebbing at Imperial College London for being personal role models, and Barb, Sharon, Anne-Marie, Chris, Rebecca (Becksteropilopidous), and Colette for their friendship and support.

Personalis EX2155

Publications relating to this thesis

Page K, Hava N, Ward B, Brown J, Guttery DS, Ruangpratheep C, **Blighe K**, Sharma A, Walker RA, Coombes RC, and Shaw JA (April, 2011), *Detection of HER2 amplification in circulating free DNA in patients with breast cancer*, British Journal of Cancer 104(8): 1342-1348

Payne R, Hava N, Page K, **Blighe K**, Ward B, Slade M, Brown J, Guttery D, Zaidi SAA, Stebbing J, Jacob J, Tat T, Yagüe E, Shaw JA, and Coombes RC (January, 2012), *The presence of disseminated tumour cells in the bone marrow is inversely related to circulating-free DNA in plasma in breast cancer dormancy*, British Journal of Cancer 106(2): 375-382

Shaw JA, Page K, **Blighe K**, Hava N, Guttery D, Ward B, Brown J, Ruangpratheep C, Stebbing J, Payne R, Palmieri C, Cleator S, Walker RA, and Coombes RC (February, 2012), *Genomic analysis of circulating cell free DNA infers breast cancer dormancy*, Genome Research 22(2): 220-231

Stebbing J, Filipovic A, Lit LC, **Blighe K**, Grothey A, Sommer A, Beckmann G, Yasuhiro M, Chow LWC, Coombes RC, Sasano H, Shaw JA, and Giamas G, *LMTK3 is implicated in intrinsic endocrine resistance via multiple signaling pathways* (manuscript in preparation)

Personalis EX2155

Abbreviations

Acronym	Definition
A	Eigenvector
aCGH	Array comparative genomic hybridisation
ACS	American Cancer Society
ADHD	Attention deficit hyperactivity disorder
ANOVA	Analysis of variance
ASD	Autism spectrum disorder
BCAS3	Breast carcinoma amplified sequence 3
BCSC	Breast Cancer Surveillance Consortium
BD	Benign breast disease
BIDD	Biomarkers and Imaging Discovery and Development
BLAST	Basic Local Alignment Search Tool
bp	Base-pair
BRLMM	Bayesian robust linear model with mahalanobis
cDNA	Complementary DNA
cfDNA	Cell-free DNA
CEPH	Utah, USA residents with ancestry from northern and western Europe
CHB	Han Chinese in Beijing, China
CIN	Chromosomal instability
CQC	Contrast QC
CO_2	Carbon dioxide
CN	Copy number
CNA	Copy number alteration
CN-LOH	Copy neutral-LOH
CNV	Copy number variation
COSMIC	Catalogue of Somatic Mutations in Cancer
cov	Covariance
CRUK	Cancer Research UK
C_T	Cycle threshold
CTC	Circulating tumour cell
DBC1	Deleted in bladder cancer 1
DCIS	Ductal carcinoma in-situ
DGV	Database of genomic variants
DIRC3	Disrupted in renal carcinoma
DLC1	Deleted in liver cancer 1
DTC	Disseminated tumour cell
DMEM	Dulbecco's Modified Eagle Medium
e	Eigenvalue
ECM	Extracellular matrix
ECMC	Experimental Cancer Medicine Centre
EDTA	Ethylenediaminetetraacetic acid
EGFR	Epidermal growth factor receptor
EM	Expectation maximisation
EMT	Epithelial-to-mesenchymal transition
ER	Oestrogen receptor

Personalis EX2155

ERBB2	HER-2/neu
ESR1	Oestrogen receptor 1-α
FAM	Family with sequence similarity
FDA	Food and Drug Administration
FDR	False discovery rate
FGF	Fibroblast growth factor
FGFR	Fibroblast growth factor receptor
FISH	Fluorescent in situ hybridisation
FRS2	Fibroblast growth factor receptor substrate 2
g	Gram
GTC	Genotyping Console
H&E	Haematoxylin and Eosin
hESC	Human embryonic stem cell
HMM	Hidden Markov model
HR	High risk
HRT	Hormone replacement therapy
htSNP	Haplotype-tagging SNP
IARC	International Agency for Research on Cancer
IGF1	Insulin-like growth factor 1
IGFBP3	IGF-binding protein-3
IGFR	Insulin-like growth factor receptor
IHC	Immunohistochemistry
HSP	Heat-shock protein
JPT	Japanese in Tokyo, Japan
Kb	Kilobase
KDa	KiloDalton
L	Litre
LCIS	Lobular carcinoma in-situ
LD	Linkage disequilibrium
LH	Luteinising hormone
LHRH	Luteinising hormone releasing hormone
LOH	Loss of heterozygosity
Lowess	Locally-weighted linear regression
LR	Low risk
M	Molar
MAPD	Median of the absolute values of all pairwise differences
MAPK	Mitogen activated protein kinase
Mb	Megabase
mGy	MilliGray
ml	Millilitre
MM	Mismatch
MRD	Minimal residual disease
MRI	Magnetic resonance imaging
MTA1	Metastasis associated 1
NaCl	Sodium chloride
NAHR	Non-allelic homologous recombination
NAIP	Neuronal apoptosis inhibitory protein
NCBI	National Centre for Biotechnology Information
NCI	National Cancer Institute
NeoCEnT	Neo-adjuvant chemo- or endocrine therapy

Personalis EX2155

ng	Nanogram
NGS	Next-generation sequencing
NHS	National Health Service
NHSBSP	NHS Breast Screening Programme
nm	Nanometre
NTN1	Netrin-1
OMIM	Online Mendelian Inheritance in Man
PAM	Partitioning around means
PCA	Principal component analysis
P/C/IAA	Phenol/chloroform/isoamylalcohol
pH	Power of the [hydrogen ion]
PI3K	Phosphoinositide kinase-3
pM	Picomolar
PM	Perfect match
PR	Progesterone receptor
PRAME	Preferentially expressed antigen in melanoma
QC	Quality control
qPCR	Quantitative PCR
REC	Research Ethics Committee
ROC	Receiver operating characteristic
RPIA	Ribose 5-phosphate isomerase A
rpm	Revolutions per minute
RPMI	Roswell Park Memorial Institute
RQ	Relative quantitation
RS	Recurrence score
SAP	Application of SNP arrays (S) to monitor the post-adjuvant (A) period (P)
SDS	Sodium dodecyl sulphate
SERM	Selective estrogen receptor modulator
SNP	Single nucleotide polymorphism
TERT	Telomerase reverse transcriptase
UCSC	University of California Santa Cruz
µl	Microlitre
var	variance
VNTR	Variable number of tandem repeats
WGA	Whole genome amplification
YRI	Yoruba in Ibadan, Nigeria
σ	Standard deviation
φ29	Phi29
℃	Degrees Celsius
$\overline{X}$	Mean

Personalis EX2155

Table of contents

ABSTRACT ..... **I**

ACKNOWLEDGEMENTS ..... **II**

PUBLICATIONS RELATING TO THIS THESIS ..... **III**

ABBREVIATIONS ..... **IV**

TABLE OF CONTENTS ..... **VII**

INDEX OF FIGURES ..... **XV**

INDEX OF TABLES ..... **XXII**

1 INTRODUCTION **1**

1.1 THE BREAST 2

1.2 BREAST CANCER: HISTORY AND CURRENT RESEARCH FOCUS 3

1.2.1 Incidence 3

1.2.1.1 Future trends 5

1.2.2 Types of breast cancer 6

1.2.3 Morphological and molecular factors 7

1.2.4 Risk factors 8

1.2.5 Detection and screening 10

1.2.5.1 Breast mammography 10

1.2.5.2 Therapy 12

1.3 MOLECULAR PROFILING AND PERSONALISED MEDICINE 13

1.3.1 Breast gene expression subtypes 16

1.3.2 Commercial profiling 18

1.3.3 Profiling with SNPs and CNV 20

1.3.4 Next-generation sequencing 21

1.3.5 Potential of DNA arrays 24

1.3.5.1 Affymetrix microarrays 25

1.3.5.2 The use of *Affymetrix SNP 6.0* 26

1.3.5.2.1 Genotyping calling algorithms 27

Personalis EX2155

1.3.5.2.2 Capabilities 28
1.4 CIRCULATING CELL-FREE DNA 28
1.4.1 Origins of cfDNA 29
1.4.1.1 Tumourigenic origin 30
1.4.1.2 Other sources of cfDNA 31
1.4.2 Biological importance of cfDNA 33
1.4.2.1 Cell-free DNA as a diagnostic and prognostic marker 33
1.4.2.2 The search for markers of minimal disease 34
1.5 BACKGROUND TO THESIS 35
1.6 HYPOTHESIS TO BE TESTED 35
1.7 AIMS 36
1.8 OBJECTIVES 36
2 MATERIALS AND METHODS 37
2.1 PROJECT FUNDING AND SPONSORSHIP DETAILS 38
2.2 ASSOCIATED CLINICAL STUDIES AND TRIALS 38
2.3 SAMPLE INFORMATION 40
2.4 SAMPLE PROCESSING 43
2.5 DNA EXTRACTION AND WHOLE GENOME AMPLIFICATION 43
2.6 MICROARRAY PROCESSING 44
2.6.1 The scanning process 45
2.7 BIOINFORMATIC DATA ANALYSIS 45
2.7.1 Normalisation 46
2.7.1.1 Referencing and derivation of $\log_2$ values 47
2.7.1.2 Variations to normalisation 48
2.7.2 Initial QC 49
2.7.3 Genotyping and generation of CN segments 49
2.7.3.1 Single nucleotide polymorphism concordance checks 49
2.7.3.2 Genomic segmentation 50
2.7.4 Principal component analysis and clustering 51
2.7.4.1 Genetic signature of breast cancer 51
2.7.5 Loss of heterozygosity 52
2.7.6 Copy number variation frequency analysis 53

Personalis EX2155

2.7.7 Genomic aberration detection 53
2.8 CULTURING CELL-LINES 54
2.9 DESIGN OF PRIMERS AND PROBES 55
2.10 REAL-TIME QUANTITATIVE PCR 56
2.11 STATISTICAL ANALYSIS 57

3 MOLECULAR CHARACTERISATION OF CIRCULATING CELL-FREE DNA IN BREAST CANCER PATIENTS USING SNP 6.0 MICROARRAYS ... 58

3.1 OVERVIEW 59
3.2 INTRODUCTION 59
3.2.1 Genome-wide characterisation 60
3.2.2 Sequential analysis 60
3.2.3 Aims 61
3.3 RESULTS 62
3.3.1 Reproducibility and sensitivity of qPCR for measuring *ERBB2* copy number in breast cancer cell-lines 62
3.3.2 *Affymetrix SNP 6.0* QC analysis of whole genome amplification cfDNA, tumour, and lymphocyte DNA samples 69
3.3.3 Single nucleotide polymorphisms and the comparison of genomic profiles 74
3.3.3.1 HapMap comparisons 74
3.3.3.2 Matched lymphocyte comparisons 74
3.3.3.3 Matched tumour comparisons 77
3.3.4 Copy number variation analysis 78
3.3.4.1 Status 78
3.3.4.2 Number of CNVs called 78
3.3.4.3 Amplification and deletion 82
3.3.4.3.1 Distribution 82
3.3.4.3.2 Thresholding amplifications by copy number 83
3.3.4.3.3 Logistic regression to reveal significant copy number variant regions 97
3.3.5 Principal component analysis and hierarchical clustering to interrogate both SNP and CN data 99

Personalis EX2155

3.3.5.1 Principal component analysis 99
3.3.5.2 Hierarchical clustering 108
3.3.6 Loss of heterozygosity analysis 110
3.4 DISCUSSION 113
3.4.1 *ERBB2* assay validation experiments using breast cell-lines 114
3.4.2 Formalin-fixed Vs fresh frozen tissue 114
3.4.3 Single nucleotide polymorphism concordance segregated cfDNA of healthy controls from breast cancer patients 115
3.4.4 Principal component analysis reveals ethnicity and may predict breast cancer 116
3.4.5 Copy number variation 117
3.4.5.1 Many novel CNVs 118
3.4.5.2 Amplifications and deletions 119
3.4.5.3 Commonly-amplified intervals in plasma cfDNA 120
3.4.5.4 Receptor status 121
3.4.5.5 Profiles of relapsed patients 121
3.4.5.6 Surgery and therapy received 122
3.4.6 Genes overlapping commonly-amplified CNVs 123
3.4.6.1 4q13.2 123
3.4.6.2 5q13.2 123
3.4.6.3 7q11.23 124
3.4.6.4 8p23.1 124
3.4.6.5 10q11.22 and 10q11.23 125
3.4.6.6 15q25.2 126
3.4.6.7 16p12.3 126
3.4.6.8 Other CNVs with potential genes of interest 126
3.4.7 Breast cancer dormancy 128
3.5 CONCLUSION 130
3.6 FUTURE STUDIES 131

4 MONITORING CIRCULATING CELL-FREE DNA IN RESPONSE TO THERAPY IN BREAST CANCER PATIENTS USING SNP 6.0 MICROARRAYS 132

Personalis EX2155

4.1 OVERVIEW 133
4.2 INTRODUCTION 133
4.2.1 Cancer therapy 133
4.2.2 Resistance to therapy 134
4.2.3 The NeoCEnT clinical trial 135
4.2.4 Aims 136
4.3 RESULTS 136
4.3.1 *Affymetrix SNP 6.0* QC analysis of cfDNA, tumour, and lymphocyte samples 136
4.3.2 Single nucleotide polymorphisms and the comparison of genomic profiles 139
4.3.3 Copy number variation analysis 141
4.3.3.1 Number of CNVS called 141
4.3.3.2 Amplifications and deletions 143
4.3.3.2.1 Frequency 143
4.3.3.2.2 Distribution 144
4.3.3.2.3 Thresholding amplifications by copy number 145
4.3.3.2.4 Logistic regression to reveal significant copy number variant regions 149
4.3.3.3 Copy number variation in growth factor receptor genes 150
4.3.3.4 Detection of genomic aberrations 151
4.3.4 Principal component analysis to interrogate both SNP and CN data 155
4.4 DISCUSSION 158
4.4.1 Microarray detects genomic aberrations 158
4.4.2 Single nucleotide polymorphism/copy number variation changes with time and therapy 159
4.4.2.1 Single nucleotide polymorphism concordance checks give mixed results 159
4.4.2.2 Plasma cfDNA CNV profile becomes less heterogeneous 161
4.4.2.3 Principal component analysis indicates subtle changes in the plasma cfDNA profile 161
4.4.2.4 Plasma cfDNA CNVs become less prevalent 162
4.4.3 Genes overlapping commonly-amplified CNVs 163

Personalis EX2155

4.4.4 *Affymetrix SNP 6.0* reproducibility 165
4.5 CONCLUSION 165
4.6 FUTURE STUDIES 166

5 PAINTING A PORTRAIT OF BREAST CANCER THROUGH PRINCIPAL COMPONENT ANALYSIS 167

5.1 OVERVIEW 168
5.2 INTRODUCTION 168
5.2.1 Principal component analysis as a tool for analysing large bodies of data 168
5.2.1.1 Limitations of PCA 169
5.2.2 Deriving a genetic 'signature' 169
5.2.3 Aims 171
5.3 RESULTS 171
5.3.1 Genes with known roles in breast cancer 171
5.3.1.1 Markers overlapping exons 179
5.3.1.2 Markers overlapping introns 180
5.3.1.2.1 Breast cancer patients on follow-up 180
5.3.1.2.2 Relapsed breast cancer patients on follow-up 180
5.3.1.2.3 Unaffected BRCA1 mutation carriers 182
5.3.1.2.4 Pre-surgical breast cancer patients on neoadjuvant therapy 183
5.3.2 Towards a breast cancer signature through cfDNA analysis 184
5.4 DISCUSSION 189
5.4.1 Exons were infrequently affected by variation 190
5.4.2 Principal component analysis revealed numerous existing and novel gene targets of oncogenic potential 191
5.4.2.1 Patients on follow-up 192
5.4.2.1.1 Lymphocyte Vs Tumour DNA 192
5.4.2.1.2 T1N0 (T≤20mm) Vs T>20mm/node-positive (first and second cfDNA) 194
5.4.2.2 Relapsed patients' final cfDNA samples 195
5.4.2.2.1 First Vs Final/second cfDNA 195
5.4.2.2.2 Tumour DNA Vs Final cfDNA 196

Personalis EX2155

5.4.2.2.3 Non-relapsed T1N0 (T≤20mm) Vs Relapsed T1N0 (T≤20mm) final cfDNA 197
5.4.2.2.4 Non-relapsed T>20mm/node-positive Vs Relapsed T>20mm/node-positive final cfDNA 198
5.4.2.3 Healthy control Vs Unaffected *BRCA1* carriers' cfDNA 199
5.4.2.4 Pre-surgical patients' cfDNA 199
5.4.3 ER receptor activity is altered in tumour and cfDNA 200
5.4.4 SNP is more conducive to variation than CNV 202
5.4.5 Principal component analysis creates breast cancer cfDNA genetic signatures 203
5.4.5.1 Genetic signatures are highly defined and narrow 204
5.5 CONCLUSION 205
5.6 FUTURE STUDIES 206

6 MICROARRAY DATA VALIDATION 207

6.1 OVERVIEW 208
6.2 INTRODUCTION 208
6.3 AIMS 208
6.4 RESULTS 209
6.4.1 Examples of known genomic regions of variance 209
6.4.2 *ERBB2* gene amplification 209
6.4.3 Analysis of commonly-amplified CNVs by qPCR 211
6.4.4 Comparison to other studies 213
6.4.4.1 *LMTK3* as a potential oncogene 213
6.4.4.2 Comparison to aCGH data of breast primary tumour and blood 216
6.4.4.3 Comparison to aCGH data of breast primary tumour and cancer cell-line 219
6.5 DISCUSSION 220
6.5.1 Real-time qPCR validates microarray results 220
6.5.2 LMTK3 is involved in resistance to therapy 220
6.5.3 Microarray can serve as a retrospective and prospective tool 221
6.6 CONCLUSION 222

7 CONCLUSIONS AND FUTURE DIRECTION 223

Personalis EX2155

7.1 ILLUMINATING THE BREAST CANCER GENOME 224
7.1.1 Common CNVs and the cfDNA genetic signatures may have utility in the clinical setting 224
7.1.2 Genes significantly-different through PCA can be used as panel of cancer biomarkers 227
7.2 MISSING HERITABILITY? 228
7.3 UNRESOLVED QUESTIONS 228
7.4 CONCLUSION AND FUTURE DIRECTION 229

8 APPENDIX I

8.1 23 COMMONLY-AMPLIFIED INTERVALS - ADDITIONAL INFORMATION II
8.2 GENOMIC ABERRATIONS IV
8.3 EIGENVECTORS AND EIGENVALUES XV
8.3.1 Derive the eigenvectors and eigenvalues XVII
8.4 PRINCIPAL COMPONENT ANALYSIS XVIII
8.4.1 End analysis XX
8.4.1.1 Selecting principal components XX
8.4.1.2 Selecting variables XXI
8.4.1.2.1 Remove and compare XXI
8.4.1.2.2 High correlation XXI
8.4.1.2.3 Rotation of principal components XXII
8.5 VARIANTS WITH ROLES IN BREAST CANCER - ADDITIONAL INFORMATION .. XXIII
8.6 REFERENCES XXX

Personalis EX2155

Index of figures

FIGURE 1-1 - BREAST ANATOMY, SHOWING THE RELATIVE LOCATIONS OF THE DUCTS, LOBES, LOBULES, AND ALVEOLI. ALSO HIGHLIGHTED IS THE PROXIMITY OF THE SURROUNDING LYMPH NODES AND VESSELS (ADAPTED FROM CRUK (2011A). 2

FIGURE 1-2 - BREAST CANCER INCIDENCE AND MORTALITY RATES IN GREAT BRITAIN PER 100,000 POPULATION, HIGHLIGHTING THE INCREASE IN INCIDENCE YET CONCURRENT DECREASE IN MORTALITY (TAKEN FROM CRUK, 2011B). 4

FIGURE 1-3 - BREAST CANCER IS PREDICTED TO ACCOUNT FOR 29% OF ALL FEMALE CANCER CASES IN THE UK IN 2030 (TAKEN FROM MISTRY *ET AL.*, 2011). 6

FIGURE 1-4 - HIERARCHICALLY-CLUSTERING THE GENE EXPRESSION DATA FROM A SUBSET OF GENES IN BREAST TUMOURS REVEALED 5 DIFFERENT SUBTYPES. A, ERBB2 OVEREXPRESSION; B, LUMINAL B; C, BASAL-LIKE; D, 'NORMAL' BREAST-LIKE; E, LUMINAL A (ADAPTED FROM SORLIE *ET AL.*, 2003). 15

FIGURE 1-5 - GENOMIC PROFILE OF 4 BREAST CANCER SUBTYPES. A, SIMPLE (1Q/16) RELATING TO LUMINAL A; B, AMPLIFIER (FIRESTORM) RELATING TO LUMINAL B; C, AMPLIFIER (FIRESTORM) RELATING TO *ERBB2*; D, COMPLEX (SAWTOOTH) RELATING TO BASAL-LIKE (ADAPTED FROM KWEI *ET AL.*, 2010). 16

FIGURE 1-6 - IMAGE HIGHLIGHTING CORRELATION BETWEEN AMPLIFICATION AND DELETION (TOP), INCREASED AND DECREASED GENE EXPRESSION (MIDDLE), AND HYPER- AND HYPOMETHYLATION (BOTTOM) (TAKEN FROM ANDREWS *ET AL.*, 2010). 18

FIGURE 1-7 - ALIGNMENT OF NGS SEQUENCE READS FOR TWO TEST SAMPLES TO A REFERENCE GENOME (HG19) IDENTIFIES A SINGLE SNP ON CHROMOSOME 4 IN AN EXON OF *FGFR3* IN BOTH SAMPLES. BLUE, ADENINE; YELLOW, THYMINE; GREEN, GUANINE; RED, CYTOSINE. 23

FIGURE 1-8 - DETERMINING THE RELATIONSHIP BETWEEN CFDNA AND CTCS IN THE BLOOD CIRCULATION (A) AND DTCS IN THE BONE MARROW (B). 32

FIGURE 2-1 – FLOW-DIAGRAM OF SNP GENOTYPE COMPARISON TO DETERMINE SNP CONCORDANCE. 50

Personalis EX2155

FIGURE 3-1 - *ERBB2* RQ VALUES FOR 6 BREAST CANCER CELL-LINES USING DIFFERENT ENDOGENOUS CONTROL COMBINATIONS (GROUPED ACCORDING TO CELL-LINE): *CNTNAP1* (A); *RPPH1* (B); AND *CNTNAP1* WITH *RPPH1* (C). 65

FIGURE 3-2 - *ERBB2* RQ VALUES FOR 6 BREAST CANCER CELL-LINES USING DIFFERENT ENDOGENOUS CONTROL COMBINATIONS (GROUPED ACCORDING TO NUMBER OF REPLICATES): *CNTNAP1* (A); *RPPH1* (B); AND *CNTNAP1* WITH *RPPH1* (C). 67

FIGURE 3-3 - SNP CALL-RATES ACROSS 273 DNA SAMPLES USED IN THE STUDY. 70

FIGURE 3-4 - CLOSE CORRELATION IN QC CALL-RATE BETWEEN REPEAT-SAMPLE EXTRACTION AND HYBRIDISATION. L, LYMPHOCYTE; P1, FIRST PLASMA CFDNA; P2, SECOND PLASMA CFDNA; T, TUMOUR. 72

FIGURE 3-5 - GENOME-VIEW REPRESENTATION OF THE FREQUENCY OF AMPLIFICATION AND DELETION IN THE 13 SAMPLES HYBRIDISED IN 2 BATCHES. THE Y-AXIS INDICATES THE PERCENTAGE OF SAMPLES HAVING A CNV AT A SPECIFIC POINT IN THE GENOME. CHROMOSOME NUMBER AT TOP; AMPLIFICATIONS IN GREEN; DELETIONS IN RED. 73

FIGURE 3-6 - HIGH % SNP CONCORDANCE BETWEEN PLASMA CFDNA AND MATCHED LYMPHOCYTE DNA IN HEALTHY CONTROLS (MEAN 89.35%, RANGE 81.10-94.08%). 75

FIGURE 3-7 - SCATTER PLOT SHOWING % CONCORDANCE BETWEEN PLASMA CFDNA AND MATCHED LYMPHOCYTE DNA FOR EACH GROUP REVEALS A SIGNIFICANT DIFFERENCE BETWEEN EACH. P1, FIRST PLASMA CFDNA; P2, SECOND PLASMA CFDNA. 76

FIGURE 3-8 - SCATTER PLOT SHOWING % CONCORDANCE BETWEEN PLASMA CFDNA AND MATCHED TUMOUR DNA REVEALS A SIGNIFICANT DIFFERENCE BETWEEN ON FOLLOW-UP AND PRE-SURGICAL PATIENTS. P1, FIRST PLASMA CFDNA; P2, SECOND PLASMA CFDNA. 77

FIGURE 3-9 - NUMBER OF CNVS CALLED BETWEEN THE FIRST (A) AND SECOND (B) PLASMA CFDNA AND MATCHED LYMPHOCYTE DNA SAMPLES BY SUBGROUP. NB - T>20MM/NODE-POSITIVE PATIENTS OF PATIENTS ON FOLLOW-UP NOT SHOWN. 79

FIGURE 3-10 - KARYOVIEW OF THE LYMPHOCYTE (A), FIRST AND SECOND PLASMA CFDNA (B AND C), AND TUMOUR (D) SAMPLES OF ONE PATIENT ON FOLLOW-UP WHO RELAPSED. A: 80.25% AMPLIFICATION; 19.75% DELETION. B: 79.12% AMPLIFICATION; 20.88% DELETION. C: 53.67% AMPLIFICATION; 46.33% DELETION.

Personalis EX2155

D: 77.41% AMPLIFICATION; 32.59% DELETION. CNVS BASED ON 50 CONSECUTIVE MARKERS (SNP OR CN) WITH A MINIMUM SEGMENT SIZE OF 50KB. GAIN, AMPLIFICATION; LOSS, DELETION. 81

FIGURE 3-11 - HIERARCHICAL CLUSTERING (INCORPORATING DENDROGRAM AND HEATMAP) OF ALL DNA SAMPLES BASED ON 750 CYTOBANDS IDENTIFIED USING A THRESHOLD OF CN>6.0 BY GAUSSIAN-SMOOTHED SIGNAL. L, LYMPHOCYTE; P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. HEATMAP: RED, HIGH OCCURRENCE OF AMPLIFICATION PER SAMPLE GROUP; BLUE, LOW OCCURRENCE. 84

FIGURE 3-12 - CNVS DETECTED IN THE NORMAL LYMPHOCYTE DNA SAMPLE OF A HEALTHY CONTROL USING A THRESHOLD OF CN>4.0, HIGHLIGHTING HOW COMMON CNVS ARE IN THE 'HEALTHY' HUMAN GENOME. GAIN, AMPLIFICATION; LOSS, DELETION. 88

FIGURE 3-13 - GENOME-VIEW REPRESENTATIONS OF THE FREQUENCY OF AMPLIFICATION AND DELETION IN THE HEALTHY CONTROLS (A), PATIENTS ON FOLLOW-UP (B), AND PRE-SURGICAL PATIENTS (C). THE Y-AXIS INDICATES THE PERCENTAGE OF SAMPLES HAVING A CNV AT A SPECIFIC POINT IN THE GENOME. CHROMOSOME NUMBER AT TOP; AMPLIFICATIONS IN GREEN; DELETIONS IN RED; L, LYMPHOCYTE; P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. 89

FIGURE 3-14 - HIGH LEVEL AMPLIFICATION IN PLASMA CFDNA AND MATCHED TUMOUR DNA OF BREAST CANCER PATIENTS ON FOLLOW-UP, SHOWING COMMONLY-AMPLIFIED CYTOBANDS 4Q13.2 (A), 5Q13.2 (B), AND 7Q11.23 (C) (BASED ON A CN>4.0 THRESHOLD BY GAUSSIAN-SMOOTHED SIGNAL). AMPLIFICATION IS MATCHED BETWEEN T AND P1 AT 4Q13.2, T AND P2 AT 5Q13.2, AND T, P1, AND P2 AT 7Q11.23. L, LYMPHOCYTE; P1, FIRST PLASMA CFDNA; P2, SECOND PLASMA CFDNA; T, TUMOUR. 93

FIGURE 3-15 - HIGH FREQUENCY OF AMPLIFICATION AT 5Q13.2 IN PATIENTS ON FOLLOW-UP. THE Y-AXIS INDICATES THE PERCENTAGE OF SAMPLES HAVING A CNV AT A SPECIFIC POINT IN THE GENOME. AMPLIFICATIONS IN GREEN; DELETIONS IN RED. L,

Personalis EX2155

LYMPHOCYTE; P1, FIRST PLASMA CFDNA; P2, SECOND PLASMA CFDNA; T, TUMOUR. 96

FIGURE 3-16 - PCA BI-PLOT COMPARISON OF LYMPHOCYTE DNA FROM CANCER PATIENTS AND HEALTHY CONTROLS (A, UNFILTERED DATA); AND CANCER PATIENTS, HEALTHY CONTROLS, AND 128 INTERNATIONAL HAPMAP SAMPLES (B, UNFILTERED DATA). CEPH, UTAH, USA RESIDENTS WITH ANCESTRY FROM NORTHERN AND WESTERN EUROPE; CHB, HAN CHINESE IN BEIJING, CHINA; JPT, JAPANESE IN TOKYO, JAPAN; YRI, YORUBA IN IBADAN, NIGERIA. 100

FIGURE 3-17 - PCA BI-PLOT COMPARISON OF PRE-SURGICAL BREAST CANCER PATIENTS AND HEALTHY CONTROLS (A, UNFILTERED DATA), RECREATED USING JUST SNP (B) AND CN MARKERS (C). TWO PLASMA CFDNA SAMPLES FROM HEALTHY WOMEN WITH A STRONG FAMILY HISTORY ARE ALSO SHOWN. 102

FIGURE 3-18 - PCA BI-PLOTS OF ON FOLLOW-UP GROUP (A) AND ON FOLLOW-UP GROUP WITH HEALTHY CONTROLS. TWO PLASMA CFDNA SAMPLES FROM HEALTHY WOMEN WITH A STRONG FAMILY HISTORY ARE ALSO SHOWN (B). 105

FIGURE 3-19 - PCA BI-PLOT OF PATIENTS ON FOLLOW-UP WITH TRIPLE-NEGATIVE PATIENTS' CFDNA SAMPLES SHADED BLACK. 106

FIGURE 3-20 - PCA BI-PLOT OF TUMOUR SAMPLES (UNFILTERED DATA) HIGHLIGHTING SEPARATION BETWEEN FF AND FFPE TISSUE. 107

FIGURE 3-21 - HIERARCHICAL CLUSTERING OF ON FOLLOW-UP GROUP REVEALED A DIVISION BETWEEN TUMOUR AND LYMPHOCYTE DNA SAMPLES. 108

FIGURE 3-22 - HIERARCHICAL CLUSTERING OF ON FOLLOW-UP GROUP'S FIRST (A) AND SECOND (B) PLASMA CFDNA SAMPLES REVEALED TWO MAJOR CLUSTERS, BUT NO APPARENT SEPARATION BASED ON SUBGROUP. CLUSTERING FOR TUMOUR DNA SAMPLES (C) NEITHER REVEALED MAJOR DIVISIONS BETWEEN SUBGROUPS. 109

FIGURE 3-23 - OVERLAP BETWEEN TWO CN SEGMENTS. BLACK LINES REPRESENT DNA SEQUENCES FROM TWO PATIENTS; CN SEGMENTS INDICATED BY RED BARS. 118

FIGURE 4-1 - PERCENT CONCORDANCE BETWEEN TUMOUR DNA AND MATCHED NORMAL LYMPHOCYTE (MEAN 88.08%; RANGE 35.47-99.75%). 139

FIGURE 4-2 - PERCENT CONCORDANCE BETWEEN CFDNA, MATCHED NORMAL LYMPHOCYTE (A), AND TUMOUR DNA (B). RED, BASELINE CFDNA; GREEN, 6 WEEK CFDNA; PURPLE, 21 WEEK CFDNA. 140

Personalis EX2155

FIGURE 4-3 - NUMBER OF CNVS CALLED PER SAMPLE AND PATIENT (A), REPEATED WITH LINES CONNECTING SAMPLES BELONGING TO THE SAME PATIENT (B). 142

FIGURE 4-4 - GENOME-VIEW REPRESENTATION OF THE FREQUENCY OF AMPLIFICATION AND DELETION IN THE NEOCENT SAMPLE GROUP. THE Y-AXIS INDICATES THE PERCENTAGE OF SAMPLES HAVING A CNV AT A SPECIFIC POINT IN THE GENOME. AMPLIFICATIONS IN GREEN; DELETIONS IN RED. 144

FIGURE 4-5 - *UGT2B15* AND *UGT2B17* GENE INTERVALS FOR 10 NEOCENT TUMOUR SAMPLES. EACH VERTICAL BAR REPRESENTS A SINGLE MICROARRAY MARKER (OR BUNCH OF CONSECUTIVE MARKERS) AND VALUES ARE AS INFERRED BY THE COLOUR-SCALE (AT TOP). THE EXONS AND INTRONS OF THE GENES ARE INDICATED AT THE BOTTOM OF THE DIAGRAM. 148

FIGURE 4-6 - GAUSSIAN-SMOOTHED (COLOURED DOTS), CN OUTPUT OF NEOCENT TUMOUR DNA SAMPLES SHOWING EVIDENCE OF GENOMIC INSTABILITY - ISOLATED AMPLIFICATION (A; CIRCLED IN RED), COMMON AMPLIFICATION AND TELOMERE ATTRITION (B; BOTH CIRCLED IN RED), BREAKAGE ON CHROMOSOME 4 (C; INDICATED BY RED LINE), AND ISOCHROMOSOME 8 SHOWING DELETION OF THE SHORT (Q) ARM AND AMPLIFICATION OF THE LONG (P) ARM (D). SAMPLES ARE COLOUR-CODED AS FOLLOWS: RED, N1; YELLOW, N2; LIGHT-GREEN, N3; DARK-BLUE, N6; BEIGE, N7; DARK-GREEN, N8; LIGHT-BLUE, N10; PINK, N11; BROWN, N12; PURPLE, N14. CN SCALE FIXED TO 0-4; CHROMOSOMES NOT TO SCALE. 152

FIGURE 4-7 - PCA BI-PLOT OF NEOCENT PLASMA CFDNA SAMPLES (UNFILTERED DATA). INDIVIDUAL PATIENT SAMPLES ARE COLOUR-CODED. 156

FIGURE 4-8 - PCA BI-PLOT OF NEOCENT GROUP'S SAMPLES HAVING FIRST FILTERED DATA FOR SIGNIFICANT DIFFERENCES BETWEEN THE BASELINE AND 6 WEEK (A) AND 21 WEEK (B) CFDNA SAMPLES. 157

FIGURE 5-1 - KEGG CANCER PATHWAY SHOWING GENES OVERLAPPED BY SIGNIFICANT SNP/CN MARKERS (RED STARS). NB - SOME GENES HAVE MORE THAN 1 PATHWAY ROLE. 178

FIGURE 5-2 - PERCENT OVERLAPPING MARKERS FOR THE DERIVED CFDNA GENETIC SIGNATURES BETWEEN THE UNAFFECTED *BRCA1* CARRIERS AND THE FIRST (A) AND SECOND (B) SAMPLES OF THE RELAPSED AND NON-RELAPSED PATIENTS ON FOLLOW-UP. 186

Personalis EX2155

FIGURE 5-3 - PERCENT OVERLAPPING MARKERS BETWEEN EACH TIME-POINT IN THE PRE-SURGICAL PATIENTS' PLASMA CFDNA SAMPLES. 187

FIGURE 5-4 - HIERARCHICAL CLUSTERING USING SIGNATURE MARKERS REVEALS DISTINCT CLUSTERS THAT DISTINGUISH THE PLASMA CFDNA OF UNAFFECTED *BRCA1* CARRIERS FROM THE FIRST (A) AND SECOND (B) PLASMA CFDNA SAMPLES OF THE RELAPSED AND NON-RELAPSED BREAST CANCER PATIENTS ON FOLLOW-UP (ALSO SHOWN ARE THE PLASMA CFDNA SAMPLES OF THE HEALTHY CONTROLS). THE CFDNA SAMPLES OF PRE-SURGICAL PATIENTS ON NEOADJUVANT THERAPY ARE ALSO SHOWN (C). 188

FIGURE 6-1 - *AFFYMETRIX SNP 6.0* CN STATE FOR THE *ERBB2* GENE INTERVAL IN PATIENTS ON FOLLOW-UP. TWO ERBB2-POSITIVE (A) AND -NEGATIVE (B) PATIENTS BY REAL-TIME QPCR AND IHC ARE SHOWN. EACH VERTICAL BAR REPRESENTS ONE CN MARKER (OR CONSECUTIVE MARKERS) AND VALUES ARE AS INFERRED BY THE COLOUR-SCALE (AT TOP). THE EXONS AND INTRONS OF THE *ERBB2* GENE ARE INDICATED AT THE BOTTOM OF THE DIAGRAM, WITH EXON 16 MARKED IN RED. L, LYMPHOCYTE; P1 AND P2, FIRST AND SECOND PLASMA CFDNA, RESPECTIVELY; T, TUMOUR. 210

FIGURE 6-2 - REAL-TIME QPCR SCATTER-PLOTS FOR COMMON CNVS AT 4Q13.2 AND 16P12.3. NORMAL DIPLOID CN WAS SEEN IN LYMPHOCYTE DNA, WHILE SOME TUMOURS (INDEPENDENT COHORT) EXHIBITED AMPLIFICATION (A). AMPLIFICATION WAS ALSO OBSERVED IN SOME CFDNA SAMPLES (B). L, LYMPHOCYTE; P1 AND P2, FIRST AND SECOND PLASMA CFDNA, RESPECTIVELY; T, TUMOUR. 212

FIGURE 6-3 - *LMTK3* GENE INTERVAL FOR TWO ON FOLLOW-UP PATIENTS. EACH VERTICAL BAR REPRESENTS ONE CN MARKER (OR CONSECUTIVE MARKERS) AND VALUES ARE AS INFERRED BY THE COLOUR-SCALE (AT TOP). THE EXONS AND INTRONS OF THE *LMTK3* GENE ARE INDICATED AT THE BOTTOM OF THE DIAGRAM. L, LYMPHOCYTE; P1 AND P2, FIRST AND SECOND PLASMA CFDNA, RESPECTIVELY; T, TUMOUR. 214

FIGURE 7-1 - COMMON AND AMPLIFIED CNVS TO BE SCREENED FOR THROUGH PLASMA CFDNA AS A DIAGNOSTIC AND PROGNOSTIC TOOL FOR CANCER. EACH CNV IS LISTED UNDER ITS RESPECTIVE CHROMOSOME. 225

FIGURE 7-2 - NUMBER OF MARKERS PREVIOUSLY SHOWN (FIGURE 5-4) TO BE CAPABLE OF DISTINGUISHING THE CFDNA OF HEALTHY CONTROLS, UNAFFECTED *BRCA1*

Personalis EX2155

MUTATION CARRIERS, AND PATIENTS ON FOLLOW-UP WHO DID AND DID NOT RELAPSE FOR THE FIRST (A) AND SECOND (B) PLASMA CFDNA. THE SECOND PLASMA CFDNA MARKERS HAVE MORE UTILITY AT DISTINGUISHING RELAPSED PATIENTS DUE TO THESE SAMPLES HAVING BEEN TAKEN CLOSER TO TIME OF RELAPSE THAN THEIR FIRST. ALSO SHOWN IS THE NUMBER OF MARKERS DERIVED FOR THE TIME-POINT PLASMA CFDNA SAMPLES OF THE PRE-SURGICAL/NEOCENT BREAST CANCER PATIENTS ON NEOADJUVANT THERAPY (C). 226

FIGURE 8-1 - GAUSSIAN-SMOOTHED (SMOOTHED POINTS SHOWN AS COLOURED DOTS), CN OUTPUT OF NEOCENT TUMOUR SAMPLES. SAMPLES ARE COLOUR-CODED AS FOLLOWS: RED, N1; YELLOW, N2; LIGHT-GREEN, N3; DARK-BLUE, N6; BEIGE, N7; DARK-GREEN, N8; LIGHT-BLUE, N10; PINK, N11; BROWN, N12; PURPLE, N14. CN SCALE FIXED TO 0-4; CHROMOSOMES NOT TO SCALE. IV

Personalis EX2155

Index of tables

TABLE 1-1 - RELATING THE CNA PATTERNS IDENTIFIED BY KWEI *ET AL.* (2010) TO PREVIOUSLY IDENTIFIED GENE EXPRESSION DATA BY SORLIE *ET AL.* (2003). 14

TABLE 1-2 - SUMMARY OF BREAST CANCER MOLECULAR SUBTYPES IDENTIFIED BY SORLIE *ET AL.* (2003), SHOWING RECEPTOR STATUS, PREVALENCE, AND DNA METHYLATION PROFILE. 17

TABLE 1-3 - COMPARISON OF ONCOTYPE-DX® AND MAMMAPRINT™. 19

TABLE 1-4 - COMPARISON OF 3 NGS TECHNOLOGIES, HIGHLIGHTING THE MAIN DIFFERENCES AND ADVANTAGES OF EACH (TABULATED FROM GLENN, 2011; WALL *ET AL.*, 2009; 454 LIFE SCIENCES; APPLIED BIOSYSTEMS™; AND ILLUMINA INC.). 22

TABLE 1-5 - COMPARISON OF *AFFYMETRIX SNP 6.0* TO THE LATEST RELEASES OF TWO OTHER LEADING ORGANISATIONS. 27

TABLE 1-6 - SNP AND CN PROBE INFORMATION FOR THE *AFFYMETRIX SNP 6.0* (AFFYMETRIX, 2008A AND LAFRAMBOISE, 2009). 27

TABLE 2-1 - OVERVIEW OF PATIENT GROUPS INVOLVED IN THIS THESIS, INCLUDING INFORMATION ON NUMBERS AND TYPES OF SAMPLES AVAILABLE. 40

TABLE 2-2 - CLINICOPATHOLOGICAL DETAILS OF BREAST CANCER PATIENTS ON FOLLOW-UP (TAKEN FROM SHAW *ET AL.*, 2012). 42

TABLE 2-3 - ORIGIN, GROWTH MEDIUM USED, AND ERBB2 STATUS FOR EACH CULTURED BREAST CANCER CELL-LINE. 55

TABLE 2-4 - SEQUENCES OF CUSTOM-DESIGNED PRIMER-PROBE SETS USED FOR MICROARRAY VALIDATION. 56

TABLE 3-1 - MEAN *ERBB2* CT VALUES PER CELL-LINE, PER NUMBER OF REPLICATES. CORRESPONDING VALUES FOR *CNTNAP1* AND *RPPH1* ARE SHOWN IN PARENTHESES AND SQUARE-BRACKETS RESPECTIVELY. THE OVERALL MEAN AND STANDARD DEVIATION FOR EACH ARE ALSO SHOWN. 63

TABLE 3-2 - *ERBB2* RQ VALUES PER CELL-LINE, PER NUMBER OF REPLICATES. 64

TABLE 3-3 - CANCER GENOME PROJECT (WELLCOME TRUST SANGER INSTITUTE) *AFFYMETRIX SNP 6.0* CN VALUES FOR *CNTNAP1*, *RPPH1*, AND *ERBB2* FOR EACH BREAST CELL-LINE EXAMINED (DATA FOR MDA-MB-436 AND SK-BR-3 NOT AVAILABLE). 69

Personalis EX2155

TABLE 3-4 - DIFFERENCES IN QUALITY BETWEEN FFPE AND FF TUMOUR DNA TISSUE WERE APPARENT WHEN LOOKING AT THE QC AND SNP CALL-RATES FOR EACH. 71

TABLE 3-5 - DISTRIBUTION BETWEEN CNV AMPLIFICATION AND DELETION, SHOWING PERCENT AMPLIFICATION AND DELETION, MEAN TOTAL CNVS, AND THE RANGE. L, LYMPHOCYTE; P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. 83

TABLE 3-6 - TWENTY-THREE INTERVALS EXHIBITING COMMON AMPLIFICATION IN PLASMA CFDNA AND TUMOUR (USING A THRESHOLD OF CN>6.0 BY GAUSSIAN-SMOOTHED SIGNAL AND ONLY DISPLAYING THOSE THAT HAD A FREQUENCY >10%). THE FREQUENCY OF AMPLIFICATION (%) AT A GIVEN INTERVAL IS INDICATED IN SQUARE-BRACKETS. L, LYMPHOCYTE; P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. APPENDIX TABLE 8-1 PROVIDES ADDITIONAL INFORMATION FOR THESE INTERVALS. 86

TABLE 3-7 - LIST OF RELATED GROUPS AND MOLECULAR FUNCTIONS OF GENES OVERLAPPING COMMON AMPLIFICATIONS FROM TABLE 3-6. 87

TABLE 3-8 - EIGHT INTERVALS EXHIBITING COMMON AMPLIFICATION IN >90% BREAST CANCER TUMOUR DNA OF ON FOLLOW-UP PATIENTS (USING A THRESHOLD OF CN>4.0 BY A GAUSSIAN-SMOOTHED SIGNAL). THE NUMBER OF SAMPLES WITH THE THRESHOLDED CNV PER SAMPLE GROUP IS INDICATED AS A PERCENTAGE, WHILE ALSO SHOWN IN SQUARE-BRACKETS IS THE NON-GAUSSIAN-SMOOTHED CN RANGE. L, LYMPHOCYTE; P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. NB - CNVS AT 10Q11.22 AND 10Q11.23 WERE TREATED AS A SINGLE CNV DUE TO CLOSE GENOMIC PROXIMITY AND HIGH CORRELATION. 92

TABLE 3-9 - SUMMARY TABLE OF BINARY LOGISTIC REGRESSION ANALYSIS. SOURCE DATA WAS BASED ON THE PRESENCE OR ABSENCE OF AN AMPLIFICATION (USING A THRESHOLD OF CN>6.0 BY GAUSSIAN-SMOOTHED SIGNAL) AT EACH OF 23 COMMON CNVS LISTED IN TABLE 3-6. SIGNIFICANT P-VALUES SHOWN; NON-SIGNIFICANT P-VALUES DENOTED BY 'NS'. P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-

Personalis EX2155

SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. 98

TABLE 3-10 - LOH DETECTED IN THE CFDNA OF PRE-SURGICAL BREAST CANCER PATIENTS AND PATIENTS ON FOLLOW-UP. FOR EACH PATIENT, NORMAL LYMPHOCYTE DNA WAS USED AS REFERENCE AND LOH WAS CALLED USING A 6-STATE HIDDEN MARKOV MODEL (HMM) (SEE MATERIALS AND METHODS 2.7.5). ONLY LOH IN EXONS WERE REPORTED AND THE DNA STRAND (+, CODING; -, NON-CODING) FOR EACH AFFECTED GENE IS INDICATED IN PARENTHESES. P, PLASMA CFDNA OF PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP. 113

TABLE 4-1 - THE SNP CALL-RATES FOR 47 NEOCENT SAMPLES HIGHLIGHTS THE HIGH QUALITY OF THE TISSUE USED. L, LYMPHOCYTE; P, PLASMA CFDNA; T, TUMOUR. SAMPLES ARE GROUPED BY RESPONSE, NON-RESPONSE, AND UNKNOWN RESPONSE TO THERAPY. 138

TABLE 4-2 - NUMBER OF CNVS CALLED PER SAMPLE AND PATIENT. A CNV WAS IDENTIFIED AS HAVING A MINIMUM OF 50 MARKERS PER SEGMENT, P-VALUE CUT-OFF OF <0.0001, SIGNAL-TO-NOISE RATIO OF 0.5, AND A MINIMUM SIZE OF 1KB. 141

TABLE 4-3 - PERCENT CNV AMPLIFICATION AND DELETION IN THE NEOCENT AND HEALTHY CONTROL GROUPS. L, LYMPHOCYTE; P, HEALTHY CONTROL OR NEOCENT PLASMA CFDNA; T, TUMOUR. 145

TABLE 4-4 - TWENTY-ONE INTERVALS EXHIBITING COMMON AMPLIFICATION IN THE NEOCENT PATIENT SAMPLES (USING A THRESHOLD OF CN>6.0 BY GAUSSIAN-SMOOTHED SIGNAL AND ONLY DISPLAYING THOSE THAT HAD A FREQUENCY >10%). THE FREQUENCY OF AMPLIFICATION (%) OF A GIVEN CNV IS INDICATED IN THE SQUARE-BRACKETS. L, LYMPHOCYTE; P, HEALTHY CONTROL OR NEOCENT PLASMA CFDNA; T, TUMOUR. *NOT APPEARING IN TABLE 3-6 AND THEREFORE SPECIFIC TO THE NEOCENT GROUP. 147

TABLE 4-5 - BINARY LOGISTIC REGRESSION ANALYSIS FOR 21 COMMONLY-AMPLIFIED CNVS IN NEOCENT PATIENT SAMPLES. SOURCE DATA WAS BASED ON THE PRESENCE OR ABSENCE OF AN AMPLIFICATION (USING A THRESHOLD OF CN>6.0 BY A GAUSSIAN-SMOOTHED SIGNAL) AT EACH OF 21 CNVS LISTED IN TABLE 4-4. SIGNIFICANT P-VALUES ARE SHOWN; NON-SIGNIFICANT P-VALUES DENOTED BY 'NS'.

Personalis EX2155

L, LYMPHOCYTE; P, HEALTHY CONTROL OR NEOCENT PLASMA CFDNA; T, TUMOUR. 150

TABLE 4-6 - COPY NUMBER VARIATION IN THE *FGFR*S AND *EGFR* HIGHLIGHTS VARIATION BETWEEN TUMOUR DNA AND PLASMA CFDNA AT DIFFERENT TIME-POINTS. THE FREQUENCY (%) OF A GIVEN CNV IS INDICATED IN THE SQUARE-BRACKETS. 'NS' DENOTES NO SIGNIFICANT DIFFERENCES BETWEEN CFDNA AND TUMOUR DNA. 151

TABLE 5-1 - PCA EXPERIMENTS CONDUCTED TO DEDUCE 'DRIVER' GENES OF VARIATION, INCLUDING AIMS FOR EACH. L, LYMPHOCYTE; P, HEALTHY CONTROL OR PRE-SURGICAL PLASMA CFDNA; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. NB - N=7 FOR 21 WEEK P OF PRE-SURGICAL GROUP PATIENTS. 172

TABLE 5-2 - GENES OVERLAPPING SIGNIFICANTLY-DIFFERENTIATED MARKERS (SNP OR CN) THROUGH POST-PCA ANALYSIS. MARKERS WERE FIRST FILTERED TO INCLUDE THOSE HAVING A CORRELATION OF >0.7 OR <-0.7 TO THE FIRST PRINCIPAL COMPONENT AND THEN OVERLAPPING THESE WITH THE HUMAN GENOME TO UNEARTH A LIST OF GENES. THESE WERE THEN FILTERED TO INCLUDE THOSE ONLY PRESENT IN OMIM® AND HAVING KEYWORDS RELATED TO 'CANCER' AND 'BREAST CANCER' (PUBMED HEALTH, HTTP://WWW.NCBI.NLM.NIH.GOV/PUBMEDHEALTH/). L, LYMPHOCYTE; P, HEALTHY CONTROL OR PRE-SURGICAL PLASMA CFDNA; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. AMPLIFICATION=CN>2.5; DELETION=CN<1.5. APPENDIX TABLE 8-3 PROVIDES ADDITIONAL INFORMATION FOR THESE MARKERS. NB - PREVALENCE AND STATE RELATE TO THE LATTER SAMPLE IN EACH COMPARISON; ASTERISK (*) DENOTES SNP/CN MARKER MAPPING TO AN EXON; GENES LISTED ALPHABETICALLY; 'NA' DENOTES 'NOT APPLICABLE' DUE TO NO SIGNIFICANT MARKERS USING THE SELECTED KEYWORDS; 'NO CALL' IMPLIES A FAILED HYBRIDISATION AND DELETED LOCUS (LAFRAMBOISE, 2009). 176

TABLE 5-3 - REDUCTION OF MARKERS FOR THE BREAST CANCER GENETIC SIGNATURES. P, PLASMA CFDNA OF PRE-SURGICAL PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA SAMPLES OF PATIENTS ON FOLLOW-UP, RESPECTIVELY. 185

TABLE 6-1 - COPY NUMBER OF *LMTK3* GENE INTERVAL IN THE PATIENTS ON FOLLOW-UP. 'UP' DENOTES CN>2.5; 'DOWN' DENOTES CN<1.5. L, LYMPHOCYTE; P1 AND P2,

Personalis EX2155

FIRST AND SECOND PLASMA CFDNA, RESPECTIVELY; T, TUMOUR; NA, NO SAMPLE AVAILABLE; SUPERSCRIPT 'R' DENOTES RELAPSED PATIENT. 215

TABLE 6-2 - FREQUENCIES OF AMPLIFICATION/DELETION (%) FOR CYTOBANDS IDENTIFIED BY BAE *ET AL.*, (2010) IN TUMOUR. AMPLIFICATION AND DELETION WAS BASED ON A THRESHOLD OF CN>6.0 AND <1.0, RESPECTIVELY, BY GAUSSIAN-SMOOTHED SIGNAL. 217

TABLE 6-3 - FREQUENCIES OF AMPLIFICATION/DELETION (%) FOR CYTOBANDS IDENTIFIED BY BAE *ET AL.*, (2010) IN BLOOD. AMPLIFICATION AND DELETION WAS BASED ON A THRESHOLD OF CN>6.0 AND <1.0, RESPECTIVELY, BY GAUSSIAN-SMOOTHED SIGNAL. P1 AND P2, FIRST AND SECOND PLASMA CFDNA, RESPECTIVELY. 218

TABLE 6-4 - TOP 10 AMPLIFIED AND DELETED TUMOUR SUPPRESSOR GENES REPORTED BY NAYLOR *ET AL.* (2005) AND THEIR CORRELATION TO FINDINGS IN THIS STUDY FROM ON FOLLOW-UP TUMOUR. FOR NAYLOR *ET AL.* (2005) DATA, FREQUENCIES OF AMPLIFICATION/DELETION (%) ARE SHOWN. 219

TABLE 8-1 - CNV SIZE RANGE (KB) AND MARKER DENSITY FOR 23 CYTOBANDS THAT EXHIBITED A COMMONLY-AMPLIFIED REGION (TABLE 3-6). AMPLIFICATIONS AND DELETIONS FOR THE ENTIRE CYTOBAND CONTAINING THE CNV WERE TAKEN INTO ACCOUNT; MEAN SHOWN IN SQUARE-BRACKETS; SINGLE VALUES INDICATE THE PRESENCE OF JUST ONE CNV; ABSENT VALUES INDICATE NO CNV. L, LYMPHOCYTE; P, PLASMA CFDNA OF HEALTHY CONTROLS OR PRE-SURGICAL BREAST CANCER PATIENTS; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. III

TABLE 8-2 - COMPLETE LISTING OF GENES HARBOURING SIGNIFICANTLY-DIFFERENTIATED MARKERS (SNP OR CN) FOR POST-PCA COMPARISONS CARRIED OUT IN TABLE 5-2. NUMBER OF GENES FOR EACH COMPARISON IN PARENTHESES. L, LYMPHOCYTE; P, HEALTHY CONTROL OR PRE-SURGICAL PLASMA CFDNA; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. XXVI

TABLE 8-3 - OVERLAPPING GENE, MARKER ID, CHROMOSOME, GENOMIC POSITION, AND CYTOBAND FOR EACH SIGNIFICANT MARKER OUTLINED IN TABLE 5-2. NB - ASTERISK (*) DENOTES SNP/COPY NUMBER PROBE MAPS TO AN EXON. L, LYMPHOCYTE; P, HEALTHY CONTROL OR PRE-SURGICAL PLASMA CFDNA; P1 AND P2, FIRST AND SECOND PLASMA CFDNA OF PATIENTS ON FOLLOW-UP, RESPECTIVELY; T, TUMOUR. Page XXVI

Personalis EX2155

NB - ASTERISK (*) DENOTES SNP/CN MARKER MAPPING TO AN EXON; GENES LISTED ALPHABETICALLY; 'NA' DENOTES 'NOT APPLICABLE' DUE TO NO SIGNIFICANT MARKERS USING THE SELECTED KEYWORDS. XXIX

Personalis EX2155



1 INTRODUCTION

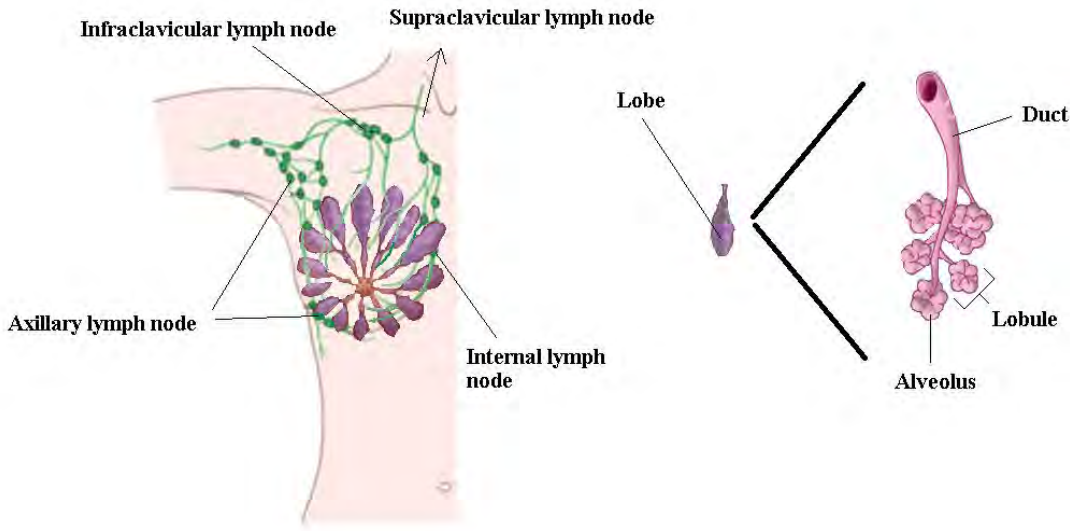
Personalis EX2155

1.1 The breast

The breast lies against the pectoralis muscle on the chest wall and is divided into four quadrants: upper inner and outer; and lower inner and outer. It is composed of 15-20 lobes that are positioned radially from the nipple, each containing many lobules and surrounded by adipose and connective tissue. The lobes are connected to the nipple via ducts. The lobules are further divided into up to 100 alveoli, which produce milk. Milk is secreted into small ducts (ductules) within each lobe before reaching the nipple through the 10-15 major ducts (Figure 1-1). The remainder of the breast is mostly composed of adipose tissue (McKian *et al.*, 2009 and CRUK, 2011a).

Blood and lymphatic vessels are also found in the breast region. Some of the lymphatic vessels lead to the internal mammary and axillary lymph nodes on either side of the breast, while above and below the clavicle rest the supra- and infra-clavicular nodes, respectively (Figure 1-1). Cancer cells that invade the lymphatic system, for example at the axillary lymph nodes, can be ultimately carried away to other parts of the body (Gajdos *et al.*, 1999).

Figure 1-1 - Breast anatomy, showing the relative locations of the ducts, lobes, lobules, and alveoli. Also highlighted is the proximity of the surrounding lymph nodes and vessels (adapted from CRUK (2011a).



Personalis EX2155

1.2 Breast cancer: history and current research focus

Cancer is a multi-step process whereby cells acquire genetic alterations/mutations that leads to their uncontrollable growth (Holm *et al.*, 2010). Breast cancer, in particular, is a major health problem worldwide (Hicks *et al.*, 2012). Research pertaining to breast cancer has received much funding, which over the past three decades has focussed on screening and adjuvant therapies (Cleator *et al.*, 2007).

In the breast, normal cells can undergo alterations that result in abnormal patterns of growth and thereafter become neoplastic. Neoplasms can be benign, such as non-proliferative lesions, proliferative lesions [without atypia], and atypical hyperplasia (Hartmann *et al.*, 2005), but can also be pre-malignant and malignant (or cancerous). Symptoms of breast cancer include: a breast lump; nipple discharge; a change in the size or shape of the breast; an inverted nipple; or redness, thickening, or dimpling of the skin (CRUK, 2011a).

For patients who have already had breast cancer, recurrent metastasis is a possibility. Metastasis is the main cause of cancer death in the developed world and involves a series of steps that the cancer must complete in order to develop a malignancy at a distant site (Farace *et al.*, 2011). It involves cell adhesion and contractility, followed by the subsequent remodelling of the extracellular matrix (ECM) (Saha Roy & Vadlamudi, in press). Another critical step is the release of material of tumour origin into the bloodstream and how well it survives there, which at the same time also provides a non-invasive source of tumour material that can provide information on the patient's disease status (Farace *et al.*, 2011). Hence, such circulating tumour nucleic acids will be the focus of this thesis.

1.2.1 Incidence

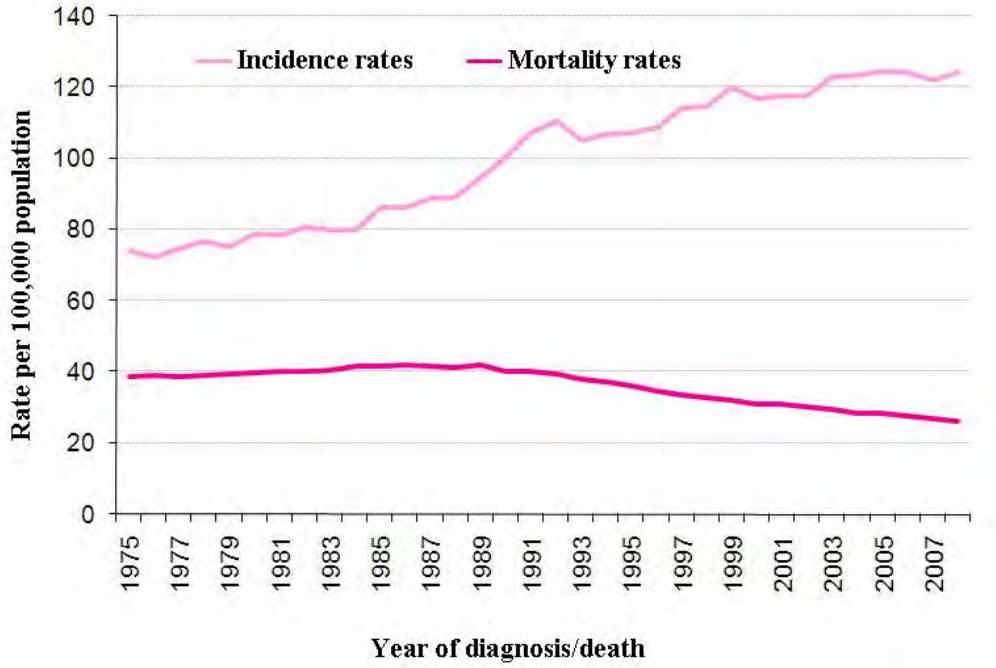
Increase or decrease in cancer incidence rates can be due to changes in the size and age-structure of a population, but also in the rates of the disease itself due to lifestyle changes (Mistry *et al.*, 2011). The incidence rates of breast cancer are higher in more

Personalis EX2155

affluent cultures and the rate has been increasing in economically-developed countries (Glass *et al.*, 2007 and Héry *et al.*, 2008). Indeed, despite mortality rates having decreased over the past few decades when looking at data for the European Union (Espinosa *et al.*, in press), breast cancer remains the most common malignant disease in women of Western industrialised nations (Zhong *et al.*, 2007).

In the UK in particular, breast cancer is the most common form of cancer despite the fact that it is rare in men. In recent years, looking solely at data for Great Britain, whilst the incidence rate has been increasing, the mortality rate has concurrently been decreasing (Figure 1-2). In 2008 alone, 48,034 new cases were diagnosed. Solely observing figures for females, 47,693 were diagnosed, which accounted for 31% of all cancers in that year. After breast, the most common cancer in females is that of the lung, with 12% (CRUK, 2011b).

Figure 1-2 - Breast cancer incidence and mortality rates in Great Britain per 100,000 population, highlighting the increase in incidence yet concurrent decrease in mortality (taken from CRUK, 2011b).



Personalis EX2155

In the USA, between 1999 and 2006, there was a decreased rate of 2% per annum in newly-diagnosed invasive breast cancer cases (Horner *et al.*, 2009). Conversely, breast cancer incidence rates are rising in Asia and other nations whose economies are still developing (Curado *et al.*, 2007). Indeed, in Shanghai and other Chinese cities, incidence rates have seen a sharp rise (Hesketh *et al.*, 2005 and Linos *et al.*, 2008) as lifestyles become more 'Westernised'. Worldwide, it is estimated that almost 1.38 million women were diagnosed with breast cancer in 2008, accounting for 23% of all female cancers (Ferlay *et al.*, 2010).

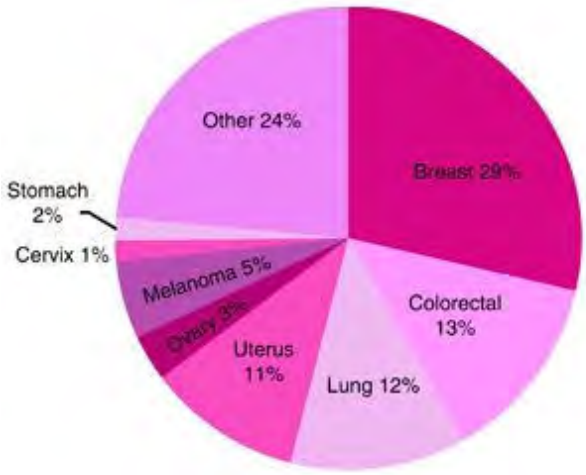
1.2.1.1 Future trends

The estimated future trend is that the number of cases in the UK will increase for the majority of cancers[1], including breast, due to a concurrent increase in population and people living longer (Mistry *et al.*, 2011). The proportion of those over 65, for example, is expected to jump from 16% in 2007 to 22% in 2030, while those over 85 from 4.5 to 7.6%. Breast cancer is expected to be the cancer with the largest increase, with 11,681 more cases predicted in 2030 than in 2007 (increase of 26%), accounting for an estimated 29% of all female cancers in that year (Mistry *et al.*, 2011) (Figure 1-3). However, the actual incidence rate will decline by 8%, according to Mistry *et al.* (2011), citing the reduction in the use of hormone replacement therapy (HRT) as a conceivable cause. Conversely, a separate study predicted breast cancer incidence rates to increase year-on-year up to 2020 (Møller *et al.*, 2007).

[1] Numbers were predicted for 28 different cancers.

Personalis EX2155

Figure 1-3 - Breast cancer is predicted to account for 29% of all female cancer cases in the UK in 2030 (taken from Mistry *et al.*, 2011).



1.2.2 Types of breast cancer

The majority (70-80%) of breast carcinomas begin in the cells lining the ducts, while only 10% begin in the cells of the lobules - these are known as ductal and lobular carcinomas, respectively. If the cancer has not spread beyond the basement membrane, it is referred to as '*in situ*' (ductal and lobular carcinomas i*n situ*), whereas cancers that have spread beyond this membrane are invasive (invasive ductal and lobular carcinomas) (CRUK, 2011a). Benign changes of inflammatory or epithelial origin that are similar to cancer also occur and can result in an increased risk of cancer (Hartmann *et al.*, 2005). Such changes are common[1] and include lumps, inflammation, and nipple discharge (ACS, 2010). There are other types of invasive breast cancers, amongst them including tubular, medullary, metaplastic, mucinous, and invasive papillary carcinoma (CRUK, 2011a).

[1] Research suggests that 80-90% of lump biopsies that are examined turn out to be benign (ACS, 2010 and CRUK, 2011a).

Personalis EX2155

1.2.3 Morphological and molecular factors

Morphological properties of the primary tumour can be both prognostic and predictive (Rakha & Ellis, 2011), and can guide decisions surrounding the treatment of patients (Szoke & Udvarhelyi, 2012). These include tumour size, lymph node stage, and histological grade (Walker, 2003). In general, the larger the tumour, the poorer the outcome: tumours with diameters≤10mm have more favourable outcomes (88% 20-year relapse free survival), whereas those with diameters>10mm are less favourable (Rosen & Groshen, 1990). Additionally, smaller tumours are less likely to have lymph node metastasis, the presence or absence of which has great predictive potential, matched by that of histological grade (Walker, 2003). In general, the longer that a tumour is present, the more likely it is to metastasise, and those who have any lymph node involvement have poor 10-year survival rates (Walker, 2003). Histological grade is based on a scoring system that accounts for 3 grades: 1, well differentiated; 2, moderately differentiated; and 3, poorly differentiated (Carriaga & Henson, 1995). The importance of histological grade was emphasised by Elston and Ellis (1991), who were in-part responsible for deriving the Nottingham prognostic index (Haybittle *et al.*, 1982).

Tumour size and lymph node stage, along with the presence or absence of distant metastases, make up the TNM staging system (International Union Against Cancer, 1960), where each cancer is given a designation of $T_{1\text{-}4}N_{0\text{-}3}M_{0\text{-}1}$: $T_{1\text{-}3}$ relate to those tumours of favourable outcome that have diameters≤10mm, whilst T_4 are those >10mm; N_0 is where no lymph node metastasis has occurred, whilst $N_{1\text{-}3}$ refers to the level of metastasis in the axillary and supra- and infra-clavicular lymph nodes; finally, M_0 and M_1 refer to the presence or absence of distant metastases (Sicher & Waterhouse, 1973).

Molecular markers are also routinely used in the clinical setting, most commonly being the presence or absence and level of expression of oestrogen receptor (ER), progesterone receptor (PR), and HER-2/neu (ERBB2), whilst Ki67 is used as a marker of proliferation for tumours with high mitotic rates (Leong & Zhuang, 2011). Also, recent molecular profiling has led to the introduction of two new predictive tests (see section 1.3.2).

Personalis EX2155

1.2.4 Risk factors

In cancer, a risk factor is any characteristic or behaviour that affects the chances of developing the disease (ACS, 2012). Risk factors act in concert, possibly having a multiplicative effect (Parkin *et al.*, 2011), and can be both genetic and environmental in origin (Key *et al.*, 2001) - lifestyle choices also play a key role (Parkin *et al.*, 2011). The stratification of patients based on risk factors can help in the tailoring of specific treatments to each group, thus saving on treatment costs. Known breast cancer risk factors include:

- **Gender and age**

  Being female is the greatest risk factor (CRUK, 2012), whilst risk also increases with higher age - 81% of female breast cancers occur in those over the age of 50 (Key *et al.*, 2001 and CRUK, 2012).

- **Family history/familiar disease**

  The inheritance of certain gene mutations is related to familiar disease (Key *et al.*, 2001) and having a first- or second-degree relative with breast cancer increases risk: one affected first-degree relative means the risk is almost doubled; whilst two and three affected relatives increases risk by almost 3x and 4x, respectively (Collaborative Group on Hormonal Factors in Breast Cancer, 2001). Conversely, having a single affected second-degree relative increases risk by half (Pharoah *et al.*, 1997).

- **Personal history**

  Having a previous diagnosis of typical hyperplasia almost doubles the risk of developing breast cancer in the same breast, whilst the risk is more than quadrupled for atypical hyperplasia (Hartmann *et al.*, 2005). Conversely, having a previous diagnosis of ductal or lobular carcinoma *in situ* doubles the risk of developing invasive breast cancer (Robinson *et al.*, 2008).

Personalis EX2155

- **Weight and diet**

  Obesity is a factor (García-Closas *et al.*, 2006 and Parkin *et al.*, 2011), potentially due to the fact that aromatase, which is used to synthesise oestrogen, is secreted by adipose tissue (O'Neill *et al.*, 1988). Also, alcohol raises oestrogen levels (Coronado *et al.*, 2011) and is a known risk factor (Key *et al.*, 2001 and Parkin *et al.*, 2011). The effect of diet is highlighted by the fact that Japanese women who migrate to the USA have an increased risk, when normally they have a lower risk if they remain in their home country (Ziegler *et al.*, 1993 and Deapen *et al.*, 2002).

- **Hormonal factors**

  Prolonged exposure to oestrogen through early onset puberty or late menopause is a known risk factor (Key *et al.*, 2001 and Parkin *et al.*, 2011). Hormone replacement therapy (Beral, 2003) and taking the contraceptive pill (Collaborative Group on Hormonal Factors in Breast Cancer, 1996) are also known to increase risk, but this diminishes once stopped in both cases. After HRT's introduction in the 1990s, an increase in breast cancer incidence rates was observed and the two are believed to be related (Beral *et al.*, 1997 and García-Closas *et al.*, 2006).

- **Genetic factors**

  Carrying germline mutations in key genes - such as *BRCA1*, *BRCA2*, *PTEN*, and *TP53* - greatly increase risk (Key *et al.*, 2001), whilst other yet unknown genes involved in breast cancer are likely (Easton, 1999). Isolating these genes is an important goal of cancer research (Smith & Ponder, 1993). Inherited mutations in *BRCA1* and *BRCA2* result in a 65 and 45% chance of developing cancer by the age of 70, respectively (Antoniou *et al.*, 2003). Due to the fact that mutations in these genes only account for a small percentage of breast cancers (Easton, 1999), standardised genetic testing is not available on the National Health Service (NHS). However, genetic testing is available for those with a strong family history of breast cancer (CRUK, 2012).

Personalis EX2155

- **Breast density**

  Having a mammographically-dense breast greatly increases risk by up to 5x (McCormack & dos Santos Silva, 2006 and Checka *et al.*, 2012).

1.2.5 Detection and screening

The purpose of cancer screening is to reduce mortality by detecting lesions that would otherwise not have been detected (Vainio *et al.*, 2002). In Great Britain, before the introduction of breast screening in 1988, the increase in incidence of breast cancer was 2% per annum, with screening itself resulting in a transient increase due to improved detection. It is estimated that 550,000 women who received a previous diagnosis of breast cancer are still alive today, based on diagnoses up to and including 2004 (CRUK, 2011b).

1.2.5.1 Breast mammography

A woman aged 50-69 whose cancer is detected early by attending a biennial mammogram is less likely to die from breast cancer than those who don't (Mandelblatt *et al.*, 2010). However, there has been much controversy about breast mammography and its utility as a screening tool (Kösters & Gøtzsche, 2003; Gøtzsche & Jørgensen, 2011; and Gøtzsche & Nielsen, 2011). In the UK, because 81% of breast cancers occur in women aged over 50 - with almost half occurring between the ages of 50 and 70 (CRUK, 2011b) - the NHS Breast Screening Programme (NHSBP) invites women with an age between 50 and 70 to attend for mammographic screening, while those aged 70+ can request an appointment. Plans are currently in place to extend the NHSBSP to those aged between 47 and 73 by the end of 2012, owing to a lack of evidence that mammograms prevent breast cancer mortality for those aged under 47 (CRUK, 2011a). However, an urgent review of the evidence surrounding the benefits of mammography is underway (Bewley, 2011 and Richards, 2011).

Personalis EX2155

The mammogram consists of a low dosage of x-rays that is directed at the breasts in order to detect abnormalities, and it is typically less than 2.5mGy[1] per breast (Vainio *et al.*, 2002). Studies have found widely-different sensitivities and specificities for mammographic detection, with the National Cancer Institute (NCI) reporting a sensitivity and specificity of 83.5 and 90.9%, respectively, in the USA alone (BCSC, 2011), whilst the International Agency for Research on Cancer (IARC) report corresponding figures of 68-90% and 82-97%, respectively (Vainio *et al.*, 2002). A genetic element, such as the genotype of an individual, can play a role in mammography sensitivity. For example, a comprehensive review by Lindström *et al.* (2011) found that SNPs in *ESR1*, *LSP1*, and *ZNF365* were associated with percent mammographic density, a higher level of which is related to a higher breast cancer risk (ACS, 2012). In addition, it was recommended by Bigenwald *et al.* (2008) that women with *BRCA1* mutations have magnetic resonance imaging (MRI) scans in addition to their mammograms, even those with low breast density.

Aids for mammography include the digital mammography, which involves digitally-recording the x-ray image taken during routine scans. Computer software can then pick out alterations and lesions in the breast tissue that could be missed by the radiologist (Vainio *et al.*, 2002). Problems with the mammogram, however, include false-positives and subsequent unnecessary treatment of healthy women (Elmore *et al.*, 1998; Hofvind *et al.*, 2004; and Gøtzsche & Nielsen, 2011), false-negatives, and overtreatment based on erroneous interpretation of results (Vainio *et al.*, 2002 and Zahl & Mæhlen, 2011). Additionally, although it would seem beneficial that the mammogram was offered to younger people, the cumulative effect of repeated radiation exposure over the years could lead an increase in mortality due to radiation-induced breast cancer (Vainio *et al.*, 2002). Additionally, younger women have a greater proportion of dense breast tissue, making mammogram less sensitive (Checka *et al.*, 2012). Thus, there is a necessity to provide a safer and more efficient alternative to mammography.

Other tests/methods that are used to either assist or confirm a diagnosis of breast cancer include a physical breast exam, breast ultrasound (uses sound waves to produce internal

[1] 1 Gray (Gy)≡absorption of 1 Joule/Kg of radiation.

Personalis EX2155

structure images of the breast), and biopsy (tissue sample removed and then further tested to determine grade and hormone receptor status) (Vainio *et al.*, 2002).

1.2.5.2 Therapy

Therapy is more effective when given at an early stage of cancer development (Zhang *et al.*, 2007) or, at best, before overt metastatic disease develops (Boér, 2010). There are many drugs used to treat breast cancer that can delay progression - such as tamoxifen and trastuzumab (Herceptin®) (Cleator *et al.*, 2007) - but tumour cells are capable of change, adaptation, and often acquire drug resistance (Coombes *et al.*, 2004; Pestrin *et al.*, 2009; and Saha Roy & Vadlamudi, in press). For example, 50% of ER-positive patients will not respond to endocrine therapy, and some of those who do respond can relapse (Hicks *et al.*, 2012). Trastuzumab is used in the treatment of ERBB2-overexpressing breast cancers, which can arise due to an amplified *ERBB2* gene or increased transcription (Sahin, 2000), and accounts for 20-25% of breast cancers (Slamon *et al.*, 1987). The encoded protein is a 185KDa cell surface glycoprotein that plays a crucial role in regulating cell growth and differentiation (Ullrich & Schlessinger, 1990). It is used as a diagnostic marker for aggressive disease and an increased risk of metastasis (Ross & Fletcher, 1998).

Tamoxifen, on the other hand, is responsible for improving the survival of patients diagnosed with early stage breast cancer and also improving the quality of life of those with metastatic disease (Tomlinson *et al.*, in press). It is the main treatment for ER-positive breast cancer patients requiring adjuvant treatment post-surgery. These account for almost two-thirds of all breast cancers. However, it is known to increase the risk of endometrial cancer and blood clotting disorders (Hutchinson, 2004 and Sheh, 2011). Additionally, the effect of this drug weakens with time (Levenson & Jordan, 1997 and Coombes *et al.*, 2004). Coombes *et al.* (2004) carried out a trial involving tamoxifen and an aromatase inhibitor known as exemestane (Aromasin®). It was found that switching patients to exemestane after 2-3 years of tamoxifen improved the disease-free survival of the patients over those who only took tamoxifen. Similar results were found for other aromatase inhibitors with tamoxifen, including anastrozole (Arimidex®) (Baum *et al.*, 2003 and Boccardo, 2008) and letrozole (Femara®) (Goss *et al.*, 2003).

Personalis EX2155

Another problem is the fact that tamoxifen is ineffective in the up to 30% of breast cancers whose ER-receptor is negative.

1.3 Molecular profiling and personalised medicine

Pathology's future lies in personalised medicine (Moch *et al.*, 2012), which aims to molecularly characterise and understand each individual's disease and to thereafter tailor therapy based on findings. In the adjuvant setting, the choice of therapy depends on the risk of recurrence (Espinosa *et al.*, in press). However, such therapies - including targeted drugs - are expensive (Moch *et al.*, 2012). Therefore, their efficacy must be assured through their application only to those patients most likely to benefit from them. For prognostic purposes, hormone receptor (ER and PR) and ERBB2 status are still the best factors in dictating the outcome and course of treatment for breast cancer patients (Espinosa *et al.*, in press). In addition, the *ERBB2* gene is used as a diagnostic marker for aggressive disease and an increased risk of metastasis (Ross & Fletcher, 1998). The two currently-accepted methods for determining the status of these are immunohistochemistry (IHC), which assesses the expression of their protein on cell surfaces, and fluorescent in situ hybridisation (FISH), which can detect amplification of the *ERBB2* gene (Pauletti *et al.*, 2000; Bartlett *et al.*, 2001; and Lebeau *et al.*, 2001). However, both methods require tumour biopsies to be taken and do not permit for the continual monitoring of patients on follow-up after surgery.

Complementary DNA (cDNA) microarrays have focussed on classifying different types of breast cancer using gene expression profiling. Perou *et al.* (2000), for example, classified breast tumours into subtypes using a hierarchical clustering method and inferred the subtypes based on how each sample clustered with another in the resulting dendrogram. Although only looking at a subset of genes (less than one-tenth of that in any human genome), the group inferred from their data the following subtypes:

- Luminal epithelial/ER gene cluster
- ERBB2 overexpression gene cluster
- Basal epithelial gene cluster (keratins 5/17)

Personalis EX2155

- Basal epithelial gene cluster (epithelial cell-enriched)

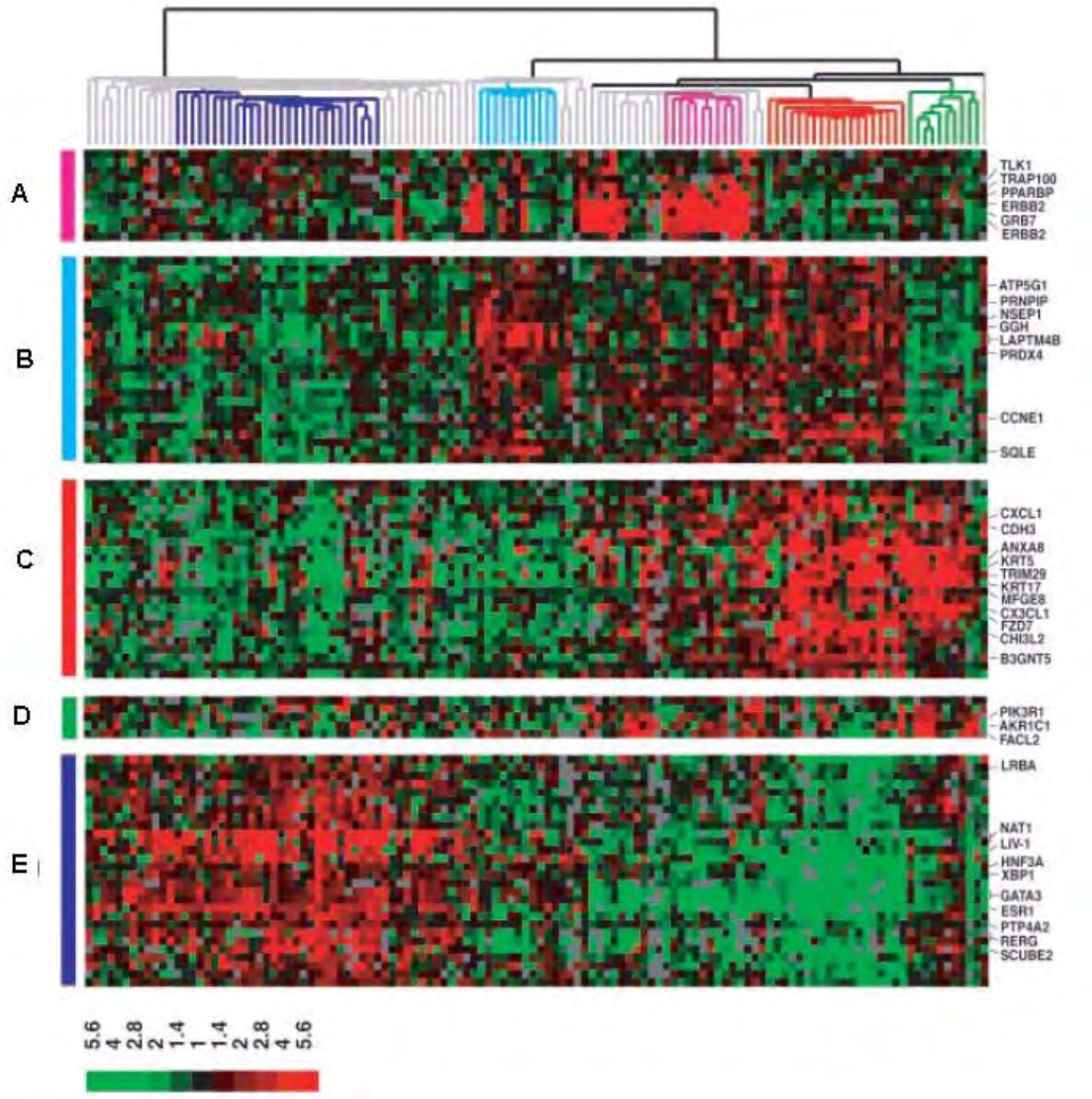
Sorlie *et al.* (2003) looked at a slightly larger subset of genes and merged the 2 basal epithelial subtypes, whilst splitting the luminal group into 2. They also noted a 'normal' breast-like subtype (Figure 1-4), which was later suggested as being the result of normal tissue contamination within the tumour (Parker *et al.*, 2009). Kwei *et al.* (2010) then associated these with copy number alteration (CNA) patterns (Figure 1-5). They noted three different CAN patterns (Table 1-1). More recently, Curtis *et al.* (in press) correlated gene expression and CN data from 997 (validated in a further 995) breast cancers to derive 10 subgroups.

CNA pattern (Kwei *et al.*, 2010)	**Previous classification (Sorlie *et al.*, 2003)**
Simple (1q/16)	Luminal A
Amplifier (firestorm)	Luminal B
Amplifier (firestorm)	ERBB2 overexpression
Complex (sawtooth)	Basal-like

Table 1-1 - Relating the CNA patterns identified by Kwei *et al.* (2010) to previously identified gene expression data by Sorlie *et al.* (2003).

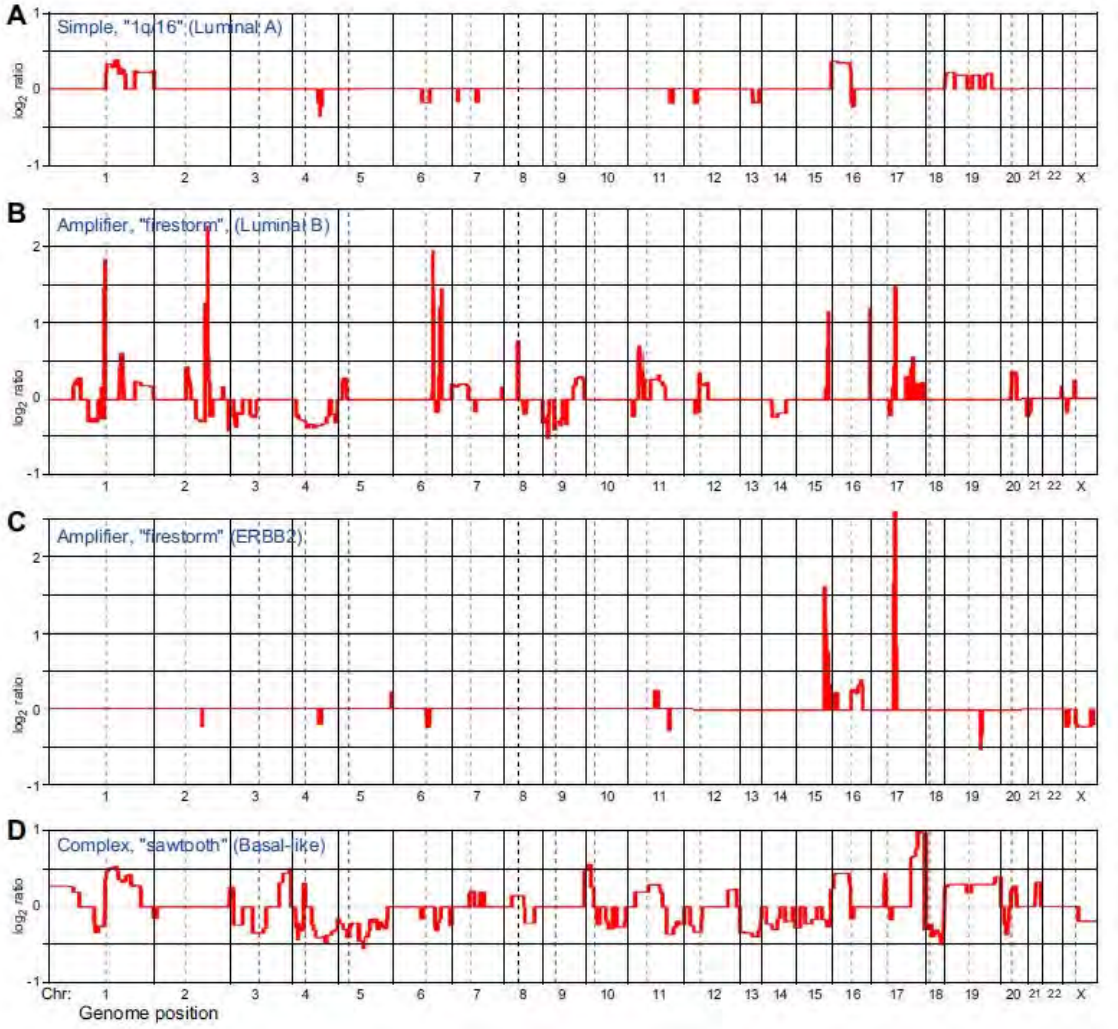
Personalis EX2155

Figure 1-4 - Hierarchically-clustering the gene expression data from a subset of genes in breast tumours revealed 5 different subtypes. A, ERBB2 overexpression; B, luminal B; C, basal-like; D, 'normal' breast-like; E, luminal A (adapted from Sorlie *et al.*, 2003).



Personalis EX2155

Figure 1-5 - Genomic profile of 4 breast cancer subtypes. A, simple (1q/16) relating to luminal A; B, amplifier (firestorm) relating to luminal B; C, amplifier (firestorm) relating to *ERBB2*; D, complex (sawtooth) relating to basal-like (adapted from Kwei *et al.*, 2010).



1.3.1 Breast gene expression subtypes

The four subtypes identified by Sorlie *et al.* (2003) have been shown to have an associated receptor status profile (Sanpaolo *et al.*, 2011), prevalence (Hannemann *et al.*, 2006 and Gaudet *et al.*, 2011), DNA methylation pattern (Andrews *et al.*, 2010 and Holm *et al.*, 2010), and risk (Gaudet *et al.*, 2011 and Yang *et al.*, 2011) (Table 1-2). ER-positive tumours include luminal A and B, with those of luminal A tending to have the best prognoses and those of luminal B tending to be highly-positive for Ki67

Personalis EX2155

(Dawood *et al.*, 2011). An additional subset of the luminal tumours that did not require adjuvant therapy was also found to have excellent clinical outcome (Geyer *et al.*, 2011).

Conversely, basal-like tumours are triple-negative and can have mutations in *BRCA1* (Tung *et al.*, 2010 and Espinosa *et al.*, in press). Wild-type BRCA1 binds to the *ESR1* gene-promoter and can restore ER expression in ER-negative breast cell-lines, conversely, the loss of BRCA1 function through mutations was shown to remove ER expression in ER-positive breast cell lines and could result in the accumulation of ER-negative cells (Hosey *et al.*, 2007). These triple-negative breast cancers can also express ERBB1 and/or cytokeratin 5/6, are aggressive, and can be difficult to treat due to their lack of ER expression (Stead *et al.*, 2009).

ERBB2-positive tumours have poorer prognoses, however, along with the basal-like tumours, have a better chance of responding to neoadjuvant therapy compared to luminal A and B (Espinosa *et al.*, in press).

Subtype	Receptor status[1]	Prevalence[2]	DNA methylation profile[3]
Luminal A	ER+ PgR+ ERBB2-	55-59.2%	Distinct
Luminal B	ER+ PgR+ ERBB2+		Distinct
Basal-like	ER- PgR- ERBB2-	13.1-18%	Distinct
ERBB2+	ER- PgR- ERBB2+	27-27.6%	None

Table 1-2 - Summary of breast cancer molecular subtypes identified by Sorlie *et al.* (2003), showing receptor status, prevalence, and DNA methylation profile.

[1] Taken from Sanpaolo *et al.* (2011).

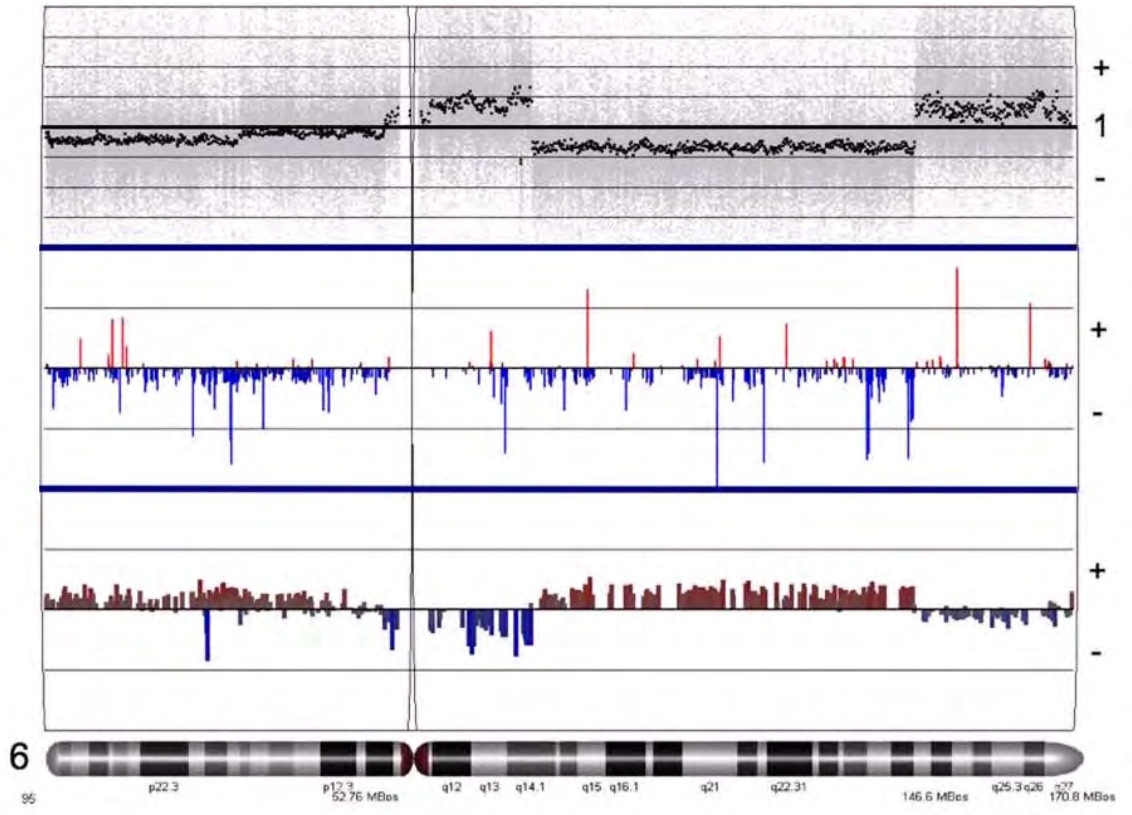
[2] Taken from Hannemann *et al.* (2006) and Gaudet *et al.* (2011).

[3] Taken from Andrews *et al.* (2010) and Holm *et al.* (2010).

Personalis EX2155

Gene expression profiles may not necessarily relate to the observed pathological state - epigenetics can play a role too (Bicciato *et al.*, 2003). For example, Andrews *et al.* (2010) found that copy number variation (CNV) in a breast cancer metastatic cell-line was superimposed by DNA methylation to promoter regions. The authors additionally found that hypermethylation correlated with decreased gene expression and deletion, whereas hypomethylation correlated with increased expression and amplification (Figure 1-6). The group also noted that hypomethylation (and, thus, amplification) increased as cells became more invasive.

Figure 1-6 - Image highlighting correlation between amplification and deletion (top), increased and decreased gene expression (middle), and hyper- and hypomethylation (bottom) (taken from Andrews *et al.*, 2010).



1.3.2 Commercial profiling

Gene 'signature' analysis, such as that used by commercially available tools, can help in the prognostic setting (Espinosa *et al.*, in press). Two such tools include

Personalis EX2155

MammaPrint™ (Agendia NV, Science Park 406, 1098 XH Amsterdam, The Netherlands, http://www.agendia.com/) and Oncotype DX® (Genomic Health Inc., 101 Galveston Drive, Redwood City, CA 94063-4700, USA, http://www.genomichealth.com/). Oncotype DX® calculates a recurrence score (RS) based on the levels of expression of 21 genes in ER-positive, lymph node-negative patients, and gives a result of low (RS<18), intermediate (RS≥18 and <31), or high (RS≥31), relating to the 10-year risk of recurrence (Sparano & Paik, 2008 and Contreras, 2009). It has proved valuable in numerous studies for predicting response to chemotherapy and relapse (Albain *et al.*, 2010 and Dowsett *et al.*, 2010). MammaPrint™, conversely, calculates the 10-year survival rate and takes a microarray-based approach to analyse the expression of 70 genes - it is designed for ER-positive or -negative, lymph node-negative patients (Ross, 2009 and Kunz, 2011).

Both assays are compared in Table 1-3.

	Oncotype-DX®	**MammaPrint™**
Tumour criteria	ER-positive Lymph node-negative No distant metastases	ER-positive or -negative Lymph node-negative No distant metastases
Tissue type	Formalin-fixed, paraffin-embedded (FFPE)	Fresh Fresh frozen (FF)
Approach	Quantitative PCR (qPCR)	cDNA microarray
Result	10-year risk of recurrence (low, intermediate, or high)	10-year survival rate

Table 1-3 - Comparison of Oncotype-DX® and MammaPrint™.

Caution must be emphasised when using gene profiling methods until they can be proven accurate through clinical trials (Sanpaolo *et al.*, 2011). Both tests mentioned, in particular, are also expensive, require an amount of tissue that might not be available from small tumours, and require that tissue be sent to a third party (Espinosa *et al.*, in press), which generates both ethical and legal issues.

Personalis EX2155

Agendia NV received 510(k) clearance for MammaPrint™ in 2007 from the US Food and Drug Administration (US FDA), which permitted the marketing of the assay and was the first test of its kind to receive such clearance (US FDA, 2007).

1.3.3 Profiling with SNPs and CNV

Single nucleotide polymorphisms (SNPs) are polymorphisms affecting a single base that have frequencies of at least 1% in the population (Haimovich, 2011) and account for some of the roughly 1% of the human genome that differs between individuals (LaFramboise, 2009). They have traditionally been used for linkage disequilibrium (LD) studies where an allele predisposing a particular disease is linked with a particular SNP (LaFramboise, 2009). Current estimates are that there are more than 10 million SNPs in the human genome that have a minor allele frequency of 5% (Kruglyak & Nickerson, 2001 and LaFramboise, 2009). Studies have been carried out on a whole-genome scale to map SNPs in different populations in normal healthy samples, such as the International HapMap Project (International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/). However, studies that analyse LD on a whole-genome scale and with a high resolution are lacking. Additionally, variance in the human genome was originally thought to be mostly consisting of SNPs, but this notion has been revised since the discovery that CNVs are common also (Redon *et al.*, 2006). The commonality and abundance of CNV can be inferred when, for example, any two healthy individuals' samples are compared to reveal gains and losses between each (Van de Wiel *et al.*, 2011).

Copy number variations are more likely than SNPs to be causal variants owing to the fact that a larger section of DNA can be affected (Jin *et al.*, 2011). Therefore, DNA copy number (CN) analysis has an important role in cancer research that is rising (Van de Wiel *et al.*, 2011). Indeed, the identification and cataloguing of CNVs should benefit the genetic and functional analysis of the human genome, and also help to advance our understanding of its dynamic nature (Wang *et al.*, 2007). Copy number variations include duplications, losses, and inversions, and they can be of variable size - ranging from a few kilobases (Kb) to megabases (Mb) (Scherer *et al.*, 2007 and Grayson & Aune, 2011) - but are generally regarded as being ≥1,000 bases in length and can vary

Personalis EX2155

in CN between individuals (Redon *et al.*, 2006 and LaFramboise, 2009). They are generated somatically throughout life, thus they contribute to genomic variation in the normal human genome (Sebat *et al.*, 2004; Piotrowski *et al.*, 2008; and Grayson & Aune, 2011).

Copy number variations detected using microarray have long-been associated with neurological disorders, including schizophrenia and schizoaffective disorder (Kirov *et al.*, 2009 and Tam *et al.*, 2009), and other conditions such as the autism spectrum disorders (ASDs) (Sebat *et al.*, 2007) and attention deficit hyperactivity disorder (ADHD) (Williams *et al.*, 2010). Kirov *et al.* (2009) found, for example, a 4-fold increase in deletions of more than 1Mb in the DNA of schizophrenia patients. Copy number variations are also regarded as susceptibility factors for infectious and autoimmune disorders (Yoshihara *et al.*, 2010). Indeed, it was noted by Schuster-Böckler *et al.* (2010) that genes affected by CNV often had roles in immunity and sensory perception. The authors also noted a high probability that CNVs were paralogues that became fixed in the human genome through evolution.

Work relating CNVs to cancer - particularly breast cancer - is lacking. Redon *et al.* (2006) looked at CNVs in healthy individuals, while Graubert *et al.* (2007) examined them in mice. Jin *et al.* (2011) aimed to identify CNV and related SNPs that could both be used in distinguishing between types of aggressiveness in prostate cancer using the process of tagging SNPs for CNVs. They identified a 32.3Kb deletion that was 'tagged' with SNP rs2209313 on 20p13. One other notable study, which compared CNV in *BRCA1*-associated ovarian cancer to that in sporadic ovarian cancer, suggested that the two were distinct from each other based on their respective CNV profiles (Yoshihara *et al.*, 2010).

1.3.4 Next-generation sequencing

High throughput approaches, such as next-generation sequencing (NGS), have great potential to aid the detection, management, and treatment of disease through the rapid and simultaneous screening of numerous gene targets (potentially resulting in a cost reduction), thereby replacing current assays that only target single genes (Meldrum *et*

Personalis EX2155

al., 2011 and Moch *et al.*, 2012). Next-generation sequencing offers greater resolution and accuracy over microarrays and is capable of detecting both translocations and inversions, which arrays currently cannot (LaFramboise, 2009). Table 1-4 provides a comparison of the latest NGS technology releases by three of the leading providers.

	454 Life Sciences *GS FLX Titanium XL+*	**Applied Biosystems™** *SOLiD™ 5500xl*	**illumina Inc.** *HiSeq 2500*
DNA required	500ng	-	50-1,000ng
Read length	Up to 1,000 base	Up to 75b	Up to 100b
Reads per run	~1,000,000	≥100,000,000	3,000,000,000
Throughput	700Mb per run	10-15Gb per day	600Gb per day
Accuracy	99.997%	Up to 99.99%	-
Sequencing method	Synthesis (Margulies *et al.*, 2005)	Ligation (Porreca *et al.*, 2007)	Synthesis
Sample barcoding (Hamady *et al.*, 2008)	Yes	Yes	Yes
Main advantage	Longest read length Lowest cost per experiment	Highest accuracy	Highest reads per run Highest throughput Lowest cost per Mb

Table 1-4 - Comparison of 3 NGS technologies, highlighting the main differences and advantages of each (tabulated from Glenn, 2011; Wall *et al.*, 2009; 454 Life Sciences[1]; Applied Biosystems™[2]; and illumina Inc.[3]).

However, the majority of current NGS platforms follow similar protocols and are prone to error, including: the incorporation of incorrect bases during the PCR step; shorter read lengths (compared to Sanger di-deoxy sequencing); and sequence detection errors (Meldrum *et al.*, 2011). To compensate, the sequencing reaction is performed numerous times (number of reads) in order to improve the precision of base-calling. Ideally, each base should be read 30-50x (coverage) in each experiment (Haimovich, 2011 and Meldrum *et al.*, 2011). For cancer genomes, greater coverage and specific

[1] 454 Life Sciences, http://454.com/.

[2] Applied Biosystems™, http://www.appliedbiosystems.com/.

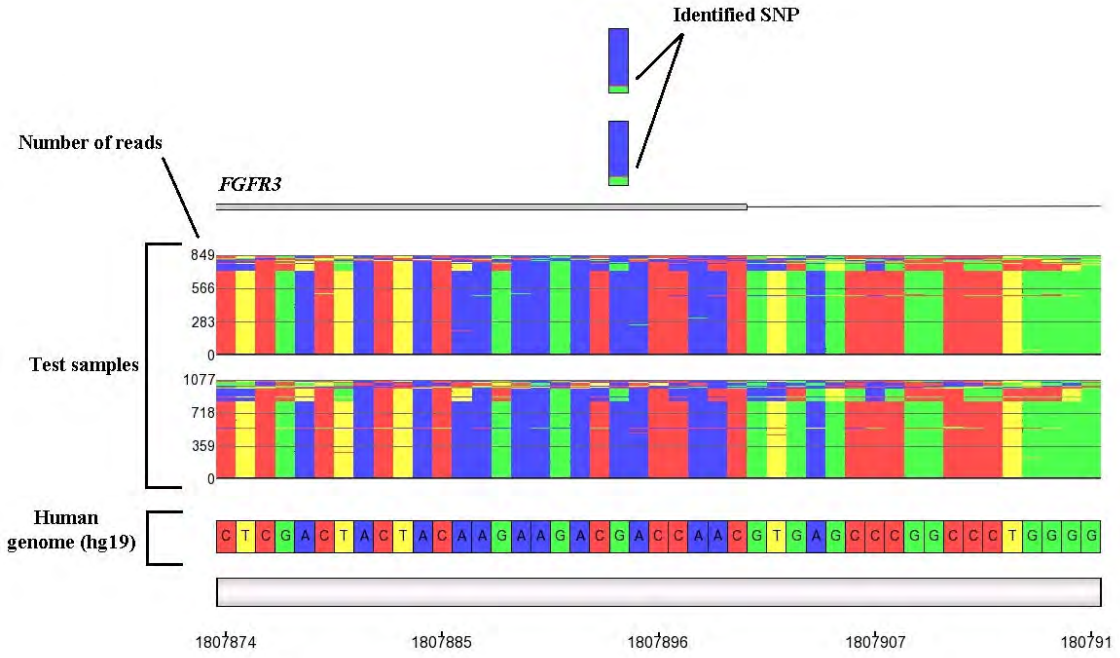
[3] illumina Inc., http://www.illumina.com/.

Personalis EX2155

algorithms may be required to compensate for aneuploidy and heterogeneity. Hence, NGS is currently considered a research tool - rather than an aid - for diagnosis and treatment (Meldrum *et al.*, 2011).

Analysis of NGS data is a challenging issue that must be considered. The management and processing of raw sequence data necessitates the availability of both high-end hardware and dedicated manpower (Meldrum *et al.*, 2011) - storage issues are also important (Haimovich, 2011). Raw sequence data must be processed in order to carry out base-calling, for which different algorithms are available (Ledergerber & Dessimoz, 2011). Each sequence of bases must then be aligned to a reference genome (Figure 1-7) in order to permit the identification of mutations and other genomic variants (Meldrum *et al.*, 2011). The methods for this purpose are still under research and development (Haimovich, 2011 and Meldrum *et al.*, 2011). The application of NGS to DNA extracted from FFPE tissue is another technical challenge (Meldrum *et al.*, 2011).

Figure 1-7 - Alignment of NGS sequence reads for two test samples to a reference genome (hg19) identifies a single SNP on chromosome 4 in an exon of *FGFR3* in both samples. Blue, adenine; yellow, thymine; green, guanine; red, cytosine.



Personalis EX2155

1.3.5 Potential of DNA arrays

DNA arrays are central to biomedical research (LaFramboise, 2009). They have the potential to characterise genomes by picking-up SNP, CNV, and other genomic variation through taking a whole-genome approach, whilst also having the potential to compare differences in these variants across different tissues and samples. They improve upon targeted/candidate gene studies by analysing many genes and other genomic regions in tandem, and they have the fundamental concept that variants of risk associated with common diseases have minor allele frequencies of 0.5-5%, while rarer variants have frequencies of less than 0.5% (Manolio *et al.*, 2009). A different chip can be used for measuring levels of each label/target, represented by a short oligonucleotide probe, across the tissues. However, a more robust, reproducible, and quicker method is to measure multiple targets on the same array chip, with each having a different and distinguishable fluorescent dye (Quackenbush, 2002).

Arrays have the potential to unlock genomic instability, which is an increased rate of genomic alteration (Kwei *et al.*, 2010) that can promote cancer development and progression (Uchida, 2006). It also results in aberrations in the genome that can occur in cancer. For example, DNA mismatch repair, chromosomal instability (CIN), and telomere dysfunction have previously been mentioned as being mechanisms of genomic alterations. Problems that can result from these include breakages, aneuploidy of chromosomes in tumour DNA, isochromosomes, and 'uncapped' telomeres (Uchida, 2006 and Kwei *et al.*, 2010). Telomeres, specifically, permit that the ends of chromosomes are distinguishable from double-strand breaks (Heidinger *et al.*, 2012) and thus help to maintain genomic stability. They consist of repetitive sequences that are non-coding but that shorten during cell division due to polymerase's inability to completely replicate the DNA strand being copied (Heidinger *et al.*, 2012).

Aneuploidy in a particular chromosome could also be inferred through microarray by the fluorescent intensity of the probes that hybridised - the resulting $\log_2$ probe ratios would then infer the aneuploid anomaly. Looking at CN could also infer it: for example, if an entire chromosome had been duplicated, then it would appear as CN 4 across both of its arms. Telomere loss could also be inferred by significant CN loss

Personalis EX2155

(Castle *et al.*, 2010) - if consecutive probes gave a CN of 0 at a telomere, for example, one could then infer that telomere attrition had occurred. Around 85% of immortal tumours have short telomeres and exhibit elevated expression of telomerase (Shawi & Autexier, 2008).

1.3.5.1 Affymetrix microarrays

Affymetrix Inc. (Affymetrix Inc., 3420 Central Expressway, Santa Clara, CA 95051, USA, http://www.affymetrix.com/) was the first to produce oligonucleotide-based arrays (Mount, 2004 and LaFramboise, 2009). The earliest application of such produced encouraging results - for example, different studies found good agreement between PCR and microarray data when analysing loss of heterozygosity (LOH) in prostate (Dumur *et al.*, 2003) and bladder (Hoque *et al.*, 2003) cancer. However, these arrays had just 1,494 SNP markers to genotype (LaFramboise, 2009). A different study made use of a higher resolution Affymetrix array, but still just utilising 10,000 SNP markers (Lieberfarb *et al.*, 2003). Further releases by Affymetrix incorporated 100,000, 250,000, and then 500,000 SNP markers to genotype (LaFramboise, 2009).

The other major providers of microarray technology in the world are Agilent Technologies Inc. (Agilent Technologies Inc., 5301 Stevens Creek Boulevard, Santa Clara, CA 95051, USA, http://www.agilent.com/) and illumina Inc. (illumina Inc., Corporate Headquarters, 5200 Research Place, San Diego, CA 92122, USA, http://www.illumina.com/). The latest release by illumina - the *HumanOmni5-Quad BeadChip* - can accommodate four samples per chip, with each being scanned for 4,301,331 variants using 50mer probes and having additional space for a further 500,000 custom markers. However, the amount of DNA required per sample is 400ng (illumina Inc., http://www.illumina.com/). Conversely, Agilent's latest release - the *SurePrint G3 Human CNV Microarray* - scans the genome for 963,331 variants using 60mer probes and has a DNA requirement of 500ng (Agilent Technologies Inc., http://www.agilent.com/).

A problem associated with arrays (especially those involving a high-density of probes such as microarrays) is that a large starting quantity of DNA is often required for

Personalis EX2155

analysis: for example, the latest release by Affymetrix Inc. requires 50ng/µl (total 250ng) DNA, but 100ng/µl is recommended (Affymetrix Inc., 2008a), whilst the latest releases by Agilent Technologies Inc. require 200ng of DNA (the most recent release requires 400ng) (Agilent Technologies Inc., http://www.agilent.com/).

The process of whole genome amplification (WGA) can be used to generate high quantities of DNA in studies where only a small number of cells are available, such as those obtained from laser capture microdissection (Rook *et al.*, 2004). This can provide sufficient template for post-analysis validation studies. However, in order to be successful, the WGA process must both accurately amplify the target DNA and also provide sufficient quantities of it (Jasmine *et al.*, 2008). Using the *Affymetrix Early Access Mendel Nsp 250K GeneChip* array platform, Jasmine *et al.* (2008) compared the SNP call-rates between genomic DNA and whole genome amplified DNA of 14 healthy controls and found excellent concordance (>99%) between the two sets of samples on a paired basis. A similar study by Tzvetkov *et al.* (2005) found the same high level of concordance as that found by Jasmine *et al.* (2008).

1.3.5.2 The use of *Affymetrix SNP 6.0*

The sixth generation of Affymetrix SNP arrays is the *Genome-Wide Human SNP Array 6.0* (*Affymetrix SNP 6.0*). It genotypes much less than the ~10 million known SNPs but is still capable of building a large picture of variation (LaFramboise, 2009). The array chip is capable of genotyping 906,600 SNPs and contains an additional 945,826 probes for detecting CN. Of the CN probes, 115,000 target previously-know CNVs. Overall, markers are distributed evenly across the genome, with a median marker spacing of 1-5Kb) (Affymetrix, 2008a).

A comparison of the *Affymetrix SNP 6.0* to the latest releases by illumina Inc. and Agilent Technologies Inc. is provided by Table 1-5.

Personalis EX2155

	Affymetrix Inc.	illumina Inc.	Agilent Technologies Inc.
Total markers scanned	1,852,426	4,301,331	963,331
Marker spacing (median)	1-5Kb	0.36Kb	2.6Kb
Probe length	25mer	50mer	60mer
DNA required	50ng/µl	400ng/µl	500ng/µl

Table 1-5 - Comparison of *Affymetrix SNP 6.0* to the latest releases of two other leading organisations.

1.3.5.2.1 Genotyping calling algorithms

Early versions of Affymetrix arrays (10K and 100K) utilised the partitioning around means (PAM) algorithm for making genotype calls. This involved comparing the signal intensities of perfect match (PM) and mismatch (MM) probes to deduce the genotype. Later (500K), the bayesian robust linear model with mahalanobis (BRLMM) algorithm was used - this eliminated the use of the MM probes. The *Affymetrix SNP 6.0* saw the adaptation of the Birdseed algorithm (Broad Institute, http://www.broadinstitute.org/mpg/birdsuite/index.html) and only incorporates 6 or 8 probes for each target (i.e. - 3 or 4 PM probes for each SNP allele genotype of A or B, respectively) (LaFramboise, 2009) (Table 1-6).

	SNP	CN
Total	906,600	945,826
Replicate probes	6 or 8 (3 or 4 for each allele of AB)	1
Probe length	25mer	25mer
Genotype-calling algorithm	Birdseed	Birdseed

Table 1-6 - SNP and CN probe information for the *Affymetrix SNP 6.0* (Affymetrix, 2008a and LaFramboise, 2009).

Personalis EX2155

1.3.5.2.2 Capabilities

As well as genotyping SNPs, Affymetrix arrays are also capable of detecting LOH (Yamamoto *et al.*, 2007 and Huggins *et al.*, 2008), which can reveal the presence of a recessive mutation by knocking out the wild-type allele (Knudson's two-hit hypothesis). However, the phenomenon of copy neutral-LOH (CN-LOH) can pose problems to this: CN-LOH can occur, for example, where a chromosomal region is subject to LOH but is then duplicated, leaving a neutral CN state of 2 (Langdon *et al.*, 2006). However, this could still prove to be a tumourigenic event (Teh *et al.*, 2005) whilst remaining undetected through other array technologies (Andersen *et al.*, 2007). The *Affymetrix SNP 6.0*, however, is equipped to unearth CN-LOH as it scans the genome with both CN and SNP probes (Andersen *et al.*, 2007). This phenomenon has already been explored in follicular lymphoma (Fitzgibbon *et al.*, 2007 and Ross *et al.*, 2007), neuroblastoma (George *et al.*, 2007), and colorectal cancer (Andersen *et al.*, 2007).

Additionally, although the *Affymetrix SNP 6.0* uses both SNP and CN probes, the SNP probe signal intensities can - along with those of the CN probes - be used to estimate CN - i.e. CNVs can be revealed by SNPs and thus be indirectly inferred from genotyping (Korn *et al.*, 2008; Jin *et al.*, 2011; and Stjernqvist *et al.*, 2011). This owes to the fact that signal intensity is merely dependent on the amount of target sequence present (LaFramboise, 2009). For example, Jin *et al.* (2011) noted that although studies had identified CNVs at three loci associated with Crohn's disease, rheumatoid arthritis, and diabetes, these CNVs had already been identified in SNP-based studies. Additionally, SNP analysis can be used to determine which allele has been either amplified or deleted (Korn *et al.*, 2008 and Irvin *et al.*, 2011) - for example, knowing that a section of DNA within a gene has been amplified does not provide information on the particular allele in question. Combining SNP and CN data can help determine this (Korn *et al.*, 2008).

1.4 Circulating cell-free DNA

Circulating cell-free DNA (cfDNA), defined as extracellular DNA found in the blood, was first described in 1948 (Mandel & Metais, 1948). In the blood of a healthy

Personalis EX2155

individual, it can be found at a concentration of roughly 1ng/ml (Jung *et al.*, 2010). However, in some disease settings, the concentration can be elevated to as much as 100ng/ml of soluble DNA (Shaw *et al.*, 2000; Goebel *et al.*, 2005; Jung *et al.*, 2010; Schwarzenbach *et al.*, 2011; and Alix-Panabières *et al.*, 2012).

Cell-free DNA analysis can be both plasma- and serum-based (Anker & Stroun, 2000; Goebel *et al.*, 2005; Gahan, 2007; and Tsang & Lo, 2007). Serum-based studies have shown cfDNA's promise as a potential predictor of therapy-response in cancer (Leon *et al.*, 1977) and also as having use in diagnosis and prognosis (Chen *et al.*, 2000; Wang *et al.*, 2004; and Umetani *et al.*, 2006), whilst plasma-based studies have also marked cfDNA's promise in diagnosis, prognosis, and follow-up studies (Anker *et al.*, 2001; Shaw *et al.*, 2000; and Pinzani *et al.*, 2011). However, there remains debate over which of the two is better to use (Jung *et al.*, 2010). Although circulating cfDNA has been found at higher concentrations in serum compared to plasma (Lee *et al.*, 2001; Holdenrieder *et al.*, 2005; Umetani *et al.*, 2006; and Board *et al.*, 2008), cfDNA isolated from plasma was suggested as a more reliable source of tumour DNA (Board *et al.*, 2008). Indeed, Lee *et al.* (2001) concluded that serum should not be used for measuring the concentration of cfDNA as higher levels of cfDNA observed in serum can result from blood cell contamination (Gormally *et al.*, 2007).

1.4.1 Origins of cfDNA

In healthy individuals, cfDNA is believed to be derived from apoptosis (Stroun *et al.*, 2001 and Suzuki *et al.*, 2008), although this has been refuted (Van der Vaart & Pretorius, 2008; Schwarzenbach *et al.*, 2011; and Alix-Panabières *et al.*, 2012). However, in cancer, cfDNA is also released as a result of the necrosis of tumour cells (Li *et al.*, 2003; Zhong *et al.*, 2007; Schwarzenbach *et al.*, 2011; and Alix-Panabières *et al.*, 2012), which have a high turnover. DNA that originates from necrosis is made-up of fragments of longer length that are above the apoptopic limit (Jung *et al.*, 2010).

Personalis EX2155

1.4.1.1 Tumourigenic origin

Studies by Leon *et al.* (1977), Shapiro *et al.* (1983), and Stroun *et al.* (1987) were amongst the first to demonstrate that the cfDNA concentration in blood was often elevated in malignant - compared to benign - diseases, and this was later confirmed in breast cancer (Huang *et al.*, 2006 and Zhong *et al.*, 2007). The study by Zhong *et al.* (2007) observed that the amount of cfDNA correlated with tumour size and was pronouncedly increased in breast cancer patients who had lymph node involvement. Additionally, the DNA integrity index (ratio of longer to shorter fragments found circulating in the blood) is known to be higher in cancer patients (Page *et al.*, 2006).

The tumourigenic origin of cfDNA was shown previously through the detection of *ERBB2* gene amplification in matching tumour and blood (Chiang *et al.*, 1999 and Page *et al.*, 2011). However, the majority of earlier studies have focussed on the analysis of LOH in microsatellite DNA and DNA methylation. For example, two studies first showed tumour-specific LOH in microsatellite DNA analysed from the plasma and serum of advanced small cell lung cancer patients (Chen *et al.*, 1996) and head and neck cancer patients (Nawroz *et al.*, 1996). Shaw *et al.* (1996 and 2000) then analysed DNA taken from the plasma of primary and metastatic breast cancer patients to search for any tumour-specific changes that were present in this DNA. Promising results were found: in more than 85% of patients, tumour-specific DNA was detected in plasma. However, the group only searched for two markers, which had previously been shown to undergo LOH or molecular instability in breast cancer patients, and admitted that more markers could have been analysed to improve the frequency of detected alterations.

Tumour-specific DNA methylation at certain loci was also observed through cfDNA analysis in different cancers, including ovarian (Balch *et al.*, 2004), lung (Bowman *et al.*, 2006), and breast (Levenson, 2007). In breast cancer, specifically, higher levels of methylation were positively-correlated with tumour progression (Hoque *et al.*, 2006), although only 4 genes were studied. Additionally, the methylation of 3 genes - *CCND2*, *RARB*, and *TWIST1* - occurred more frequently in cfDNA collected from the mammary ducts of women with ductal carcinoma (34%) when compared to those without (11%). However, two-fifths of those without subsequently received a diagnosis of cancer

Personalis EX2155

(Evron *et al.*, 2001). Later, in a metastatic breast cell-line, Andrews *et al.* (2010) noted that hypermethylation correlated with decreased gene expression and deletion, whereas hypomethylation correlated with increased expression and amplification.

1.4.1.2 Other sources of cfDNA

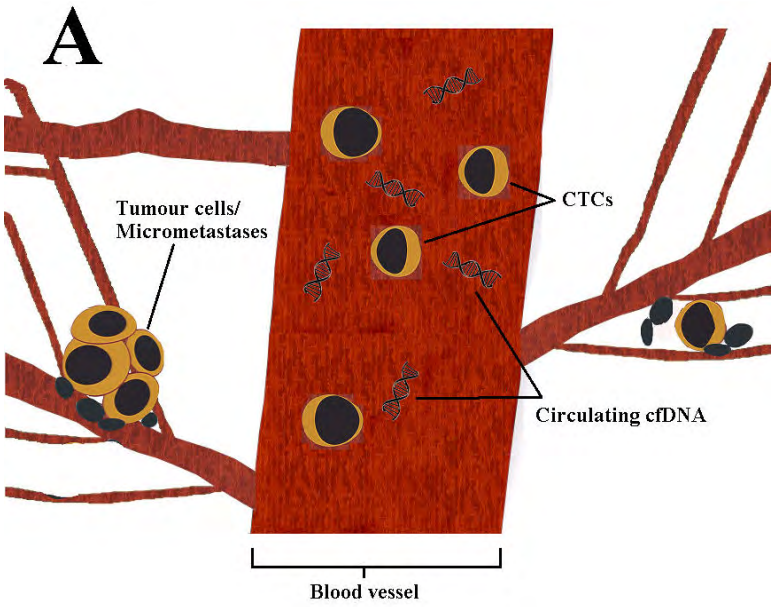
Possible sources of circulating cfDNA, other than from the tumour and surrounding tissue, include lysis of circulating tumour cells (CTCs) and disseminated tumour cells (DTCs). Alterations in cfDNA do not always match those of the primary tumour and it has been suggested that this can be explained by the genomic differences between CTCs and DTCs (Alix-Panabières *et al.*, 2012). Circulating tumour cells are cells of tumourigenic origin that are associated with a higher risk of recurrence (Zippelius & Pantel, 2000; Meng *et al.*, 2004; and Payne *et al.*, 2012) and are very heterogeneous (Alix-Panabières *et al.*, 2012). Separate studies showed that the lifetime of CTCs in the body ranged from just 1-2.4 hours (Meng *et al.*, 2004), while that of cfDNA ranged from 15 minutes to several hours (Schwarzenbach *et al.*, 2011). Organs involved in the disappearance of both CTCs and cfDNA have been suggested as being the liver, spleen, and kidney (Minchin *et al.*, 2001; Korabecna *et al.*, 2008; Tamkovich *et al.*, 2008; and Schwarzenbach *et al.*, 2011).

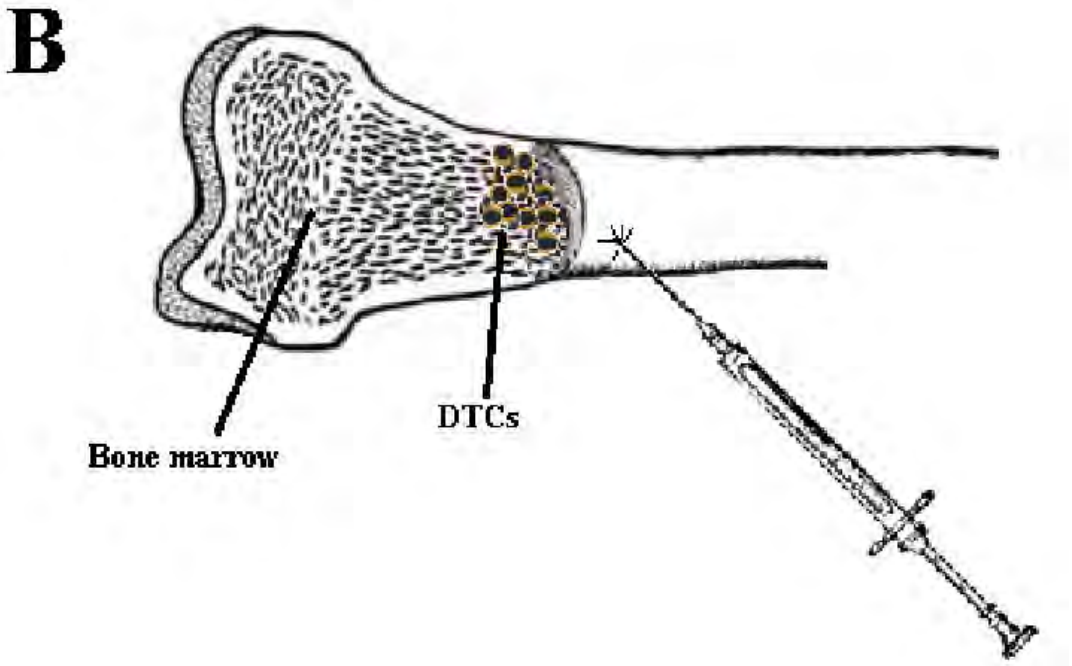
The detection of CTCs in blood could help in optimising therapy for patients, while its quantification could help in the adjustment of the amount of such therapy received (Zhong *et al.*, 2007 and Lianidou & Markou, 2011). Disseminated tumour cells, on the other hand, can seed micrometastases that can remain dormant for many years, even after the complete resection of the primary tumour (Alix-Panabières *et al.*, 2012). It has recently been demonstrated that the presence of viable DTCs in the bone marrow was inversely related to that of circulating cfDNA in blood plasma of patients in the post-adjuvant setting (Payne *et al.*, 2012).

Although the true source of cfDNA is not known, Alix-Panabières *et al.* (2012) report how research groups are continuing the investigation of the relationship(s) between cfDNA and both CTCs and DTCs (Figure 1-8).

Personalis EX2155

Figure 1-8 - Determining the relationship between cfDNA and CTCs in the blood circulation (A) and DTCs in the bone marrow (B).





Personalis EX2155

1.4.2 Biological importance of cfDNA

The biological importance and significance of cfDNA in relation to metastasis in cancer has been addressed previously (Jung *et al.*, 2010). Metastases can form from only a small subset of cells that arrive at a distant organ and can exist in 3 states:

- Actively-growing with angiogenic capabilities
- Growing, but growth is balanced by apoptosis
- Dormant cells

(Goss & Chambers, 2010)

However, to be capable of successfully metastasising, a tumour cell first needs to be capable of leaving the primary site. It then also needs to relocate itself to the distant organ whilst resisting apoptosis and an immune system attack. Once at the distant site, it has to have the capability of growing into a secondary tumour (Goss & Chambers, 2010). It was shown by a previous study that cfDNA could be taken-up by cells and thereafter incorporated into their respective genetic material (García-Olmo & García-Olmo, 2001). However, the study did not involve human DNA, nor was it performed *in vivo*. Nevertheless, if tumour-derived DNA were taken-up by cells at distant metastatic sites - and thereafter incorporated into the cell's genome, thus conferring oncogenic potential upon it - it would serve as a plausible mechanism for how metastasis involving cfDNA occurred in cancer patients. This theory has been termed 'genometastasis' (García-Olmo *et al.*, 1999 and 2000).

1.4.2.1 Cell-free DNA as a diagnostic and prognostic marker

Cancer is a chronic disease in patients (Meng *et al.*, 2004) and the development of a simple blood test for cancer screening is highly desirable (Sunami *et al.*, 2008). However, the finding by Gangnus *et al.* (2004) and others (Schmidt-Kittler *et al.*, 2003 and Vecchi *et al.*, 2008) that different genetic changes can occur with progressive micrometastases from solid tumours poses a barrier to this. It highlights the importance of characterising, not just the primary tumour, but also the cells and cellular material that are disseminated from it and how such cells genetically evolve over time.

Personalis EX2155

Analysis of cfDNA is one enticing approach to interrogate the 'circulating cancer genome'. The main advantages of using cfDNA as a biomarker are its availability through non-invasive means and its stability (Gormally *et al.*, 2007). It can also be used for genetic polymorphism studies (Blömeke *et al.*, 1997). The concentration of cfDNA in relation to tumour size and metastasis has been suggested as a diagnostic tool (Wang *et al.*, 2003; Zhong *et al.*, 2007; and García-Olmo *et al.*, 2008). However, a pitfall to using the cfDNA concentration as a biomarker for early cancer detection is that the concentration can be affected by exercise, benign disease, and other more serious illnesses not related to cancer, including diabetes and problems related to the hepatic and respiratory systems (Jung *et al.*, 2010 and Alix-Panabières *et al.*, 2012). The usefulness of the DNA integrity index as a biomarker in cancer has also been examined through cfDNA (Wang *et al.*, 2003; Sunami *et al.*, 2008; and Pinzani *et al.*, 2011), with promising results in rectal (Agostini *et al.*, 2011) and head and neck cancer (Jiang *et al.*, 2006), and also renal cell carcinoma (Gang *et al.*, 2009).

1.4.2.2 The search for markers of minimal disease

The fundamental principle of detecting cancer at an early stage is to look for changes from what is considered normal or healthy, such as altered levels of gene expression (Srivastava *et al.*, 2001). Early detection is key as it can result in a better chance of survival (Alix-Panabières *et al.*, 2012). Until now, breast cancer patients were simply monitored until they developed overt metastases, the onset of which can ultimately lead to the death of the patient. Thus, a method to predict metastatic disease progression is needed, as recent results of adjuvant therapy trials have shown that long-term eradication of micro-metastases can be achieved by adjuvant chemotherapy or endocrine therapy (Boér, 2010 and Stickeler, 2011).

Morphological factors such as tumour type and size and lymph node-stage - and molecular factors such as hormone receptor status - are used to determine individual risk, but these are assessed in the primary tumour that is often removed during surgery. Thus, they are not useful for detecting minimal residual disease. For cfDNA to serve as a useful circulating biomarker, it will first have to be proven that DNA extracted from

Personalis EX2155

the plasma of blood can be used as a surrogate for the primary tumour. Previous studies showed this in specific loci (Shaw *et al.*, 2000; Papadopoulou *et al.*, 2006; and Shao & Nguyen, 2006) but the whole genome has thus far not been compared between tumour and cfDNA.

1.5 Background to thesis

Identifying disease-related variants is the prime target of human genetics and association studies must be capable of distinguishing between variants associated with disease and those that are unrelated (LaFramboise, 2009). Previous pilot data from our research group using Affymetrix's 10K array showed good concordance between tumour and plasma DNA from the same patient when looking at LOH, confirming previous findings with individual microsatellite loci (Shaw *et al.*, 2000). In the pilot study, LOH was detected in the paired tumour and plasma DNA samples from 5 patients. The common frequency of LOH between any two matched samples (i.e. - from the same patient) ranged from 17-71% across all patients. Interestingly, alterations unique to plasma DNA were found in all cases, suggesting disease progression and relating to the earlier finding by Gangnus *et al.* (2004) that genetic changes occur with progressive micrometastases from solid tumours.

These encouraging results prompted the characterisation of cfDNA in detail in the on follow-up of breast cancer in this thesis, with a view to monitoring SNPs and CNVs using *Affymetrix SNP 6.0* for characterising the changing nature of the circulating cancer genome over time in a single patient.

1.6 Hypothesis to be tested

The hypothesis to be tested in this thesis is that cfDNA is a useful surrogate tumour marker that can be analysed to determine a suite of useful biomarkers for diagnostic and prognostic purposes. A secondary hypothesis is that the circulating cancer genome 'evolves' over time and can give indications, through cfDNA, of an impending metastasis.

Personalis EX2155

1.7 Aims

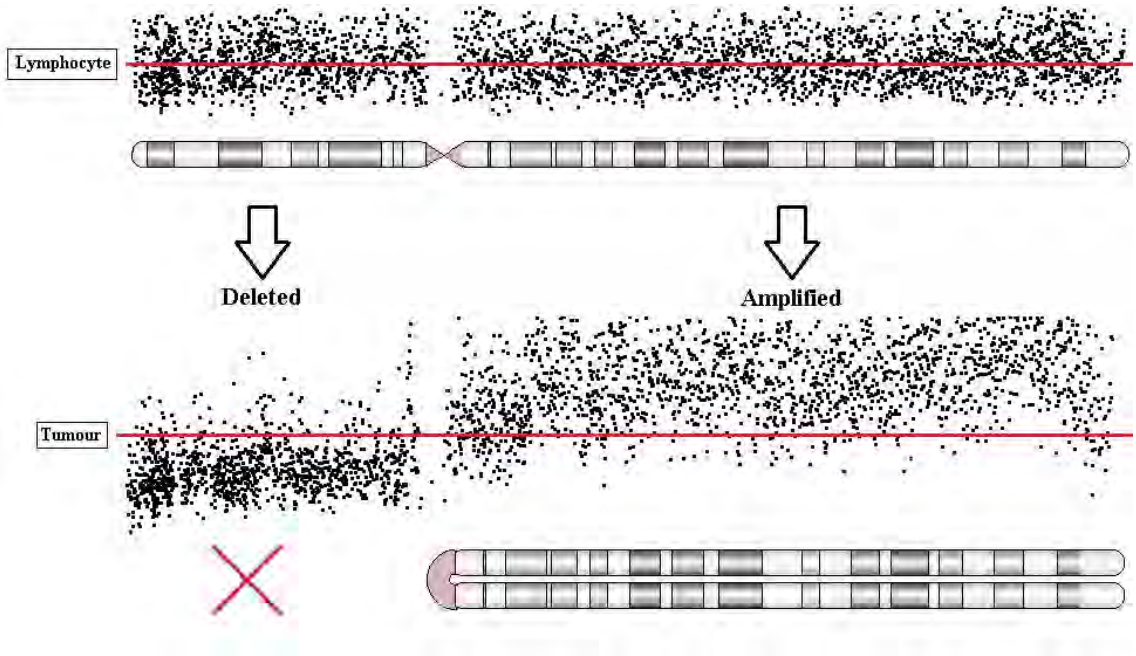
The aim of this thesis was to test these two related hypotheses using bioinformatics to analyse/interrogate *Affymetrix SNP 6.0* microarray data and compare cfDNA extracted from blood plasma with matched lymphocyte and tumour DNA (where available).

1.8 Objectives

The specific objectives were:

1. Molecular characterisation of circulating cell-free DNA in breast cancer patients using SNP 6.0 microarrays (Chapter 3).
2. Monitoring circulating cell-free DNA in response to therapy in breast cancer patients using SNP 6.0 microarrays (Chapter 4).
3. Painting a portrait of breast cancer through principal component analysis (Chapter 5).
4. Microarray data validation (Chapter 6).

Personalis EX2155



2 MATERIALS AND METHODS

Personalis EX2155

2.1 Project funding and sponsorship details

The work carried out as part of this thesis supported a project grant funded by Cancer Research UK's (CRUK) Biomarkers and Imaging Discovery and Development (BIDD) committee, entitled 'Development of a cell-free plasma DNA-based method to predict metastatic progression in breast cancer patients: the application of SNP arrays in the post-adjuvant period'. Funding was also provided by the CRUK/Department of Health's Experimental Cancer Medicine Centres (ECMC). The protocols were approved by the Riverside Research Ethics Committee (REC) (now the West London REC) and were conducted in accordance with the Declaration of Helsinki. Patients gave written informed consent prior to participation. Samples were blinded for analysis and patients understood that the results would not be made available to them. The project was sponsored by University Hospitals of Leicester.

2.2 Associated clinical studies and trials

The main clinical study with which my research was associated was the minimal residual disease (MRD) clinical study. It was set-up with the aim of investigating the utility of bone marrow immunocytochemistry and CK19 real-time quantitative PCR (qPCR) for the detection of minimal residual disease in breast cancer patients, both at diagnosis and follow-up. The trial was reviewed in 2005 and - out of a total of 170 patients that were originally recruited - 65 consented to have repeat bone marrow samples taken at regular intervals during follow-up. Six years post the trial's inception, our research group continued to take samples every 6 months, and from two categories of patients: 29 high risk (HR) (>3 positive nodes at diagnosis) and 28 low risk (LR) (T_1 or lymph node-negative).

Early results from the bone marrow samples were promising, with a positive status reported in 52 and 26% of HR and LR patients, respectively, by immunocytochemistry. For real-time qPCR, the figures for CK19 transcripts were 47 and 32% for HR and LR, respectively. Both categories of patients constitute one of the most intensively studied cohorts of breast cancer patients. In the LR group, only 1 patient was found with a

Personalis EX2155

single circulating tumour cell (CTC) in the peripheral blood, while all other repeat bone marrow and blood tests were negative by immunocytochemistry and real-time qPCR. As a contrast, more than 2 abnormal bone marrow or blood tests were found in each of 15 out of the 29 patients in the HR group (Slade *et al.*, 2009).

The samples provided for my thesis have come from the 'Application of SNP arrays (S) to monitor the post-adjuvant (A) period (P) (SAP)' study, where blood samples were taken from the patients in the MRD clinical trial for DNA extraction and analysis by *Affymetrix SNP 6.0* microarray and real-time qPCR.

Samples from the pilot neoadjuvant chemo- or endocrine therapy (NeoCEnT) clinical trial were also used in this study. NeoCEnT is a multi-centre, randomised, parallel group, comparative, phase III, neoadjuvant study of chemotherapy versus endocrine therapy in postmenopausal patients who had strongly ER-positive, primary breast cancer with no evidence of overt metastases, and for whom pre-operative systemic therapy was indicated (chief investigator, Professor RC Coombes, Imperial College London). The study aimed to help clarify the optimal treatment - chemo- (fluorouracil/epirubicin/cyclophosphamide) or endocrine (letrozole) therapy - for this patient group in the neoadjuvant setting. The pilot study to assess the feasibility of patient recruitment and tissue collection (20 patients per treatment; 40 patients in total) has been completed. As part of my thesis, the aim was to determine whether or not changes in the single nucleotide polymorphism (SNP) and copy number (CN) microarray profiles correlated with resistance to therapy using the NeoCEnT samples (Chapter 4).

An additional 5 patients provided samples through the benign breast disease (BD) study. In the study, a blood sample was taken before and after surgery for each patient and the aim was to compare the blood samples of those with benign to invasive disease. The 5 patients whose samples were analysed in my thesis had invasive breast cancer. The healthy control samples used in my thesis also derived from this [BD] study.

Personalis EX2155

2.3 Sample information

All patient samples investigated in this thesis were taken from three different clinical studies: MRD; NeoCEnT; and BD. Samples from healthy individuals and women with familial breast cancer are also included. Due to the similarity in time of sample, the BD and NeoCEnT groups will be referred to as the 'pre-surgical' group, while the MRD group will be referred to as the 'on follow-up' group.

Information on each group and the samples obtained there-with are found in Table 2-1.

Group		Number	Description	Samples[1]
Healthy control		8	Healthy volunteers	▪ Lymphocyte ▪ Blood plasma
Familial breast cancer		2	Strong family history of breast cancer	▪ Lymphocyte ▪ Blood plasma
		5	Unaffected women with *BRCA1* mutations	▪ Lymphocyte ▪ Blood plasma
On follow-up		50	Breast cancer patients who have had surgery and are subsequently on follow-up, either with or without endocrine therapy	▪ Lymphocyte ▪ Tumour (n=40) ▪ First blood plasma ▪ Second blood plasma
Pre-surgical	NeoCEnT	10	Primary breast cancer patients who have had either chemo- or endocrine therapy prior to surgery, and who are strongly ER-positive	▪ Lymphocyte ▪ Tumour ▪ Baseline blood plasma ▪ 6 week blood plasma ▪ 21 week blood plasma (n=7)
	BD	5	Primary breast cancer patients	▪ Tumour DNA ▪ Pre-surgery blood plasma ▪ Post-surgery blood plasma

Table 2-1 - Overview of patient groups involved in this thesis, including information on numbers and types of samples available.

[1] For all groups, blood plasma DNA was obtained as cfDNA.

Personalis EX2155

In the pre-surgical group, the 10 NeoCEnT tumour samples were taken as biopsy cores at the beginning of treatment, 2-4 days, and 21 weeks before surgery, and stored as fresh frozen (FF) tissues. The blood plasma samples were taken at the beginning of treatment and then after 6 and 21 weeks. The 5 BD tumour samples were taken during the surgical removal of the primary tumour and stored as formalin-fixed, paraffin-embedded (FFPE) tissues, while these 5 patients' blood plasma samples were taken at first hospital visit (2 weeks prior to surgery) and also after surgery (2 weeks).

In the on follow-up group, patients had been operated-on for breast cancer, with tumour samples taken during surgery and stored as FFPE tissue. The paired blood plasma samples for each of these patients were as follows: first blood plasma (P1) was taken an average of 6 years after surgery (maximum 10 years); second blood plasma (P2) was taken an average of 9 years after surgery (maximum 12 years). Within the group, 18 patients had small, ≤20mm tumours, with no nodal involvement (T_1N_0 (T≤20mm)). The remainder either had nodal-involvement (n=30) (node-positive) or were node-negative while having a tumour size >20mm (n=2) (T>20mm). Eight patients (4 node-positive; 4 node-negative) had relapsed between 3 and 8 years after diagnosis.

Details of the on follow-up patients are shown in Table 2-2.

Personalis EX2155

Menopausal status	**Number (%)**
Pre-menopausal	28 (56)
Post-menopausal	19 (38)
Data not available	*3 (6)*
Type of surgery	
Wide Local Excision (WLE)	30 (60)
Mastectomy	19 (38)
Data not available	*1 (2)*
Histology	
Invasive Ductal Carcinoma (IDC)	34 (68)
Invasive Lobular Carcinoma (ILC)	7 (14)
Other/mixed invasive	6 (12)
Data not available	*3 (6)*
Lymph node status	
Positive	31 (62)
Negative	16 (32)
Data not available	*3 (6)*
Tumour size	
Equal to or greater than 20mm	24 (48)
Less than 20mm	21 (42)
Data not available	*5 (10)*
Tumour Grade	
I	5 (10)
II	18 (36)
III	23 (46)
Data not available	*4 (8)*
Hormone receptor status	
ER positive	37 (74)
ER negative	13 (26)
ER positive on endocrine therapy	30 (60)
PgR positive	24 (48)
PgR negative	23 (46)
PgR data not available	*3 (6)*
Growth factor status	
ERBB2 positive	14 (28)
ERBB2 negative	35 (70)
Data not available	*1 (2)*
Triple negative (ER, PR and ERBB2)	10 (20)
Total number of patients	**50 (100%)**

Table 2-2 - Clinicopathological details of breast cancer patients on follow-up (taken from Shaw *et al.*, 2012).

Personalis EX2155

2.4 Sample processing

Blood samples were taken at Charing Cross Hospital, London and then processed within 2 hours of collection by trained research nurses, as follows: blood plasma was separated by centrifugation at 2,000rpm for 10 minutes (x2) in ethylenediaminetetraacetic acid (EDTA) tubes; plasma and cell pellets were then separated and stored at -80°C. Plasma samples were then shipped to Leicester Royal Infirmary on dry-ice. On thawing, plasma was centrifuged at 13,500rpm in a smaller centrifuge for 5 minutes; and the supernatant was removed to a clean tube prior to DNA isolation. The only blood samples that did not follow this protocol were those from the familial breast cancer group, which were taken at Leicester Royal Infirmary, and therefore sample processing was shorter for these.

For the analysis of tumour samples, haematoxylin- and eosin-stained (H&E) tissue sections were reviewed by a consultant pathologist (Professor RA Walker) to confirm the presence of cancer cells. Foci of tumour cells were marked out and isolated by manual microdissection.

2.5 DNA extraction and whole genome amplification

DNA extraction and whole genome amplification (WGA) were conducted, with assistance (Dr. Karen Page), as follows: DNA was extracted from blood-cell pellets, 1ml blood plasma, and foci of tumour cells as previously described by Page *et al.* (2006); WGA was performed using the Illustra™ GenomiPhi™ V2 DNA Amplification Kit (GE Healthcare UK Ltd, Amersham Place, Little Chalfont, Buckinghamshire, HP7 9NA, United Kingdom, http://www.gelifesciences.com/) as described by Rook *et al.* (2004), with minor modifications.

The WGA process was conducted as follows: In triplicate (pooled), DNA (5µl) and sample buffer (20µl) were incubated at 95°C for 3 minutes and then transferred to ice for 5 minutes; reaction buffer (22.5µl) and φ29 (phi29) DNA polymerase (2.5µl) were added to each tube and then incubated at 30°C for 16 hours; the polymerase was then inactivated by incubation at 65°C for 10 minutes. DNA concentration was measured

Personalis EX2155

using a NanoDrop® ND-1000 Spectrophotometer (Thermo Fischer Scientific Inc., 81 Wyman Street, Waltham, MA 02454, USA, http://www.thermofisher.com/) and subsequently diluted to 100ng/µl with sterile, ultra-pure water.

2.6 Microarray processing

In total, 283 WGA DNA samples (blood plasma cfDNA, lymphocyte, and tumour) from 65 breast cancer patients, 8 healthy female controls, and 7 women with familial breast cancer (2 women with a strong family history and 5 unaffected *BRCA1* mutation carriers) were hybridised to the *Genome-Wide Human SNP Array 6.0* (*Affymetrix SNP 6.0*) using the *Affymetrix® Genome-Wide Human SNP Nsp/Sty 6.0* protocol (Affymetrix Inc., 2008a). This - and QC checks - were carried out as a service by Almac Diagnostics (Almac Group Ltd., Diagnostics Durham, 4238 Technology Drive, Durham, NC 27704, USA, http://www.almacgroup.com/). The process proceeded using the 96-well format.

In brief, WGA DNA was digested using a restriction enzyme (*Nsp*/*Sty*) using a thermal cycler. During the process, adaptor molecules (specific to the restriction enzyme) were ligated to the ends of the digested DNA fragments. The ligated molecules were then amplified using a standard PCR reaction involving a generic primer that amplifies fragments ranging in size from 250-1100bp. The products were then concentrated using a cleaning and purification process before being quantified by a spectrophotometric plate reader based on optical density, and volumes were thereafter adjusted accordingly for the fragmentation process. During fragmentation, fragments of less than 180bp in size were generated. These were subsequently labelled using terminal deoxynucleotidyl transferase before being hybridised to individual microarray plates (one for each sample). After hybridisation, the plates were washed, stained, and finally scanned. A summary of each step with an estimate for the amount of time each takes (to the nearest hour) follows (Affymetrix Inc., 2008a):

Personalis EX2155

Restriction enzyme digestion (~3 hours)

|

Ligation (~4 hours)

|

PCR (including concentration) (~9 hours)

|

Fragmentation (~2 hours)

|

Labelling (~5 hours)

|

Hybridisation (~19 hours)

|

Scanning

2.6.1 The scanning process

Post-hybridisation, the chips were scanned and an image produced. Images were then analysed to determine the fluorescent intensity of each probe. Such images can be of very high resolution for microarrays involving the analysis of many probes, such as the *Affymetrix SNP 6.0* (Quackenbush, 2002). Each point of light in an image corresponded to an individual probe and fluorescence value was calculated as a raw probe intensity value.

2.7 Bioinformatic data analysis

Analysis of raw data microarray CEL files was performed using Partek® Genomics Suite™ 6.5, copyright © 2010, build 6.10.1129 (Partek Inc., 12747 Olive Boulevard, Suite 205, St. Louis, MO 63141, USA, http://www.partek.com/), Genotyping Console™ (GTC) 4.0 (Affymetrix Inc., http://www.affymetrix.com/), and Nexus Copy Number™ 5.1 Discovery Edition (BioDiscovery Inc., 2221 Rosecrans Avenue, Suite 227 - El Segundo, CA 90245, USA, http://www.biodiscovery.com/). Initially, all analysis was performed using Partek® Genomics Suite™, but this was repeated using Genotyping

Personalis EX2155

Console™ and Nexus Copy Number™ 5.1 Discovery Edition to confirm reproducibility across software platforms.

What follows are the steps taken during normalisation, QC, genotyping and the generation of CN, and all other bioinformatic procedures carried-out.

2.7.1 Normalisation

Data normalisation comprises all of the operations carried out in order to arrive at a quantity of data that is of interest (Van de Wiel *et al.*, 2011) by removing artefactual data and diminishing effects from both biological and non-biological sources (LaFramboise, 2009), including:

- Unequal quantities of starting DNA/RNA
- DNA/RNA quality
- GC content
- Probe-labelling inefficiencies
- Fluorescence detection inefficiencies

(Quackenbush, 2002 and Van de Wiel *et al.*, 2011)

These were corrected-for by normalising all derived raw probe intensity values in relation to those of control probes (for whose own values are regarded as constant). In the *Affymetrix SNP 6.0*, such control probes are located on-chip and thus a certain degree of normalisation is carried-out as standard. Additional procedures were also carried-out: 1) probe intensity values were adjusted for fragment length, probe sequence, and GC content, as these can affect the kinetics of later PCR experiments aimed at validating the array (LaFramboise, 2009); and 2) values were adjusted through $\log_2$ transformation with a reference.

Personalis EX2155

2.7.1.1 Referencing and derivation of $\log_2$ values

The International HapMap (build 270 na30 r1 a5) (International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/) was used as the initial reference model file. The referencing process involves comparing each probe value in the test sample with that of its corresponding probe in the reference. Thus, for each probe in each sample, a ratio (I) is obtained after comparing the reference (R) to the test (T) sample:

$$I = \frac{T}{R}$$

In a biological context, however, this is inappropriate. A better and more common approach is to derive $\log_2$ values (Quackenbush, 2002). For example, a probe that has an intensity value twice that of the reference will be $\log_2 2 = 1$, whereas one with the equivalent ratio but in the reverse (i.e. half-intensity) will be $\log_2 0.5 = -1$, thus keeping all $\log_2$-derived values symmetrical. If only ratios were used, then any symmetry between higher and lower values would be removed as the equivalent values would be:

$$I = \frac{2}{1} = 2 \text{ and } I = \frac{0.5}{1} = 0.5$$

Thus, through $\log_2$ transformation and in comparison to the International HapMap, all normalised values were adjusted such that the overall mean value of each array was 1 (equivalent to $\log_2$ of 0)

$$\sum_{i=1}^{Total}\left(\frac{\log_2(T_i^{'})}{\log_2(R_i^{'})}\right)$$

Where:

Total=Total number of markers used in the array

T'=Adjusted test sample intensity

R'=Adjusted reference sample intensity

Personalis EX2155

Log_2 values were then plotted according to physical position to result in the formation of genomic profiles where positive and negative values retained symmetry.

2.7.1.2 Variations to normalisation

A variation to normalisation is quantile normalisation, which involves replacing the highest signal intensity value of each array in the project with the average of the highest values across all arrays, and repeating for the second and third highest, etc. (LaFramboise, 2009). In this way, outlier values are 'pulled' closer to the main group of values. However, this approach was not adopted due to the different tissue origins of the samples - i.e. lymphocyte, plasma cfDNA, and tumour (FFPE and FF). Thus, any outlier values in tumour that were an actual reflection of true biology, for example, could be misrepresented through quantile normalisation due to lower overall values in lymphocyte and plasma cfDNA.

Other variations to the normalisation process include lowess (locally-weighted linear regression), log-centering, and Chen's ratio statistics (Quackenbush, 2002), and it can also be global or local. Global and local normalisation are especially useful if an entire genome is being analysed, considering that probe density varies from region to region in the genome (in *Affymetrix SNP 6.0*, for example, the short arms of some chromosomes have no probes mapped to them). However, no algorithm/process can completely escape the problem of background noise and type I and II error (false-positive and -negative respectively) (Grayson & Aune, 2011).

Once data had been normalised, artefactual data that necessitated filtering could still persist (Van de Wiel *et al.*, 2011). Examples include analysing each probe against the surrounding (or 'background') signal[1] and using other statistical measures such as t-test and ANOVA. Such measures were employed during analysis.

[1] If its standard deviation was more than 1.96 times that of the background, then it would be regarded as a significant probe and therefore included in the end-analysis (Quackenbush, 2002).

Personalis EX2155

2.7.2 Initial QC

Initial QC was carried out using contrast QC (CQC) (Affymetrix Inc., 2008b) and SNP call-rate (generally, cut-offs of ≥0.4 and ≥97% were used for in-bounds, respectively). QC call-rate was also used as a general indicator of sample quality, but wasn't used as a definitive indicator of in- or out-of-bounds per sample. Gender-calling (also performed during later genotyping) was made using the cn-probe-chrXYratio_gender algorithm (Affymetrix Inc., 2009).

2.7.3 Genotyping and generation of CN segments

Genotyping analysis was performed using the Birdseed v2 algorithm (Broad Institute, Harvard-Massachusetts Institute of Technology, 7 Cambridge Center, Cambridge, MA 02142, USA, http://www.broadinstitute.org/mpg/birdsuite/index.html/), incorporating the expectation maximisation (EM) algorithm (Dempster *et al.*, 1977). Post genotyping, the median of the absolute values of all pairwise differences (MAPD) was used as an indication of sample quality by assessing noise levels. The *CN5* and *BRLMM-P+* algorithms were subsequently used for CN generation from probe intensity values, incorporating CN-GC waviness correction (Affymetrix Inc., 2009). Regional GC correction was also applied (Partek Inc., 2009).

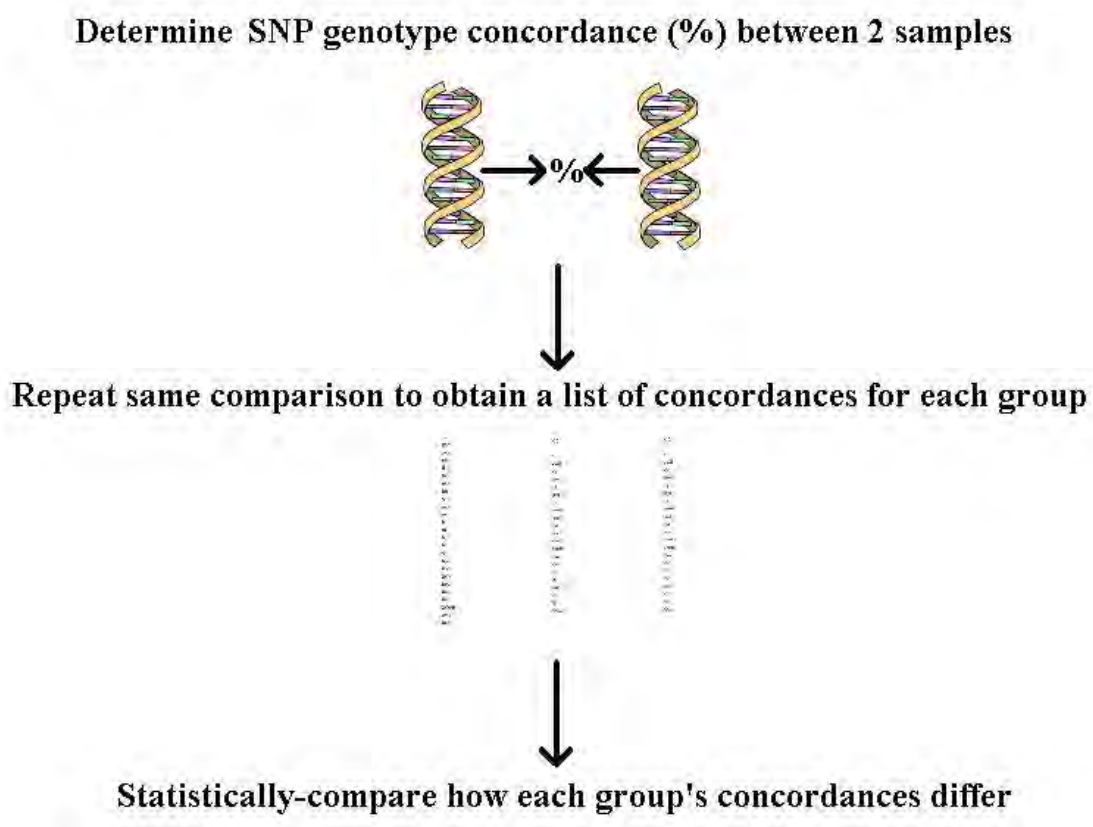
2.7.3.1 Single nucleotide polymorphism concordance checks

Subsequent SNP concordance[1] analysis performed using 15 healthy females from the CEPH (Utah residents with ancestry from northern and western Europe) population contained within the HapMap project. Additional comparisons were also performed using the 8 healthy female volunteers who had donated a lymphocyte and plasma sample for analysis. Where the case arose of one sample being compared to multiple others, the average concordance value was obtained. A flow-diagram of how these comparisons were performed follows:

[1] Based on the number of SNP markers across any two samples that have the same genotype call.

Personalis EX2155

Figure 2-1 – Flow-diagram of SNP genotype comparison to determine SNP concordance.



2.7.3.2 Genomic segmentation

Genomic segmentation (i.e. - the generation of CN segments) was then performed to identify amplified and deleted segments in the data using a minimum of 50 markers per segment, p-value cut-off of <0.0001, and a signal-to-noise ratio of 0.5 (low ratios result in many break-points being reported; higher results in fewer). Each patient/control's own lymphocyte sample was used as the reference during segmentation, with minimum segment sizes of 1, 50, 100, and 1,000Kb being used for viewing different-sized amplifications and deletions across different samples. Other studies have explored smaller segment sizes, such as 0.5Kb (Conrad *et al.*, 2009), however 1Kb was chosen as the minimum due to the definition of a CN variant by Scherer *et al.* (2007).

Personalis EX2155

2.7.4 Principal component analysis and clustering

Principal component analysis (PCA) was performed using Partek® Genomics Suite™ (Partek Inc., http://www.partek.com/) (see Appendix 8.4 for detailed information on how principal components were derived). In brief, principal components were determined using a covariance matrix method (used for mean-centred data, which in this case is CN 2), with normalised eigenvector scaling. Unless otherwise stated, an ANOVA $p<0.0001$ or Bonferroni-corrected $p<0.0001$ for multiple comparisons was used to filter-out markers of insignificance before determining the principal components and construction of the bi-plot. In addition, a fold-change of >4 or <-4 was applied to further filter data. False discovery rate (FDR) was not used as it was considered too lenient and unsuitable for the amount of comparisons involved.

The same filtering was used before the generation of dendrograms using the average or complete linkage method and incorporating a distance matrix / heatmap.

Samples used in PCA from the International HapMap included those from the following groups: CEPH (Utah, USA residents with ancestry from northern and western Europe); CHB (Han Chinese in Beijing, China); JPT (Japanese in Tokyo, Japan); and YRI (Yoruba in Ibadan, Nigeria).

2.7.4.1 Genetic signature of breast cancer

For generating the genetic signature of breast cancer, new PCA experiments were carried out. The principal components in each experiment were rotated using varimax rotation with Kaiser normalisation on a UNIX-based system (Hewlett Packard ProLiant DL580 G5, running 4 64-bit quad-core Intel® Xeon® X7350 CPUs and 32GB memory) using R (R Core Development Team, 2008).

To arrive at a suite of significant genes (the genetic signature), significant markers from PCA were first chosen by selecting those SNP and CN markers that had correlations to the first principal component of >0.7 or <-0.7 (other values were used in a select few cases). These markers were then overlapped with genes listed in the *Affymetrix SNP 6.0* libraries using a custom-designed script written in Visual Basic (Microsoft Corporation,

Personalis EX2155

Microsoft Campus, Thames Valley Park, Reading, Berkshire RG6 1WG, UK, http://www.microsoft.com/) (see Appendix 8.4.1.2.2 for further information). The list of genes was subsequently cross-compared to the KEGG (Kanehisa & Goto, 2000) and BioCarta pathways using DAVID (Huang *et al.*, 2009a & 2009b), and also to OMIM® (Online Mendelian Inheritance in Man, McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University, Baltimore, MD, http://omim.org/) using keywords related to 'cancer' and 'breast cancer' (PubMed Health, US National Library of Medicine, 8600 Rockville Pike, Bethesda, MD 20894, USA http://www.ncbi.nlm.nih.gov/pubmedhealth): "carcinoma", "malignant tumor/tumour", "breast cancer", "ductal carcinoma", "lobular carcinoma", "DCIS", "LCIS", "HER2/ERBB2-positive breast cancer", "ER-positive breast cancer", ductal carcinoma *in situ*", "lobular carcinoma *in situ*"[1].

For deriving the genetic signature, separate PCA experiments were conducted. Prior to the derivation of principal components, markers of insignificance were first filtered-out through ANOVA comparison to healthy control cfDNA. After PCA, a suitable number of rotated principal components were selected to account for a large proportion of the variation (>85% variation) exhibited in each experiment. Then, the correlation of original variables to the newly-rotated principal components was examined using the haplotype-tagging SNP (htSNP) method described by Meng *et al.* (2003) in order to create a haplotype of the data through SNP and CN markers (see Appendix 8.4.1.2.3 for information on this method).

2.7.5 Loss of heterozygosity

Loss of heterozygosity (LOH) using Partek® Genomics Suite™ (Partek Inc., http://www.partek.com/) was determined on a paired basis (matched to normal lymphocyte) using a hidden Markov model (HMM). The level of noise disturbance and 'emission density' (computed using the EM algorithm (Dempster *et al.*, 1977)) between each 2 points in the Markov chain were additionally factored-in. Markov models are used to measure the dependence of one marker's state on that of a nearby marker and

[1] Keywords "estrogen/oestrogen", "progesterone", and "growth factor" were also used.

Personalis EX2155

are employed because the CN between adjacent markers can be correlated. For example, if marker 'i' was in one state, then there would be a certain probability that adjacent marker 'j' was in the same state. However, the distance between the 2 markers should be factored into the probability (Stjernqvist *et al.*, 2011).

The following parameters were used: genomic decay of 1Mb; max probability of 0.98; genotype error of 0.02; and filtered using a Hardy-Weinberg equilibrium $p<0.001$ or <0.0001. Additionally, the model was configured to convert the sequence of $\log_2$ ratios into a chain of 6 possible states:

- CN 0
- CN 1
- CN 2 (i.e. - normal)
- Copy-neutral LOH (CN-LOH)
- CN 3
- CN 4

(Wang *et al.*, 2007)

2.7.6 Copy number variation frequency analysis

Frequency analysis for copy number variations (CNVs) was performed using Nexus Copy Number™ 5.1 Discovery Edition (BioDiscovery Inc., http://www.biodiscovery.com/). This software program was also used alongside the University of California Santa Cruz's (UCSC) table browser (Karolchik *et al.*, 2004) to find genes overlapping CN segments, both being based on the RefSeq gene database (Pruitt *et al.*, 2005).

2.7.7 Genomic aberration detection

For the detection of genomic aberrations, genomic segmentation was first performed, as previously described, using a minimum of 50 markers per segment, p-value cut-off of <0.0001, and a signal-to-noise ratio of 0.5. The $\log_2$ values were also Gaussian-

Personalis EX2155

smoothed and visualised using Partek® Genomics Suite™. Large amplifications and deletions were defined as those greater than or equal to 1Mb, while uncapped chromosomes/telomere attrition was defined as a deletion at a telomere of 5Kb or more.

2.8 Culturing cell-lines

Cell-lines and culture conditions used in this thesis are shown in Table 2-3. Both media types contained L-glutamine and 10% foetal calf serum, while at the same time lacked phenol red, as this can have a weak oestrogenic effect on breast cancer cells and thus interfere with growth (Levenson & Jordan, 1997). Cells were incubated at 37°C with 5% CO_2, with DNA being subsequently extracted using the phenol/chloroform/isoamylalcohol (P/C/IAA; ratio 25:24:1) protocol, briefly as follows: cell-pellets were suspended in 0.05M tris (pH8)/0.1% SDS and incubated with proteinase K (500µg/ml) for 1 hour at 37°C; an equal volume of P/C/IAA was added, followed by centrifugation at 13,000rpm for 3 minutes; the supernatant was removed, an equal volume of C/IAA was added to it, and the mixture was centrifuged at 13,000rpm for 3 minutes; the supernatant was again removed and had 3x volume -20°C ethanol followed by 1/10th volume 1M NaCl added to it, and was left at -20°C overnight; and finally, the mixture was centrifuged at 13,000rpm for 15 minutes at 4°C, the supernatant discarded to keep the DNA pellet, which was then washed with 70% ethanol, again centrifuged at 13,000rpm for 15 minutes, air-dried in a fume-hood, and finally suspended in sterile, ultra-pure water.

DNA concentrations and 260/280nm absorbance ratios were measured using the Nanodrop® ND-1000 Spectrophotometer (Thermo Fischer Scientific Inc., http://www.thermofisher.com/), as follows: 1µl ultra-pure water was pipetted onto the electrode to initialise the instrument, which was then 'blanked' to set the baseline; the electrode was then cleaned before adding and measuring the concentration and 260/280nm of 1µl sample (repeated thrice for each sample). Information on each cell-line is provided by Table 2-3.

Personalis EX2155

Cell-line	Culture medium	ERBB2 status[1]
MCF-7 (Soule *et al.*, 1973)	DMEM 1g/L glucose[2]	Negative
MDA-MB-231 (Cailleau *et al.*, 1974)	DMEM 1g/L glucose	Negative
MDA-MB-436 (Cailleau *et al.*, 1978)	RPMI 1640[3]	Negative
MDA-MB-468 (Cailleau *et al.*, 1978)	DMEM 1g/L glucose	Negative
SK-Br-3 (Fogh & Trempe, 1975)	RPMI 1640	Positive
T-47D (Keydar *et al.*, 1979)	DMEM 1g/L glucose	Negative

Table 2-3 - Origin, growth medium used, and ERBB2 status for each cultured breast cancer cell-line.

2.9 Design of primers and probes

Primers and TaqMan® MGB probes (Applied Biosystems™, Life Technologies Corporation, 5791 Van Allen Way, PO Box 6482, Carlsbad, CA 92008, http://www.appliedbiosystems.com/) for the validation of CNVs were custom-designed using Primer3 0.4.0 (Rozen & Skaletsky, 2000) and Primer Express® 3.0 (Applied Biosystems™). Target sequences were first screened for previous mutations (including amplifications, deletions, and LOH) using the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes *et al.*, 2008), and for SNPs and uniqueness in the human genome using NCBI SNP and Nucleotide BLAST (Altschul *et al.*, 1990), respectively. Once the primer and probe sequences were designed, the possible amplicons resulting from each primer pair were screened for uniqueness in the human genome using UCSC *in-silico* PCR (Fujita *et al.*, 2011). The robustness of each assay was tested through real-time qPCR on human genomic (HG) DNA (Roche Diagnostics Ltd., Charles Avenue, Burgess Hill, West Sussex, RH15 9RY, UK, https://www.roche-applied-

[1] Taken from American Type Culture Collection (ATCC), PO box 1549, Manassas, VA 20108, USA, http://www.atcc.org/.

[2] PAA Laboratories GmbH, PAA-Strasse 1, 4061 Pasching, Austria, http://www.paa.com/.

[3] BioWhittaker®, Lonza Group Ltd., Muenchensteinerstrasse 38, CH-4002 Basel, Switzerland, http://www.lonza.com/.

Personalis EX2155

science.com/). Amplicons were measured relative to a custom-designed individual reference, or to the mean of more than one reference, by relative quantitation (RQ).

Table 2-4 provides a list of the primer-probe sets that were designed, including both target and reference assays.

	Target sequences	
Target interval	**Primers**	**Probe**
4q13.2	5`-GTGGGATCTAATTAAACTAACTTTTGCA-3` 5`-TCCCACTCTGTGGGTTGTCTG-3`	5`-CAGGCAGCAGAGTAA-3`
16p12.3	5`-CCTTTCTGTCTTAACGGGAGCAT-3` 5`-CCTGTATTACGGGGTGCTCAAT-3`	5`-CGGAGAGAATGATGCTGTGA-3`

	Reference sequences	
Target designed for	**Primers**	**Probe**
4q13.2	5`-CAGGATGTGGTCTAGCTTGATGA-3` 5`-CACTCCATCCAACAACAGCAAA-3`	5`-CATGTGAGCTTGAGAAGA-3`
16p12.3	5`-TGCTGACTGGGGACATCCA-3` 5`-CCTCGTGATCCACCCGTAAA-3`	5`-AGGGAGAGATTCCTGGGA-3`

Table 2-4 - Sequences of custom-designed primer-probe sets used for microarray validation.

2.10 Real-time quantitative PCR

Real-time qPCR reactions were run on the StepOnePlus™ Real-Time PCR System (Applied Biosystems™, http://www.appliedbiosystems.com/). A 10µl reaction volume was used in each well, consisting of 5µl TaqMan® Fast Universal PCR Master Mix x2 (Applied Biosystems™), 1.4µl primer-probe at 200pM, and 10ng of sample DNA in a volume of 3.6µl. Additionally - in each experiment - HG DNA and ultra-pure water samples were also included to serve as negative and 'no template' controls, respectively. All samples were run in triplicate. Data analysis was subsequently conducted with StepOne Software v2.1 (Applied Biosystems™, http://www.appliedbiosystems.com/) and Microsoft® Excel (Microsoft Corporation, http://www.microsoft.com/) using the $\Delta\Delta C_T$ method: $\Delta\Delta C_T$ = [cell-line ΔC_T - HG DNA ΔC_T]; ΔC_T = [target mean C_T - endogenous control(s) mean C_T].

Personalis EX2155

For measuring the reproducibility and sensitivity of the *ERBB2* qPCR assay in breast cancer cell-lines, the endogenous controls used were contactin associated protein 1 (*CNTNAP1*) and ribonuclease P RNA component H1 (*RPPH1*), both supplied by Applied Biosystems™ (Applied Biosystems™, http://www.appliedbiosystems.com/).

2.11 Statistical analysis

For statistical analyses, GraphPad Prism 5 for Windows (GraphPad Software Inc., 2236 Avenida de la Playa, La Jolla, CA 92037, USA, http://www.graphpad.com/) and PASW Statistics 18.0.2 (SPSS Inc., Floor 11, 233 South Wacker Drive, Chicago, IL 60606-6412, USA, http://www.spss.com/) were used. For all statistical analyses, the alpha (α) value was varied between 0.0001 and 0.05.

Personalis EX2155



3 MOLECULAR CHARACTERISATION OF CIRCULATING CELL-FREE DNA IN BREAST CANCER PATIENTS USING SNP 6.0 MICROARRAYS

Personalis EX2155

3.1 Overview

In this chapter, the utility of circulating cell-free DNA (cfDNA) analysis for patient monitoring in the post-adjuvant period in breast cancer is the main focus. Each patient underwent surgery and subsequently gave serial blood samples whilst on follow-up. For the purpose of this chapter, only 2 blood samples were analysed for each patient, taken an average of 6 and 9 years after surgery, and the genomic profiles of these were related back to matched tumour DNA taken during resection to elucidate findings.

3.2 Introduction

In general, research in this thesis is focussed on the characterisation of cfDNA in breast cancer patients. Previous plasma-based analyses showed the promise of using cfDNA in diagnostics, prognostics, and follow-up studies in cancer (Anker *et al.*, 2001; Shaw *et al.*, 2000; and Pinzani *et al.*, 2011), whilst the measurement of levels of cfDNA were also suggested for breast cancer diagnosis (Huang *et al.*, 2006). However, elevated levels are sometimes seen in benign disease (Zanetti-Dällenbach *et al.*, 2008). Hence, our group elected to focus on the detection of tumour-specific alterations - including single nucleotide polymorphisms (SNPs), amplifications (initially focussing on *ERBB2*), deletions, and loss of heterozygosity (LOH) - by real-time qPCR approaches.

In order to achieve this, the standardisation of PCR-based assays is required in the clinical setting (Schwarzenbach *et al.*, 2011) and the reproducibility and sensitivity of any new assay must be demonstrated before being applied (Alix-Panabières *et al.*, 2012). For *ERBB2*, we first demonstrated tumour-specific amplification in the cfDNA of breast cancer patients (Page *et al.*, 2011) and suggested that there is benefit to using gene-based analyses in the future. Patients are more likely to respond to trastuzumab treatment for ERBB2-positive breast cancer if they also have *ERBB2* amplification (Vogel *et al.*, 2002). Thus, focussing on methods that detect amplification of the gene will help in tailoring therapy to patients with ERBB2-positive breast cancer. However, to molecularly characterise the entire circulating breast cancer genome should elucidate many more targets for analysis.

Personalis EX2155

3.2.1 Genome-wide characterisation

Copy number variations (CNVs) are amplified or deleted regions of the genome that are variable in size (Scherer *et al.*, 2007 and Grayson & Aune, 2011) and are a principle source of natural variation in the human genome (Redon *et al.*, 2006 and LaFramboise, 2009). Specific CNVs may be characteristic of different tumour types, whilst LOH is also common in many tumours and can reveal recessive alleles (Wang *et al.* 2004). The *Genome-Wide Human SNP Array 6.0* (*Affymetrix SNP 6.0*) (Affymetrix Inc., 2008a) analyses ~1.8 million markers across the genome for SNPs and CNV and is one of the leading microarray technologies at representing genetic variation. Through the analysis of *Affymetrix SNP 6.0* data, SNP genotypes, CNV, and LOH can be generated per sample in a single hybridisation experiment.

3.2.2 Sequential analysis

Due to the inherent problems in obtaining sequential samples in the adjuvant setting, little is known about the nature of dynamic changes in the cancer genome over time. However, sequential analyses are paramount for the following-up of cancer patients (Alix-Panabières *et al.*, 2012). In breast cancer, gene expression analysis revealed that multiple genetic changes from the primary tumour can occur in micrometastases in the bone marrow (Gangnus *et al.*, 2004). It was therefore hypothesised in this thesis that in patients on follow-up who are otherwise disease free, evidence of tumour DNA detected in cfDNA would suggest the presence of micrometastases, possibly deriving from - or related to - disseminated tumour cells (DTCs) in the bone marrow. Hence, it would be hugely advantageous to be able to detect these specific changes indicative of progression in cfDNA.

The cfDNA samples investigated as part of this thesis permitted such sequential analysis through the characterisation of the 'evolving' circulating breast cancer genome, and allow for the search for markers of minimal disease. Normal lymphocyte DNA samples were used as reference to help discriminate benign and pathogenic SNPs and CNVs (Conrad *et al.*, 2009); tumour DNA samples helped to define the 'mutation state'

Personalis EX2155

of the primary tumour (Schwarzenbach *et al.*, 2011); and, finally, cfDNA samples provided a tool for analysis of circulating, tumour-specific markers.

3.2.3 Aims

The aims of this chapter were to:

1. Demonstrate that amplified *ERBB2* is detectable in the cfDNA of ERBB2+++ breast cancer patients.
2. Compare *Affymetrix SNP 6.0* whole genome profiles of the primary tumour with paired plasma cfDNA samples taken from breast cancer patients on follow-up, and relate findings to cfDNA profiles of samples taken pre-surgically from breast cancer patients and from those of healthy female controls.

In order to achieve these aims, the specific objectives were to:

1. Test the reproducibility and sensitivity of the previously designed *ERBB2* qPCR assay (3.3.1).
2. Determine post-hybridisation sample quality through QC checks (3.3.2).
3. Genotype samples and compare SNP profiles of samples belonging to different tissue-types and groups (3.3.3).
4. Characterise CNV across all samples and distinguish between those that are benign and pathogenic (3.3.4).
5. Use principal component analysis (PCA) and hierarchical clustering to separate samples into logical groupings (3.3.5).

Personalis EX2155

3.3 Results

3.3.1 Reproducibility and sensitivity of qPCR for measuring *ERBB2* copy number in breast cancer cell-lines

The *ERBB2* copy number (CN) assay was applied to DNA isolated from 6 breast cancer cell-lines with known ERBB2 status: MCF-7 (Soule *et al.*, 1973); MDA-MB-231, -436, and-468 (Cailleau *et al.*, 1974 and 1978); SK-Br-3 (Fogh & Trempe, 1975); and T-47D (Keydar *et al.*, 1979).

Mean real-time qPCR cycle threshold (C$_T$) values obtained using 3, 5, and 9 replicates for *ERBB2* (target), and *CNTNAP1* and *RPPH1* as reference genes, in the 6 cell-lines are shown in Table 3-1.

Personalis EX2155

Cell-line	Replicates			Mean ($\overline{X}$)	Standard deviation (σ)
	3	5	9		
MCF-7	28.77 (28.33) [26.74]	28.89 (28.24) [26.69]	28.89 (28.15) [26.72]	28.85 (28.24) [26.72]	0.07 (0.09) [0.03]
MDA-MB-231	30.97 (30.26) [29.72]	31.04 (30.30) [29.71]	31.03 (30.28) [29.79]	31.01 (30.28) [29.74]	0.04 (0.02) [0.04]
MDA-MB-436	29.50 (28.81) [27.63]	29.61 (28.80) [27.36]	29.69 (28.85) [27.32]	29.60 (28.82) [27.44]	0.10 (0.03) [0.17]
MDA-MB-468	27.31 (27.64) [24.96]	27.64 (27.41) [25.00]	27.90 (27.28) [25.07]	27.62 (27.44) [25.01]	0.30 (0.18) [0.06]
SK-Br-3	26.85 (29.98) [28.14]	26.87 (30.01) [28.38]	26.92 (30.09) [28.49]	26.88 (30.03) [28.34]	0.04 (0.06) [0.18]
T-47D	28.42 (27.49) [28.48]	28.48 (27.41) [28.47]	28.41 (27.30) [28.50]	28.44 (27.40) [28.48]	0.04 (0.10) [0.02]

Table 3-1 - Mean *ERBB2* CT values per cell-line, per number of replicates. Corresponding values for *CNTNAP1* and *RPPH1* are shown in parentheses and square-brackets respectively. The overall mean and standard deviation for each are also shown.

Relative quantitation (RQ) values for *ERBB2* relative to each endogenous control (and the mean of these) was then calculated for each cell-line for 3, 5, and 9 replicates (Table 3-2, Figure 3-1, and Figure 3-2).

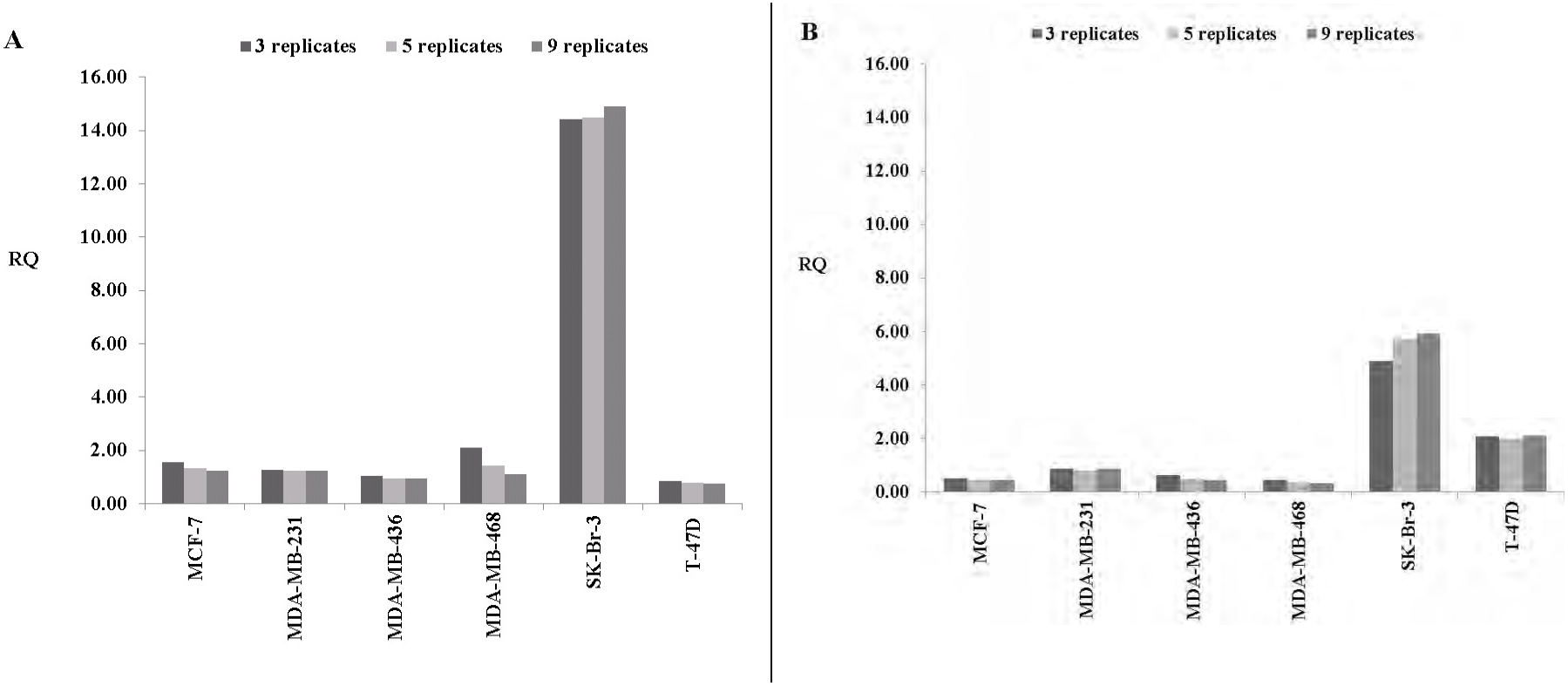
Personalis EX2155

	CNTNAP1			***RPPH1***			***CNTNAP1* and *RPPH1***		
	3	**5**	**9**	**3**	**5**	**9**	**3**	**5**	**9**
MCF-7	1.55	1.34	1.25	0.51	0.45	0.46	0.88	0.78	0.76
MDA-MB-231	1.28	1.25	1.25	0.87	0.82	0.88	1.06	1.01	1.05
MDA-MB-436	1.05	0.96	0.94	0.64	0.49	0.45	0.82	0.69	0.65
MDA-MB-468	2.12	1.43	1.10	0.46	0.37	0.33	0.98	0.73	0.60
SK-Br-3	14.41	14.47	14.89	4.89	5.70	5.93	8.40	9.08	9.40
T-47D	0.86	0.79	0.76	2.09	2.00	2.13	1.34	1.25	1.28

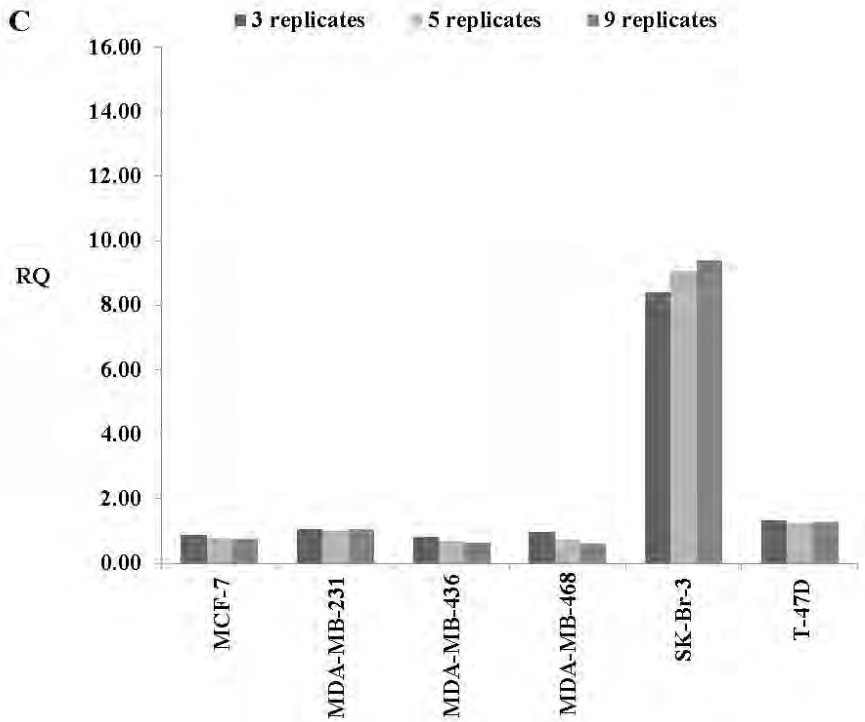
Table 3-2 - *ERBB2* RQ values per cell-line, per number of replicates.

Personalis EX2155

Figure 3-1 - *ERBB2* RQ values for 6 breast cancer cell-lines using different endogenous control combinations (grouped according to cell-line): *CNTNAP1* (A); *RPPH1* (B); and *CNTNAP1* with *RPPH1* (C).

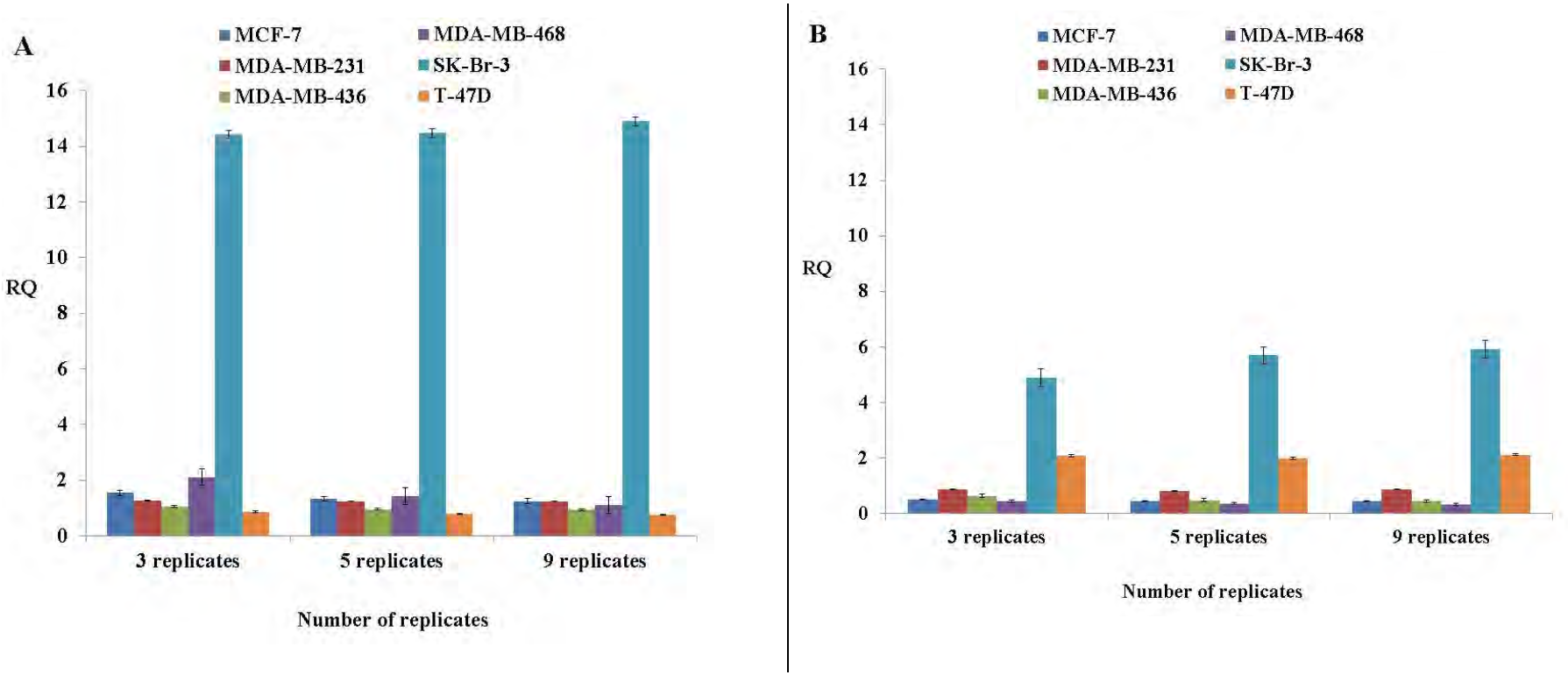


Personalis EX2155

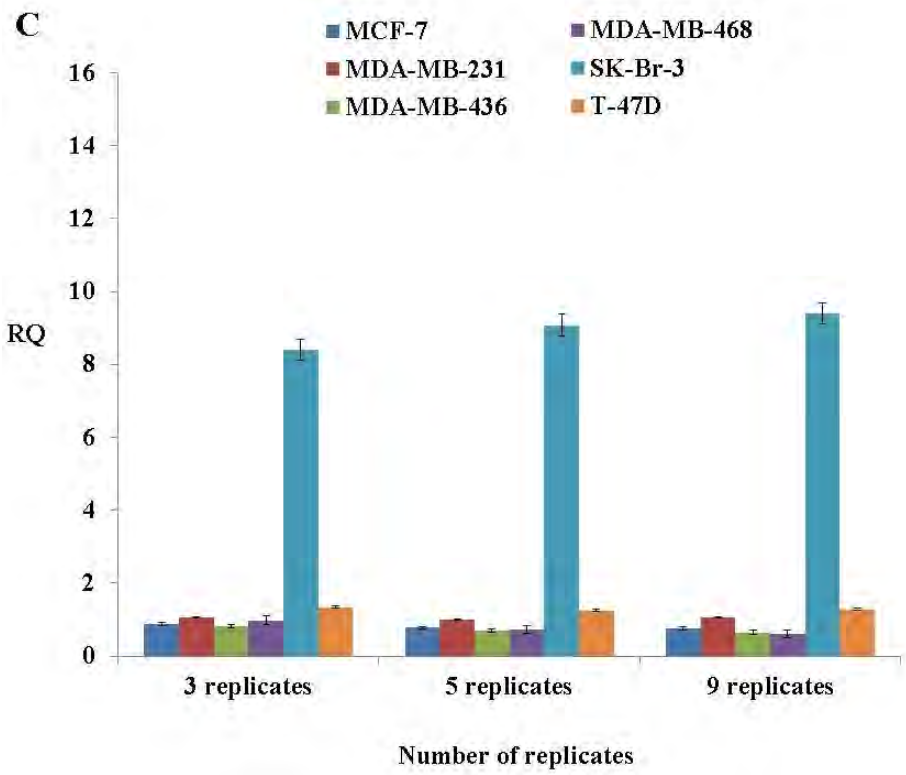


Personalis EX2155

Figure 3-2 - *ERBB2* RQ values for 6 breast cancer cell-lines using different endogenous control combinations (grouped according to number of replicates): *CNTNAP1* (A); *RPPH1* (B); and *CNTNAP1* with *RPPH1* (C).



Personalis EX2155



Personalis EX2155

Amplification was observed in the ERBB2-positive SK-Br-3 cell-line and was highest using *CNTNAP1* as an unamplified reference, and lowest when using *RPPH1*. The ERBB2-negative cell-line, T-47D, also exhibited low-level amplification using *RPPH1* as the control.

A 1-way repeated measures ANOVA revealed that there was no statistically-significant difference between using 3, 5, or 9 replicates with either reference locus or the mean of these ($p<0.0001$). Table 3-3 shows corresponding *Affymetrix SNP 6.0* derived CN values for each of the 6 cell-lines taken from the Cancer Genome Project (Wellcome Trust Sanger Institute, Genome Campus, Hinxton, Cambridge, CB10 1SA, UK, http://www.sanger.ac.uk/). These data highlight how the use of the two reference genes will affect RQ results in MCF-7 and T-47D cells, as demonstrated by qPCR (Figure 3-1 and Figure 3-2).

Gene	Cell-line					
	MCF-7	**MDA-MB-231**	**MDA-MB-436**	**MDA-MB-468**	**SK-Br-3**	**T-47D**
CNTNAP1	2	3	-	2	-	4
RPPH1	3	3	-	2	-	2
ERBB2	2	3	-	2	-	4

Table 3-3 - Cancer Genome Project (Wellcome Trust Sanger Institute) *Affymetrix SNP 6.0* CN values for *CNTNAP1*, *RPPH1*, and *ERBB2* for each breast cell-line examined (data for MDA-MB-436 and SK-Br-3 not available).

3.3.2 *Affymetrix SNP 6.0* QC analysis of whole genome amplification cfDNA, tumour, and lymphocyte DNA samples

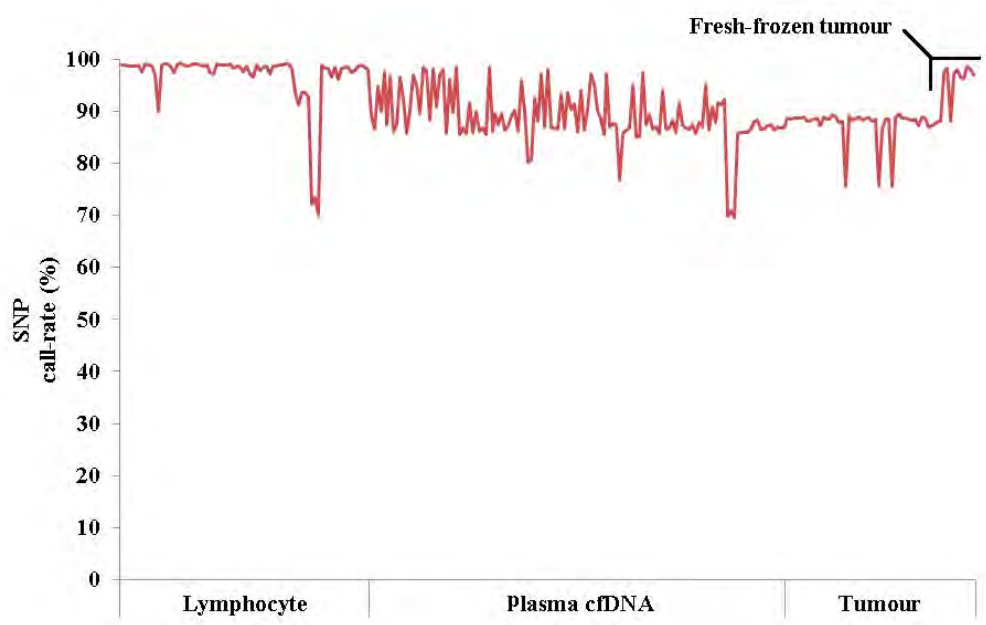
In total, 273 whole genome amplification (WGA) DNA samples (plasma, lymphocyte, and tumour) from 65 breast cancer patients and 8 healthy female controls were analysed using the *Affymetrix SNP 6.0* microarray technology. The WGA DNA samples were hybridised in two batches in order to reduce inter-assay variability, and subsequent

Personalis EX2155

analysis was performed independently using Partek® Genomics Suite™, Genotyping Console™, and Nexus Copy Number™, with each platform identifying the same CNVs (Shaw *et al.*, 2012).

High SNP call-rates were achieved for the majority of samples (Figure 3-3), highlighting the efficiency of the array technology. Of note, lymphocyte DNA across all patients and controls had the highest overall mean SNP call-rate (96.89%, range 70.14-99.32%), compared to the other samples: in the on follow-up group, first plasma cfDNA, second plasma cfDNA, and tumour DNA had means of 91.37% (range 85.04-98.51%), 88.62% (range 85.21-98.33%), and 88.52% (range 86.65-89.62%), respectively. The means for pre-surgical breast cancer patients' plasma cfDNAs (86.80%; range 85.85-88.36%) and those of the healthy controls (83.06%, range 69.62-92.33%) were comparable. Finally, the pre-surgical patients' tumour DNA samples exhibited a mean of 93.61% (range 87.06-98.70%). These high values for SNP call-rate, similar to those seen for lymphocyte DNA, reflect the fact that 10 of these tumour samples were from the fresh frozen (FF) tissue of patients in the pre-surgical group (the remainder being formalin-fixed, paraffin-embedded (FFPE)).

Figure 3-3 - SNP call-rates across 273 DNA samples used in the study.



Personalis EX2155

On close inspection, there was a striking difference in the QC data for the 10 FF and 5 FFPE tumour DNA samples of the pre-surgical breast cancer patients (Table 3-4): the FFPE tissues had a mean QC call-rate of 57.43%, while that of the FF tissues was 89.75%. The SNP call-rates were closer (FFPE mean of 87.63%; FF mean of 96.6%) but there was still a noticeable difference between these.

Tissue type	QC call-rate (%) (Affymetrix Inc., 2009)	SNP call-rate (%)
FFPE	57.81	87.06
	58.41	87.35
	58.57	87.68
	55.72	88.00
	56.62	88.04
FF	91.30	97.61
	95.70	98.32
	54.80	88.00
	93.58	97.32
	96.10	97.99
	93.18	96.60
	91.89	96.48
	97.39	98.70
	94.08	98.02
	89.44	96.96

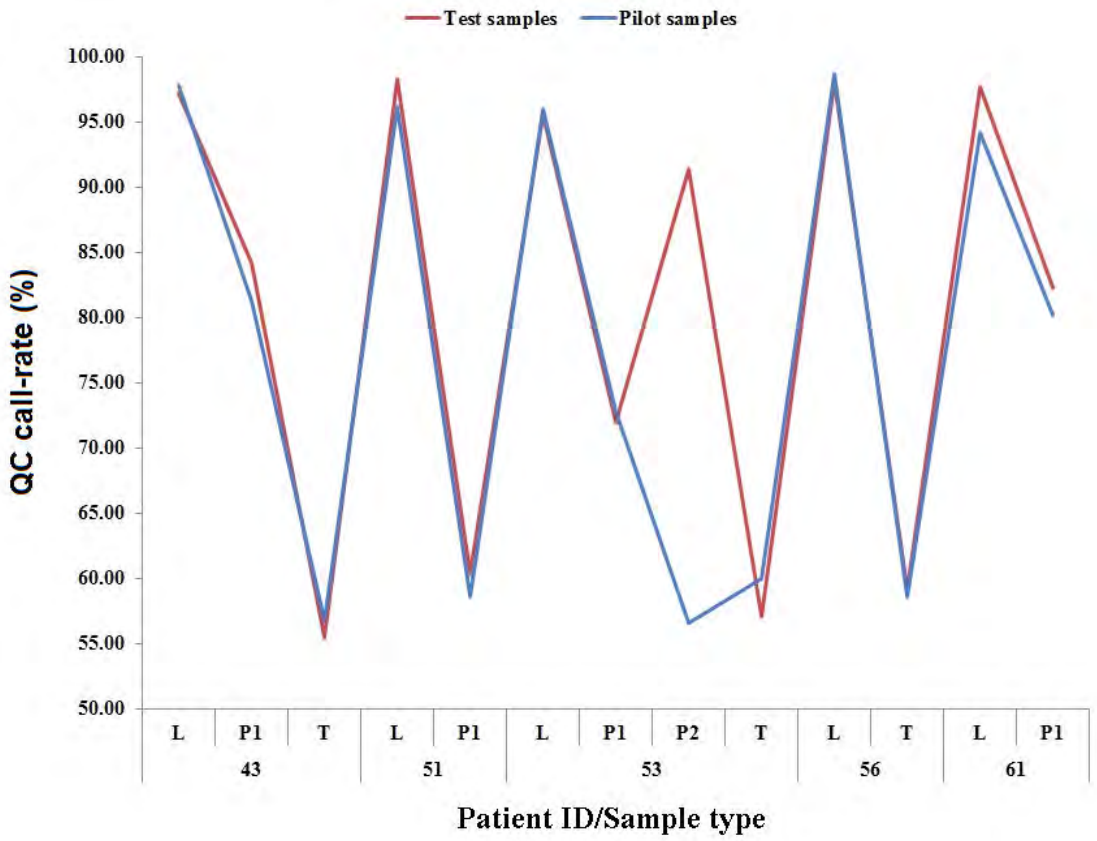
Table 3-4 - Differences in quality between FFPE and FF tumour DNA tissue were apparent when looking at the QC and SNP call-rates for each.

As all samples in the study were processed in two batches, it was possible to determine intra-assay reproducibility. To this end, 13 samples were repeated for the entire process - from DNA extraction to array hybridisation - across the 2 batches. These showed excellent correlation between CQC ($p<0.0001$), QC call-rate ($p=0.0001$) (Figure 3-4),

Personalis EX2155

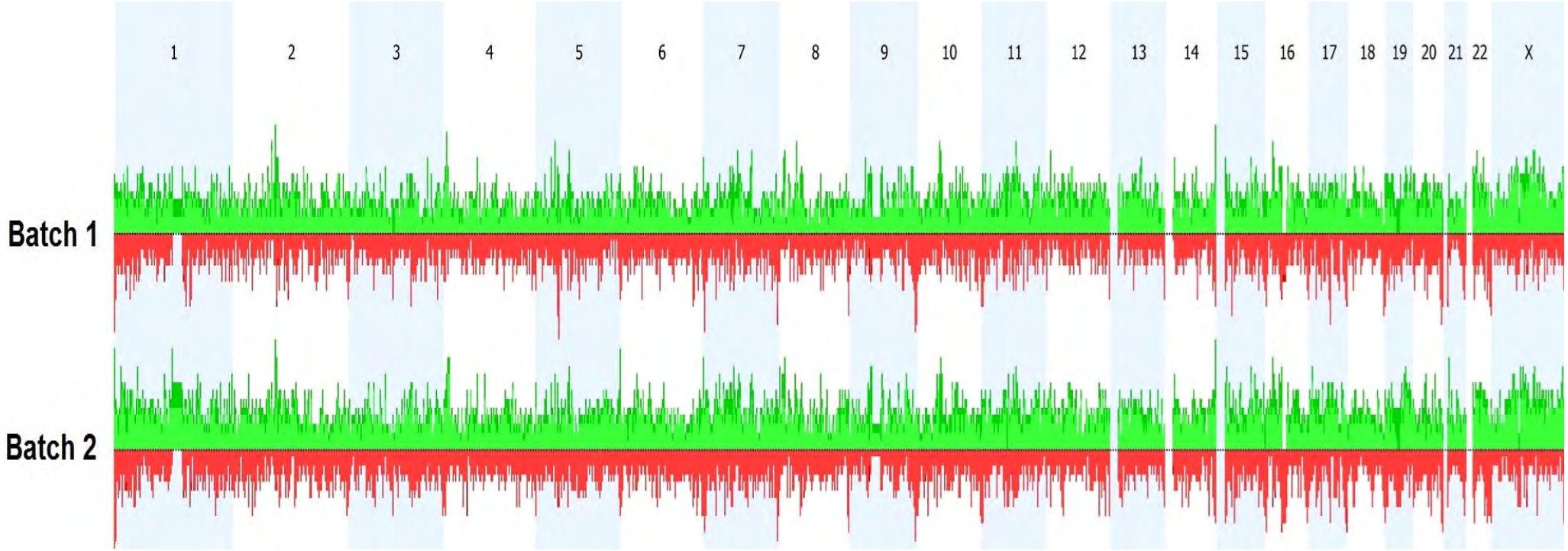
MAPD ($p=0.0005$), and in the number of CNVs derived ($p<0.05$), using paired t-tests. The results also exhibited good correlation (average correlation 0.727, Spearman's rank) when comparing unfiltered probe intensity values between replicate samples. The frequency of each specific CNV detected between the two batches also showed good agreement (Figure 3-5). One sample (53, P2) exhibited significant variability between the replicate extractions and hybridisations (Figure 3-4).

Figure 3-4 - Close correlation in QC call-rate between repeat-sample extraction and hybridisation. L, lymphocyte; P1, first plasma cfDNA; P2, second plasma cfDNA; T, tumour.



Personalis EX2155

Figure 3-5 - Genome-view representation of the frequency of amplification and deletion in the 13 samples hybridised in 2 batches. The y-axis indicates the percentage of samples having a CNV at a specific point in the genome. Chromosome number at top; amplifications in green; deletions in red.



Personalis EX2155

3.3.3 Single nucleotide polymorphisms and the comparison of genomic profiles

3.3.3.1 HapMap comparisons

Various comparisons were carried-out between patients and samples within - and between - groups based on the concordance of SNP genotype calls[1]. When the SNP genotype calls of each lymphocyte DNA sample from 8 healthy controls was compared to 15 randomly selected healthy females from the CEPH HapMap (International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/samples), results were consistent: a mean concordance of 64.23% (range 62.75-66.13%) was obtained, which highlights the validity of the normal controls. Similar results were obtained for cfDNA from the plasma of the healthy controls (mean 63.50%, range 60.87-65.38%) and the normal lymphocyte samples from 15 breast cancer patients (mean 61.08%, range 55.00-62.22%).

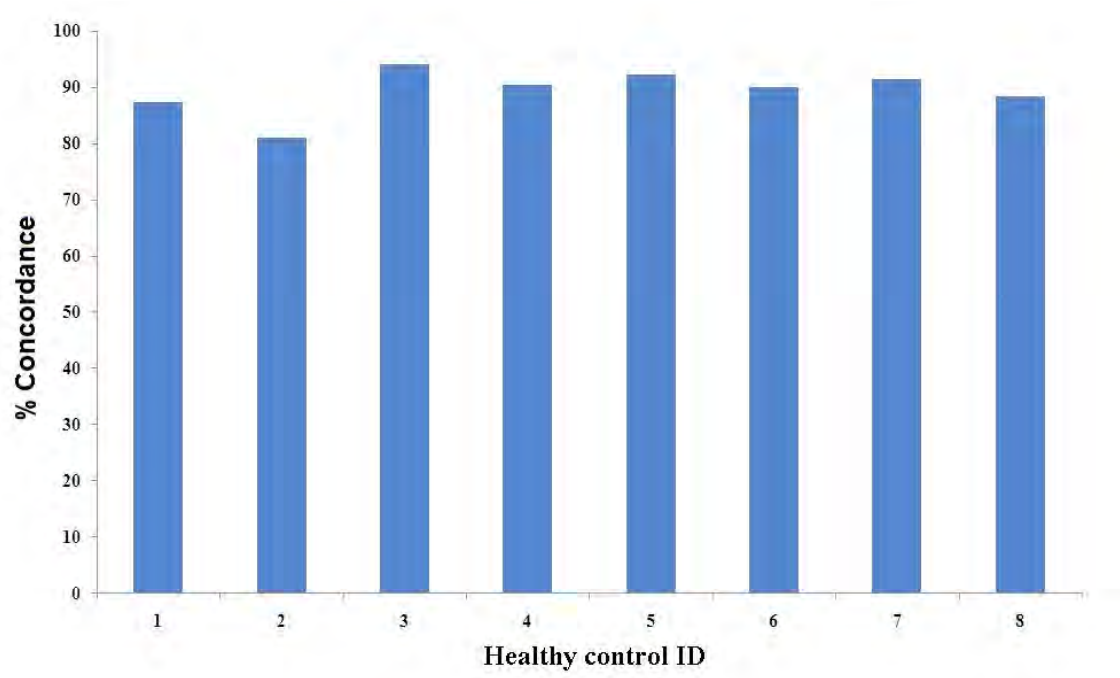
3.3.3.2 Matched lymphocyte comparisons

Next, the concordance of SNP genotype calls between each cfDNA and matched lymphocyte DNA sample for each cancer patient and healthy control were compared. The highest mean concordance existed in the healthy control group (Figure 3-6), exhibiting a mean concordance of 89.35% (range 81.10-94.08%) and suggesting that the 2 genomes were closely related, as would be expected if the cfDNA was derived from apoptosis of normal cells. In addition, the level of concordance was not significantly correlated to sample quality (using the SNP call-rate as indicator of quality) (0.0007, P1; -0.173, P2 (Spearman's rank)).

[1] Based on the number of SNP markers across any two samples that have the same genotype call.

Personalis EX2155

Figure 3-6 - High % SNP concordance between plasma cfDNA and matched lymphocyte DNA in healthy controls (mean 89.35%, range 81.10-94.08%).



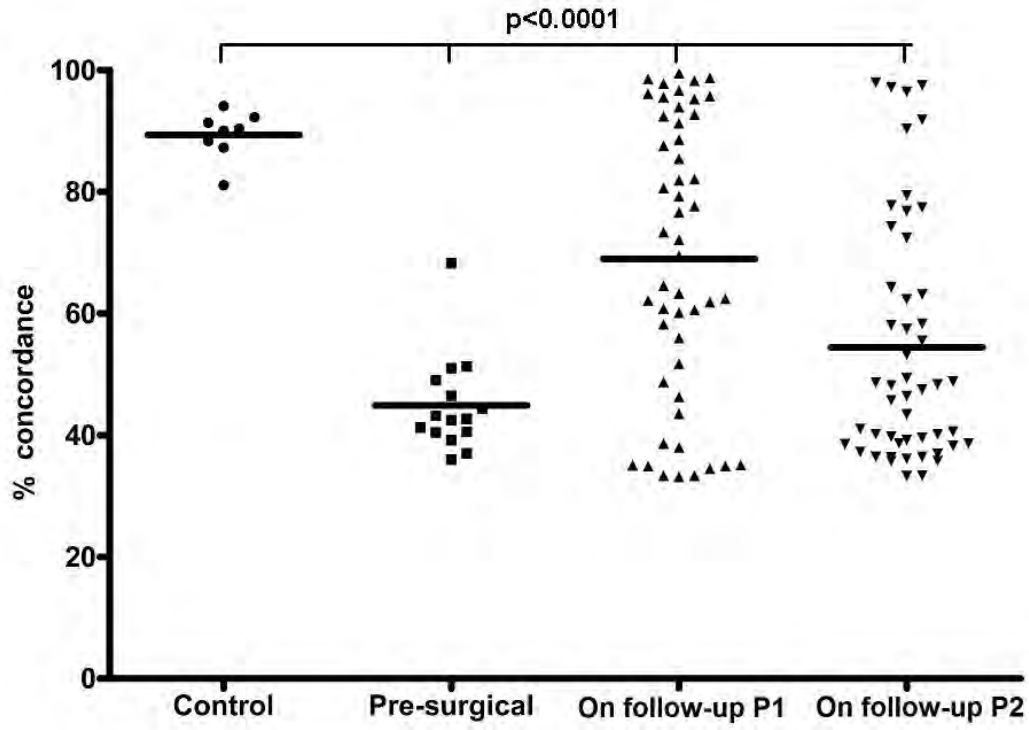
By group, the SNP concordance was lowest for the 15 pre-surgical primary breast cancer patients (mean 44.88%, range 36.00-68.27%), but also remained low for the 50 patients on follow-up in both the first (P1, mean 69.10%, range 33.17-99.44%) and second (P2, mean 54.22%, range 33.31-97.96%) cfDNA samples. A point to note is that 28% of the first - and 12% of the second - plasma samples from the on follow-up group exhibited a concordance at or above the mean value obtained for the healthy control group (89.35%). This potentially indicates a 'normal' genomic plasma cfDNA profile for these patients. Of the first plasma cfDNA samples, 6 were T_1N_0 ($T \leq 20mm$) and 8 were T>20mm/node-positive at diagnosis. Of the second plasma cfDNA samples, 3 were T_1N_0 ($T \leq 20mm$) and 3 were T>20mm/node-positive at diagnosis.

The mean SNP concordance by group was also compared, which revealed a significant difference between the healthy controls and the pre-surgical patients (p=0.0001, Mann-Whitney unpaired t-test). In the on follow-up group, the mean SNP concordance of the first cfDNA samples (P1) was not significantly different to that of the healthy controls (p=0.0612), but those of the second cfDNA (P2) were (p=0.0004) (Figure 3-7). When

Personalis EX2155

all groups were compared to each other by Kruskal-Wallis ANOVA, there was also a significant difference ($p<0.0001$). This demonstrates that there were differences in these patients' SNP calls of plasma cfDNA samples compared to their matched lymphocyte DNA samples.

Figure 3-7 - Scatter plot showing % concordance between plasma cfDNA and matched lymphocyte DNA for each group reveals a significant difference between each. P1, first plasma cfDNA; P2, second plasma cfDNA.



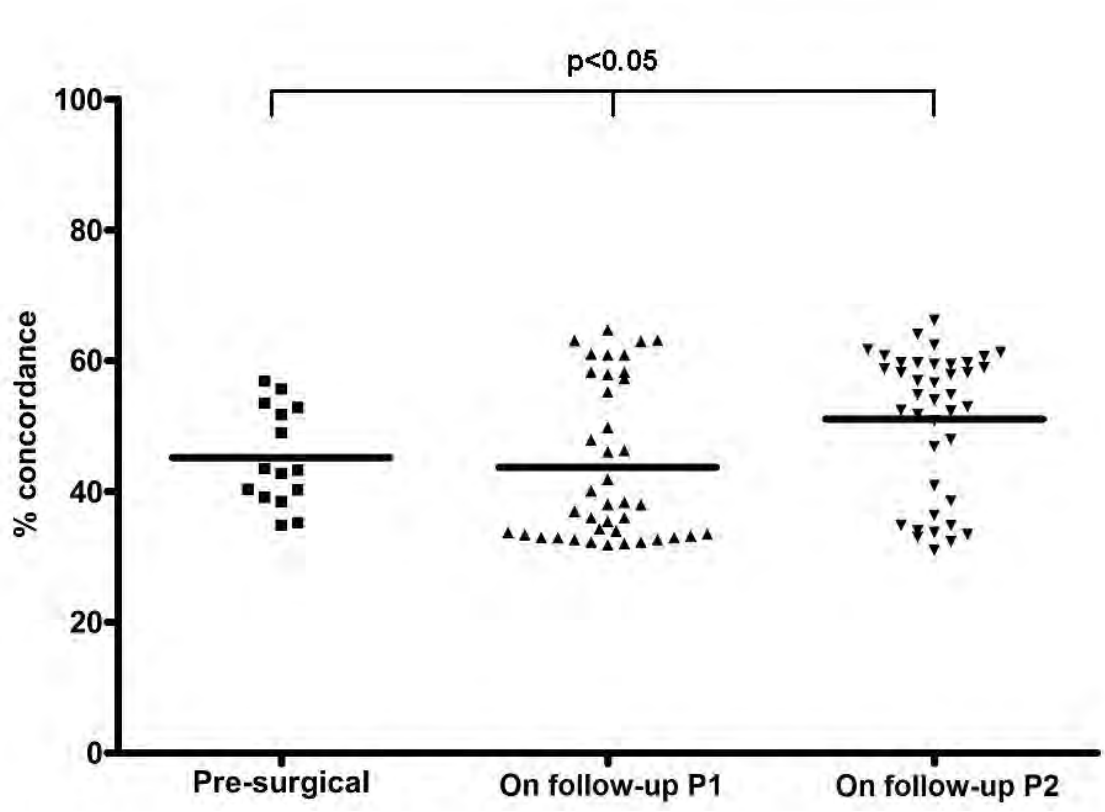
The level of SNP concordance was also compared between each patient's cfDNA sample and each of the healthy controls (n=8) (i.e. - each cancer patient's cfDNA sample compared to each of the 8 healthy control cfDNA samples and an average obtained). This revealed a significant difference between the first and second cfDNA samples of the patients on follow-up ($p=0.0006$, paired Wilcoxon signed-rank t-test), and between the cfDNAs of the patients on follow-up and pre-surgical patients ($p=0.0016$, Kruskal-Wallis ANOVA).

Personalis EX2155

3.3.3.3 Matched tumour comparisons

The SNP genotype calls were also compared between each cfDNA sample and matched tumour, where available (available for 15 pre-surgical and 40 patients on follow-up), which revealed low concordance (overall mean 46.89%, range 31.04-66.20%). This highlights genomic difference between cfDNA and matched tumour DNA. However, comparing these concordances again revealed a significant difference between the pre-surgical and on follow-up groups ($p<0.05$, Kruskal-Wallis ANOVA) (Figure 3-8).

Figure 3-8 - Scatter plot showing % concordance between plasma cfDNA and matched tumour DNA reveals a significant difference between on follow-up and pre-surgical patients. P1, first plasma cfDNA; P2, second plasma cfDNA.



Personalis EX2155

3.3.4 Copy number variation analysis

3.3.4.1 Status

Using 1Kb as the minimum size, the number of CNVs called for each sample/group was identified, and the percentage of overlap with the Toronto database of genomic variants (DGV) (Iafrate *et al.*, 2004) was determined. A total of 9,301 CNVs were identified in the lymphocyte DNA samples and 9,341 in the plasma cfDNA of the healthy controls (n=8). Of these CNVs, 57.07% and 55.95% respectively were found to overlap - either partially or completely - with the Toronto DGV, and are thus known/confirmed genomic variants. The remainder - 42.93% and 44.05%, respectively - did not overlap and are thus novel. When looking at the lymphocyte DNA samples in the pre-surgical patients and those on follow-up, 40.38% and 40.52% respectively were identified as novel. However, in the plasma cfDNA samples, 50.05% of the pre-surgical plasma cfDNA CNVs were identified as novel, compared to 45% and 46.48% in the on follow-up first plasma and healthy control plasma cfDNA. Finally, in the tumour DNA samples, 50.05% of CNVs called in the FFPE tumour DNA of patients on follow-up (n=40) were novel, and 40.12% were novel in the pre-surgical group (n=15).

3.3.4.2 Number of CNVs called

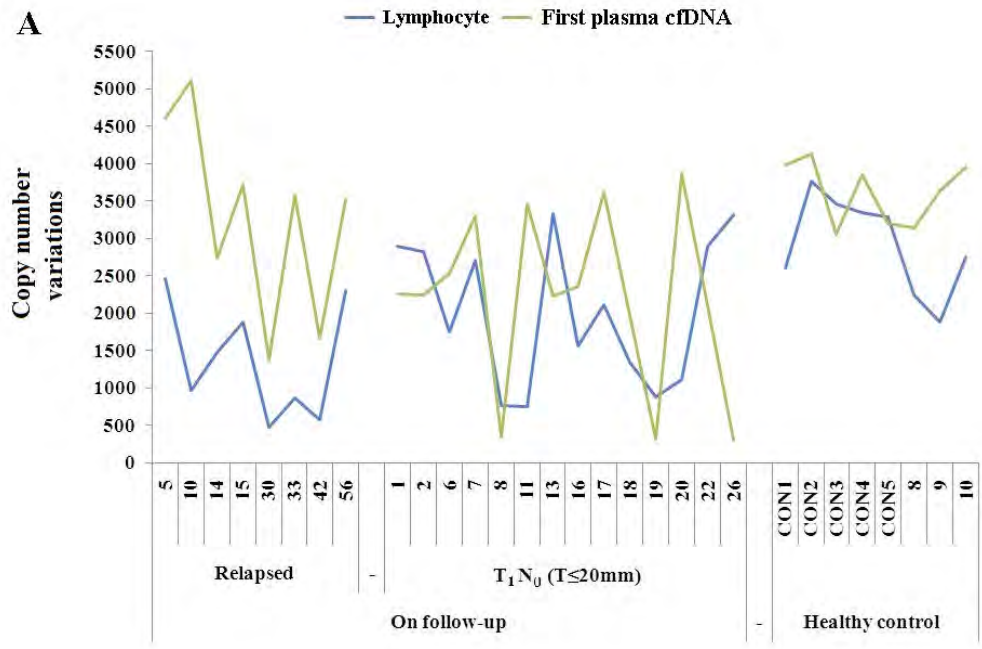
Next, the number of CNVs called in each sample was compared between paired plasma and matched lymphocyte DNA samples. In the patients on follow-up, there were 8 patients who relapsed and these showed a significant difference between the first cfDNA and matched lymphocyte ($p<0.01$, paired Wilcoxon signed-rank t-test) (Figure 3-9A). A statistically-significant difference ($p<0.05$) was also found for relapsed second cfDNA compared to matched lymphocyte (Figure 3-9B). When results for the other 42 patients on follow-up were stratified by size and node-status (i.e. - T_1N_0 ($T\leq20$mm) and T>20mm/node-positive), there were no significant differences for either the first or second cfDNA sample. The healthy control group also showed no significant differences.

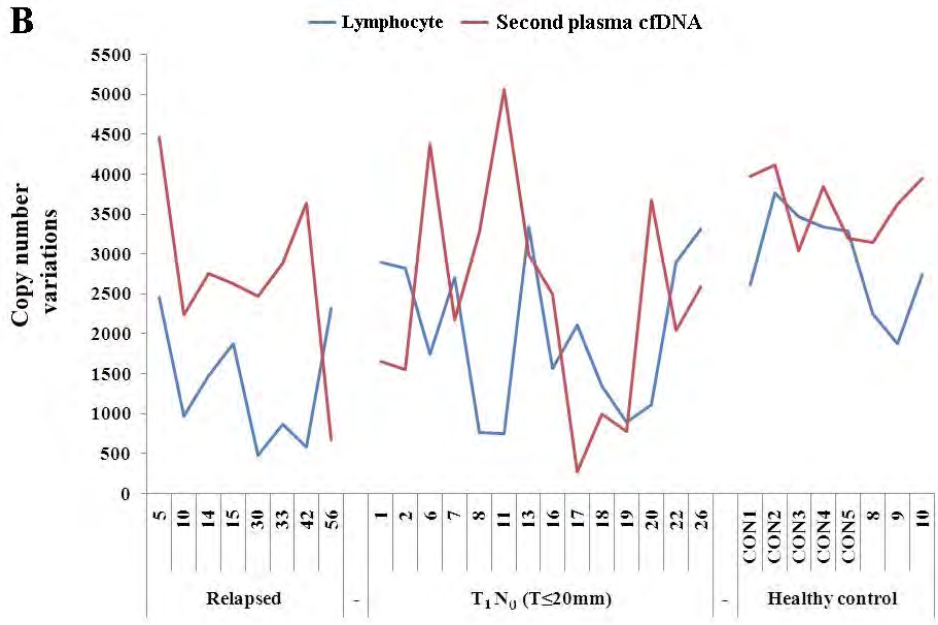
The difference in the number of CNVs called between the first and second cfDNA samples of the patients on follow-up was also analysed: overall, 30 patients showed a

Personalis EX2155

decrease in CNVs from their first to second cfDNA, and 20 showed an increase; while, of the relapsed patients, 5 showed a decrease and 3 an increase. Hence, there were no obvious trends with time.

Figure 3-9 - Number of CNVs called between the first (A) and second (B) plasma cfDNA and matched lymphocyte DNA samples by subgroup. NB - T>20mm/node-positive patients of patients on follow-up not shown.



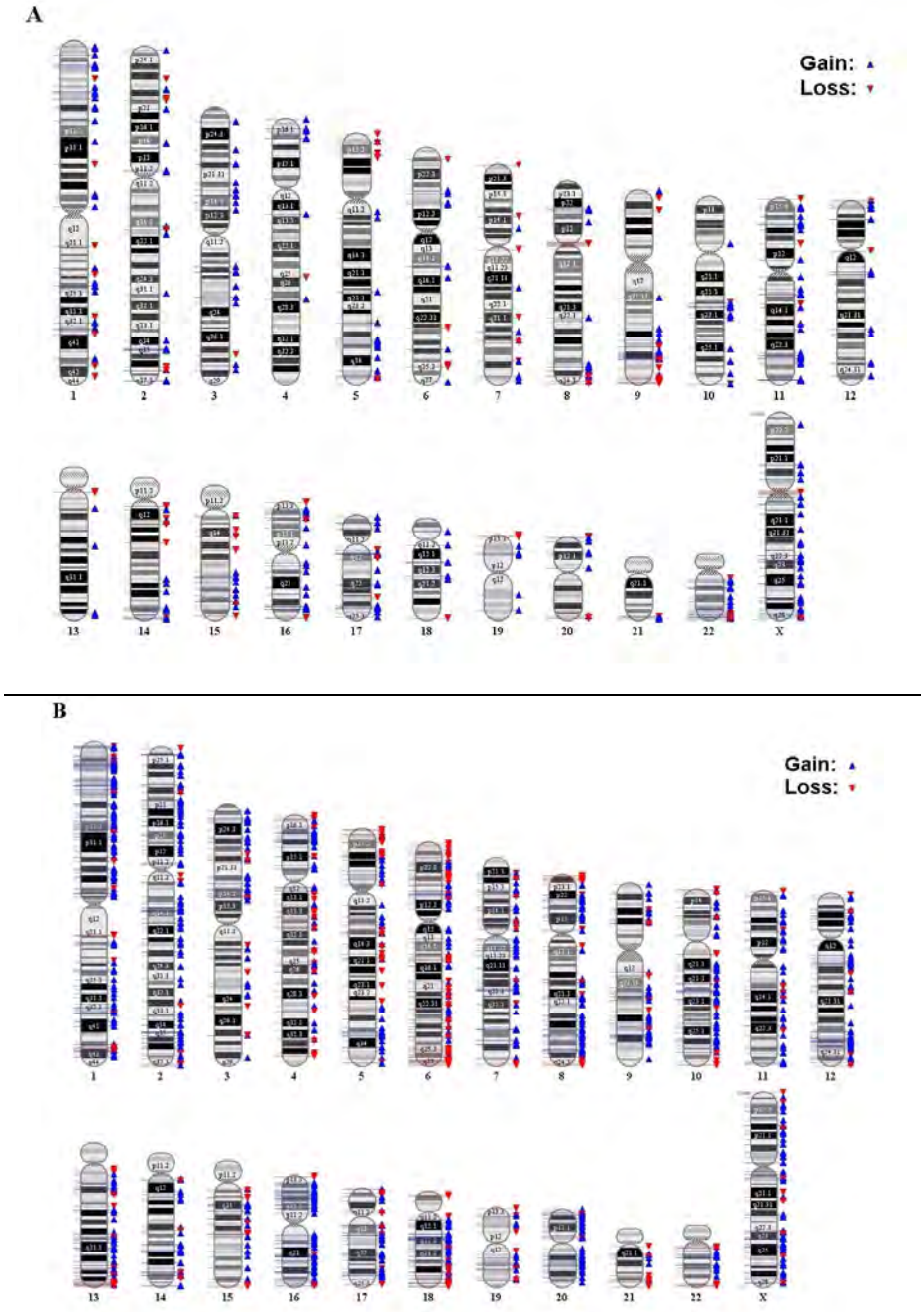


Personalis EX2155

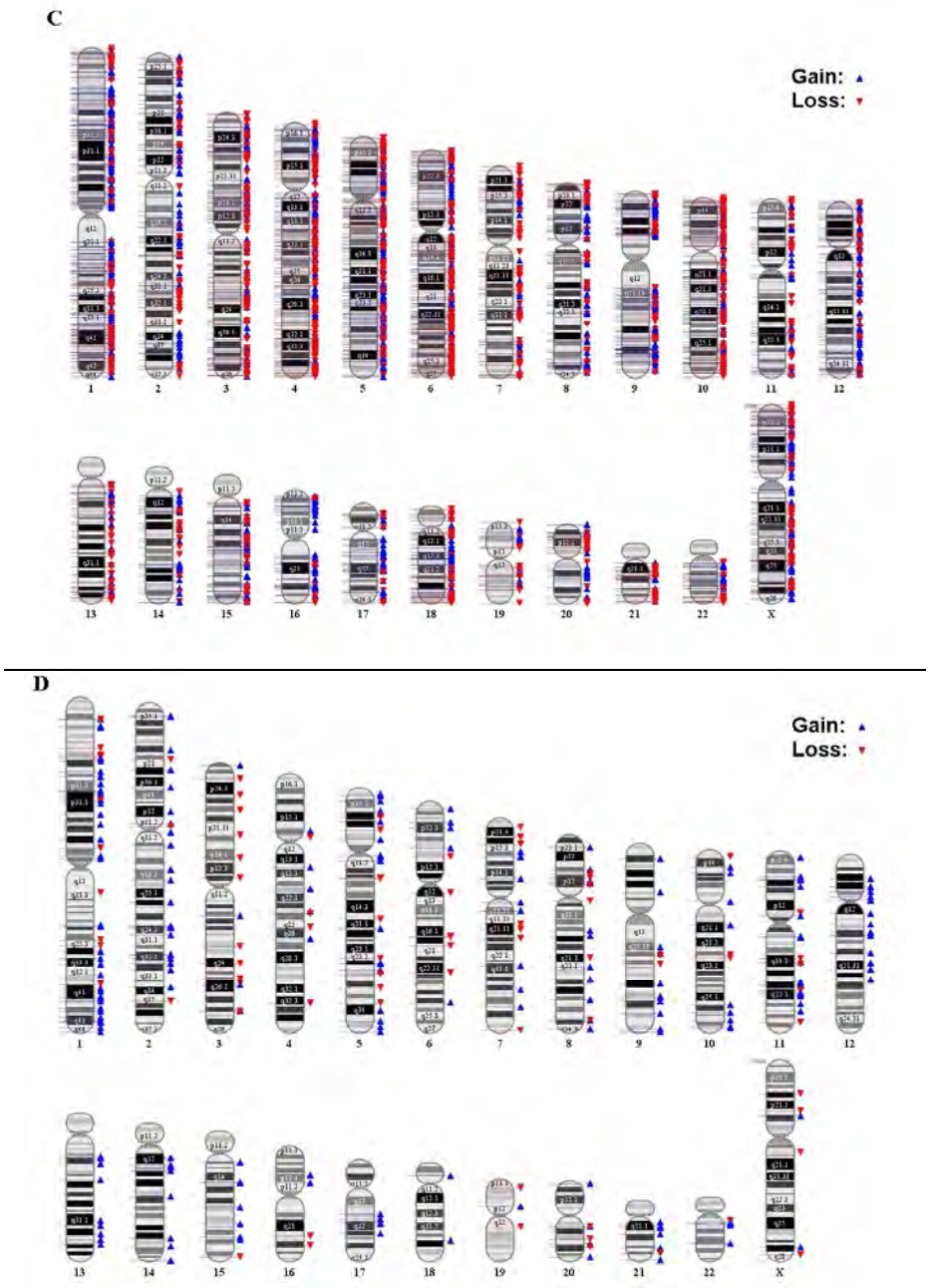
In the 8 patients who relapsed, the second cfDNA analysed was from the final blood sample taken prior to the time of relapse. Hence, genetic 'signatures' of relapse might be detected by looking at these samples. These patients had the greatest difference in the number of CNVs called between lymphocyte and plasma cfDNA (Figure 3-9), and some of these differences were striking. Figure 3-10 illustrates CN amplifications and deletions in one patient who relapsed. In this particular patient, the first blood plasma sample was taken 6 years after initial diagnosis, while the second blood plasma was taken 2 months before the patient was diagnosed with metastatic disease. This second blood plasma cfDNA sample showed more deletion than amplification.

Personalis EX2155

Figure 3-10 - Karyoview of the lymphocyte (A), first and second plasma cfDNA (B and C), and tumour (D) samples of one patient on follow-up who relapsed. A: 80.25% amplification; 19.75% deletion. B: 79.12% amplification; 20.88% deletion. C: 53.67% amplification; 46.33% deletion. D: 77.41% amplification; 32.59% deletion. CNVs based on 50 consecutive markers (SNP or CN) with a minimum segment size of 50Kb. Gain, amplification; Loss, deletion.



Personalis EX2155



3.3.4.3 Amplification and deletion

3.3.4.3.1 Distribution

Next, the distribution between amplification and deletion was analysed (Table 3-5). The majority of CNVs detected were amplifications, which was highest in the lymphocyte DNA samples (healthy controls, 67%; patients on follow-up, 74%; and pre-

Personalis EX2155

surgical breast cancer patients, 77%), with a mean of 72.80% amplification across these samples. Tumour DNA showed a slightly lower proportion of amplification (patients on follow-up, 65%; and pre-surgical breast cancers, 70%), with a mean of 67.5%. A lower proportion of amplification was detected in the cfDNA samples: the proportions were similar (59% and 57%, respectively) in the healthy control and first plasma cfDNA, however, the second plasma cfDNA of the patients on follow-up and the cfDNA of the pre-surgical patients both showed more deletion than amplification (55 and 62% deletion, respectively).

Group	Sample	All CNVs		Amplifications			Deletions		
		Mean total	Range	%	Mean total	Range	%	Mean total	Range
Healthy Control	L	2922	1885-3769	67	1961	1421-2325	33	962	464-1686
	P	3619	3052-4126	59	2134	1858-2421	41	1485	1091-2033
Pre-surgical	L	1549	426-3137	77	1260	370-2382	23	362	56-1050
	P	2090	1035-4272	38	786	369-2134	62	1304	661-2241
	T	1799	296-4336	70	1229	201-3057	30	551	95-1279
On follow-up	L	1709	134-4062	74	1261	113-2729	26	448	21-1333
	P1	2365	279-5110	57	1354	177-2698	43	1016	39-2587
	P2	2265	282-5096	45	1136	163-2702	55	1186	66-2789
	T	389	201-950	65	311	125-1999	35	156	38-695

Table 3-5 - Distribution between CNV amplification and deletion, showing percent amplification and deletion, mean total CNVs, and the range. L, lymphocyte; P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour.

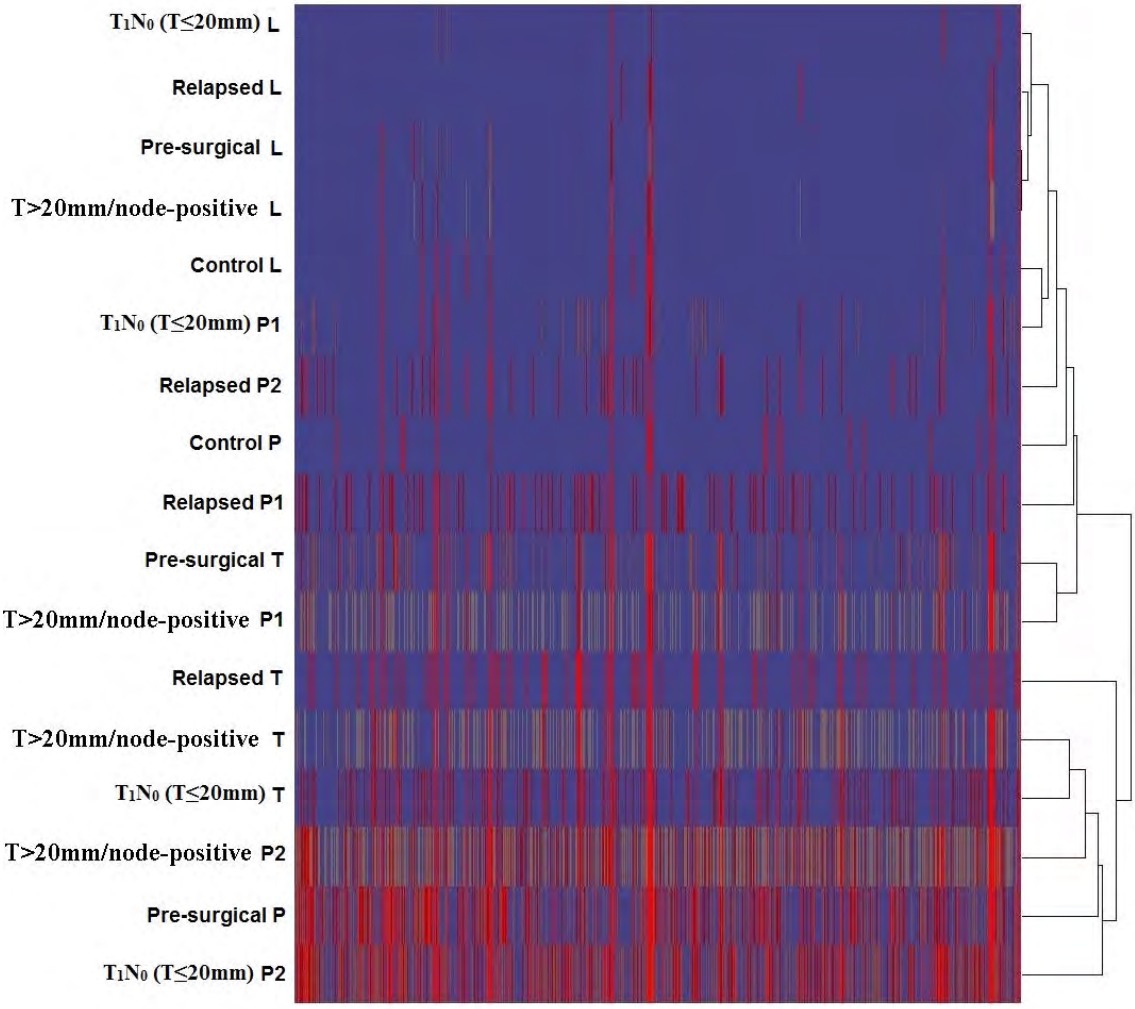
3.3.4.3.2 Thresholding amplifications by copy number

In relation to amplifications and deletions, a 'sweep' of all 273 samples in the study revealed 750 different cytobands that exhibited amplification using a high but arbitrary threshold of CN>6.0 by Gaussian-smoothed signal. Many of these amplifications, however, were unique to a single sample. The presence or absence of amplification at each of the 750 cytobands by sample type and patient group was revealed through hierarchical clustering (Figure 3-11). Two main clusters were identified: the top cluster

Personalis EX2155

contained all lymphocyte DNA samples and both the first cfDNA samples of patients on follow-up and the cfDNA of pre-surgical patients, whereas the bottom cluster contained all tumour DNA samples and the second cfDNA samples of patients on follow-up. However, of interest, the second plasma cfDNA samples from the relapsed patients also clustered at the left with the normal lymphocytes and first plasma samples.

Figure 3-11 - Hierarchical clustering (incorporating dendrogram and heatmap) of all DNA samples based on 750 cytobands identified using a threshold of CN>6.0 by Gaussian-smoothed signal. L, lymphocyte; P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour. Heatmap: red, high occurrence of amplification per sample group; blue, low occurrence.



Personalis EX2155

Next, results were filtered to remove uniquely-affected cytobands (i.e. - cytobands showing amplification in a single DNA sample), which reduced the total 634. These 634 were filtered further to show amplification in >10% of samples only (per sample group), giving 23 cytobands (Table 3-6; Appendix Table 8-1 provides additional information for these 23 cytobands, including mean CNV size and marker density).

Interval	Overlapping genes	Healthy control		Pre-surgical		On follow-up		
		L n=8	P n=8	P n=15	T n=15	P1 n=5	P2 n=50	T n=40
1p36.33	*LOC100132287*	4 [50]	5 [62.5]	15 [100]	10 [66.667]	35 [70]	44 [88]	39 [98]
1p36.21	*C1orf158* *LOC [440563 and 649330]* *PRAMEF [2 3 4 5 7 8 9 10 12 13 14 15 16 19 20 21 and 22]*	0 [0]	0 [0]	8 [53]	1 [6.667]	2 [4]	12 [24]	3 [8]
1q21.1	*PPIAL4 [A and C]*	3 [37.5]	4 [50]	12 [80]	9 [60]	33 [66]	37 [74]	38 [95]
2p11.2	*ANKRD36BP2* *EIF2AK3* *RPIA*	0 [0]	0 [0]	9 [60]	3 [20]	14 [28]	33 [66]	10 [25]
4q13.2	*TMPRSS11 [B and BNL]* *UGT2B [15 and 17]* *YTHDC1*	0 [0]	0 [0]	4 [27]	4 [26.667]	9 [18]	13 [26]	37 [93]
5p11	-	0 [0]	0 [0]	0 [0]	2 [13.333]	4 [8]	2 [4]	12 [30]
5q13.2	*SERF1 [A and B]* *SMA5* *SMN [1 and 2]* *NAIP*	0 [0]	0 [0]	9 [60]	6 [40]	18 [36]	20 [40]	36 [90]
7p11.2	*MIR4283-2*	0 [0]	0 [0]	2 [13]	3 [20]	2 [4]	7 [14]	18 [45]
7q11.21	-	0 [0]	0 [0]	6 [40]	3 [20]	13 [26]	14 [28]	14 [35]
7q11.23	*CLIP2* *GATSL1* *GTF2IRD1 2* *LOC100093631* *NCF [1 and 1C]* *PMS2P [5 and 6]* *SPDYE8P* *STAG3L2* *WBSCR16*	0 [0]	2 [25]	13 [87]	6 [40]	21 [42]	38 [78]	40 [100]
8p23.1	*DEFB130* *DUB3* *FAM66 [A and D]* *FAM85A* *FAM86B1* *LOC392196* *ZNF705D*	0 [0]	0 [0]	4 [27]	1 [6.667]	12 [24]	13 [26]	15 [38]
9p11.2	*FAM27A* *LOC100132167*	2 [25]	1 [12.5]	14 [93]	10 [66.667]	26 [52]	40 [80]	39 [98]
9q12	-	1 [12.5]	1 [12.5]	13 [87]	8 [53.333]	20 [40]	37 [74]	40 [100]
10q11.22	*BMS1P5* *FRMPD2P1* *GDF [2 and 10]*	1 [12.5]	0 [0]	14 [93]	6 [40]	25 [50]	33 [66]	37 [93]

Personalis EX2155

	PTPN20B *RBP3* *ZNF488*							
10q11.23	*AGAP [7 and 8]* *C10orf53* *CHAT* *OGDHL* *PARG*	2 [25]	3 [37.5]	14 [93]	6 [40]	30 [60]	36 [72]	40 [100]
14q11.1	-	1 [12.5]	1 [12.5]	8 [53]	6 [40]	23 [46]	27 [54]	28 [70]
14q11.2	-	0 [0]	0 [0]	15 [100]	6 [40]	0 [0]	36 [72]	40 [100]
14q32.33	*AHNAK2* *BRF1* *C14orf79* *CDCA4* *CRIP [1 and 2]* *GPR132* *JAG2* *KIAA0284* *LOC647310* *MGC23270* *MTA1* *NUDT14* *PACS2*	1 [12.5]	0 [0]	10 [67]	3 [20]	5 [10]	15 [30]	13 [33]
15q11.2	*NF1P1* *POTEB*	0 [0]	0 [0]	5 [33]	1 [6.667]	8 [16]	16 [32]	23 [58]
15q25.2	*GOLGA6L [4 5 and 9]* *LOC [440297 727849 and 80154]* *RPS17*	1 [12.5]	0 [0]	10 [67]	7 [46.667]	21 [42]	30 [60]	34 [85]
16p12.3	*ABCC6P1* *NOMO2*	2 [25]	5 [62.5]	12 [80]	6 [40]	19 [38]	30 [60]	35 [88]
16p11.2	*ALDOA* *C16orf92* *DOC2A* *FAM57B* *HIRIP3* *INO80E* *KCTD13* *TAOK2* *TMEM219*	0 [0]	0 [0]	6 [40]	5 [33.333]	11 [22]	30 [60]	20 [50]
19p13.3	*FAM138F* *FLJ45445* *OR4F17* *STK11* *WASH5P*	2 [25]	3 [37.5]	12 [80]	7 [46.667]	39 [76]	41 [82]	36 [90]

Table 3-6 - Twenty-three intervals exhibiting common amplification in plasma cfDNA and tumour (using a threshold of CN>6.0 by Gaussian-smoothed signal and only displaying those that had a frequency >10%). The frequency of amplification (%) at a given interval is indicated in square-brackets. L, lymphocyte; P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour. Appendix Table 8-1 provides additional information for these intervals.

The lower frequencies of these CNVs in the pre-surgical tumour DNA were due to a different origin and nature of these samples: 10 of the 15 samples were FF samples and taken at the end of 21 weeks neoadjuvant therapy, while the remaining 5 were FFPE.

Personalis EX2155

Eighteen of the 23 cytobands showed amplification in plasma and tumour DNA of breast cancer patients, but little or no amplification in the cfDNA of healthy controls. The remaining 5 cytobands (i.e. - 1p36.33, 1q21.1, 10q11.23, 16p12.3 and 19p13.3) - whilst showing amplification in plasma and tumour of breast cancer patients - also showed some amplification in the lymphocyte and/or plasma cfDNA of healthy controls. Eighteen of the 23 regions have known overlapping genes and the remaining 5 have none as defined by the RefSeq gene database (Pruitt *et al.*, 2005). Some interesting groupings and molecular functions were revealed when the overlapping genes were grouped together using DAVID (Huang *et al.*, 2009a & 2009b) (Table 3-7).

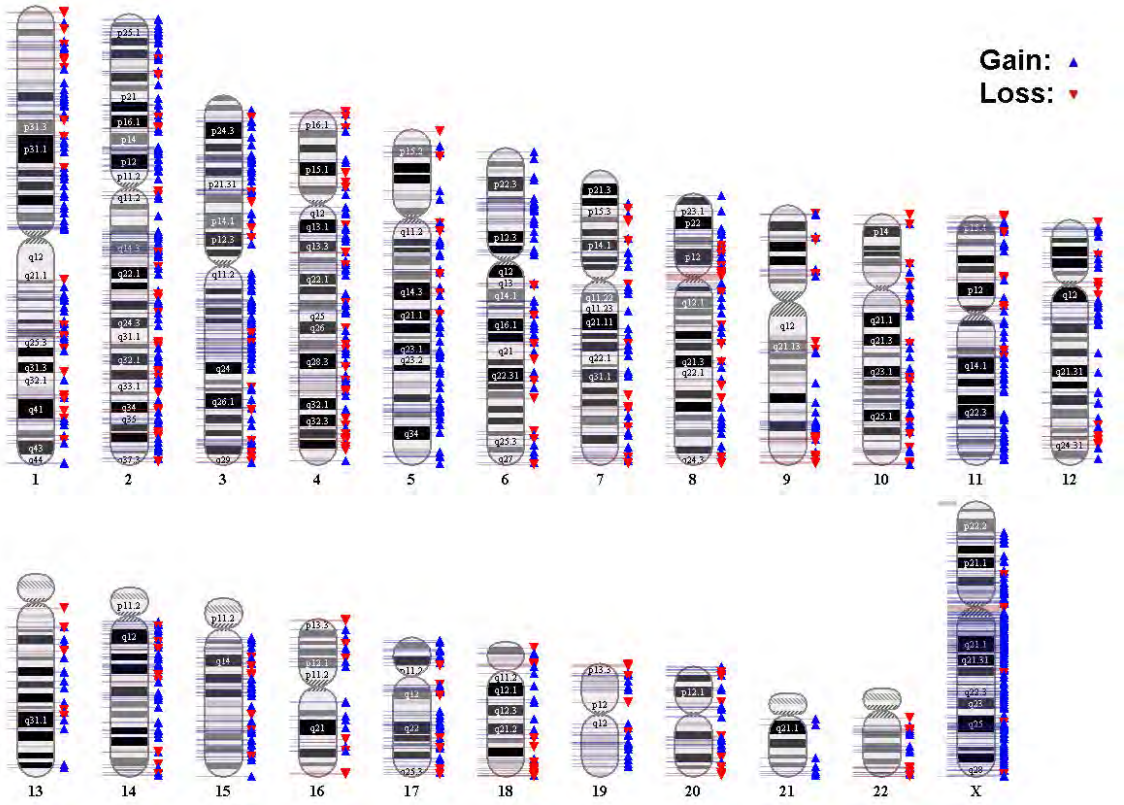
Group/Molecular function	Gene count
Nucleus	16
Leucine-rich repeat / preferentially-expressed antigen of melanoma family protein	15
Metal ion binding	14
Membrane	14
Transmembrane	13
Cytoplasm	11
DNA binding / regulation of transcription	8
Endoplasmic reticulum	6
Cell surface receptor-linked signal transduction	6
Cytoskeleton	5
Apoptosis	5
Regulation of RNA metabolic process	4
Carbohydrate catabolic process	3
Amino acid phosphorylation	3
Growth factor activity	3
Downregulation of *MTA3* in ER-negative breast tumours	2
Androgen and oestrogen metabolism	2

Table 3-7 - List of related groups and molecular functions of genes overlapping common amplifications from Table 3-6.

Personalis EX2155

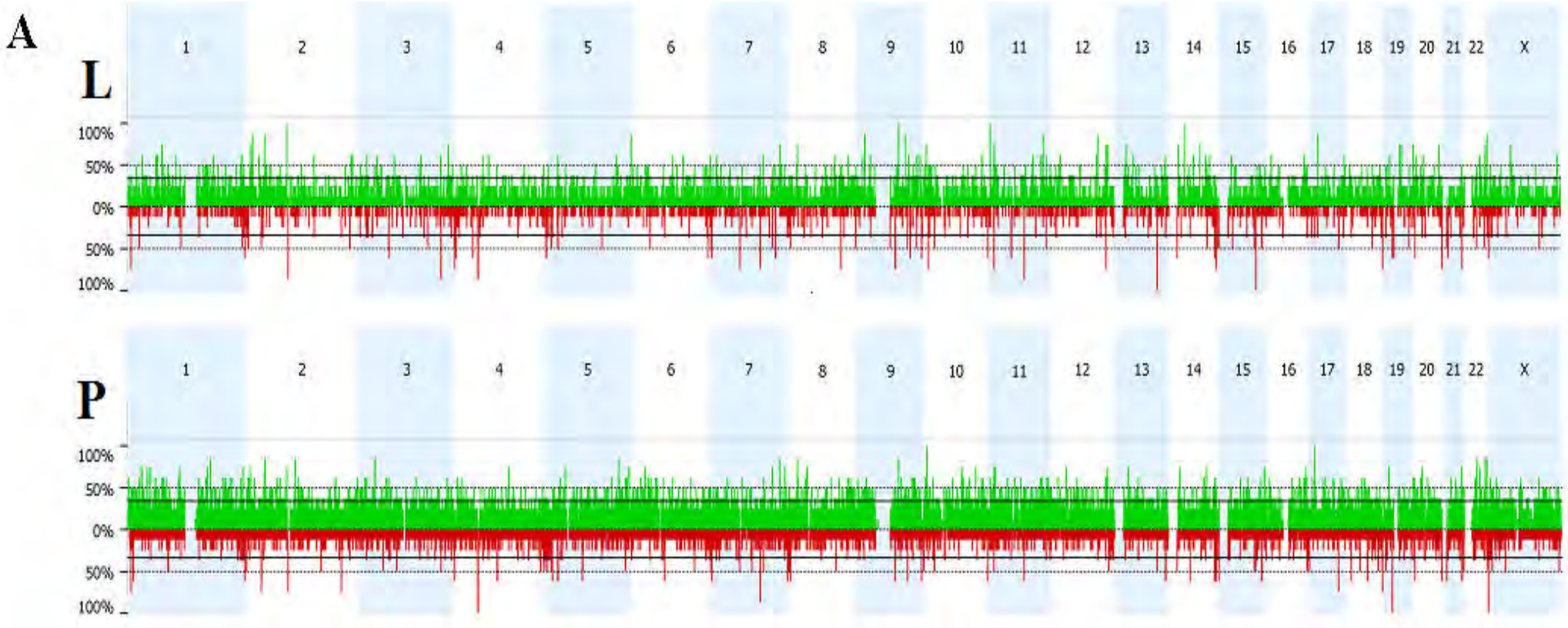
Further analysis, performed by raising the threshold to CN>8.0 revealed 352 regions, while lowering the threshold to CN>4.0 revealed amplification in at least one sample at virtually every known cytoband. These CNV levels support results of Redon *et al.* (2006) that CNV is common in the human genome. Figure 3-12 highlights this by showing the extent of CNV in the lymphocyte DNA sample of one healthy control, while Figure 3-13 shows a genome-view representation of the frequency of both amplification and deletion in all samples in this study. There was a low frequency of CNV in the tumour DNA samples of the patients on follow-up and the pre-surgical breast cancer patients.

Figure 3-12 - CNVs detected in the normal lymphocyte DNA sample of a healthy control using a threshold of CN>4.0, highlighting how common CNVs are in the 'healthy' human genome. Gain, amplification; loss, deletion.



Personalis EX2155

Figure 3-13 - Genome-view representations of the frequency of amplification and deletion in the healthy controls (A), patients on follow-up (B), and pre-surgical patients (C). The y-axis indicates the percentage of samples having a CNV at a specific point in the genome. Chromosome number at top; amplifications in green; deletions in red; L, lymphocyte; P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour.



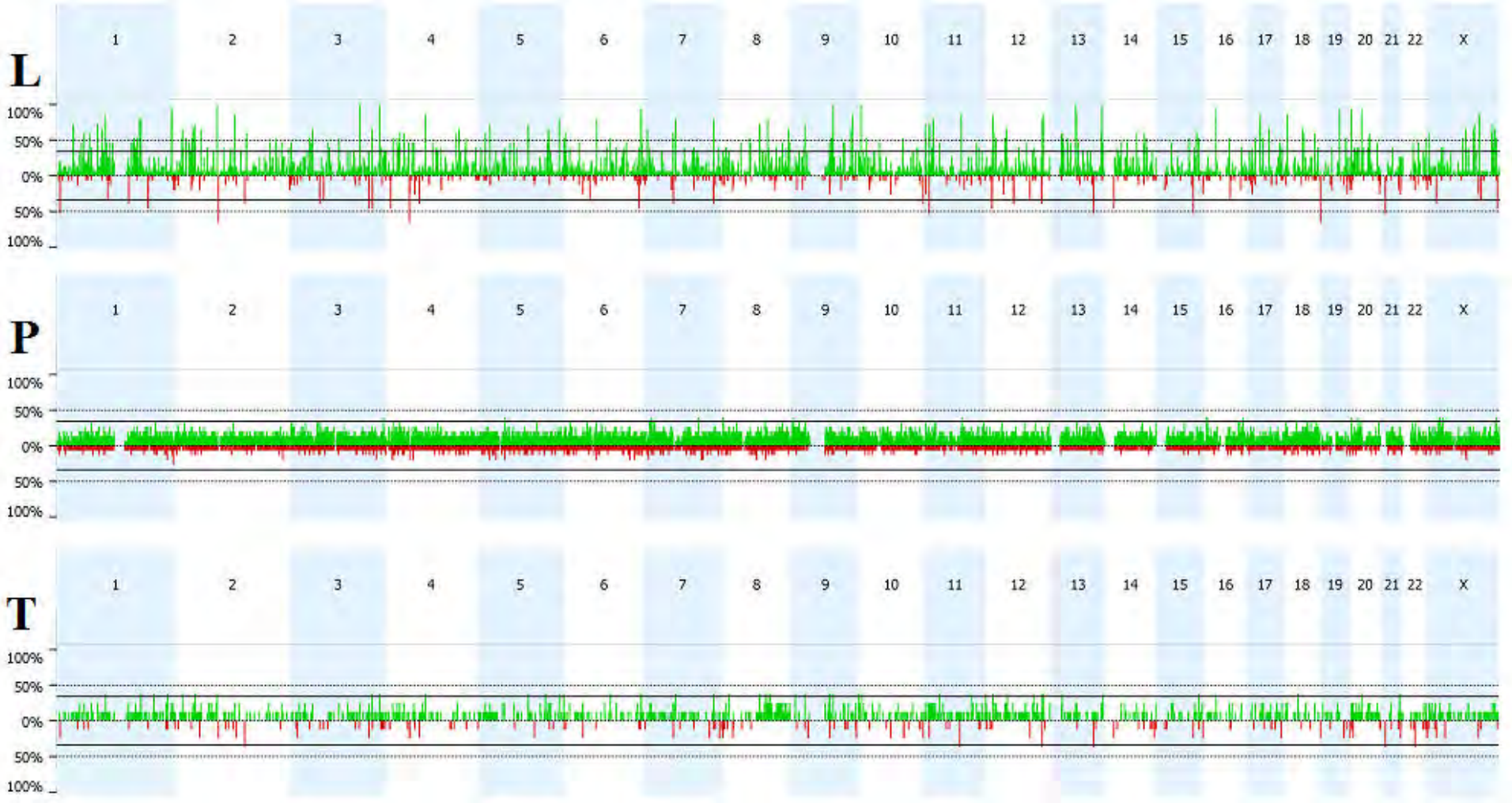
Personalis EX2155

B



Personalis EX2155

C



Personalis EX2155

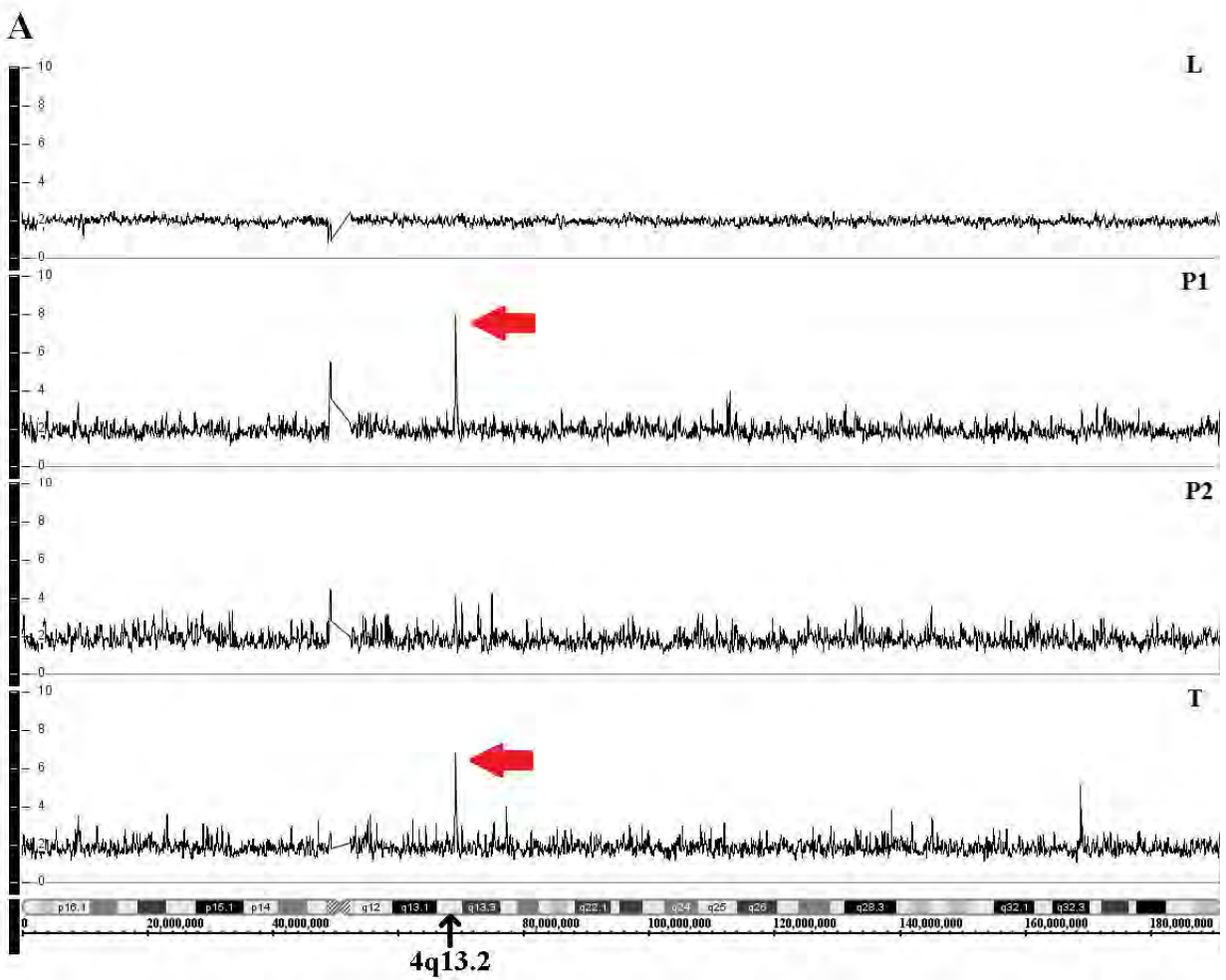
Interestingly, at this lower threshold (CN>4.0), 8 cytobands in the human genome exhibited high-frequency (>90%) amplification in the tumour DNA samples of patients on follow-up. For some of these, amplification was seen in every (i.e. - 100%) tumour sample (Table 3-8). These 8 cytobands also showed amplification in breast cancer plasma cfDNA, but at a lower frequency (>25%). Figure 3-14 shows examples of amplification detected at 3 of these cytobands (4q13.2, 5q13.2, and 7q11.23) in patients on follow-up. There was clear evidence of local amplification in cfDNA and tumour DNA as revealed by a 'spike', while Figure 3-15 highlights the frequency of amplifications at one cytoband (5q13.2) in the patients on follow-up. Regarding these CNVs, the majority resemble the amplifier (firestorm) or complex (sawtooth) CN alteration (CNA) patterns previously identified by Kwei *et al.* (2010) (Figure 1-5).

	Healthy control		**Pre-surgical**			**On follow-up**			
	L	**P**	**L**	**P**	**T**	**L**	**P1**	**P2**	**T**
	n=8	**n=8**	**n=15**	**n=15**	**n=15**	**n=50**	**n=50**	**n=50**	**n=40**
4q13.2	0% [2]	12.5% [1-6]	33.3% [1-4]	26.6% [2-7]	40% [1-8]	26% [2-4]	28% [2-9]	42% [2-10]	100% [6-10]
5q13.2	12.5% [0-3]	12.5% [2-9]	46.6% [2-4]	93.3% [2-13]	40% [2-15]	50% [2-6]	68% [2-10]	79% [2-10]	100% [6-10]
7q11.23	12.5% [2-4]	62.5% [0-9]	33.3% [2-4]	93.3% [2-10]	60% [2-10]	24% [2-8]	62% [2-10]	82% [2-10]	100% [9-10]
8p23.1	0% [2]	25% [2-4]	20% [2-3]	46.6% [2-6]	33.3% [2-7]	22% [2-4]	32% [2-10]	34% [2-8]	95% [2-10]
10q11.22 / 23	87.5% [2-10]	100% [4-10]	80% [2-10]	100% [6-10]	40% [2-10]	86% [2-24]	92% [2-13]	86% [2-30]	100% [6-28]
15q25.2	37.5% [2-6]	0% [2]	46.67% [2-5]	73.33% [2-9]	40% [2-9]	22% [2-5]	54% [2-10]	96% [2-10]	100% [4-20]
16p12.3	62.5% [2-6]	75% [2-7]	66.67% [2-6]	100% [5-11]	80% [2-9]	54% [2-6]	66% [2-10]	90% [2-15]	100% [5-12]

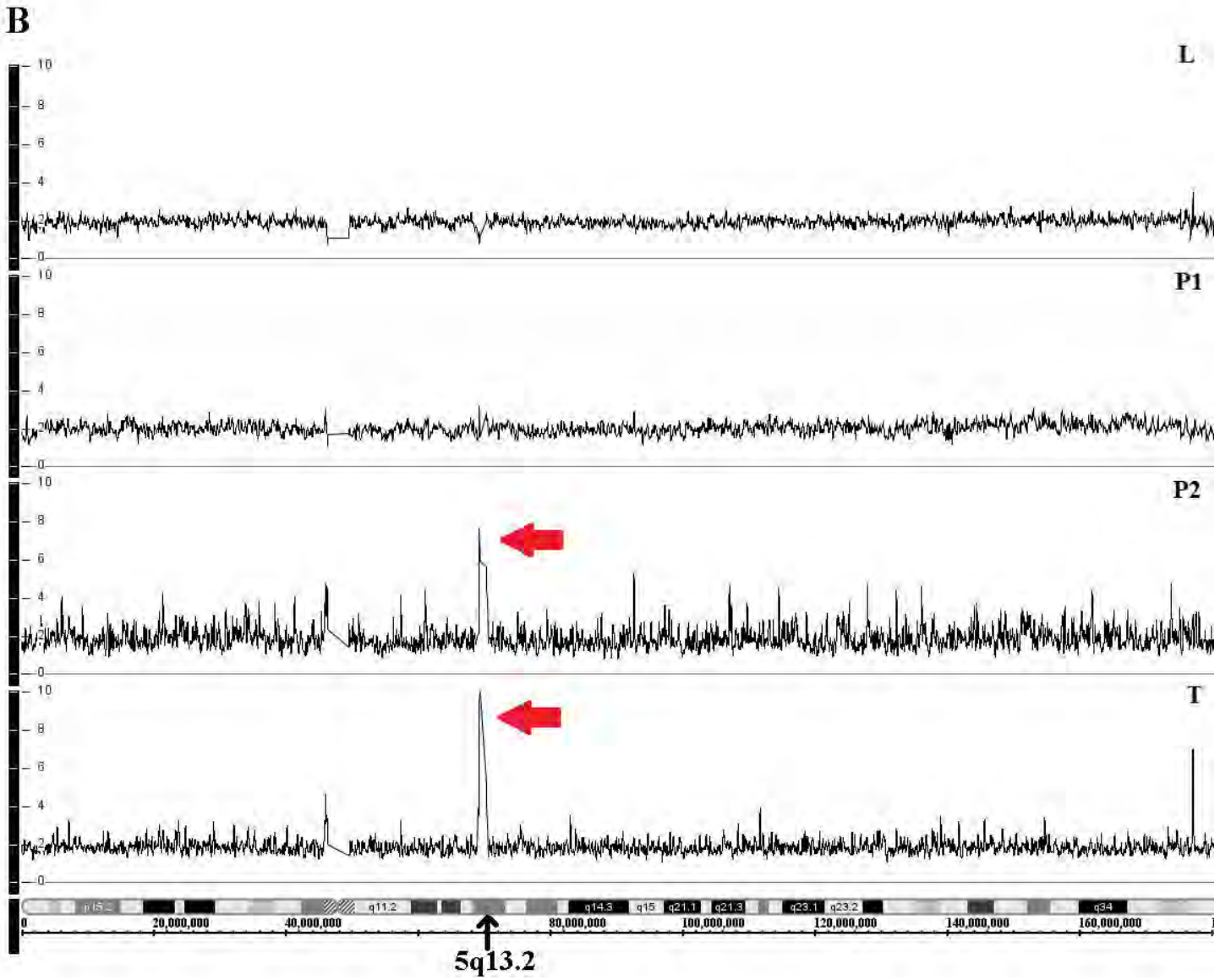
Table 3-8 - Eight intervals exhibiting common amplification in >90% breast cancer tumour DNA of on follow-up patients (using a threshold of CN>4.0 by a Gaussian-smoothed signal). The number of samples with the thresholded CNV per sample group is indicated as a percentage, while also shown in square-brackets is the non-Gaussian-smoothed CN range. L, lymphocyte; P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour. NB - CNVs at 10q11.22 and 10q11.23 were treated as a single CNV due to close genomic proximity and high correlation.

Personalis EX2155

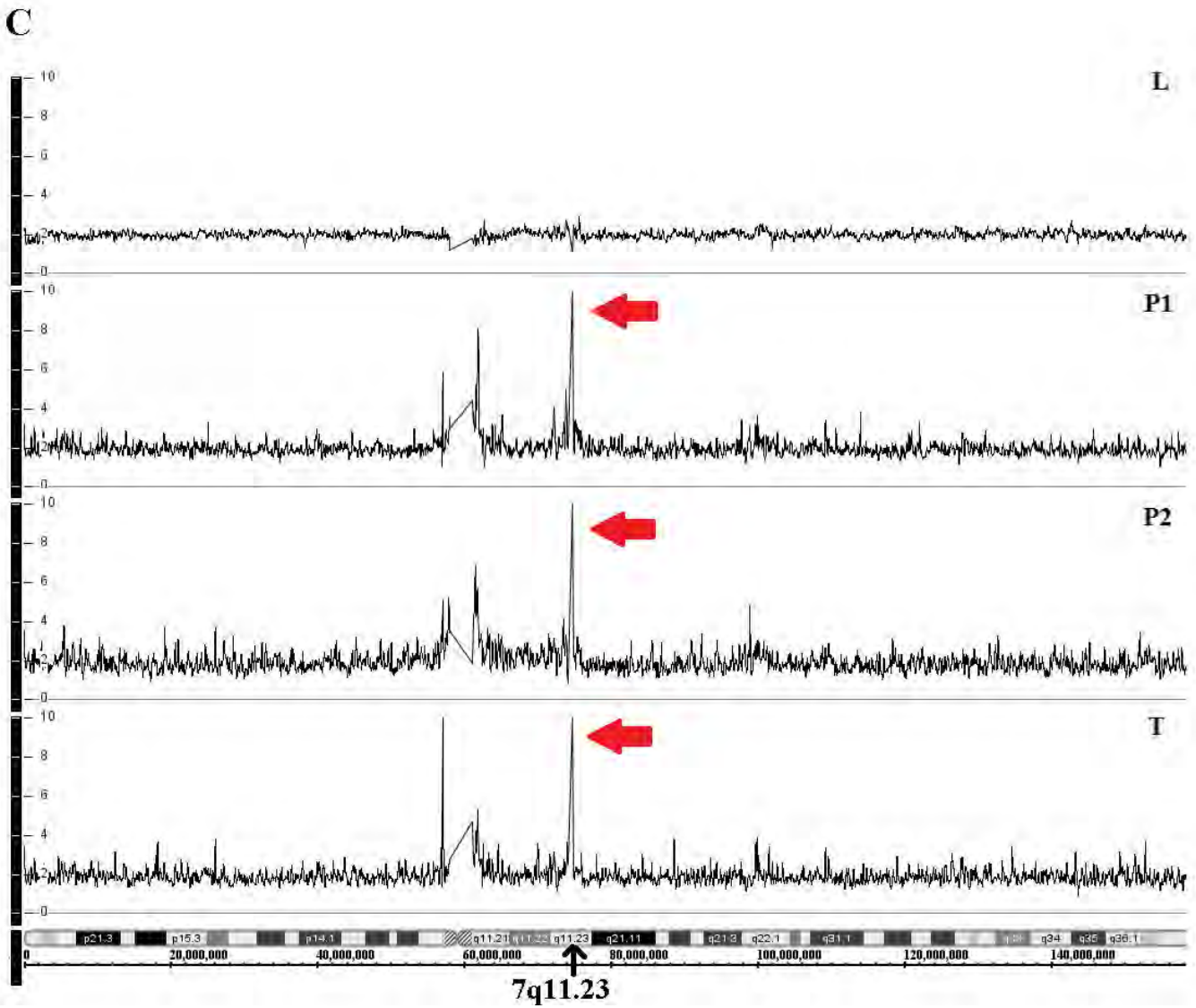
Figure 3-14 - High level amplification in plasma cfDNA and matched tumour DNA of breast cancer patients on follow-up, showing commonly-amplified cytobands 4q13.2 (A), 5q13.2 (B), and 7q11.23 (C) (based on a CN>4.0 threshold by Gaussian-smoothed signal). Amplification is matched between T and P1 at 4q13.2, T and P2 at 5q13.2, and T, P1, and P2 at 7q11.23. L, lymphocyte; P1, first plasma cfDNA; P2, second plasma cfDNA; T, tumour.



Personalis EX2155

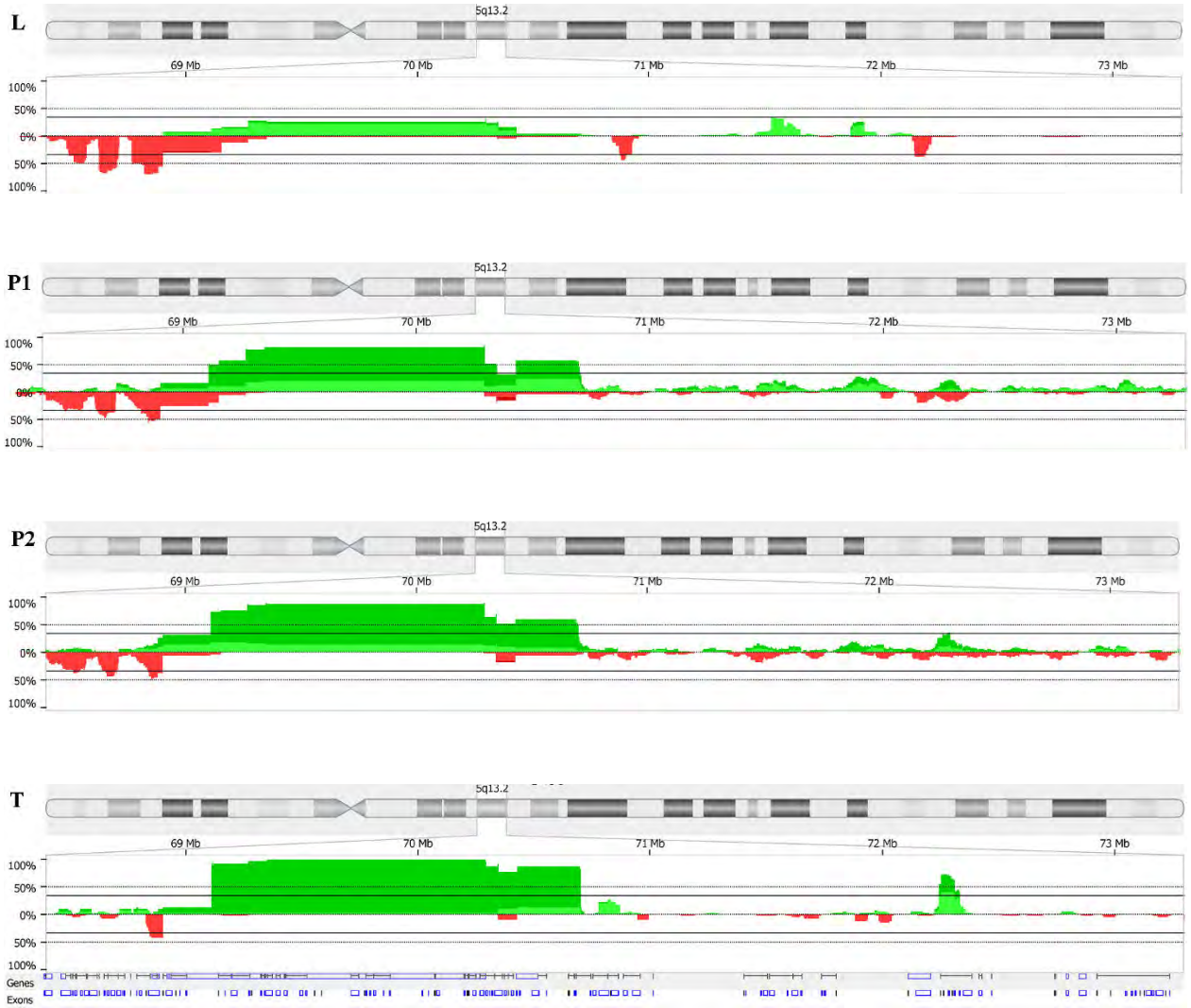


Personalis EX2155



Personalis EX2155

Figure 3-15 - High frequency of amplification at 5q13.2 in patients on follow-up. The y-axis indicates the percentage of samples having a CNV at a specific point in the genome. Amplifications in green; deletions in red. L, lymphocyte; P1, first plasma cfDNA; P2, second plasma cfDNA; T, tumour.



Personalis EX2155

3.3.4.3.3 Logistic regression to reveal significant copy number variant regions

Binary logistic regression analysis was used to correlate the prevalence of the 23 common CNVs (Table 3-6) to the molecular phenotype of the primary tumour (Table 3-9). Based on cancer status (i.e. - cancer or healthy), 5 CNVs significantly-contributed to the disease status in the first cfDNA samples in the patients on follow-up (5q13.2, 9p11.2, 10q11.22, 14q11.2, and 15q25.2), while 9 CNVs significantly-contributed in the second cfDNA samples (5q13.2, 9p11.2, 9q12, 10q11.22, 14q11.1, 14q11.2, 15q25.2, 16p11.2, and 19p13.3). Whereas, in the cfDNA samples of the pre-surgical patients, 12 CNVs significantly-contributed to the cancer status. Of these, 6 were unique to the pre-surgical cfDNA samples (1p36.33, 1p36.21, 2p11.2, 7q11.21, 10q11.23, and 14q32.33).

When comparing patients who relapsed with those who did not in the on follow-up group, two CNVs (1p36.33 and 9q12) were correlated significantly with relapsed status based on the first cfDNA samples, and 11 (1p36.33, 1q21.1, 9p11.2, 9q12, 10q11.22, 10q11.23, 14q11.1, 14q11.2, 15q11.2, 16p11.2, 19p13.3) based on the second. Whereas, in the tumours, six CNVs contributed significantly to relapse, 4 of which (1p36.33, 1q21.1, 9p11.2, and 9q12) were significant at a value of $p<0.0001$. These 6 cytobands also distinguished between T>20m/node-positive tumours compared to those that were T_1N_0 ($T\leq20mm$).

There were no significant differences by tumour type, based on the type of surgery (mastectomy or wide local excision), but there were significant differences in the subsequent plasma samples taken after surgery: first cfDNA, 5q13.2; second cfDNA, 4q13.2, 5q13.2, and 14q32.33. Significant differences also arose based on endocrine therapy received (tamoxifen, anastrozole (Arimidex®), Goserelin (Zoladex®), or tamoxifen/anastrozole combined) compared to no therapy received, predominantly in the second cfDNA comparison.

Finally, based on receptor status in the on follow-up group of patients, no CNVs contributed significantly to either ERBB2- or PR-positive status in tumour or cfDNA. However, 5p11 contributed significantly to ER-positive status in the second plasma cfDNA samples, and to the triple-negative status of the primary tumour.

Personalis EX2155

Interval	CANCER STATUS: Control (P) Vs On follow-up (P1/P2/T)			CANCER STATUS: Control (P) Vs Pre-surgical (P)	RELAPSED STATUS: Relapsed Vs Non-relapsed			SIZE AND NODAL STATUS: T>20mm/node-positive Vs T_1N_0 (T≤20mm)			SURGERY TYPE: Mastectomy Vs Wide local excision		ENDOCRINE THERAPY RECEIVED: Therapy Vs None			RECEPTOR STATUS: ERBB2 positive Vs negative	RECEPTOR STATUS: ER positive Vs negative	RECEPTOR STATUS: PR positive Vs negative	RECEPTOR STATUS: Triple negative Vs Receptor positive		Interval
	P1	P2	T	-	P1	P2	T	P1	P2	T	P1	P2	P1	P2	T	P1/P2/T	P2	P1/P2/T	P1	T	
1p36.33	NS	NS	NS	0.0446	0.0019	0.0002	<0.0001	NS	NS	0.0087	NS	NS	NS	NS	NS		NS		NS	NS	1p36.33
1p36.21	NS	NS	NS	0.0088	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.0322	NS		NS		NS	NS	1p36.21
1q21.1	NS	NS	NS	NS	NS	0.0047	<0.0001	0.0019	0.0105	0.0087	NS	NS	NS	NS	NS		NS		NS	NS	1q21.1
2p11.2	NS	NS	NS	0.0286	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	2p11.2
4q13.2	NS	NS	0.0002	NS	NS	NS	NS	NS	NS	NS	NS	0.0309	NS	NS	NS		NS		NS	NS	4q13.2
5p11	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.0324	NS		0.0143		NS	0.0164	5p11
5q13.2	0.0417	0.0271	<0.0001	0.0033	NS	NS	NS	NS	NS	NS	0.0457	0.0368	NS	NS	NS		NS		NS	NS	5q13.2
7p11.2	NS	NS	0.0338	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	7p11.2
7q11.21	NS	NS	NS	0.0386	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	7q11.21
7q11.23	NS	NS	NS	NS	NS	NS	NS	NS	0.0116	NS	NS	NS	NS	NS	NS		NS		NS	NS	7q11.23
8p23.1	NS	NS	0.0099	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	0.0170		NS		NS	NS	8p23.1
9p11.2	0.0293	<0.0001	0.0004	<0.0001	NS	0.0203	<0.0001	NS	NS	0.0031	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	9p11.2
9q12	NS	0.0088	0.0032	0.0083	0.0284	0.0173	<0.0001	NS	NS	0.0105	NS	NS	NS	NS	NS		NS		0.0306	NS	9q12
10q11.22	0.0381	0.0057	0.0004	<0.0001	NS	0.0115	NS	0.0487	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	10q11.22
10q11.23	NS	NS	NS	0.0150	NS	0.0287	NS	0.0109	NS	NS	NS	NS	0.0115	NS	NS		NS		NS	NS	10q11.23
14q11.1	NS	0.0032	<0.0001	NS	NS	0.0001	0.0372	NS	0.0105	0.0357	NS	NS	NS	NS	NS		NS		NS	NS	14q11.1
14q11.2	0.0271	<0.0001	<0.0001	NS	NS	0.0009	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	14q11.2
14q32.33	NS	NS	NS	0.0116	NS	NS	NS	NS	NS	NS	NS	0.0320	NS	NS	NS		NS		NS	NS	14q32.33
15q11.2	NS	NS	0.0039	NS	NS	0.0268	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	15q11.2
15q25.2	0.0215	0.0002	0.0002	0.0010	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	15q25.2
16p12.3	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS	NS		NS		NS	NS	16p12.3
16p11.2	NS	0.0012	0.0271	NS	NS	0.0022	NS	NS	NS	NS	NS	NS	0.0275	NS	NS		NS		NS	NS	16p11.2
19p13.3	NS	0.0057	0.0169	0.0432	NS	0.0095	0.0376	NS	0.0347	0.0308	NS	NS	NS	0.0373	NS		NS		NS	NS	19p13.3

Table 3-9 - Summary table of binary logistic regression analysis. Source data was based on the presence or absence of an amplification (using a threshold of CN>6.0 by Gaussian-smoothed signal) at each of 23 common CNVs listed in Table 3-6. Significant p-values shown; non-significant p-values denoted by 'NS'. P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour.

Personalis EX2155

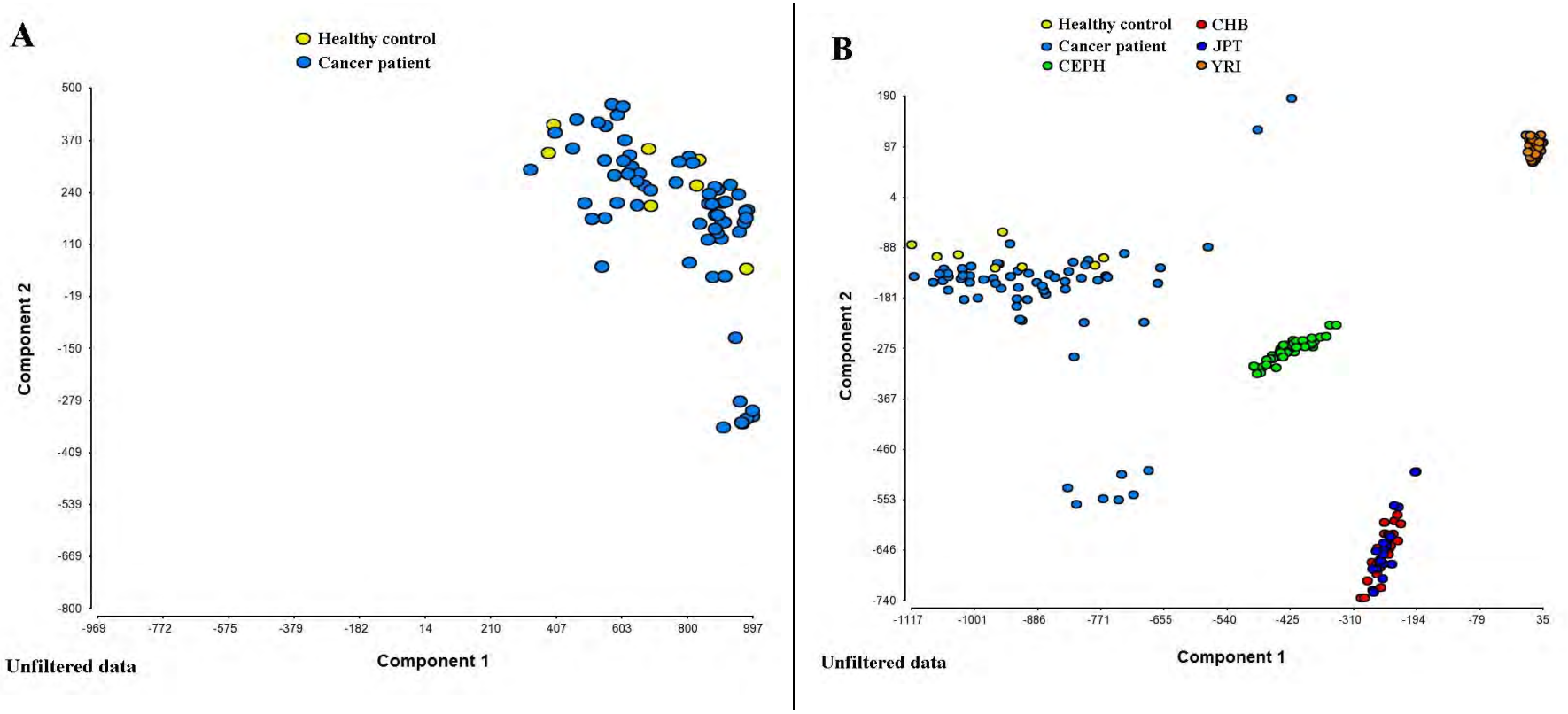
3.3.5 Principal component analysis and hierarchical clustering to interrogate both SNP and CN data

3.3.5.1 Principal component analysis

Comparison of both SNP and CNV data was investigated through PCA, which is a method used for separating a set of related data into groups (or principal components). In the lymphocyte DNA samples, PCA analysis of SNP/CN data revealed no stratification between cancer patients and healthy controls (Figure 3-16A). However, stratification was evident based on ethnicity when these samples were compared to 128 healthy female lymphocyte samples from the International HapMap (International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/), consisting of: 46 CEPH (Utah, USA residents with ancestry from northern and western Europe); 23 CHB (Han Chinese in Beijing, China); 22 JPT (Japanese in Tokyo, Japan); and 37 YRI (Yoruba in Ibadan, Nigeria) (Figure 3-16B).

Personalis EX2155

Figure 3-16 - PCA bi-plot comparison of lymphocyte DNA from cancer patients and healthy controls (A, unfiltered data); and cancer patients, healthy controls, and 128 International HapMap samples (B, unfiltered data). CEPH, Utah, USA residents with ancestry from northern and western Europe; CHB, Han Chinese in Beijing, China; JPT, Japanese in Tokyo, Japan; YRI, Yoruba in Ibadan, Nigeria.



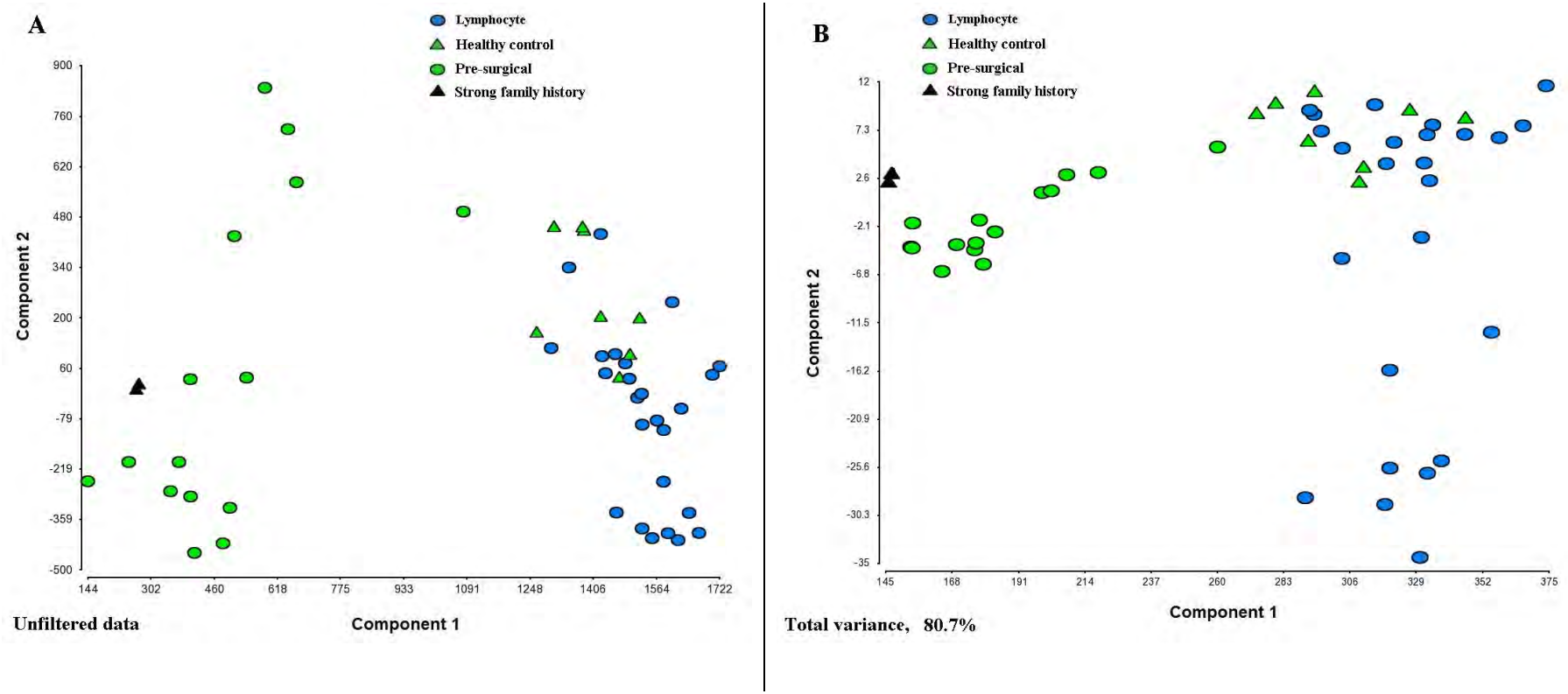
Personalis EX2155

PCA revealed a clear division between the plasma cfDNA samples of the pre-surgical patient group compared to those of the healthy controls (Figure 3-17A). This was also evident using SNP and then CN markers separately (Figure 3-17B,C). The 3 resulting plots were very similar and subsequent examination of the component loadings for each sample from each plot revealed a Spearman's rank correlation of 0.972, suggesting similar SNP and CN profiles.

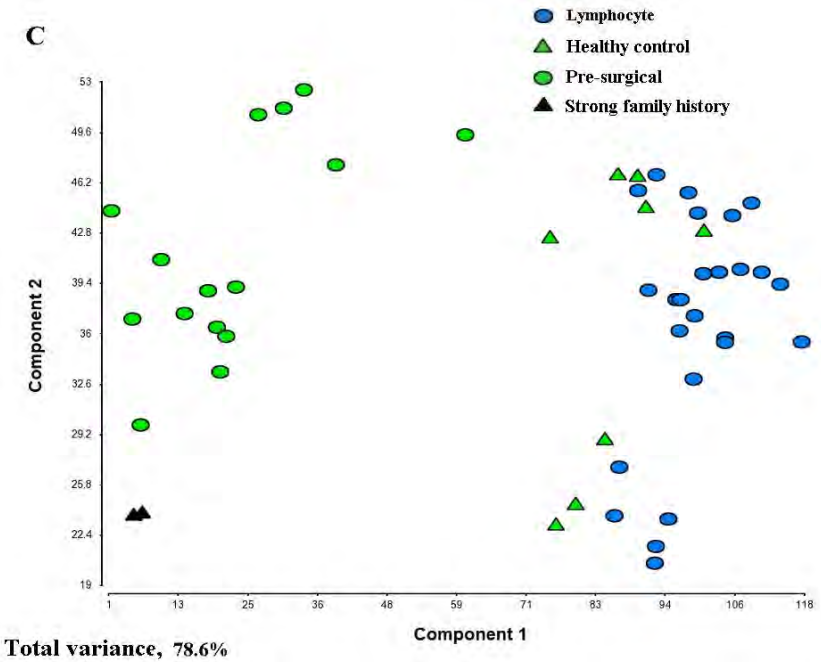
Of note, two plasma cfDNA samples from women with a strong family history of breast cancer (but who had normal mammography) also clustered with cfDNA from the pre-surgical group by PCA (Figure 3-17A). These two cfDNA samples exhibited low SNP concordance with their paired lymphocytes (55.05% and 55.02%), within the range of those of the 15 pre-surgical cfDNAs (mean 44.88%, range 36.00-68.27%) and the first and second cfDNAs of the group on follow-up (mean 54.22%, range 33.31-97.96% and 69.10%; range 33.17-99.44% respectively). However, they differed significantly from those of the healthy controls (mean 89.35%, range 81.10-94.08%).

Personalis EX2155

Figure 3-17 - PCA bi-plot comparison of pre-surgical breast cancer patients and healthy controls (A, unfiltered data), recreated using just SNP (B) and CN markers (C). Two plasma cfDNA samples from healthy women with a strong family history are also shown.



Personalis EX2155



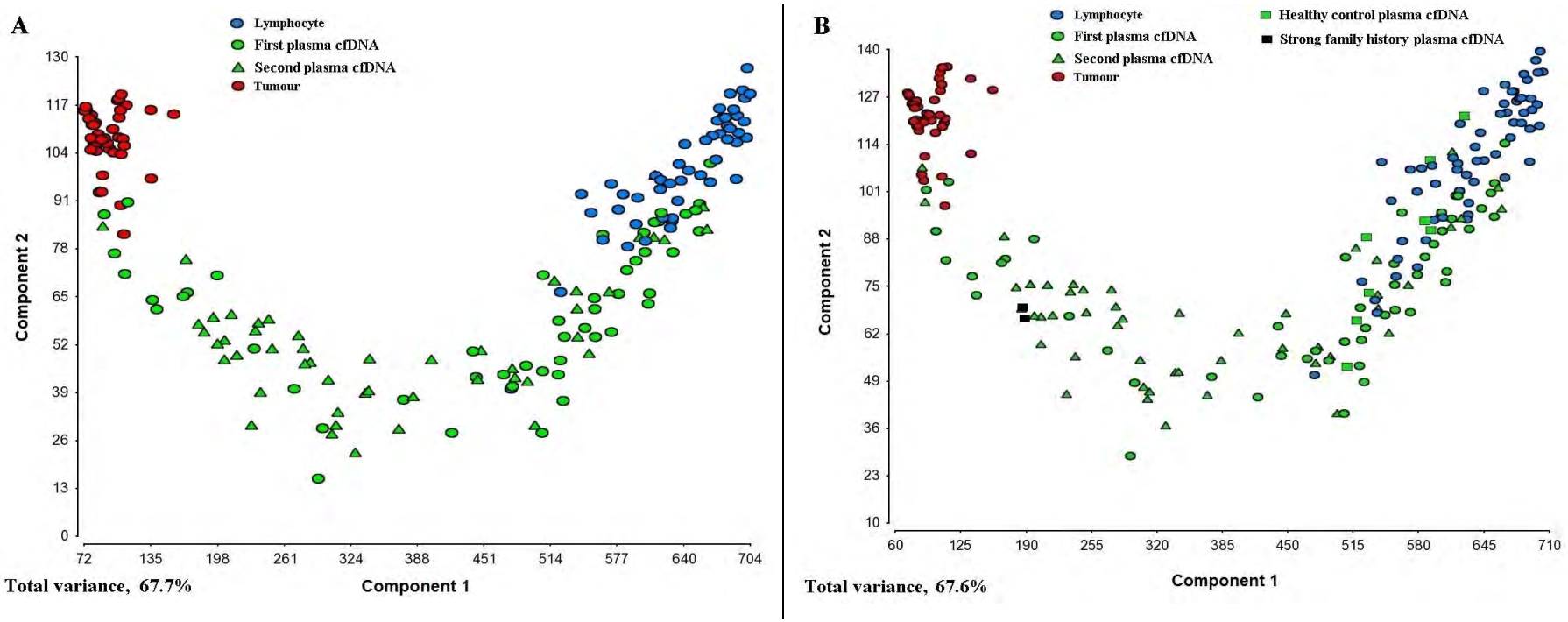
Personalis EX2155

In the patients on follow-up, PCA revealed clear separation between tumour and lymphocyte DNA samples (Figure 3-18A). However, plasma cfDNA samples were widely distributed: some clustered close to the tumours; others close to the lymphocytes; while others were scattered in-between. Overall, there were more second cfDNA samples clustered close to the tumour samples, and cfDNA samples that clustered with the lymphocyte cluster had high SNP concordances when compared to their matched lymphocyte.

When results for the healthy controls (including the two samples from women with a strong family history of breast cancer) were added to the plot, again healthy control cfDNAs clustered with the normal lymphocytes, while the two samples from women with a strong family history of breast cancer grouped separately toward the tumour samples (Figure 3-18B).

Personalis EX2155

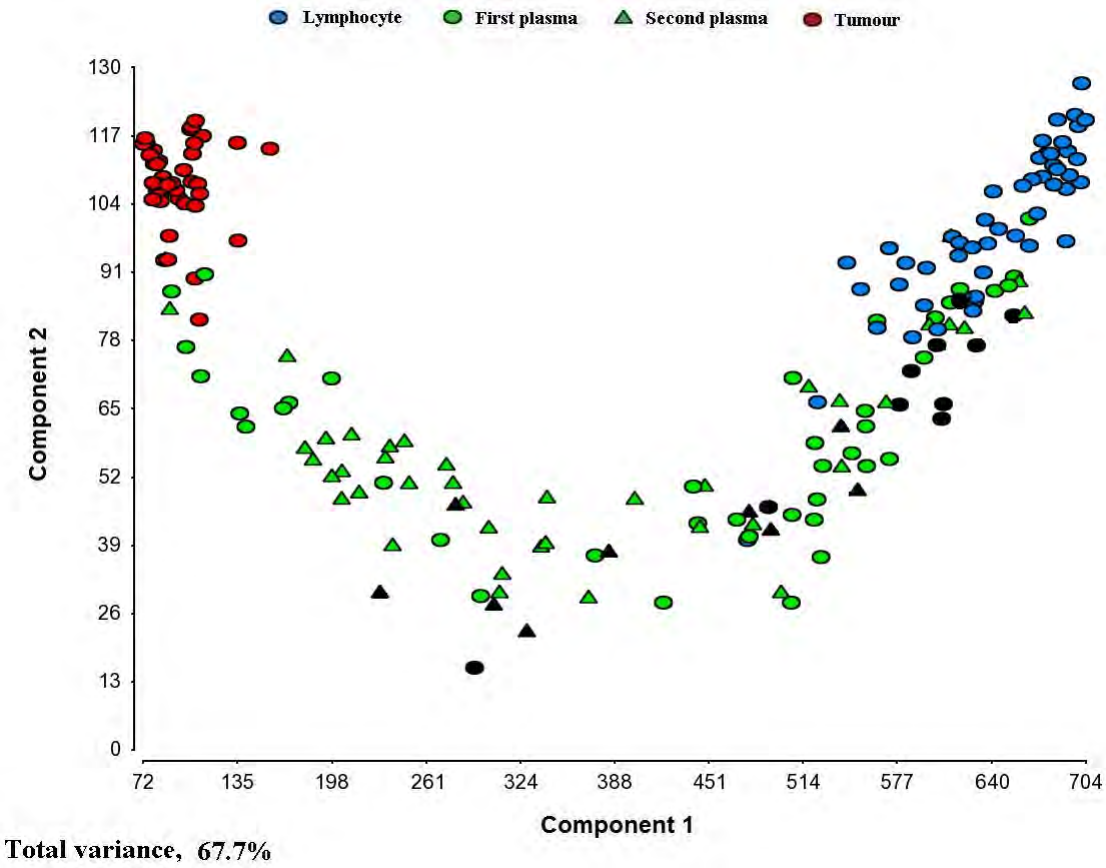
Figure 3-18 - PCA bi-plots of on follow-up group (A) and on follow-up group with healthy controls. Two plasma cfDNA samples from healthy women with a strong family history are also shown (B).



Personalis EX2155

Using data for the patients on follow-up, other PCA bi-plots were generated to elucidate differences between the type of surgery and therapy received, and hormone receptor status. There were no differences identified between these. However, when a PCA bi-plot was generated to observe the distribution of samples in relation to triple-negative (n=10) and receptor-positive (n=40) patients (Figure 3-19), neither the first nor second cfDNA samples clustered near the tumour DNA samples, instead clustering towards the normal lymphocytes.

Figure 3-19 - PCA bi-plot of patients on follow-up with triple-negative patients' cfDNA samples shaded black.



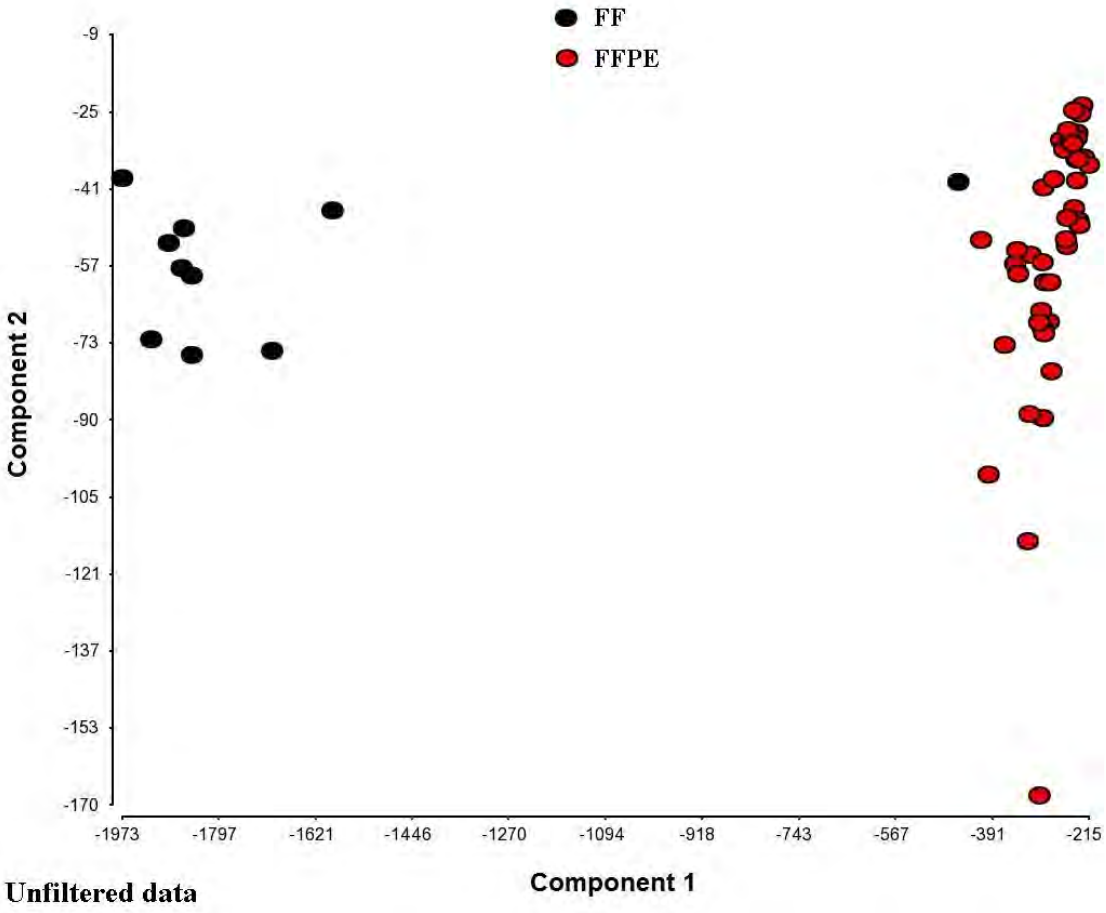
The same PCA bi-plot revealed the location of the relapsed cfDNA samples in relation to those of the non-relapsed patients for the first and second plasma cfDNA samples. The first and second cfDNA from non-relapsed patients spanned the width of the bi-

Personalis EX2155

plot, whereas those of the relapsed patients remained grouped together, toward the lymphocytes, in both cases.

It was additionally decided to analyse through PCA the differences that existed between all FF and FFPE tumour DNA samples in the study. Clear separation was evident (Figure 3-20[1]).

Figure 3-20 - PCA bi-plot of tumour samples (unfiltered data) highlighting separation between FF and FFPE tissue.



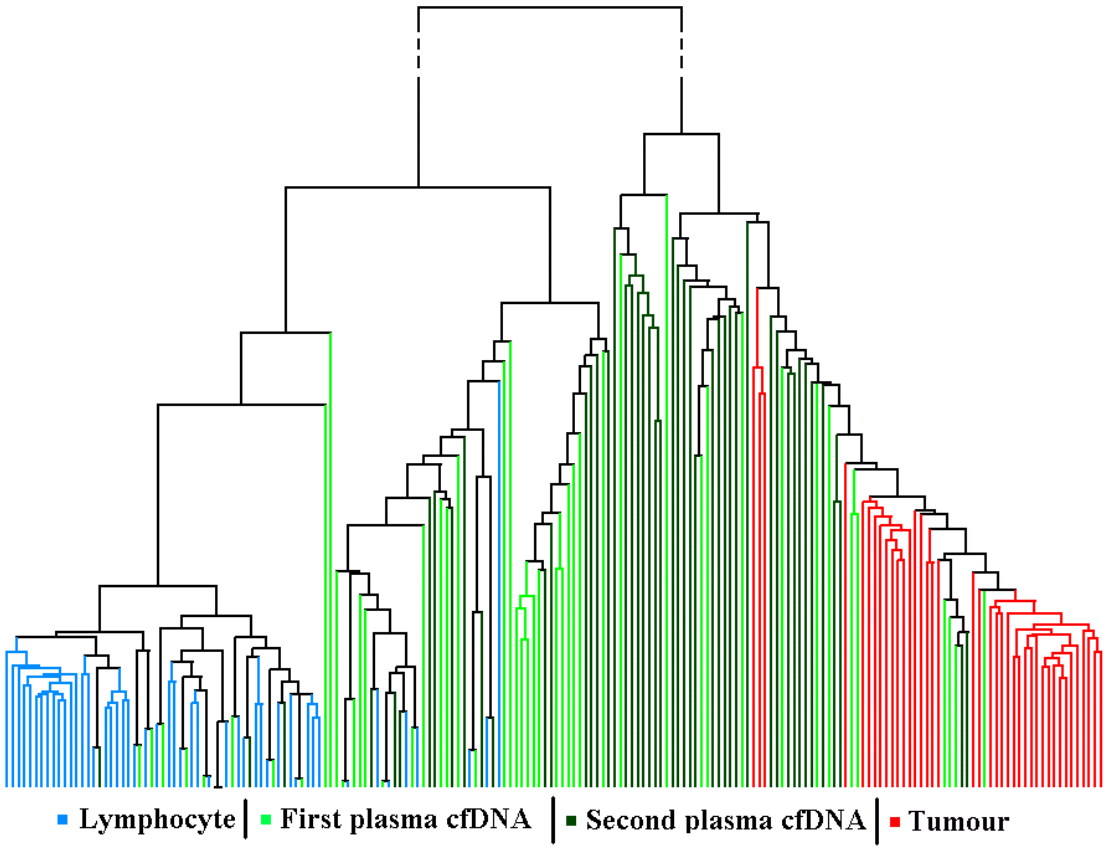
[1] Note - outlier FFPE tumour 61T excluded.

Personalis EX2155

3.3.5.2 Hierarchical clustering

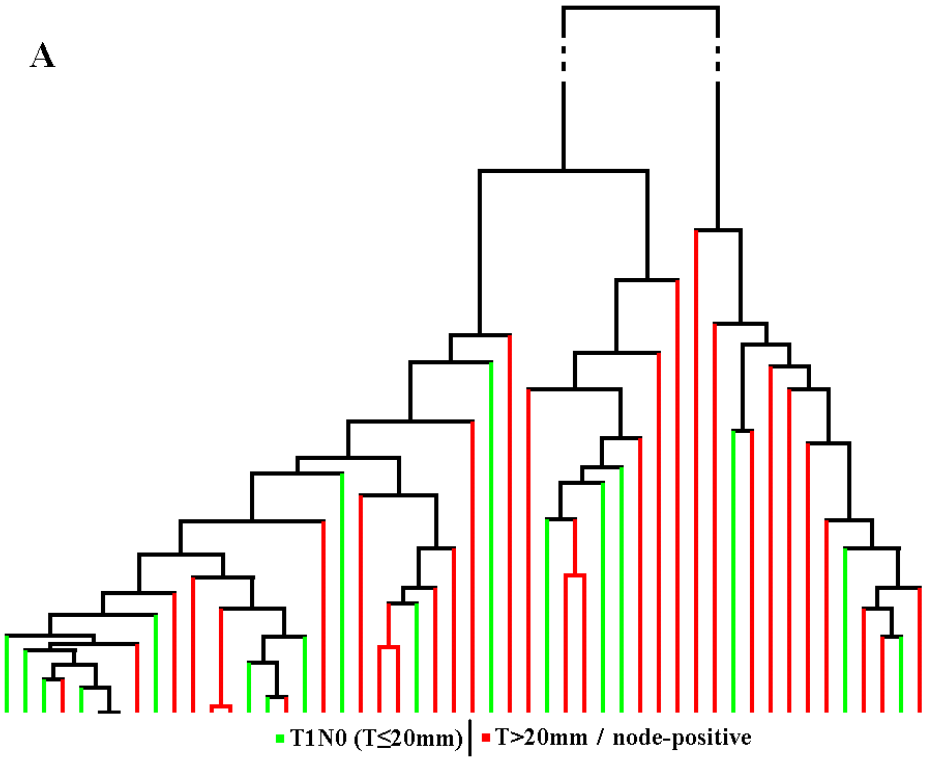
Hierarchical clustering of data for the 50 patients on follow-up revealed a similar separation of samples as PCA, with the tumour and lymphocyte DNA samples clearly separated, and cfDNA samples mostly between these (Figure 3-21). Two main clusters were identified: all tumour DNA with mostly second cfDNA samples; and all lymphocyte with mostly first cfDNA samples. When observing just the first and second cfDNA samples (Figure 3-22A and B), there was a clear division in the dendrogram structure in each case, but no apparent differences between the T_1N_0 ($T \leq 20mm$) and T>20mm/node-positive patients were evident. When observing all tumour DNA samples, the dendrogram lacked any high-level branching structure, indicating that there were no apparent differences between the subgroups (Figure 3-22C).

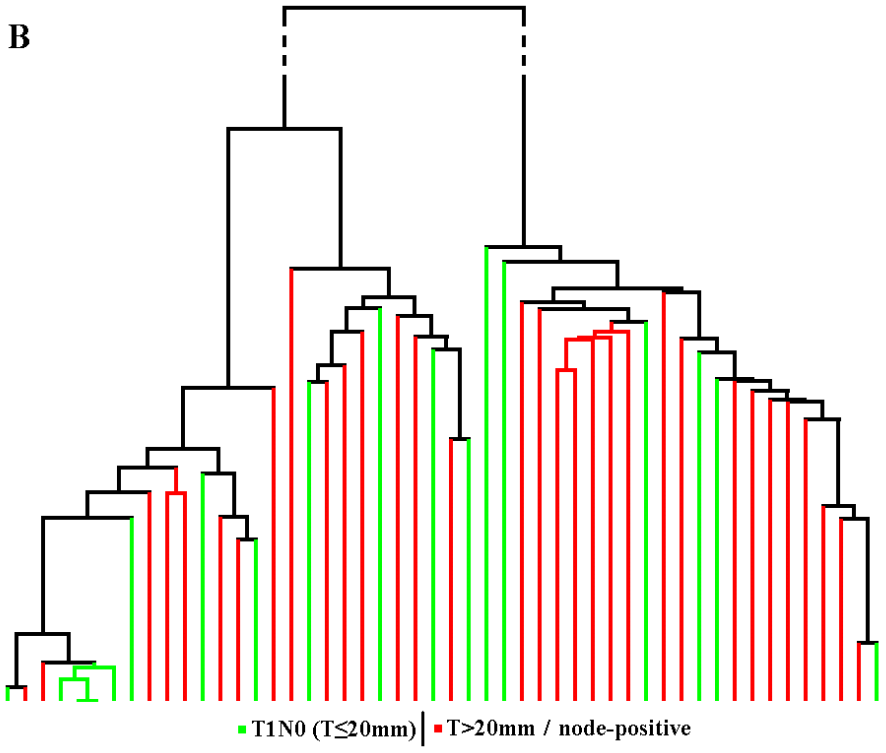
Figure 3-21 - Hierarchical clustering of on follow-up group revealed a division between tumour and lymphocyte DNA samples.



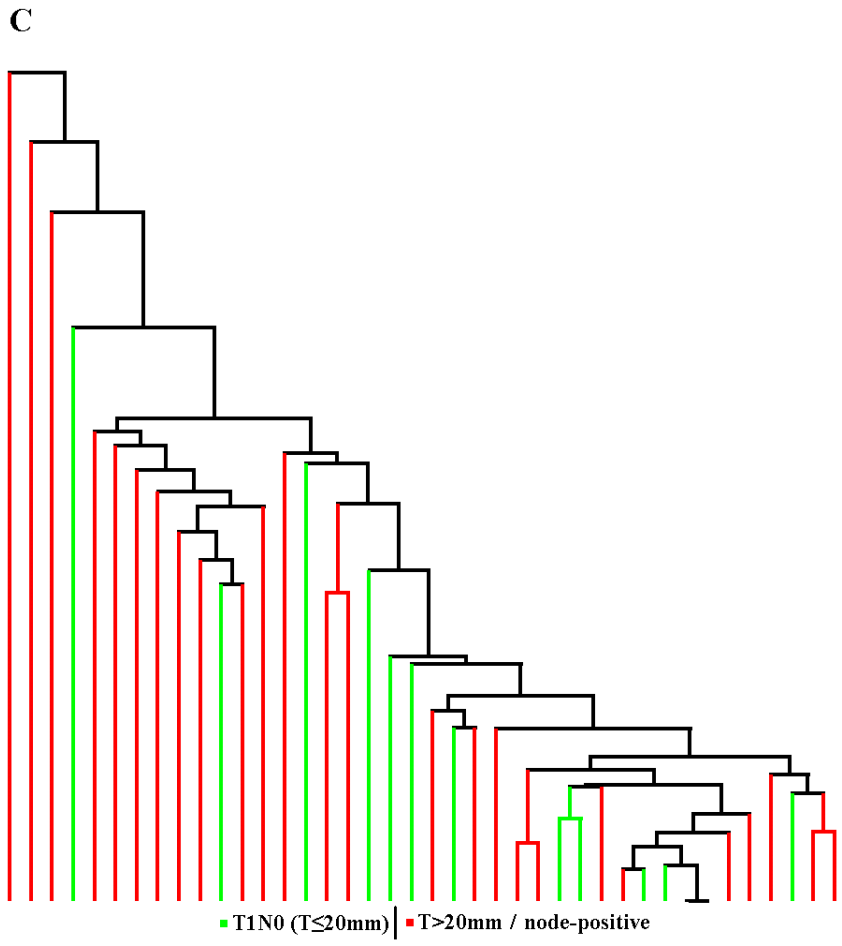
Personalis EX2155

Figure 3-22 - Hierarchical clustering of on follow-up group's first (A) and second (B) plasma cfDNA samples revealed two major clusters, but no apparent separation based on subgroup. Clustering for tumour DNA samples (C) neither revealed major divisions between subgroups.





Personalis EX2155



3.3.6 Loss of heterozygosity analysis

There was heterogeneity in LOH amongst patients and samples. Between plasma cfDNA and matched tumour, the extent of overlapping LOH ranged from just 10-35%. When solely focussing on exons in cfDNA, there were 34 LOH events that occurred in 2 or more of the 50 patients on follow-up (first and second cfDNA samples combined into one group), whilst 36 occurred in 2 or more of the 15 pre-surgical patients' cfDNA samples (Table 3-10). There was generally more LOH detected in T≥20mm/node-positive patients than T_1N_0 (T≤20mm) patients, and an overall increase in LOH detected between the first and second cfDNA samples. Through analysing both CN and LOH data concurrently revealed that 1.47% of CNVs exhibited copy neutral-LOH (CN-LOH).

Personalis EX2155

On follow-up P1					
Chromosome	**Start position**	**End position**	**Cytoband**	**Length (bp)**	**Affected genes**
1	21047616	21114833	1p36.12	67217	*HP1BP3* (-) Overlaps with 17.14% of *SH2D5* (-)
1	47055993	47068677	1p33	12684	Contained within *MKNK1* (-)
1	182542082	183804609	1q25.3	1262527	*APOBEC4* (-), *ARPC5* (-), *DHX9* (+), *LAMC1* (+), *LAMC2* (+), *LOC284648* (+), *NCF2* (-), *NMNAT2* (-), *NPL* (+), *RGS16* (-), *RGS8* (-), *RNASEL* (-), *SMG7* (+), *SHCBP1L* (-) Overlaps with 15.79% of *RGL1* (+)
2	96787899	97172125	2q11.1-2q11.2	384226	*ASTL* (-), *CIAO1* (+), *DUSP2* (-), *ITPRIPL1* (+), *LOC285033* (+), *NCAPH* (+), *SNRNP200* (-), *STARD7* (-), *TMEM127* (-) Overlaps with 2.28% of *NEURL3* (-)
4	138229548	139308745	4q28.3	1079197	*PCDH18* (-), *SLC7A11* (-)
5	37604905	37616554	5p13.2	11649	Contained within *WDR70* (+)
6	71556566	71565478	6q13	8912	Contained within *SMAP1* (+)
7	52095655	53331026	7p12.1	1235371	*POM121L12* (+)
7	137005809	137059416	7q33	53607	Overlaps with 42.42% of *PTN* (-)
8	79028009	84510629	8q21.12-8q21.13	5482620	*CHMP4C* (+), *FABP12* (-), *FABP4* (-), *FABP5* (+), *FABP9* (-), *FAM164A* (+), *HEY1* (-), *IL7* (-), *IMPA1* (-), *MRPS28* (-), *PAG1* (-), *PKIA* (+), *PMP2* (-), *SLC10A5* (-), *SNX16* (-), *STMN2* (+), *TPD52* (-), *ZBTB10* (+), *ZFAND1* (-), *ZNF704* (-)
8	120825226	120883743	8q24.12	58517	*DSCC1* (-) Overlaps with 33.92% of *TAF2* (-)
9	30995741	32798727	9p21.1	1802986	*ACO1* (+), *DDX58* (-), *LOC100129250* (+), *NDUFB6* (-), *TAF1L* (-), *TMEM215* (+), *TOPORS* (-)
9	32798728	33473893	9p21.1-9p13.3	675165	*APTX* (-), *AQP3* (-), *AQP7* (-), *B4GALT1* (-), *BAG1* (-), *CHMP5* (+), *DNAJA1* (+), *NFX1* (+), *SMU1* (-), *SPINK4* (+) Overlaps with 1.86% of *NOL6* (-)
9	137610423	137666414	9q34.3	55991	Contained within *COL5A1* (+)
9	137807802	137834125	9q34.3	26323	Overlaps with 7.62% of *FCN1* (-)
10	16905500	17520442	10p13-10p12.33	614942	*LOC100128098* (+), *ST8SIA6* (-), *TRDMT1* (-), *VIM* (+) Overlaps with 43.31% of *CUBN* (-)
11	44283722	44302311	11p11.2	18589	Contained within *ALX4* (-)
12	58010163	58010164	12q13.3	1	Contained within *ARHGEF25* (+)
13	49559791	50791372	13q14.2	1231581	*ARL11* (+), *C13orf1* (-), *CAB39L* (-), *CDADC1* (+), *CTAGE10P* (+), *DLEU1* (+), *DLEU2* (-), *EBPL* (-), *FNDC3A* (+), *KCNRG* (+), *KPNA3* (-), *MIR15A* (-), *MIR16-1* (-), *MIR3613* (-), *MLNR* (+), *PHF11* (+), *RCBTB1* (-), *SETDB2* (+), *ST13P4* (+), *TRIM13* (+)
13	50791373	52422211	13q14.2-13q14.3	1630838	*DHRS12* (-), *DLEU7* (-), *FAM124A* (+), *GUCY1B2* (-), *INTS6* (-), *NCRNA00282* (-), *RNASEH2B* (+), *SERPINE3* (+), *WDFY2* (+)
16	49821501	49825482	16q12.1	3981	Contained within *ZNF423* (-)
16	84925626	85139173	16q24.1	213547	*KIAA0513* (+), *ZDHHC7* (-) Overlaps with 3.38% of *FAM92B* (-) Overlaps with 8.19% of *CRISPLD2* (+)
21	39235347	39367443	21q22.13	132096	Overlaps with 40.39% of *KCNJ6* (-)

Personalis EX2155

On follow-up P2					
Chromosome	**Start position**	**End position**	**Cytoband**	**Length (bp)**	**Affected genes**
2	50034686	50977979	2p16.3	943293	*NRXN1* (-)
3	123019763	123435754	3q21.1	415991	*MYLK* (-), *PTPLB* (-) Overlaps with 27.75% of *ADCY5* (-)
5	95084131	95084132	5q15	1	Contained within *RHOBTB3* (+)
6	38140658	38140659	6p21.2	1	Contained within *BTBD9* (-)
6	78066685	78366842	6q14.1	300157	*HTR1B* (-)
9	134612965	135042885	9q34.13	429920	*MED27* (-) Overlaps with 1.29% of *NTNG2* (+)
11	103792043	103808152	11q22.3	16109	Contained within *PDGFD* (-)
13	56872356	59648355	13q21.1-13q21.2	2775999	*PCDH17* (+), *PRR20A* (+), *PRR20B* (+), *PRR20C* (+), *PRR20D* (+), *PRR20E* (+)
17	30114328	31319764	17q11.2	1205436	*ARGFXP2* (-), *C17orf75* (-), *C17orf79* (-), *CDK5R1* (+), *LRRC37B* (+), *MIR632* (+), *MYO1D* (-), *PSMD11* (+), *RHBDL3* (+), *RHOT1* (+), *SH3GL1P1* (+), *SUZ12* (+), *TMEM98* (+), *UTP6* (-), *ZNF207* (+) Overlaps with 0.07% of *SPACA3* (+)
22	19643245	19775287	22q11.21	132042	*GP1BB* (+), *SEPT5* (+), *GP1BB* (+), *TBX1* (+)
22	38594668	38678564	22q13.1	83896	*MAFF* (+), *TMEM184B* (-)

Pre-surgical P					
Chromosome	**Start position**	**End position**	**Cytoband**	**Length (bp)**	**Affected genes**
1	19698026	19707302	1p36.13	9276	*CAPZB* (-)
1	39355371	39360730	1p34.3	5359	*RHBDL2* (-)
1	79024688	79126050	1p31.1	101362	*IFI44L* (+)
1	89472197	89558103	1p22.2	85906	*GBP1* (-), *GBP3* (-)
1	237978001	237997087	1q43	19086	*RYR2* (+)
2	38281096	38298139	2p22.2	17043	Overlaps with 19.91% of *CYP1B1* (-) Overlaps with 77.39% of *FAM82A1* (+)
2	38298139	38306935	2p22.2	8796	Overlaps with 58.95% of *CYP1B1* (-)
2	62413989	62495209	2p15	81220	*B3GNT2* (+)
2	127446853	127464384	2q14.3	17531	Overlaps with 42.18% of *GYPC* (+)
3	38350519	38350543	3p22.2	24	*SLC22A14* (+)
3	65530707	65641243	3p14.1	110536	*MAGI1* (-)
3	132533091	132548645	3q22.1	15554	*NCRNA00119* (+)
3	132548646	132594820	3q22.1	46174	Overlaps with 96.17% of *NCRNA00119* (+)
4	10528226	10559817	4p16.1	31591	*CLNK* (-)
4	79347625	79443850	4q21.21	96225	*FRAS1* (+)
5	177977679	178084843	5q35.3	107164	*CLK4* (-) Overlaps with 37.21% of *COL23A1* (-)
5	178084843	178179234	5q35.3	94391	*ZNF354A* (-)
6	12654540	13005284	6p24.1	350744	Overlaps with 81.95% of *PHACTR1* (+)
6	13025819	13063824	6p24.1	38005	*PHACTR1* (+)
6	37595546	37609305	6p21.2	13759	Overlaps with 65.57% of *MDGA1* (-)
6	152521244	152826222	6q25.2	304978	*SYNE1* (-)

Personalis EX2155

8	17532573	17535058	8p22	2485	*MTUS1* (-)
8	87221488	87241319	8q21.3	19831	Overlaps with 75.80% of *SLC7A13* (-)
10	14519892	14638939	10p13	119047	Overlaps with 65.84% of *FAM107B* (-)
10	70502288	70502289	10q21.3	1	*CCAR1* (+)
11	19351515	19401124	11p15.1	49609	Overlaps with 58.16% of *NAV2* (+)
11	47969152	55683985	11p11.2-11q11	7714833	*FOLH1* (-), *LOC440040* (+), *LOC441601* (-), *LOC646813* (+), *OR4A15* (+), *OR4A16* (+), *OR4A47* (+), *OR4A5* (-), *OR4B1* (+), *OR4C11* (-), *OR4C12* (-), *OR4C13* (+), *OR4C15* (+), *OR4C16* (+), *OR4C3* (+), *OR4C45* (-), *OR4C46* (+), *OR4C6* (+), *OR4P4* (+), *OR4S1* (+), *OR4S2* (+), *OR4X1* (+), *OR4X2* (+), *OR5D13* (+), *OR5D14* (+), *OR5D16* (+), *OR5D18* (+), *OR5L1* (+), *OR5L2* (+), *OR5W2* (-), *PTPRJ* (+), *SPRYD5* (+), *TRIM48* (+)
13	74054816	74327204	13q22.1	272388	Overlaps with 24.62% of *KLF12* (-)
13	74327205	74474851	13q22.1	147646	*KLF12* (-)
14	31895975	32247665	14q12	351690	*C14orf126* (-), *GPR33* (-) Overlaps with 61.72% of *NUBPL* (+)
16	2330169	2336394	16p13.3	6225	*ABCA3* (-)
16	19646235	19691514	16p12.3	45279	*C16orf62* (+)
16	31102589	31102590	16p11.2	1	*VKORC1* (-)
16	81223646	81237214	16q23.2	13568	*PKD1L2* (-)
16	81237214	81242637	16q23.2	5423	*PKD1L2* (-)
22	30450660	30486826	22q12.2	36166	Overlaps with 28.68% of *HORMAD2* (+)

Table 3-10 - LOH detected in the cfDNA of pre-surgical breast cancer patients and patients on follow-up. For each patient, normal lymphocyte DNA was used as reference and LOH was called using a 6-state hidden Markov model (HMM) (see Materials and Methods 2.7.5). Only LOH in exons were reported and the DNA strand (+, coding; -, non-coding) for each affected gene is indicated in parentheses. P, plasma cfDNA of pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up.

3.4 Discussion

The aims of this chapter were to

1. Demonstrate that amplified *ERBB2* is detectable in the cfDNA of ERBB2+++ breast cancer patients.
2. Compare *Affymetrix SNP 6.0* whole genome profiles of the primary tumour with paired plasma cfDNA samples taken from breast cancer patients on follow-up, and relate findings to cfDNA profiles of samples taken pre-surgically from breast cancer patients and from those of healthy female controls.

Personalis EX2155

3.4.1 *ERBB2* assay validation experiments using breast cell-lines

The results of analysing *ERBB2* CN in different cell-lines showed that RQ values can vary widely depending on the endogenous control used. The purpose of using an endogenous control is to ensure consistency in results across different DNA samples using a known CN locus. A control on the same chromosome as that of the target gene should ideally be used. *CNTNAP1* is located on chromosome 17 (like *ERBB2*), while *RPPH1* is on chromosome 14. *RPPH1* results varied by cell-line as several harboured a deletion of this gene, thus *CNTNAP1* seemed to be the more stable gene. Similar RQ values were obtained when analysing *ERBB2* using either *CNTNAP1* or *GADPH*, another reference that maps to chromosome 12 (Page *et al.*, 2011).

The results of *ERBB2* CN analysis showed reproducibility when using 3, 5, and 9 replicates in qPCR ($p<0.0001$), thus proving the reproducibility and sensitivity of the *ERBB2* qPCR assay (Page *et al.*, 2011). The amplification of *ERBB2* in tumour and matching cfDNA was shown previously (Chiang *et al.*, 1999 and Page *et al.*, 2011). However, due to low yield of DNA from blood (Xue *et al.*, 2009), it is important to ensure that an assay can reproducibly and sensitively detect amplification in as low amounts of sample as possible. Thus, as there were no significant differences between 3, 5, and 9 replicates, it was deduced that triplicate reactions should be pursued in further studies in order to spare limiting DNA.

3.4.2 Formalin-fixed Vs fresh frozen tissue

One important finding of this study was the difference in quality of *Affymetrix SNP 6.0* data that is generated between FFPE and FF tumour tissue. Genomic instability is a feature of cancer (Gilley *et al.*, 2005 and Naylor *et al.*, 2005) and can drive disease progression (Uchida, 2005). This instability poses additional problems for analysing lower integrity and quality DNA, such as that which is typically extracted from FFPE tumour tissues, by PCR (Howe *et al.*, 1997) and microarray (Hostetter *et al.*, 2010 and Oikawa *et al.*, 2011). A major problem that can occur when tissue undergoes the FFPE process is the cross-linking of nucleic acids that can leave DNA molecules of less than 200bp in length (Blow, 2007).

Personalis EX2155

In this study, only 389 CNVs, on average, were called per sample in the FFPE on follow-up tumours (n=40). However, in the pre-surgical FF tumours (n=10), the average was 2,515 per sample. The QC data for these also revealed differences (Table 3-4). It is documented that DNA fragments from FF tumour DNA are of a higher molecular weight than those from their matched FFPE counterparts (Tuefferd *et al.*, 2008). Hence, better quality DNA fragments appear to result in greater efficiency, with regard to CNV calling. All data generated from FFPE tissues was therefore regarded as preliminary, with careful comparison to FF and cfDNA results. It is thus important to determine the optimum methods for using FFPE tissue that can give the greatest confidence in results. For example, WGA might not accurately amplify FFPE DNA (Jasmine *et al.*, 2008).

Another study also involving the *Affymetrix SNP 6.0* microarray showed good concordance between FF and matched FFPE tissues, with the highest concordance existing between those pairs with the highest SNP call-rates (Tuefferd *et al.*, 2008). In the study, in order to improve the quality of the FFPE DNA, it was necessary to alter the DNA extraction and PCR amplification steps (Tuefferd *et al.*, 2008).

Overall, FF is a useful reference against which FFPE tissue can be compared in future studies aimed at finding effective ways of using FFPE tissue.

3.4.3 Single nucleotide polymorphism concordance segregated cfDNA of healthy controls from breast cancer patients

The results have clearly shown that microarray technology can significantly distinguish between women with breast cancer and healthy controls ($p<0.0001$; Figure 3-7) through the analysis of plasma cfDNA SNP profiles. This could potentially pave the way for a simple blood test to diagnose breast cancer if these results were validated in another patient cohort.

In healthy controls, the high % SNP concordance between plasma cfDNA and matched lymphocyte DNA (Figure 3-6) was taken as indicating a status of health or as having a 'normal' genomic profile for plasma cfDNA. Also, using healthy control plasma

Personalis EX2155

cfDNA samples as baseline, the first and second cfDNA samples of the patients on follow-up significantly differed from each other ($p=0.0006$), highlighting that the plasma cfDNA profiles of breast cancer patients on follow-up change with time.

Additionally, the results demonstrated a significant difference ($p<0.05$; Figure 3-8) between the cfDNA profiles of pre-surgical breast cancer patients and the first and second cfDNA samples of patients on follow-up. This is a remarkable finding, as the stage of the disease differs between these groups, and appeared to be reflected by different cfDNA SNP profiles. In a previous study by Närvä *et al.* (2010), it was found that SNP profiles could also be used to distinguish different human embryonic stem cell (hESC) lines. In our study, it has been shown that they can be used to distinguish different stages of breast cancer: pre-surgery when the tumour is still in the body and post-surgery when the person is purportedly disease-free.

Taking these findings together, these data support other research that indicates that the cancer genome can evolve with metastatic disease (Gangnus *et al.*, 2004). It is therefore possible that cfDNA analysis can function as a useful way of monitoring the evolving cancer genome.

3.4.4 Principal component analysis reveals ethnicity and may predict breast cancer

The PCA bi-plot incorporating all lymphocyte DNA samples investigated in this study revealed no obvious stratification between patients and controls (Figure 3-16A). This confirms that lymphocyte DNA can be used as a normal reference, even when from a patient with cancer. The 'homogeneity' of the profiles between samples used in this study was again highlighted through Figure 3-16B. However, ethnical differences that may influence breast cancer development (Palmieri *et al.*, 2008) were also evident. There were three main groupings elucidated: 1) Nigerian Yoruba; 2) Chinese Han and Japanese; and 3) American residents of north- and western-European descent. The majority of lymphocyte samples from this study grouped with the North- and Western-Europeans. However, as they were from Korea, 7 NeoCEnT samples grouped with the Chinese/Japanese samples. Two other samples grouped toward the Nigerian samples,

Personalis EX2155

possibly indicating African descent in these women. Overall, the spread of profiles between lymphocyte samples investigated in this study was wider than for the specific ethnic groups observed from Nigeria, China, Japan, and America, suggesting a more ethnically diverse population of women who provided samples to this study.

Principal component analysis also revealed a striking contrast between tumour and lymphocyte DNA samples in the patients on follow-up. The tumour DNA samples of these 50 patients clustered tightly, suggesting homogeneity (Figure 3-18). However, as these samples were all from FFPE tissue, the homogeneity observed may also reflect the lower amount of CNV data generated from these samples.

The finding that two cfDNA samples from healthy women with a strong family history of breast cancer (but had normal mammography) clustered with cfDNAs from the pre-surgical group by PCA (Figure 3-17) is of high importance. These data support a recent study by Rennstam *et al.* (2010) who showed that genomic 'aberrations' can be found in *BRCA1* mutation carriers with histopathologically-normal breast tissue. Clinical data has been provided recently that confirms that these 2 cfDNA samples were from women with a germ-line *BRCA1* mutation. Hence, analysis of further samples from known mutation carriers by PCA could help to confirm or refute this finding. Preliminary analysis focussing on the specific differences between the 2 healthy *BRCA1* carriers and healthy controls revealed striking separation.

3.4.5 Copy number variation

Copy number variation is common in the healthy human genome (Redon *et al.*, 2006) (Figure 3-12 and Figure 3-13). However, in cancer, instability at a particular CNV could confer advantage and promote disease progression. Thus, it is important to analyse microarray data against suitable references or controls that are known to be free of the disease being analysed. By this approach, it is possible to identify CNVs that are common to the disease and that are not found in healthy individuals.

In this study, sample data were initially referenced against the International HapMap samples (International HapMap Project, http://hapmap.ncbi.nlm.nih.gov/), which

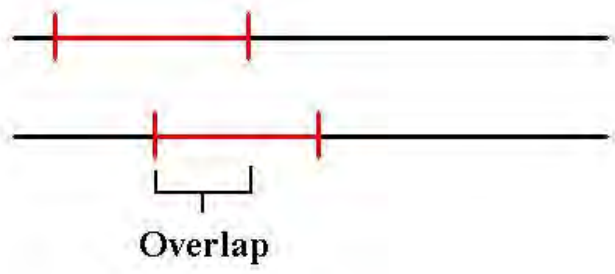
Personalis EX2155

consists of 270 human DNA samples. Subsequent comparisons between the healthy controls (n=8) and each patient's matched lymphocyte sample were further carried-out in order to filter-out data that was considered 'normal', thereby leaving data that related solely to the disease. Importantly, the lymphocyte samples were shown to be normal by PCA (Figure 3-16) and SNP concordance checks, thereby permitting their use as useful references.

3.4.5.1 Many novel CNVs

Through the overlapping of CNVs discovered in this study with known variants listed in the Toronto DGV, 45% and 46.48% of the first and second plasma cfDNA samples, respectively, of the patients on follow-up were not listed previously. For the tumour samples, the figure was slightly greater than half (50.05%). However, when interpreting these results, it is important to note that a group of patients was considered as a whole/single entity. Hence, CNVs that overlapped with each other across samples in the same group would have been counted separately. The phenomenon occurred due to the method of genomic segmentation making calls on a per sample basis and not correlating segments intra-patient (Figure 3-23). However, the extent of overlap can be elucidated better with more defined breakpoints between segments (Conrad *et al.*, 2009), which would also help the development of targeted approaches such as qPCR. Defining a minimum overlap between segments could also be a solution. Nevertheless, it was not anticipated that these would be problematic once results were averaged and examined proportionally (as percentage values).

Figure 3-23 - Overlap between two CN segments. Black lines represent DNA sequences from two patients; CN segments indicated by red bars.



Personalis EX2155

The amount of novel variation found in this study was surprising. Much of the data in the Toronto DGV is derived from projects that have used single technologies with no reproducible, parallel studies to confirm findings. Additionally, there is evidence that the size of a CNV is less than that recorded by the database (Conrad *et al.*, 2009 and Duclos *et al.*, 2011). These 2 notions combined imply that much of the data listed in the Toronto DGV is potentially erroneous (Duclos *et al.*, 2011). Couple this with the fact that genomic positions of genes can differ with progressive releases of the human genome makes for a database that is in need of standardisation and updating. As array technology improves in resolution and high throughput sequencing becomes more common, the extent of error in the Toronto DGV and other such databases will continue to be updated and errors in the past recording of genomic variation will be exposed.

3.4.5.2 Amplifications and deletions

The proportion and distribution of amplifications and deletions detected by sample group were also compared in this thesis. A recent study by Närvä *et al.* (2010) showed that amplifications comprised the majority of CNVs in hESCs, a finding also noted by Naylor *et al.* (2005) who looked in breast tissues. The same was true for the majority of samples analysed in this study. Deletions were more common in the second cfDNA samples of the patients on follow-up and in the pre-surgical patients' cfDNA samples, which may reflect the nature and quality of cfDNA or the evolving circulating breast cancer genome.

One critical finding from these data, which demonstrate amplified and deleted areas of the genome, is that plasma cfDNA characterisation may provide important information for clinicians in choosing subsequent therapies through the identification of which gene products to target. Amplifications and deletions pose different types of problems: amplified genes can represent candidate oncogenes, while deleted genes can represent the loss of tumour suppressors (LaFramboise, 2009). Amplification can be used as a clinical prognostic factor and is a mechanism for oncogenesis, tumour progression, and can also give rise to drug resistance (Uchida, 2006). The dosage level of a gene could be increased, in the case of a high level amplification (Grayson & Aune, 2011), and

Personalis EX2155

have consequences on the resulting phenotype (Uchida, 2006); conversely, a large deletion could result in no gene transcription.

However, there are conflicting results in relation to how CNV affects gene expression or 'dosage' sensitivity (Schuster-Böckler *et al.*, 2010). For example, Schuster-Böckler *et al.* (2010) report that some genes were 'expressionally-silenced' due to gene amplification in previous studies, yet their own results contrasted these. They also reported that deletions have resulted in higher gene expression and suggested a compensatory mechanism at the cellular level.

However, the exact mechanisms that lead to gene amplification are uncertain, and the extent to which gene amplification plays a causal role in cancer is equally uncertain. Profiling/mapping amplifications in cancer can help to identify putative genes involved (Uchida, 2006). Duplications are more likely than deletions to be formed by non-allelic homologous recombination (NAHR), variable number of tandem repeats (VNTRs), and retrotranspositions, and these events are likely to be sequence dependent (Conrad *et al.*, 2009). Throughout this study, amplifications were found to be more prevalent than deletions, and the implication is that certain regions of the genome are predisposed to be affected by CNV (due to their sequence) indirectly through events such as NAHR, VNTR formation, and retrotranspositions. In such a mechanism, increased genomic instability would result in an increased number of amplifications. Could a variant SNP be the driver of such events? - mapping SNPs to CNVs could provide insight into what role SNPs have in driving CNV.

3.4.5.3 Commonly-amplified intervals in plasma cfDNA

Through the analysis of the 8 commonly-amplified CNVs (Table 3-8), it was shown that tumour-specific alterations are able to persist in the cfDNA extracted from blood plasma up to 12 years after surgery to remove the primary tumour (Figure 3-14). The fact that these alterations were identical in >90% of the tumour DNA samples indicates their importance in breast cancer. These intervals could therefore provide novel diagnostic information to be used in the future, especially as some of these CNVs were

Personalis EX2155

shown to contribute significantly to cancer status by binary logistic regression analysis in both the first and second cfDNA samples (Table 3-9).

Moreover, some of these CNVs were found to be significant in discriminating between disease and healthy, relapsed and non-relapsed, and size and nodal status (Table 3-9). A previous study by Naylor *et al.* (2005)[1] tried to do the same through the analysis of the frequency of CN alterations in breast tumour DNA but failed to demonstrate any significant results. However, in this study, it was shown that this could be achieved conclusively through focussing on certain altered regions in the genome.

3.4.5.4 Receptor status

Based on receptor status of primary tumour of patients on follow-up, 5p11 contributed significantly to ER-positive status in the second cfDNA samples and to triple-negative status. There were no significant differences when comparing PR and ERBB2 receptor statuses (Table 3-9). Naylor *et al.* (2005) also attempted to correlate receptor status to frequency of CN alterations observed in breast tumour DNA but showed no significant differences.

3.4.5.5 Profiles of relapsed patients

When taking the number of CNVs into account (amplifications and deletions combined), there were few differences between groups. The number of CNVs called in a sample could be a representation of sample quality, with an average of 1,163 and 1,168 being called in healthy control lymphocyte and plasma cfDNA. Nevertheless, when statistically comparing the numbers of CNVs between sample types, a significant difference existed in the patients who relapsed, when comparing their lymphocytes to first and second cfDNA samples ($p=0.0016$ and <0.05 respectively) on a paired sample basis.

[1] Comparison between the CNVs found by Naylor *et al.* (2005) and those in this thesis is made in section 6.4.4.3.

Personalis EX2155

Differences between relapsed and non-relapsed patients' plasma cfDNA samples were also evident by PCA. The relapsed first cfDNA samples grouped with the normal lymphocyte DNA samples, whilst the second cfDNA samples remained anchored in the same location on the bi-plot. This was in contrast to the cfDNA samples of patients who did not relapse, which spanned the bi-plots. These findings confirm that the cancer genome can change with the evolution of metastatic disease and need to be followed-up in other studies.

Importantly, binary logistic regression analysis of the 23 commonly-amplified CNVs (Table 3-6) revealed significant results (Table 3-9): six of these CNVs contributed significantly to relapsed status based on tumour comparisons (4 were at a significance of $p<0.0001$). The same 6 cytobands also contributed significantly to the T>20mm/node-positive status compared to T_1N_0 ($T\leq20$mm). The CNVs at 9q12 and 1p36.33 contributed significantly to relapsed status in the first cfDNA samples, while both of these and 9 others contributed significantly in the second cfDNA samples, again suggesting evolving genomic instability. This suggests that these 11 intervals/CNVs may be involved in the development of metastasis. Moreover, they show that metastatic events can potentially be picked-up by analysing serial plasma cfDNA samples in patients on follow-up and stratifying differences based on disease progression.

3.4.5.6 Surgery and therapy received

Results of the binary logistic regression analysis also showed that the type of surgery and therapy received could potentially affect how a patient on follow-up progresses. For example, when comparing tumour CNV data, there were no significant differences between patients who underwent mastectomy and those who underwent wide local excision. However, significantly-altered intervals detected in plasma cfDNA later appeared to separate the groups: in the first cfDNA samples, 5q13.2; in the second cfDNA samples, 5q13.2 plus 4q13.2 and 14q32.33.

Based on endocrine therapy received or not, a single interval was significantly different when comparing tumour CNV data. This would be expected as differences in tumour histology and receptor status dictate the course of adjuvant therapy received on follow-

Personalis EX2155

up. In addition, further intervals exhibited significant differences in first (10q11.23 and 16p11.2) and second (1p36.21, 5p11, and 19p13.3) cfDNA samples.

3.4.6 Genes overlapping commonly-amplified CNVs

The 8 commonly-amplified CNVs (Table 3-8) and some of the other [in total] 23 intervals (Table 3-6) harbour interesting gene targets that warrant further research.

3.4.6.1 4q13.2

The CNV identified at 4q13.2 includes *TMPRSS11B*, *TMPRSS11BNL*, *UGT2B15*, *UGT2B17*, and *YTHDC1* genes. Copy number variation in these genes has been linked to osteoporosis (Yang *et al.*, 2008), notably *UGT2B17*. Higher *UGT2B17* gene CN was associated with lower bone mineral density and an increased risk of osteoporotic fractures. Additionally, both *UGT2B15* and *-17* have been shown to be up-regulated by 17-β-estradiol in MCF-7 breast cancer cells (Hu & Mackenzie, 2009) and the authors conjectured that these two genes may be involved in maintaining normal steroid hormone homeostasis and preventing excessive oestrogen signalling. In support of this, Yang *et al.* (2008) found that individuals who had a deletion in the *UGT2B17* gene had higher concentrations of testosterone and estradiol, which is important in breast cancer progression, in their blood (Saha Roy & Vadlamudi, in press).

3.4.6.2 5q13.2

Genes overlapping the 5q13.2 CNV include *SERF1A*, *SERF1B*, *SMA5*, *SMN1*, *SMN2*, and *NAIP*. The interval is a known 500Kb inverted duplication that is prone to rearrangements and deletions (Wang & Dreyfuss, 2001 and Huang *et al.*, 2010), with *SERF1A* and *SERF1B*, and *SMN1* and *SMN2*, being gene duplications. Deletion in the region has also been found in *BRCA*-mutated breast tumours (Johannsdottir *et al.*, 2006). NAIP (neuronal apoptosis inhibitory protein) belongs to a family of proteins that inhibit apoptosis (Jedrzejowska *et al.*, 2009). The level of expression of NAIP has been shown to be increased in prostate cancer samples taken from patients receiving androgen deprivation therapy, which is a mechanism used to treat advanced prostate

Personalis EX2155

cancer. This increased expression could be associated with the enhanced survival of the cancer cells (Chiu *et al.*, 2010).

3.4.6.3 7q11.23

Duplication of 7q11.23 is a recognised syndrome that is associated with neurodevelopmental disorders (Mervis, 2008), while 3 genes that overlap this CNV (*CLIP2*, *GTF2IRD1*, and *WBSCR16*) are associated with a deletion that causes Williams-Beuren Syndrome (Merla *et al.*, 2002). However, there are no known associations of these genes with cancer.

3.4.6.4 8p23.1

Genes within this amplified region include *DEFB130*, which encodes a zinc finger protein (ZNF705D), and members of the 'family with sequence similarity' (FAM) group of genes. Of interest, however, the region also includes *DUB3*. DUB3 was shown to be a regulator of the CDC25A phosphatase, which is a protein overexpressed in many human cancers (Pereg *et al.*, 2010). The authors of the study noted that knocking-down the *DUB3* gene resulted in a reduction in the growth of breast tumour xenografts in nude mice; conversely, overexpression of DUB3 resulted in oncogenic activation. Hence, amplification of *DUB3* may lead to CDC25A overexpression and increased oncogenesis.

This cytoband has also been found to be LOH-affected in hepatocellular carcinoma (Lu *et al.*, 2007), but also to have an associated duplication syndrome (Barber *et al.*, 2010). However, none of the genes showing amplification in this study were reported as duplicated/amplified in the duplication syndrome. Additionally, 8p12 has been shown to be commonly-amplified in breast cancer and a recent study identified a potential oncogene within this amplicon (*ZNF703*) using various methods including *Affymetrix SNP 6.0* technology (Holland *et al.*, 2011). In the samples pertaining to this study, *ZNF703* was amplified in 92.5% of the tumours of patients on follow-up and 75% of the tumours of pre-surgical patients. However, there are no markers that map within this

Personalis EX2155

gene on the *Affymetrix SNP 6.0* and the flanking region contains just 4 markers that span a distance of 10Kb.

3.4.6.5 10q11.22 and 10q11.23

The distal portion of chromosome 10q is known to have a very rare associated duplication syndrome that results in partial trisomy (Aalfs *et al.*, 1995 and Aglan *et al.*, 2008), which encompassed both 10q11.22 and 10q11.23 in a patient who had mental retardation (Koivisto *et al.*, 1981). Although these 2 amplifications were seen in healthy controls in this study, it was found at much higher CNs in the patients' samples at both 10q11.22 and 10q11.23.

Of the overlapping genes at 10q11.22, *PTPN20B* results from a duplication of *PTPN20A* in a process that can result in multiple copies of the gene (Pruitt *et al.*, 2005). Other genes overlapping this CNV include *GDF2* and *GDF10*, which are growth factors that regulate cell growth and differentiation. *GDF10* was identified as a paralogue of *BMP3* on chromosome 4 (Katoh & Katoh, 2006). Conversely, deletion of this cytoband has been associated with early onset 'extreme' obesity by *Affymetrix SNP 6.0* technology (Jarick *et al.*, 2011), although the precise affected region did not overlap with the CNV identified in this study.

At 10q11.23, alterations in *PARG* have been linked to thyroid papillary carcinoma, whilst LOH in this gene has also been demonstrated previously in glioblastoma (Shimokawa *et al.*, 1998). More recently, Zhou *et al.* (2011) concluded that the inhibition of the PARG glycohydrolase in combination with the addition of a DNA-alkylating agent could induce cancer cell-death. A separate study concluded that PARG's recruitment to sites of DNA damage contributes to control of the balance between DNA repair and the death of the cell (Mortusewicz *et al.*, 2011). Thus, amplification and overexpression of *PARG* could have the opposite effect by preventing cancer cells apoptosing.

A second gene overlapping this CNV is *CHAT*, which encodes a protein involved in the biosynthesis of the acetylcholine neurotransmitter. Studies have previously searched for

Personalis EX2155

associations of variant alleles within this gene in both Alzheimer's disease and depression (Grünblatt *et al.*, 2009) and nicotine dependence (Ray *et al.*, 2010), with positive associations in both cases. No association with cancer has yet been found.

3.4.6.6 15q25.2

Trisomy of a region including 15q25.2 was found previously in autism (Bonati *et al.*, 2005). This cytoband also has an associated deletion syndrome that can result in an increased risk of birth defects (Wat *et al.*, 2010). However, there are no reported associations of genes overlapping this CNV with cancer.

3.4.6.7 16p12.3

Amplification at 16p12.3 has been found previously in high-grade prostatic intraepithelial neoplasia (Hughes *et al.*, 2006) but has not been reported in breast cancer. One gene that maps within the amplified CNV is *NOMO2*, which is similar to two other genes on the p arm of chromosome 16 that lie in another region of known duplication (Pruitt *et al.*, 2005).

3.4.6.8 Other CNVs with potential genes of interest

The amplified interval at 1p36.21 includes 17 *PRAME* family member genes. However, *PRAME* (preferentially expressed antigen in melanoma) is located on chromosome 22. An epitope of the antigen is recognised by a type of cytotoxic T-lymphocyte (Quintarelli *et al.*, 2011). Although originally discovered in relation to melanoma, *PRAME* has since been linked to other cancers (De Carvalho *et al.*, 2011), while *PRAME* family members have been linked to non-small-cell lung cancer (Bankovic *et al.*, 2010). Also on chromosome 1 (at 1q21.1) is *PPIAL4A* and *PPIAL4C*, which were reported by Nilsson *et al.* (2004) as being candidate oncogenes and are in a region where amplifications were common in tumour samples.

RPIA (ribose 5-phosphate isomerase A) maps within the amplified CNV at 2p11.2. It is essential for carbohydrate metabolism (Wamelink *et al.*, 2008), has a pseudogene on

Personalis EX2155

chromosome 18 (Pruitt *et al.*, 2005), and is hypermethylated in breast tumours (Kim *et al.*, 2011). *EIF2AK3* also maps to this interval. EIF2AK3 is an endoplasmic reticulum transmembrane protein that acts as a cellular stress sensor (Liu *et al.*, in press). Mutations of this gene cause Wolcott-Rallison Syndrome, which is associated with neonatal or early infancy diabetes mellitus (Senée *et al.*, 2004), and have also been associated with a risk for low bone mineral density (Liu *et al.*, in press). However, the gene has not been associated with cancer.

The amplified CNV identified at 14q32.33 includes a number of genes, including *BRF1*, *CDCA4*, *JAG2*, *MTA1*, and *NUDT14*. *BRF1* encodes a subunit of RNA polymerase III's transcription factor complex, thus indirectly affecting transcription initiation by this polymerase (Berns, 2008), whilst *CDCA4* also encodes a transcription factor that has been shown to regulate the transcription of the *TP53* tumour suppressor gene, *E2F1*, and the *JUN* oncogene (Tategu *et al.*, 2008). *E2F1* encodes a transcription factor of the E2F family of genes and mediates cell proliferation in TP53-deficient cancer cells (Donati *et al.*, 2011). Finally, NUDT14 is a pyrophosphatase that converts substrate uridine diphosphate glucose into products glucose-1-phosphate and uridine monophosphate (Yagi *et al.*, 2003) - it has previously been linked to colorectal carcinoma (Choi *et al.*, 2011), while expression of the NOTCH receptor ligand, JAG2, has been correlated recently with vascular development and angiogenesis (Pietras *et al.*, 2011).

However, arguably the most interesting of the genes overlapping the CNV at 14q32.33 is *MTA1* (metastasis associated 1). It was originally-identified as being associated with metastasis in mammary adenocarcinoma but has since been positively associated with metastasis in endometrial cancer (Yang *et al.*, 2010), non-small-cell lung cancer (Zhu *et al.*, 2010), prostate cancer (Kai *et al.*, 2011), and invasive ductal breast carcinoma (Sharma *et al.*, 2011). MTA1 may also function as a corepressor of the oestrogen receptor (Mazumdar *et al.*, 2001 and Gururaj *et al.*, 2006). In this study, the highest level of amplification within *MTA1* was found in a patient who suffered multiple metastatic events.

Personalis EX2155

Finally, the amplified interval identified at 16p11.2 harbours 3 genes: *HIRIP3*, *TAOK2*, and *TMEM219*. *HIRIP3* shares sequence similarity with corepressors of histone gene transcription in *Saccharomyces cerevisiae*. However, its function in humans is unclear (Lorain *et al.*, 1998). Conversely, TAOK2 is involved in cell-signalling, cell-division, and apoptosis (Moore *et al.*, 2000 and Zihni *et al.*, 2007), whilst TMEM219 is a receptor for IGF-binding protein-3 (IGFBP3) that may play an important role in apoptosis and tumour suppression via IGFBP3 (Ingermann *et al.*, 2010).

Many of the other genes listed in Table 3-6 have not yet been characterised and have unknown functions, whilst others have a 'predicted' status according to the RefSeq gene database (Pruitt *et al.*, 2005).

3.4.7 Breast cancer dormancy

According to clinical records, 42 of the 50 patients in the on follow-up group remained 'disease-free' some 10-13 years after diagnosis (the other 8 patients relapsed). However, data generated in this study showed evidence of disease persistence in their plasma cfDNA (through tumour-specific alterations) many years after surgery and adjuvant treatment (Figure 3-14). Additionally, through PCA, the fact that the second cfDNA samples of the 8 patients who relapsed (i.e. - the cfDNA samples taken closest to the time of relapse) clustered with the normal lymphocyte DNA and matched first cfDNA samples (Figure 3-11) raises questions pertaining to dormancy in breast cancer.

Relapse in breast cancer can occur up to 20 years after primary treatment (Karrison *et al.*, 1999), a length of time that suggests a period of apparent dormancy (Demicheli *et al.*, 1998 and Chambers & Goss, 2008) where there is growth restriction of unseen micrometastases (Murray, 1995). Although dormancy is common in breast cancer, the associated mechanisms behind it are unknown. However, it is well-established that treatment is more effective when given before overt metastatic disease develops (Boér, 2010).

It is possible that the data obtained from plasma cfDNA in this study is elucidating dormancy in some patients, which is potentially very useful considering that the ability

Personalis EX2155

to find the location of dormant cells in patients is difficult (Goss & Chambers, 2010). However, it is important to remember that dormancy does not distinguish between those with metastatic and non-metastatic cancer. For example, a study by Goodison *et al.* (2003) using green fluorescent protein-labelled derivatives of the MDA-MB-436 breast cancer cell-line showed how both metastatic and non-metastatic cells were disseminated to distant organs where they remained dormant. Reasons why tumours remain in a dormant state have been hypothesised as being due to:

- Immunosuppression
- Inability to invade blood and/or lymph vessels
- Inability to induce angiogenesis and/or vascularisation
- Apoptosis predominates over growth
- Cell-cycle regulatory proteins function normally
- Balance between growth-inducing and -inhibiting factors in the tumour microenvironment favours inhibition

(Karrison *et al.*, 1999; Schirrmacher, 2001; Meng *et al.*, 2004; and Goss & Chambers, 2010)

Other studies have shown previously that bone-marrow DTCs and circulating tumour cells (CTCs) can persist for many years after end of treatment or mastectomy (Meng *et al.*, 2004 and Slade *et al.*, 2009). Additionally, Page *et al.* (2011) showed it was possible to detect *ERBB2* amplification in the circulating cfDNA of on follow-up patients up to 6 years or more after the removal of their primary tumour - this was despite the fact that the patients showed no evidence of metastatic disease. The results unearthed in this study correlate with those of Page *et al.* (2011) and others. In direct relation to this - and of critical importance - are the findings by Meng *et al.* (2004), that the half-life of CTCs in blood is 1-2.4 hours, whilst Schwarzenbach *et al.* (2011) report that cfDNA's half-life in the same medium of healthy individuals is between 15 minutes and several hours. Therefore, detection of tumour specific alterations in cfDNA suggests release of micrometastatic cells rather than persistence of this in circulation.

Personalis EX2155

Long-term follow-up studies can provide useful information relating to dormancy and give an indication as to the limits of it in breast cancer (and indeed other cancers). However, better methods of detecting the residual signatures of cancer cells in patients on follow-up are necessary (Karrison *et al.*, 1999). Results in this thesis have shown a possible way of doing this. However, is there ever a point where follow-up should cease? - Karrison *et al.* (1999) concluded that this point should be several years after the majority of recurrences have occurred and tentatively ask whether or not the patients are 'cured' at such a point, whilst others suggest that the limit of dormancy appears to be 20-25 years (Meng *et al.*, 2004 and Goss & Chambers, 2010).

3.5 Conclusion

In conclusion, it has been shown that tumour-specific alterations are able to persist in the blood plasma of breast cancer patients up to 12 years after surgery to remove the primary tumour. Further, it has been shown that by focussing on SNP profiles of cfDNA in blood plasma alone, one can distinguish between breast cancer patients and healthy individuals, but also between patients who have breast cancer but are at different stages of treatment (either pre- or post-surgery to remove the primary tumour).

Microarray technology can detect subtle alterations and was shown to exhibit good inter-assay reproducibility ($p<0.0001$, 0.0001, and 0.0005 for CQC, QC call-rate, and MAPD respectively; and an average Spearman's rank correlation of 0.727). When considering the 8 common CNV amplifications identified in tumour DNA, each had a prevalence of >90% in the 50 patients on follow-up group, suggesting their importance in breast cancer.

As the majority (80%) of patients in the on follow-up group remained 'disease-free', data generated in this study has given evidence of the persistence of the micrometastatic deposits years after surgery to remove the primary tumour - from this it can be inferred that the cancer in these women was in a dormant state. However, what does all of this say about long term prognosis? The fact that tumour-specific alterations remain detectable in the blood of patients on follow-up years after surgery infers that their

Personalis EX2155

blood is constantly being replenished by replicating tumour cells and/or fragments (Meng *et al.*, 2004).

Genomic instability and dormancy can be seen as having a certain level of interplay: genomic instability drives oncogenesis and metastasis (Uchida, 2006), whilst a dormant micrometastatic tumour clone concurrently persists in the body without any apparent destructive effects on the individual. However, tumour cells are under selective pressure to acquire increasing mutations (through genomic instability) that permits them to grow uncontrollably and metastasise. Thus, the following question must be proposed: Is a dormant micrometastatic clone merely accumulating mutations until such a time when it has acquired the right balance of mutations that permit it to spread/metastasise?

3.6 Future studies

Future studies will involve using bioinformatic methods to search for additional markers that can distinguish cancer from normal/healthy and that could potentially be used in a diagnostic and prognostic setting. Research will also focus on the 8 common CNV amplifications identified in this study (Table 3-8) (and also the additional 15 listed in Table 8-1). Some of these regions were found to significantly contribute to both the T>20mm/node-positive and relapsed status of the 50 patients on follow-up who were investigated in this study. Custom assays will be developed for these regions and then screened in other cohorts. If successful, these regions could potentially be used as a way of screening patients for breast cancer, which would be a great leap towards the overall goal of developing an assay that could screen individuals for breast cancer using a simple blood test.

More research also has to be carried-out to understand the nature of the cfDNA that was analysed in this study. A potential source of it is CTCs, however, for this to be determined, CTCs would have to be extracted from blood and then compared to cfDNA extracted from the same patient.

Personalis EX2155



4 MONITORING CIRCULATING CELL-FREE DNA IN RESPONSE TO THERAPY IN BREAST CANCER PATIENTS USING SNP 6.0 MICROARRAYS

Personalis EX2155

4.1 Overview

In this chapter, the search for a response to therapy in the neoadjuvant setting through circulating cell-free DNA (cfDNA) analysis is the focus. Endocrine therapy or chemotherapy was applied to each patient neoadjuvantly. Each provided serial blood samples during treatment - at baseline (0 weeks), 6 weeks, and 21 weeks - with each of these time-points being compared to elucidate any alterations that could be reflective of a response to therapy over time. A tumour biopsy core was also taken during resection at the end of treatment and its genomic profile compared back to serial cfDNA samples.

4.2 Introduction

4.2.1 Cancer therapy

Cancer therapy involves the use of drugs to destroy cancer cells and is typically used adjuvantly either in metastatic disease or where there is high risk of metastasis. Therapy given before surgery is known as neoadjuvant therapy and is applied in an attempt to reduce tumour bulk before surgical resection (Sachelarie *et al.*, 2006 and Le Guellec *et al.*, 2011).

Endocrine therapy is used to treat cancers sensitive to hormones, such as ER- (oestrogen receptor) and PR- (progesterone receptor) positive breast cancers. For example, the SERMs (selective estrogen receptor modulators) act by blocking oestrogen from binding to the oestrogen receptor, thereby slowing tumour growth. Conversely, aromatase inhibitors block aromatase, the enzyme that converts androgen steroids to oestrogen, and include anastrozole (Arimidex®), letrozole (Femara®), and exemestane (Aromasin®). Another drug called fulvestrant (Faslodex®) is given when tamoxifen cannot be used or when its effectiveness has decreased (Tomlinson *et al.*, in press).

Biological therapies include trastuzumab (Herceptin®) and lapitinib (Tykerb®), which are monoclonal antibodies that target ERBB2 and are used to treat metastatic disease. Bevacizumab (Avastin®) inhibits angiogenesis, which tumours utilise to create new blood vessels for survival and growth, by binding to vascular endothelial growth factor (Arteaga *et al.*, 2011).

Personalis EX2155

Over the past few decades, breast cancer mortality has decreased in the European Union due to the availability of these adjuvant/neoadjuvant chemo- and endocrine therapies (Espinosa *et al.*, in press). Importantly, the ability to monitor response to therapy can help in the detection of emerging resistance (Meldrum *et al.*, 2011).

4.2.2 Resistance to therapy

In cancer therapy, a proportion of patients will exhibit intrinsic (innate) resistance to the applied therapy (Saha Roy & Vadlamudi, in press), whereas others will develop acquired resistance whereby they initially respond to the therapy but then later fail to (Tomlinson *et al.*, in press). Both of these limit the success of therapies in the clinic. Tumour heterogeneity and the presence of sub-populations of cells within the tumour mass are thought to play an underlying role in resistance (Dexter & Leith, 1986).

The mechanisms behind resistance are complex and have been explored: for example, in cancer cells, the ER and its co-regulatory proteins can be activated through mechanisms involving both IGFR (insulin-like growth factor receptor) and EGFR (epidermal growth factor receptor). Also, the MAPK (mitogen activated protein kinase) and PI3K (phosphoinositide kinase-3) signalling cascades are involved in the adaptation of ER-positive breast cancer cells to oestrogen deprivation, creating ER 'hypersensitivity' (Gee *et al.*, 2001 and Tomlinson *et al.*, in press).

A study carried out by Gee *et al.* (2005), noted an increase in EGFR expression in patients who had acquired tamoxifen resistance. It was then speculated by Tomlinson *et al.* (in press) that other growth factor receptors were involved in resistance, namely the (FGFR) fibroblast growth factor receptor. Mutations of fibroblast growth factors (FGFs) and their receptors have been linked to breast cancer (Tomlinson *et al.*, in press), for example, over-expression of FGFs was shown to drive tumourigenesis (Daphna-Iken *et al.*, 1998). Fibroblast growth factors are mitogens and the activation of their receptors can stimulate downstream signalling pathways similar to those of the EGFRs (Daphna-Iken *et al.*, 1998 and Eswarakumar *et al.*, 2005) - targeting FGFs induced apoptosis and prevented tumour growth and metastasis in a study by Dey *et al.*

Personalis EX2155

(2010). Higher expression of FGFR3 was also found in tamoxifen-resistant breast tumours and it was also shown to reduce sensitivity to fulvestrant (which is normally used to degrade the ER) whilst stimulating both the MAPK and PI3K signalling pathways (Tomlinson *et al.*, in press). Amplification of the *FGFR1* gene was also linked to resistance to endocrine therapy (Turner *et al.*, 2010). Conversely, anti-oestrogen resistant cell growth was prevented in a cell-line model by overexpressing the binding domain of FRS2 (fibroblast growth factor receptor substrate 2), thus identifying FRS2 as a possible target for new therapies (Manuvakhova *et al.*, 2006).

These studies highlight the molecular complexity that lies behind resistance to therapy. To design novel therapeutic targets, the identification of genes and proteins critical to the survival of the tumour is essential. Hence, changes at the molecular level that are correlated with tumourigenesis have been the focus of recent efforts to characterise genetic variation in the hope of developing novel therapeutic targets (Brough *et al.*, 2011). However, this is made difficult by the heterogeneity of tumour DNA, which can exhibit between 1,000 and 10,000 genetic variations but only a few of which may be critical to survival [of the tumour] (Brough *et al.*, 2011).

Thus, finding biomarkers that are deregulated through therapy and are conducive to resistance is an important research goal (Hicks *et al.*, 2012).

4.2.3 The NeoCEnT clinical trial

As mentioned in Chapter 2, the neoadjuvant chemo- or endocrine therapy (NeoCEnT clinical trial is a multi-centre, randomised, parallel group, comparative, phase III, neoadjuvant study of chemotherapy versus endocrine therapy in postmenopausal patients who had strongly ER-positive, primary breast cancer with no evidence of overt metastases, and for whom pre-operative systemic therapy was indicated (chief investigator, Professor RC Coombes, Imperial College London). The pilot study aimed to help clarify the optimal treatment - chemo- (fluorouracil/epirubicin/cyclophosphamide) or endocrine (letrozole) therapy - for a group of 40 patients in the neoadjuvant setting with the aim of reducing tumour bulk before surgery. However, it was not known which treatment each patient received. For

Personalis EX2155

each patient, serial cfDNA samples were taken at the beginning of treatment, 6 weeks, and end of treatment (21 weeks). Serial tumour core biopsies were also taken at the beginning of treatment, 2-4 days, and 21 weeks, and then stored as fresh frozen (FF) tissue. However, as the study was on-going at the time of this thesis, only a single tumour sample was analysed from each of the 10 patients due to availability of samples (all taken from the beginning of treatment except for 3: one at 2-4 days and two at 21 weeks).

4.2.4 Aims

The aim of this chapter was to test the hypothesis that changes in the single nucleotide polymorphism (SNP) and copy number (CN) microarray profiles correlate with resistance to therapy in plasma cfDNA of patients on neoadjuvant therapy. The objectives were to:

1. Determine post-hybridisation sample quality through QC checks (4.3.1)
2. Genotype samples and compare SNP profiles in different samples, searching for alterations between serial cfDNA samples (4.3.2)
3. Characterise copy number variation (CNV) across all samples - distinguishing benign from pathogenic - and observing any differences between serial cfDNA samples (4.3.3)
4. Use principal component analysis (PCA) to separate tissues into logical groupings (4.3.4)

4.3 Results

4.3.1 *Affymetrix SNP 6.0* QC analysis of cfDNA, tumour, and lymphocyte samples

In total, 47 whole genome amplification (WGA) DNA samples (lymphocyte, plasma cfDNA, and tumour) from 10 breast cancer patients were analysed using *Affymetrix SNP 6.0* microarrays. The 47 samples were hybridised in a single batch and high SNP call-rates were achieved for the majority of samples (Table 4-1). The lymphocyte DNA samples were comparable to the tumour (mean 98.09%, range 96.20-98.87% and mean

Personalis EX2155

96.60%, range 88.00-98.70% respectively), and the cfDNA samples also had a high quality (baseline, mean 87.15%, range 86.48-88.36%; 6 weeks, mean 86.80%, range 86.04-87.77%; and 21 weeks, 89.03%, range 87.00-97.04%).

Preliminary review of clinical information revealed that two patients responded to treatment, three failed to respond, and the response status of the remaining five was unknown at the time of writing.

Personalis EX2155

Response status	Patient ID	Sample	SNP call-rate (%)
Unknown	N1	L	96.20
		Baseline P	88.17
		6 week P	87.77
		21 week P	88.81
		T	97.61
	N2	L	98.40
		Baseline P	88.36
		6 week P	86.67
		21 week P	87.03
		T	98.32
	N3	L	98.42
		Baseline P	86.58
		6 week P	86.04
		21 week P	87.40
		T	88.00
	N6	L	98.43
		Baseline P	86.60
		6 week P	86.45
		21 week P	98.04
		T	97.32
	N14	L	98.19
		Baseline P	86.78
		6 week P	87.03
		T	96.96
Non-response	N7	L	97.44
		Baseline P	87.27
		6 week P	86.95
		21 week P	87.77
		T	97.99
	N8	L	97.87
		Baseline P	87.46
		6 week P	86.93
		21 week P	87.00
		T	96.60
	N12	L	98.41
		Baseline P	86.73
		6 week P	86.51
		T	98.02
Response	N10	L	98.64
		Baseline P	86.48
		6 week P	86.64
		21 week P	87.14
		T	96.48
	N11	L	98.87
		Baseline P	87.10
		6 week P	87.01
		T	98.70

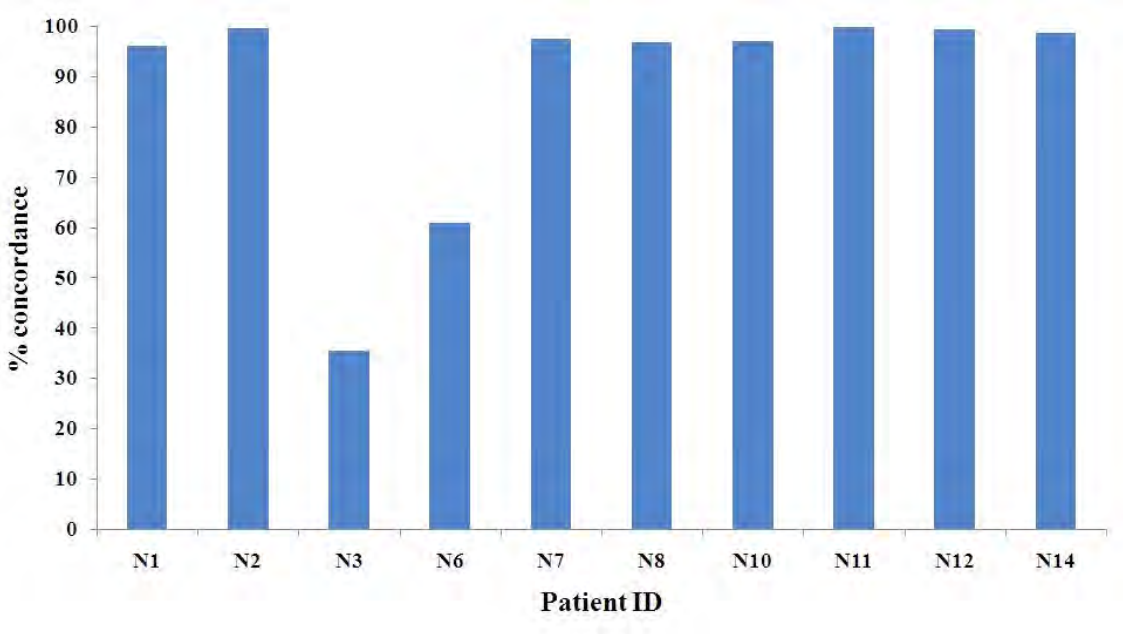
Table 4-1 - The SNP call-rates for 47 NeoCEnT samples highlights the high quality of the tissue used. L, lymphocyte; P, plasma cfDNA; T, tumour. Samples are grouped by response, non-response, and unknown response to therapy.

Personalis EX2155

4.3.2 Single nucleotide polymorphisms and the comparison of genomic profiles

There was a high % concordance of SNP genotype calls[1] between paired lymphocyte and tumour DNA for all but two samples - N3 and N6 (Figure 4-1). Independent reviews of tissue histology showed that many samples contained a large amount of stroma as well as invasive tumour cells (Professor RA Walker, Dr. C Ruangpratheep, and Dr. D Moore).

Figure 4-1 - Percent concordance between tumour DNA and matched normal lymphocyte (mean 88.08%; range 35.47-99.75%).



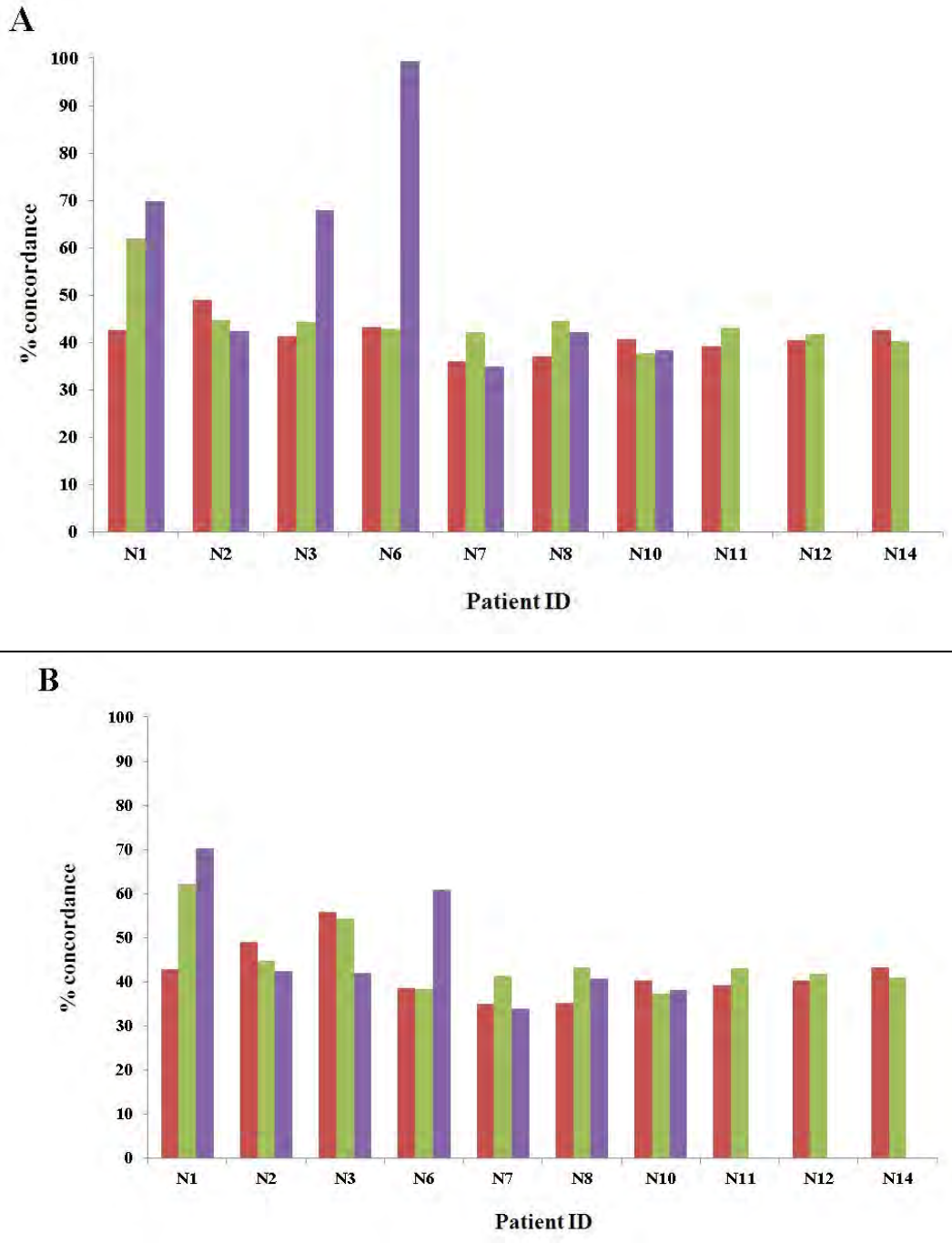
In contrast, SNP concordance was low for the majority of cfDNA samples when compared to either matched lymphocyte or tumour (Figure 4-2). The mean concordance with lymphocyte DNA was 44.21% (range 33.83-70.17%), with means of 41.19, 44.3, and 56.4% for each of the baseline, 6 week, and 21 week samples, respectively. Despite the increase in mean % concordance with time, there were no statistically significant differences between the 3 time-points. Also, when these comparisons were reversed - i.e. using tumour as the reference instead of lymphocyte -

[1] Based on the number of SNP markers across any two samples that have the same genotype call.

Personalis EX2155

a similar result was achieved due to the highly-similar SNP profile between lymphocyte and tumour DNA samples in this cohort.

Figure 4-2 - Percent concordance between cfDNA, matched normal lymphocyte (A), and tumour DNA (B). Red, baseline cfDNA; green, 6 week cfDNA; purple, 21 week cfDNA.



Personalis EX2155

4.3.3 Copy number variation analysis

4.3.3.1 Number of CNVS called

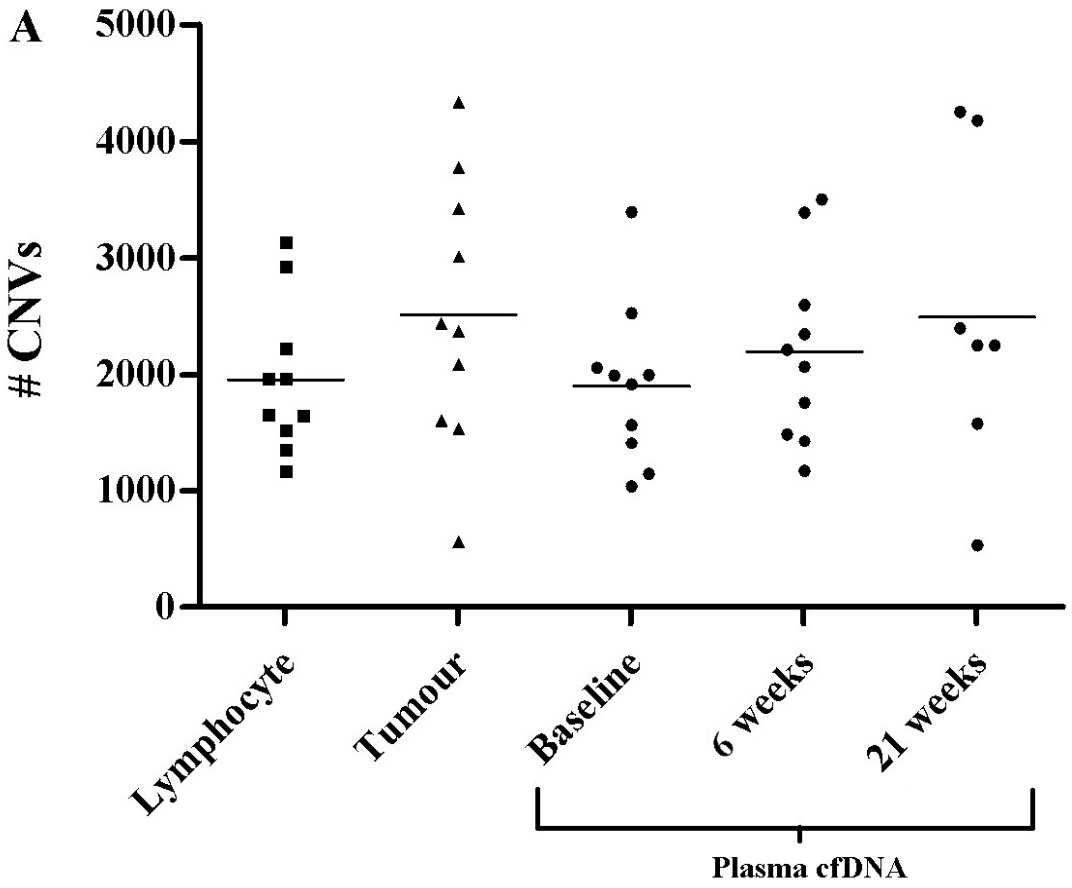
When comparing the number of CNVs called between the NeoCEnT patients' samples, there were evidential differences between both patients and samples (Table 4-2 and Figure 4-3). However, overall, there were no significant differences by sample type and time point (p=0.3865, Kruskal-Wallis ANOVA).

Response status	Patient ID	Lymphocyte	Tumour	Plasma cfDNA		
				Baseline	6 week	21 week
Unknown	N1	3137	2371	1567	3504	4178
	N2	1966	3011	2059	1429	2251
	N3	1529	562	2525	2067	4253
	N6	2921	3779	3393	2599	2399
	N14	1167	1605	1992	1762	-
Non-response	N7	1353	2090	1035	1491	532
	N8	1660	4336	1142	2348	2252
	N12	2224	2441	1998	3389	-
Response	N10	1967	3422	1918	1167	1580
	N11	1651	1541	1411	2215	-

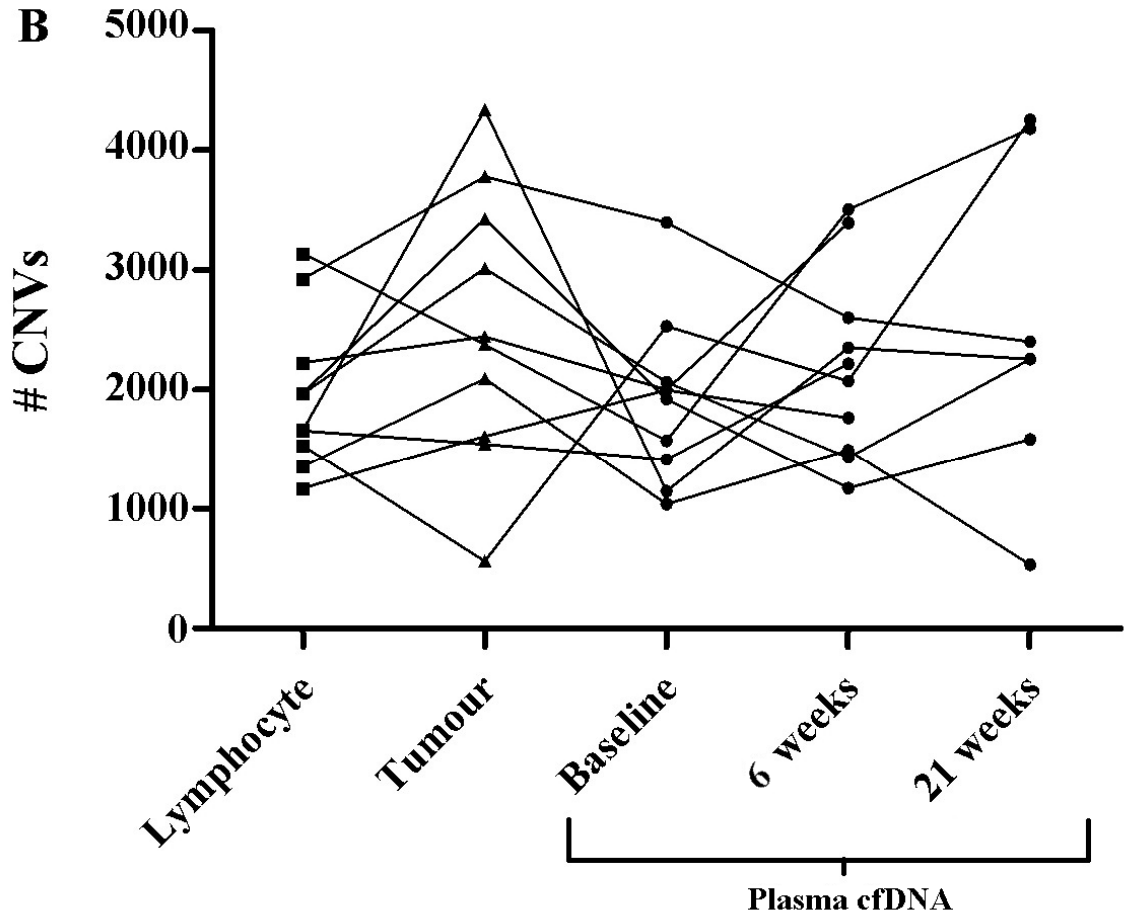
Table 4-2 - Number of CNVs called per sample and patient. A CNV was identified as having a minimum of 50 markers per segment, p-value cut-off of <0.0001, signal-to-noise ratio of 0.5, and a minimum size of 1Kb.

Personalis EX2155

Figure 4-3 - Number of CNVs called per sample and patient (A), repeated with lines connecting samples belonging to the same patient (B).



Personalis EX2155



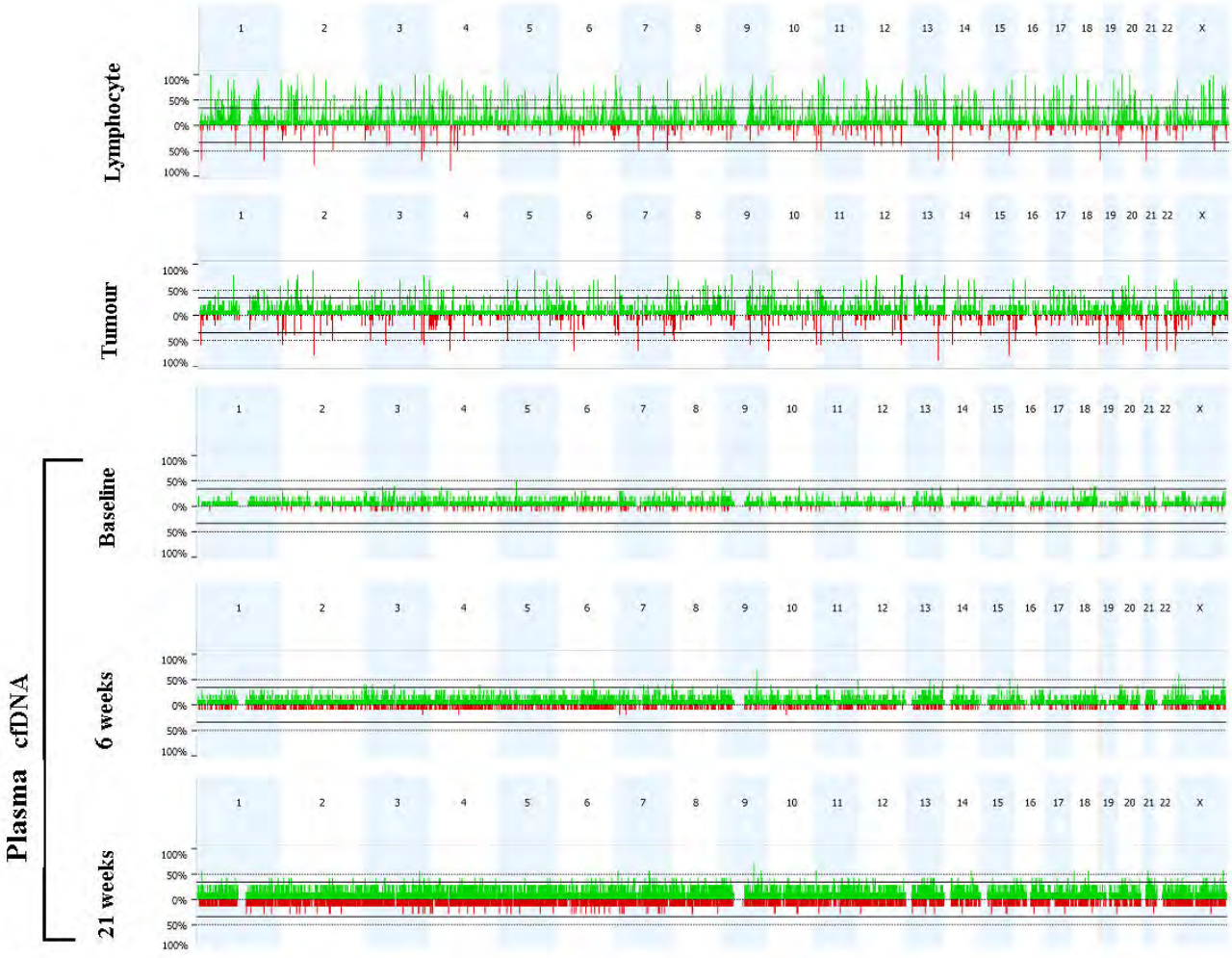
4.3.3.2 Amplifications and deletions

4.3.3.2.1 Frequency

Analysis of the frequency of CNVs in the serial plasma cfDNA samples revealed visual differences between the baseline, 6 week, and 21 week samples (Figure 4-4). There were more CNVs identified in plasma cfDNA as time passed from the baseline to 6 and 21 week samples. Clear differences were also seen between the frequency profiles of these samples and those of the lymphocyte and tumour DNA samples. Many of the CNVs identified in normal lymphocyte and tumour DNA had high frequencies and the profiles resembled each other. For example, the same deletions at frequencies greater than 50% in both were detected on chromosomes 1, 2, 4, 13, 14, 15, 18, and 21. In cfDNA, CNVs were identified at lower frequencies and few were found at a frequency greater than 50%.

Personalis EX2155

Figure 4-4 - Genome-view representation of the frequency of amplification and deletion in the NeoCEnT sample group. The y-axis indicates the percentage of samples having a CNV at a specific point in the genome. Amplifications in green; deletions in red.



4.3.3.2.2 Distribution

The distribution between amplification and deletion was analysed (Table 4-3). In lymphocyte and tumour samples, the majority of CNVs detected were amplifications (healthy control lymphocytes, 67%; NeoCEnT lymphocytes and tumours, 77% and 70% respectively). In the plasma cfDNA samples of the healthy controls, again amplifications were more common than deletions (59% and 41% respectively for amplifications and deletions). However, in the NeoCEnT plasma cfDNA samples, the

Personalis EX2155

majority of CNVs detected were deletions, with the amount of deletions decreasing with time-point (67, 64, and 55% respectively for baseline, 6 week, and 21 week samples).

Group	Tissue	Amplification (%)	Deletion (%)
Healthy control	L	67%	33%
	P	59%	41%
NeoCEnT	L	77%	23%
	Baseline P	33%	67%
	6 week P	36%	64%
	21 week P	45%	55%
	T	70%	30%

Table 4-3 - Percent CNV amplification and deletion in the NeoCEnT and healthy control groups. L, lymphocyte; P, healthy control or NeoCEnT plasma cfDNA; T, tumour.

4.3.3.2.3 Thresholding amplifications by copy number

Applying a threshold of CN>6.0 by Gaussian-smoothed signal (as for Table 3-6) in NeoCEnT samples revealed 21 cytobands (Table 4-4) that appeared in >10% of samples (per sample group).

Personalis EX2155

Interval	Overlapping genes	Healthy control		NeoCEnT				
		L n=8	P n=8	L n=10	Baseline P n=10	6 week P n=10	21 week P n=7	T n=10
1p36.33	*LOC100132287*	4 (50)	5 (62.5)	6 (60)	10 (100)	10 (100)	6 (86)	5 (50)
1p36.21	*C1orf158* *LOC [440563 and 649330]* *PRAMEF [2 3 4 5 7 8 9 10 12 13 14 15 16 19 20 21 and 22]*	0 (0)	0 (0)	0 (0)	5 (50)	3 (30)	2 (29)	0 (0)
1q21.1	*PPIAL4 [A and C]*	3 (37.5)	4 (50)	1 (10)	2 (20)	2 (20)	6 (86)	1 (10)
2p11.2	*ANKRD36BP2* *EIF2AK3* *RPIA*	0 (0)	0 (0)	3 (30)	10 (100)	10 (100)	6 (86)	4 (40)
5q13.2	*SERF1 [A and B]* *SMA5* *SMN [1 and 2]* *NAIP*	0 (0)	0 (0)	0 (0)	6 (60)	7 (70)	2 (29)	1 (10)
7p11.2	*MIR4283-2*	0 (0)	0 (0)	0 (0)	2 (20)	5 (50)	1 (14)	1 (10)
7q11.23	*CLIP2* *GATSL1* *GTF2IRD1 2* *LOC100093631* *NCF [1 and 1C]* *PMS2P [5 and 6]* *SPDYE8P* *STAG3L2* *WBSCR16*	0 (0)	2 (25)	0 (0)	9 (90)	10 (100)	4 (57)	1 (10)
8p23.1	*DEFB130* *DUB3* *FAM66 [A and D]* *FAM85A* *FAM86B1* *LOC392196* *ZNF705D*	0 (0)	0 (0)	0 (0)	1 (10)	3 (30)	1 (14)	0 (0)
9p12*	-	0 (0)	1 (12.5)	0 (0)	9 (90)	7 (70)	5 (71)	1 (10)
9p11.2	*FAM27A* *LOC100132167*	2 (25)	1 (12.5)	5 (50)	10 (100)	10 (100)	5 (71)	5 (50)
9q12	-	1 (12.5)	1 (12.5)	0 (0)	8 (80)	9 (90)	5 (71)	2 (20)
10q11.22	*BMS1P5* *FRMPD2P1* *GDF [2 and 10]* *PTPN20B* *RBP3* *ZNF488*	1 (12.5)	0 (0)	0 (0)	10 (100)	9 (90)	3 (43)	1 (10)
10q11.23	*AGAP [7 and 8]* *C10orf53* *CHAT* *OGDHL* *PARG*	2 (25)	3 (37.5)	1 (10)	10 (100)	9 (90)	6 (86)	1 (10)
14q11.1	-	1 (12.5)	1 (12.5)	0 (0)	10 (100)	10 (100)	6 (86)	1 (10)
14q11.2	-	0	0	0	10	10	6	1

Personalis EX2155

		(0)	(0)	(0)	(100)	(100)	(86)	(10)
14q32.33	*AHNAK2* *BRF1* *C14orf79* *CDCA4* *CRIP [1 and 2]* *GPR132* *JAG2* *KIAA0284* *LOC647310* *MGC23270* *MTA1* *NUDT14* *PACS2*	1 (12.5)	0 (0)	0 (0)	6 (60)	7 (70)	3 (43)	0 (0)
15q11.2	*NF1P1* *POTEB*	0 (0)	0 (0)	0 (0)	3 (30)	3 (30)	3 (43)	0 (0)
15q25.2	*GOLGA6L [4 5 and 9]* *LOC [440297 727849 and 80154]* *RPS17*	1 (12.5)	0 (0)	1 (10)	9 (90)	9 (90)	6 (86)	3 (30)
16p12.3	*ABCC6P1* *NOMO2*	2 (25)	5 (62.5)	0 (0)	9 (90)	7 (70)	6 (86)	4 (40)
17q12*	*CCL3L1*	0 (0)	0 (0)	2 (20)	4 (40)	3 (30)	2 (20)	1 (10)
19p13.3	*FAM138F* *FLJ45445* *OR4F17* *STK11* *WASH5P*	2 (25)	3 (37.5)	2 (20)	8 (80)	10 (100)	4 (57)	3 (30)

Table 4-4 - Twenty-one intervals exhibiting common amplification in the NeoCEnT patient samples (using a threshold of CN>6.0 by Gaussian-smoothed signal and only displaying those that had a frequency >10%). The frequency of amplification (%) of a given CNV is indicated in the square-brackets. L, lymphocyte; P, healthy control or NeoCEnT plasma cfDNA; T, tumour. *Not appearing in Table 3-6 and therefore specific to the NeoCEnT group.

None of the 21 CNVs had a prevalence of 100% in tumour DNA. However, some were found in 100% of the baseline, 6 week, or 21 week cfDNA samples (1p36.33, 2p11.2, 7q11.23, 9p11.2, 10q11.22, 10q11.23, 14q11.1, 14q11.2, and 19p13.3). The only CNVs that had a prevalence of ≥30% in tumour DNA were found at intervals 1p36.33, 2p11.2, 9p11.2, 15q25.2, and 16p12.3. Of these, only 15q25.2 and 16p12.3 had a low prevalence in matched normal lymphocyte DNA (10 and 0%, respectively).

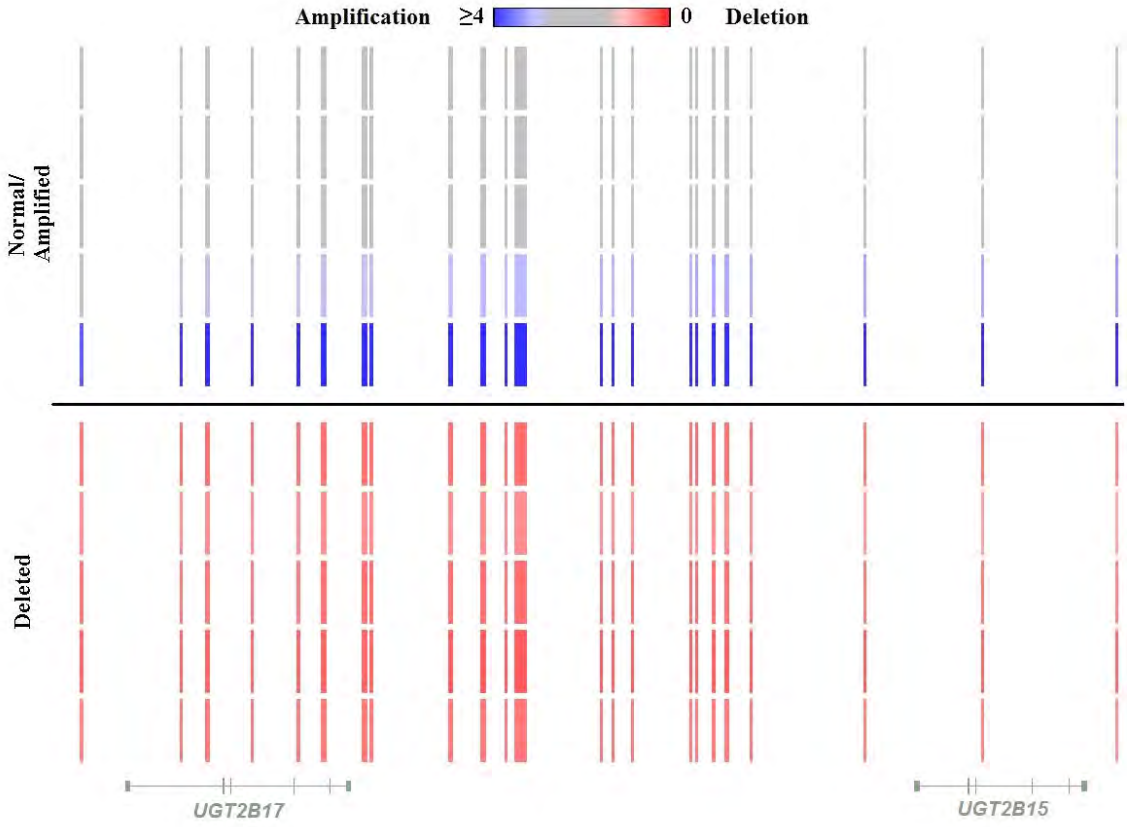
Sixteen of these cytobands showed amplification in the NeoCEnT samples but had little or no amplification in the healthy controls. The remaining 5 cytobands (i.e. - 1p36.33, 1q21.1, 10q11.23, 16p12.3 and 19p13.3), while showing amplification in NeoCEnT samples, also showed amplification in >25% of healthy control samples (lymphocyte

Personalis EX2155

and/or cfDNA). Seventeen of the intervals have known overlapping genes while the remaining 4 have none as defined by the RefSeq gene database (Pruitt *et al.*, 2005).

Copy number variations found previously in Table 3-6 but not identified in these patients were 4q13.2, 5p11, 7q11.21, and 16p11.2. 4q13.2 harbours *UGT2B15* and *UGT2B17*, two genes involved in the metabolism of estradiol. A closer look at the CN of these genes in the NeoCEnT tumours revealed deletion in 50%, normal CN in 40%, and slight amplification in 10% (Figure 4-5). Only CN markers in the *Affymetrix SNP 6.0* microarray target these genes.

Figure 4-5 - *UGT2B15* and *UGT2B17* gene intervals for 10 NeoCEnT tumour samples. Each vertical bar represents a single microarray marker (or bunch of consecutive markers) and values are as inferred by the colour-scale (at top). The exons and introns of the genes are indicated at the bottom of the diagram.



Personalis EX2155

A further scan of the samples for deletions, by altering the threshold to CN<1, revealed 4 further intervals in tumour DNA that only appeared in ≤10% of normal lymphocyte DNA samples: 2q13, 6q27, 16q23.1, and 20p11.23. Of these, only 2q13 and 16p23.2 have overlapping genes of known function according to the RefSeq gene database (Pruitt *et al.*, 2005). The deletion at 2q13 occurred in 40% of the tumour DNA samples, and then 70, 70, and 43% of the baseline, 6 week, and 21 week cfDNA samples. The deletion at 16p23.1 occurred in 20% of tumour DNA samples and none of the cfDNAs.

4.3.3.2.4 Logistic regression to reveal significant copy number variant regions

The diagnostic potential of each of the 21 CNVs listed in Table 4-4 was explored through binary logistic regression analysis based on cancer status (i.e. - cancer or healthy) (Table 4-5). A total of 12 CNVs significantly contributed to disease when comparing the baseline cfDNA samples to those of the healthy controls. For the 6 week cfDNA samples, there were again 12, but only 9 of which overlapped with baseline samples. For the 21 week cfDNA samples, only 7 intervals contributed significantly to the disease status. In total, 4 intervals were significant in all 3 cfDNA samples (9p12, 9p11.2, 14q11.1, and 15q25.2).

Binary logistic regression was also performed by comparing CNVs separately to results for the matched normal lymphocyte samples. Fourteen CNVs were significant in the baseline cfDNA, 13 in the 6 week, and 11 in the 21 week; whilst only one (16p12.3) was significant in the tumour, suggesting a close relationship between these.

Personalis EX2155

		Cancer status						
		Healthy control (P) Vs NeoCEnT (Baseline P/6 week P/21 week P)			NeoCEnT (L) Vs NeoCEnT (Baseline P/6 week P/21 week P/T)			
		Baseline P	6 week P	21 week P	Baseline P	6 week P	21 week P	T
Interval	1p36.33	NS	NS	NS	0.0077	NS	NS	NS
	1p36.21	0.0169	NS	NS	0.0248	0.0248	NS	NS
	1q21.1	0.0088	NS	NS	0.0002	0.0002	0.0224	NS
	2p11.2	NS	NS	NS	NS	NS	NS	NS
	5q13.2	0.0049	0.0009	NS	0.0017	0.0002	NS	NS
	7p11.2	NS	0.0169	NS	NS	0.0077	NS	NS
	7q11.23	NS	NS	NS	<0.0001	NS	0.0037	NS
	8p23.1	NS	NS	NS	NS	NS	NS	NS
	9p12	0.0002	0.0125	0.0019	<0.0001	<0.0001	0.0004	NS
	9p11.2	<0.0001	0.0002	0.0019	NS	NS	<0.0001	NS
	9q12	0.018	0.0027	NS	NS	NS	<0.0001	NS
	10q11.22	<0.0001	0.0002	NS	0.0077	0.0077	NS	NS
	10q11.23	0.0088	0.0088	NS	<0.0001	0.0002	<0.0001	NS
	14q11.1	<0.0001	0.0027	0.0164	0.0017	0.0002	0.0212	NS
	14q11.2	NS	<0.0001	<0.0001	NS	NS	0.0212	NS
	14q32.33	0.0417	0.0417	NS	<0.0001	<0.0001	0.0004	NS
	15q11.2	NS	NS	0.0401	<0.0001	0.0002	<0.0001	NS
	15q25.2	<0.0001	<0.0001	<0.0001	<0.0001	<0.0001	0.0003	NS
	16p12.3	NS	NS	NS	<0.0001	<0.0001	NS	0.0248
	17q12	0.0448	NS	NS	NS	NS	NS	NS
	19p13.3	NS	0.0172	0.4829	0.0052	<0.0001	NS	NS

Table 4-5 - Binary logistic regression analysis for 21 commonly-amplified CNVs in NeoCEnT patient samples. Source data was based on the presence or absence of an amplification (using a threshold of CN>6.0 by a Gaussian-smoothed signal) at each of 21 CNVs listed in Table 4-4. Significant p-values are shown; non-significant p-values denoted by 'NS'. L, lymphocyte; P, healthy control or NeoCEnT plasma cfDNA; T, tumour.

4.3.3.3 Copy number variation in growth factor receptor genes

Observing CNV differences in the *FGFR*s (*FGFR1-4*) and *EGFR* between the different NeoCEnT samples revealed that each *FGFR* was significantly different ($p<0.001$) when comparison was made between tumour and matched baseline and 6 week cfDNA. However, no significant differences existed between tumour DNA and 21 week cfDNA samples (Table 4-6). *FGFR1* had a mixture of normal/amplified CN in tumour, but the entire gene was amplified in all of the baseline and 6 week cfDNA samples. *FGFR2* was normal in the majority of tumour samples, but again amplified in all baseline and 6 week cfDNA samples. *FGFR3*, conversely, was deleted or inferred LOH in all tumour DNA samples, whilst having normal CN in cfDNA. However, two of the baseline cfDNA samples exhibited tumour-specific deletion that then disappeared in the 6 week samples. *FGFR4* followed a similar CNV pattern to *FGFR2* - i.e. normal in tumour

Personalis EX2155

DNA and then amplified in baseline and 6 week cfDNA. Finally, *EGFR* had normal CN in all tumour DNA samples but was amplified in only the baseline cfDNA and showed no CNV in either the 6 or 21 week cfDNA samples.

Gene	Tumour	Plasma cfDNA		
		Baseline	6 week	21 week
FGFR1	Normal/Amplification (50/50)	Amplification (100)	Amplification (100)	NS
FGFR2	Normal (90)	Amplification (100)	Amplification (100)	NS
FGFR3	Deletion/LOH (100)	Normal (80)	Normal (100)	NS
FGFR4	Normal (100)	Amplification (100)	Amplification (90)	NS
EGFR	Normal (80)	Amplification (100)	NS	NS

Table 4-6 - Copy number variation in the *FGFR*s and *EGFR* highlights variation between tumour DNA and plasma cfDNA at different time-points. The frequency (%) of a given CNV is indicated in the square-brackets. 'NS' denotes no significant differences between cfDNA and tumour DNA.

4.3.3.4 Detection of genomic aberrations

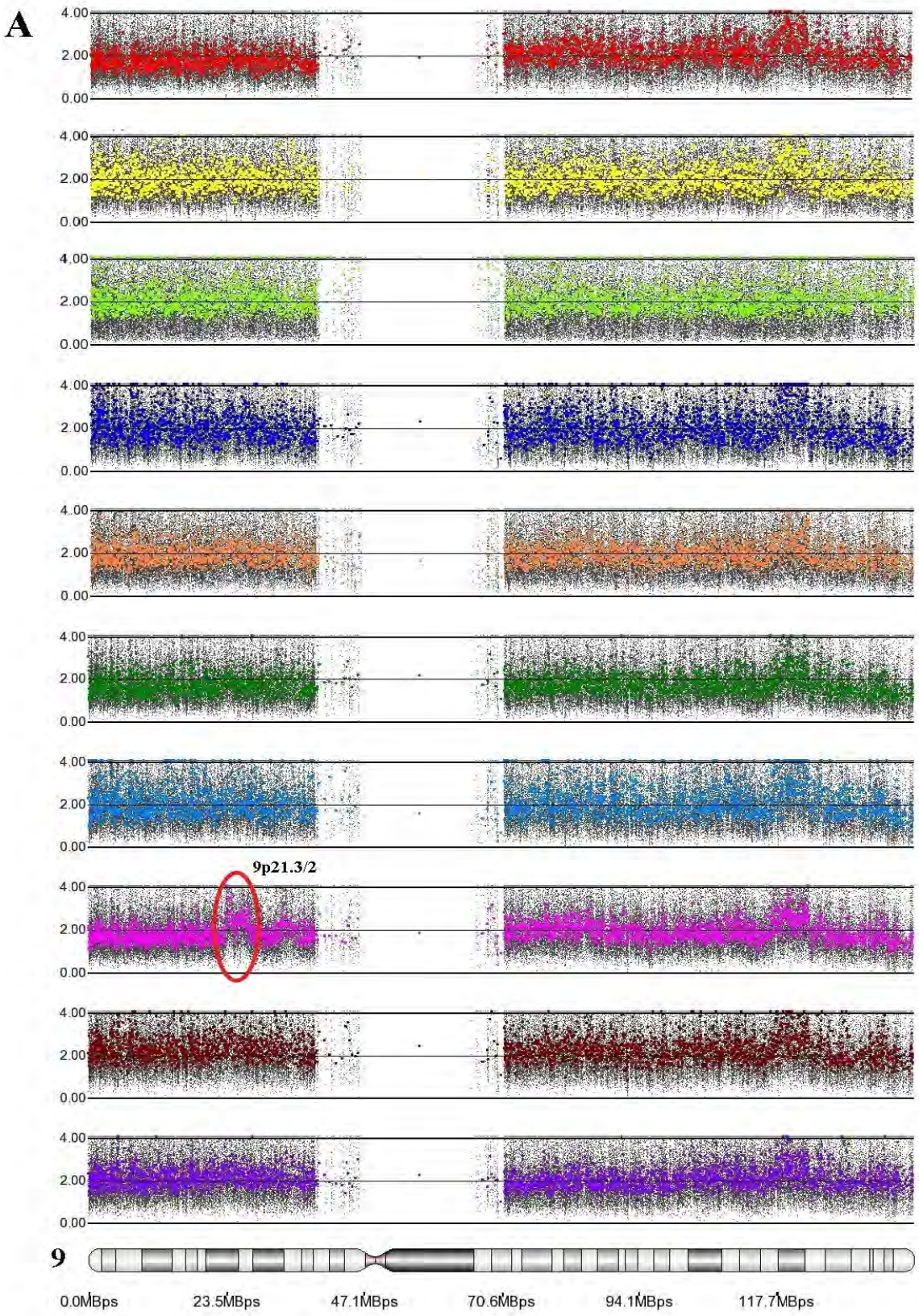
Based on derived, Gaussian-smoothed CN data, specific genomic aberrations were detected in all NeoCEnT tumour DNA samples, including evidence of both isolated (Figure 4-6A) and common (Figure 4-6B) local amplifications and deletions, uncapped chromosomal/telomere attrition (Figure 4-6B), breakages (Figure 4-6C), and isochromosomes (Figure 4-6D) (images for each sample and chromosome can be found in Appendix Figure 8-1). These phenomena were less clear in results for cfDNA due to poorer quality DNA, but many were considered tumour-specific as they were mostly absent from the paired normal lymphocyte DNA samples. For example, there was no evidence of either isochromosomes or breakages in normal lymphocyte DNA. Telomere attrition (defined in this thesis as deletion at a telomere of 5Kb or more) was

Personalis EX2155

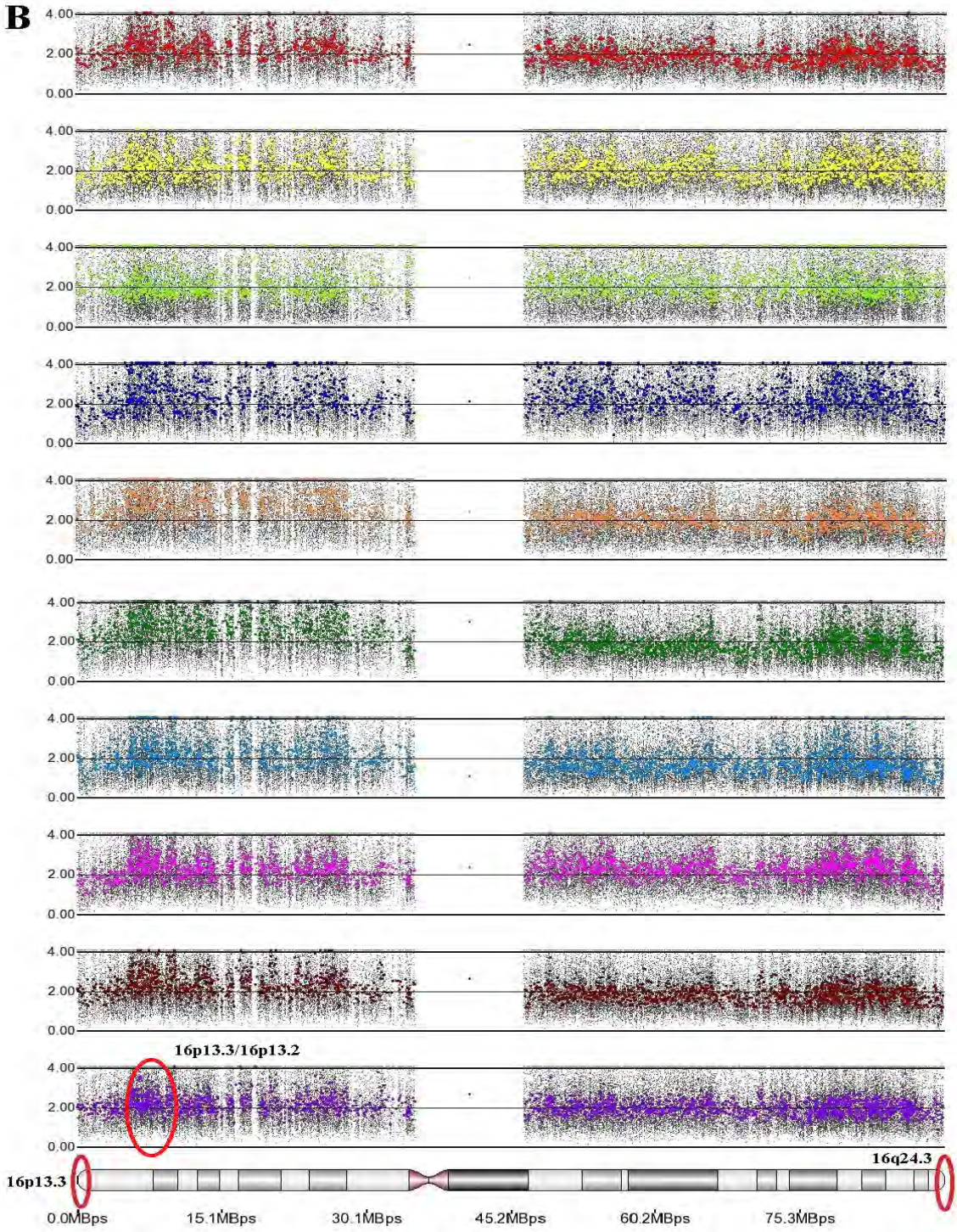
also not detected in normal lymphocytes, whereas in tumour DNA it was detected on every chromosome.

Figure 4-6 - Gaussian-smoothed (coloured dots), CN output of NeoCEnT tumour DNA samples showing evidence of genomic instability - isolated amplification (A; circled in red), common amplification and telomere attrition (B; both circled in red), breakage on chromosome 4 (C; indicated by red line), and isochromosome 8 showing deletion of the short (q) arm and amplification of the long (p) arm (D). Samples are colour-coded as follows: red, N1; yellow, N2; light-green, N3; dark-blue, N6; beige, N7; dark-green, N8; light-blue, N10; pink, N11; brown, N12; purple, N14. CN scale fixed to 0-4; chromosomes not to scale.

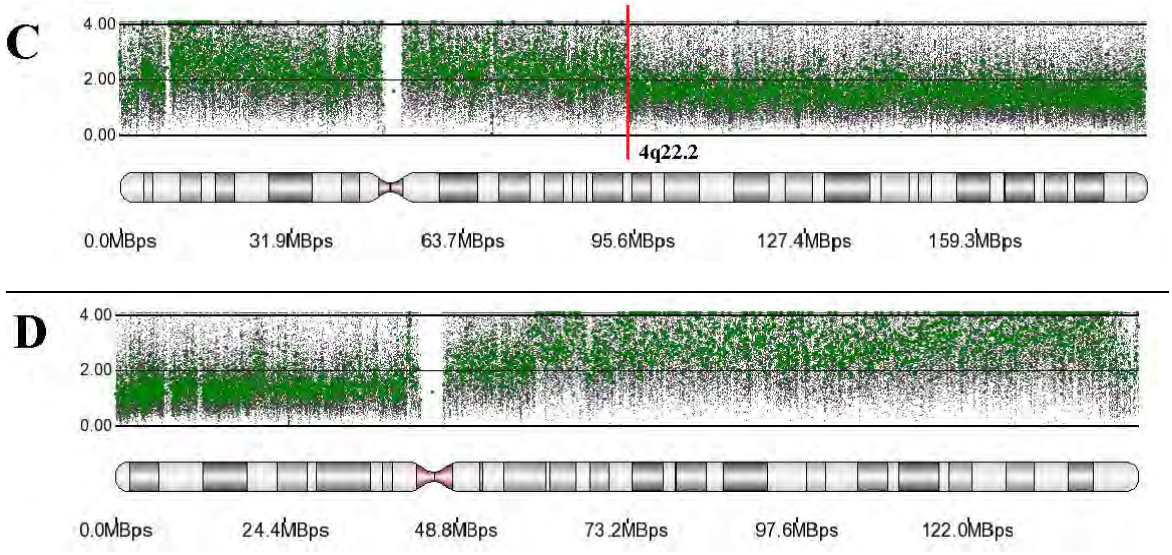
Personalis EX2155



Personalis EX2155



Personalis EX2155



4.3.4 Principal component analysis to interrogate both SNP and CN data

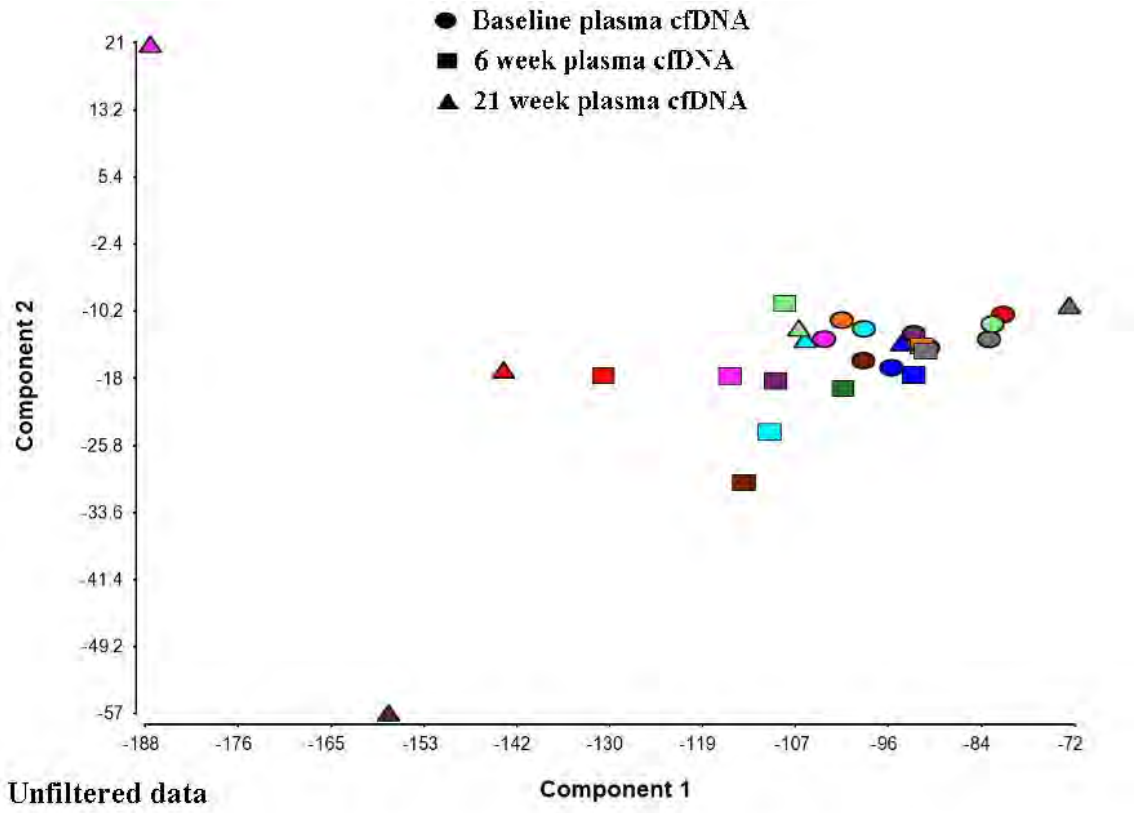
Principal component analysis was used to elucidate differences between the plasma cfDNA samples of the NeoCEnT patients, looking both within individual patients and between time-points. Firstly, looking within patients, Figure 4-7 highlights how the majority of each individual patient's cfDNA samples clustered closely by PCA, indicating little variation. The 21 week cfDNA samples of two patients (pink and brown) did not cluster with the other samples, indicating more variation within these. All markers were included in this comparison (i.e. - unfiltered data) but, overall, only 28% of the variation between the samples was represented by the bi-plot, highlighting overall little variation between these cfDNA samples.

By first filtering the data for significantly-different markers (p-value cut-off of 0.05[1]), comparisons between baseline and 6 week cfDNA (Figure 4-8A) and baseline and 21 week cfDNA (Figure 4-8B) were made. These suggested differences between the different time-points.

[1] PCA bi-plots could not be created using lower p-values as too many markers were filtered out.

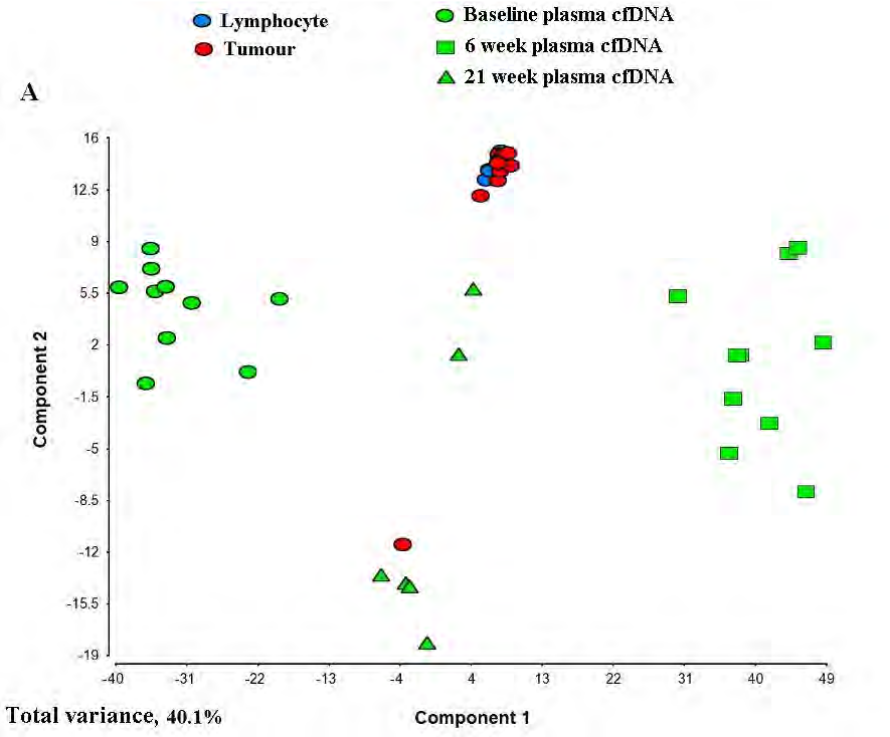
Personalis EX2155

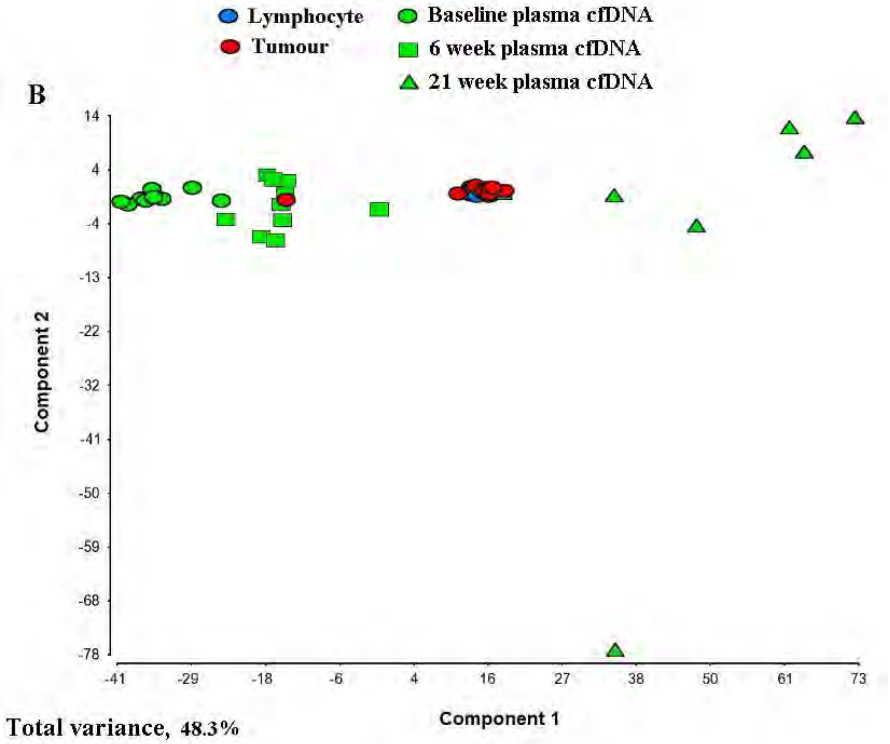
Figure 4-7 - PCA bi-plot of NeoCEnT plasma cfDNA samples (unfiltered data). Individual patient samples are colour-coded.



Personalis EX2155

Figure 4-8 - PCA bi-plot of NeoCEnT group's samples having first filtered data for significant differences between the baseline and 6 week (A) and 21 week (B) cfDNA samples.





Personalis EX2155

Finally, each of the 3 plasma cfDNA time-points was compared, by group, with results for cfDNA of healthy female controls. Markers were filtered-out through PCA using a p-value cut-off of 0.0001. Results showed that there were 6,044 significantly-different markers when comparing the baseline cfDNAs to those of the healthy controls, 4,708 for the 6 week cfDNA, and 1,378 for the 21 week cfDNA, suggesting that the blood plasma profiles became more 'normal' with time on neoadjuvant treatment.

4.4 Discussion

The aim of this chapter was to test the hypothesis that changes in the SNP and CN microarray profiles correlate with resistance to therapy in plasma cfDNA of patients on neoadjuvant therapy.

4.4.1 Microarray detects genomic aberrations

Genomic aberrations were evident in the NeoCEnT cohort of patients, including evidence of local amplifications and deletions (both isolated and common), uncapped chromosomal/telomere attrition, breakages, and isochromosomes. These phenomena were detected by observing CN results. Due to the tumour-specific nature of many of these, CN was regarded as more informative than SNP at distinguishing the NeoCEnT tumour samples, as earlier results had shown high SNP concordance between tumour and matched normal lymphocyte DNA samples.

DNA breakage - including single and double strand breakage - can lead to deletions involving large sections of chromosomal arms extending from the break-point to the telomere. Such breakages are a feature of early cancer development (Ozeri-Galai *et al.*, 2011) and clear evidence of such was seen in the NeoCEnT cohort. The exact mechanism of DNA breakage is unknown, but perhaps cataloguing break-points could help in the understanding of the origins of these. For example, DNA sequences that perturb/slow DNA fork progression during replication could be involved (Ozeri-Galai *et al.*, 2011).

Personalis EX2155

Telomere attrition has been documented in cancer (Gilley *et al.*, 2005 and Naylor *et al.*, 2005) and it has been reported that 85% of immortal tumours have short telomeres (Shawi & Autexier, 2008). In the NeoCEnT tumour DNA samples telomere attrition was evident on every chromosome. The dysfunction of telomeres, which are maintained by TERT (telomerase reverse transcriptase), was hypothesised by Gilley *et al.* (2005) as being responsible for genomic instability in cancers. Telomeres protect the ends of chromosomes from DNA repair proteins that would otherwise recognise them as double strand breaks (Shawi & Autexier, 2008), but they shorten with repeated cell-division due to the inability of polymerase to completely replicate the DNA strand being copied (Heidinger *et al.*, 2012). The Hayflick limit - i.e. the amount of times that a cell can divide - is itself limited by the length of telomeres and, subsequently, the TP53 tumour suppressor. TP53 'captures' the irreparable state of the cell once it has reached the Hayflick Limit through telomere-shortening and then carries it through the senescence process (Patil *et al.*, 2005 and Shawi & Autexier, 2008), thereby avoiding tumourigenesis.

Finally, isochromosomes, which arise due to a transverse splitting of the centromere during replication and result in the loss of one arm and duplication of the other, were evident on chromosomes 1, 8, and 10. Both isochromosomes 1 and 8 have been linked to melanoma (Radhakrishnan *et al.*, 2009). Isochromosome 8 was also reported in breast cancer where it was deduced as having significant pathogenicity (Rummukainen *et al.*, 2001 and Venter *et al.*, 2005).

4.4.2 Single nucleotide polymorphism/copy number variation changes with time and therapy

4.4.2.1 Single nucleotide polymorphism concordance checks give mixed results

The high % concordance noted between lymphocyte and tumour DNA (mean of 88.08%; range 35.47-99.75%) suggested a close similarity between these paired genomes, at least in terms of SNP calls. One other explanation for this is that there was significant contamination from normal cells in the tumour DNA samples, which was suggested by independent pathological review of haematoxylin- and eosin-stained

Personalis EX2155

(H&E) tissue sections (Professor RA Walker, Dr. C Ruangpratheep, and Dr. D Moore). However, conversely, the presence of tumour-derived DNA in these samples was confirmed by the detection of numerous genomic aberrations (Figure 4-6), even though in some of the samples not all DNA was from epithelial tumour cells. The explanation for this lies in the microarray technology itself and how CN is calculated: fluorescence detection is quantitative (Rème *et al.*, 2008), thus, amplification or deletion in tumour DNA at a particular locus will have an effect on the amount of fluorescence detected during image processing after hybridisation has occurred, even in the presence of contaminating normal DNA. Calling the amplification or deletion then rests on the algorithms employed, such as in genomic segmentation or the use of a hidden Markov model (HMM). However, the amount of normal contamination is important, as a high amount will dilute the effect of the amplification or deletion. A previous study, also using *Affymetrix SNP 6.0*, showed that tumour-specific CN could be detected even in the presence of 70% normal tissue contamination (Stjernqvist *et al.*, 2011).

Looking at the concordance of SNP genotype calls for plasma cfDNA and DNA from matched normal lymphocyte, the average concordances increased from 41.19% to 44.3% to 56.4% for the baseline, 6 week, and 21 week plasma cfDNA samples, respectively. This suggests that the plasma cfDNA became more 'normal' with time, possibly due to there being less tumour-derived DNA at the latter time points. This finding matched results found through PCA analysis for these samples. Further intriguing evidence for one patient (N6) exhibited a high SNP concordance to matched lymphocyte for the 21 week cfDNA sample (99.28% concordance) (Figure 4-2), potentially indicating a healthy cfDNA profile at this time-point. Once the study is complete, this result will be confirmable through determining this patient's response to therapy (unknown at the time of thesis writing). Two other patients (N1 and N2) also exhibited SNP concordances above the average (69.76 and 67.88%, respectively) for their 21 week cfDNA sample, but again it was not known whether these patients had clinically responded to therapy.

Personalis EX2155

4.4.2.2 Plasma cfDNA CNV profile becomes less heterogeneous

Based on the number of CNVs called in each sample, there were no significant differences between the blood plasma samples taken at baseline, 6, and 21 weeks (Figure 4-3A). For each patient, there was no obvious pattern between the number of CNVs called in each sample. For example, a low number in lymphocyte DNA did not imply a low number in tumour DNA or cfDNA (Figure 4-3B). However, when observing the distribution between amplifications and deletions, there was a noticeable decrease in the number of deletions with time-point (Table 4-3).

Analysing the frequency of CNV can help to provide insight into which alterations are of biological significance and which are not (Naylor *et al.*, 2005). A possible change in response to therapy was revealed in relation to the frequency of amplification and deletion in the different sample groups (Figure 4-4). The lymphocyte and tumour DNA had visually similar CNV frequency profiles, while the cfDNA samples showed more variation in CNVs. For example, numerous CNVs (both amplifications and deletions) had a prevalence of more than 50% in the lymphocyte and tumour DNA, while there were many more CNVs - but with lower prevalence - in the cfDNA samples, suggesting evolving genomic heterogeneity. Overall, there was a visual difference between the plasma cfDNA profiles at each time-point, appearing to become less heterogeneous with time as certain CNVs gained a higher frequency (Figure 4-4).

4.4.2.3 Principal component analysis indicates subtle changes in the plasma cfDNA profile

Through PCA, differences between individual patients' plasma cfDNA samples with time were also illustrated. For example, the 21 week cfDNA sample for 3 patients (colours brown, pink, and red) migrated away from the main cluster of samples (Figure 4-7). These 3 samples also showed the highest percent concordance of SNP calls with matched normal lymphocyte DNA (Figure 4-2). Other patients' samples did not migrate from the main cluster, potentially suggesting no therapeutic response. Additionally, there were a substantial amount of markers that were different between each of the 3 time-points: first, comparing samples taken at baseline and 6 weeks (Figure 4-8A), the amount of variance represented by the plot was 40.1%. In this plot,

Personalis EX2155

differences between each cfDNA time-point were observed as even the 21 week samples were positioned away from the others. Thus, differences between the baseline and 6 week samples that were highlighted through ANOVA were reflective of differences between each time-point. However, comparing samples at baseline and 21 weeks, the level of variance increased to 48.3% (Figure 4-8B), indicating that there were 8.2% more SNP or CN markers (of the entire ~1.8 million) that were statistically different in the 21 week samples than there were in the 6 week samples. In this plot, the baseline and 6 week samples are grouped closely, highlighting more similarity between these over the 21 week samples. Overall, however, the low amount of variation represented by these plots infers that the observed differences are subtle in origin.

When plasma cfDNA samples at each time-point were compared to those of the healthy controls through PCA, some differences were also seen. For example, the number of significantly-different markers between the healthy control and baseline plasma cfDNA samples was 6,044. This reduced to 4,708 for the 6 week samples, and again to 1,378 for the 21 week samples. Thus, concurrently with time and treatment application, the SNP and CN blood plasma profiles appeared to be becoming more like those of the healthy controls, which may indicate treatment success.

4.4.2.4 Plasma cfDNA CNVs become less prevalent

Of the CNVs showing common amplification that had a prevalence of ≥30% in cfDNA at any time-point (Table 4-4) - and that was absent or quiet in the matched lymphocyte and healthy controls - none was found to increase in prevalence from the baseline to the 21 week sample. The majority increased or remained the same from the baseline to the 6 week sample, but then dropped to below that of the baseline cfDNA prevalence at the end of treatment. Those with the biggest decreases from baseline to 21 weeks included 1p36.21, 5q13.2, 10q11.22, 14q32.33, and 17q12, with decreases of 60, 66.666̇, 70, and 50% respectively). Hence, it is possible that changes in these markers indicate genomic changes in response to therapy.

At 21 weeks and the end of treatment, CNVs that still showed common amplification were predominantly those found in pericentromeric regions (9p12, 9q12, 14q11.1, and

Personalis EX2155

14q11.2), which are regions known for having a higher than average density of CNV (Schuster-Böckler *et al.*, 2010).

Through focussing on the *FGFR*s and *EGFR*, evidence of alteration with time and therapy was also present in the plasma cfDNA samples. The important role that the FGFRs and EGFR play in cancer has been highlighted previously (Tomlinson *et al.*, in press). There were significant differences between the tumour DNA CN at each of these genes and the baseline and 6 week cfDNA samples. However, no significant differences were present at end of treatment (21 weeks), suggesting a change in the plasma cfDNA profile. It is possible that the effect of therapy in these patients resulted in increased tumour-derived DNA (from destroyed cells) circulating in the blood of these patients, which would explain the lack of significant differences at end of treatment.

4.4.3 Genes overlapping commonly-amplified CNVs

The 21 intervals listed in Table 4-4 all showed amplification with a CN>6.0 by Gaussian-smoothed signal and appeared in >10% of samples (on a per sample group). Hence, they represented the most prevalent high level CN gains observed in the series of patients and samples. Of these 21 intervals, 9p12 and 17q12 were novel to the NeoCEnT patient group as they had not been identified previously in the patients on follow-up in Chapter 3 (Table 3-6). Comparing other intervals identified in the patients on follow-up revealed that four other intervals (4q13.2, 5p11, 7q11.21, and 16p11.2) did not exhibit amplification in the NeoCEnT group based on this level of thresholding. Some of these 21 intervals harboured a CNV that appeared in 100% of either the baseline, 6 week, or 21 week cfDNA samples (1p36.33, 2p11.2, 7q11.23, 9p11.2, 10q11.22, 10q11.23, 14q11.1, 14q11.2, and 19p13.3), potentially highlighting their importance in the response to therapy.

The two intervals at 9p12 and 17q12 were unique to the NeoCEnT cohort, in addition to regions already discussed in Chapter 3 (Table 3-6). Only 17q12 harboured a gene, *CCL3L1*, whose transcribed protein is a chemokine receptor ligand and lies in a known segmental duplication (Modi, 2004). *CCL3L1* is linked to HIV1/AIDS susceptibility

Personalis EX2155

and this locus was found to have CN ranging between 1 and 10 in healthy individuals (Field *et al.*, 2009).

Of the intervals that were previously CN-altered but not in the NeoCEnT group (i.e. - 4q13.2, 5p11, 7q11.21, and 16p11.2), only 4q13.2 and 7q11.21 harboured genes. At 4q13.2, which harboured amplification above a CN of 6.0 in 93% of the tumours of patients on follow-up in Chapter 3, are *UGT2B15* and *UGT2B17*. These 2 genes roused much interest as they are up-regulated by 17-β-estradiol and could be involved in maintaining steroid hormone homeostasis and the prevention of excessive oestrogen production (Hu & Mackenzie, 2009). This was evidenced by the fact that deletion in *UGT2B17* was linked to higher levels of estradiol in the blood (Yang *et al.*, 2008), which could promote breast cancer progression (Saha Roy & Vadlamudi, in press). These 2 genes were found to be deleted in 50% of the NeoCEnT tumours, which was of interest because all of the patients in the NeoCEnT group were ER-positive.

Of the intervals that harboured a deleted CNV, *LIMS3* and *RGPD5* overlapped the deletion 2q13. Research is lacking on these genes, but it is reported that *RGPD5* arose from a duplication event (Pruitt *et al.*, 2005). At 16p23.2, *CHST5* is a homologue of *CHST6* (Dastani *et al.*, 2010) and is expressed in mainly the small and large intestine (Pruitt *et al.*, 2005).

Binary logistic regression analysis showed that many of the 21 CNVs common to the NeoCEnT group were significantly contributing to the disease group when compared to the cfDNA of the healthy controls. For example, 12 intervals were significant in both the baseline and 6 week plasma cfDNA samples (but these 12 were not identical; only 9 overlapped), and only 7 were significant in the 21 week cfDNA. The four intervals that were common to all three time-points (9p12, 9p11.2, 14p11.1, and 15q25.2) are possible targets for future studies. The fact that there were 5 less significant CNVs between the 8 and 21 week cfDNA samples could again point to a response to treatment.

Each patient's matched lymphocyte DNA was also used as a reference in binary logistic regression. Through this, many of the CNVs that were already identified as significant to the disease group (when compared to healthy control cfDNA) were again returning

Personalis EX2155

significant values. There was again a reduction in the number of significant CNVs from the baseline, to 6 week, to 21 week cfDNA samples (14, 13, and 11 respectively). Interestingly, only 16p12.3 was significantly to disease in the tumour samples. This particular CNV had already been identified in the patients on follow-up as a common amplification in tumour and a potential biomarker for breast cancer.

4.4.4 *Affymetrix SNP 6.0* reproducibility

The 21 commonly-amplified CNVs presented in Table 4-4 additionally highlight the reproducibility of the microarray technology adopted, as the majority (19) of these intervals were previously observed in the patients on follow-up in Chapter 3. Two intervals that were novel to the NeoCEnT group were 9p12 and 17q12, while the 4 that were not re-encountered were 4q13.2, 5p11, 7q11.21, and 16p11.2.

4.5 Conclusion

Analysis of genomic aberrations in NeoCEnT tumour DNA confirmed the intrinsic instability of the cancer genome, with evidence of numerous types of abnormalities that affected 1,000s of genes. Also, through cfDNA analysis, there was strong evidence of alteration with time and therapy between the baseline, 6 week, and 21 week/end of treatment cfDNA samples. This was inferred by subtle changes in the SNP and CN profiles of these samples through different analyses:

1. Through SNP concordance checks, the average concordance of plasma cfDNA compared to matched lymphocyte DNA increased from 41.19 to 44.30% for the baseline and 6 week cfDNA samples, and then to 56.40% the end of treatment samples.
2. Change was also observed by observing the distribution between types of CNVs, with amplifications increasing from 33 to 45% (beginning to end of treatment) and deletions decreasing from 67 to 55%.

Personalis EX2155

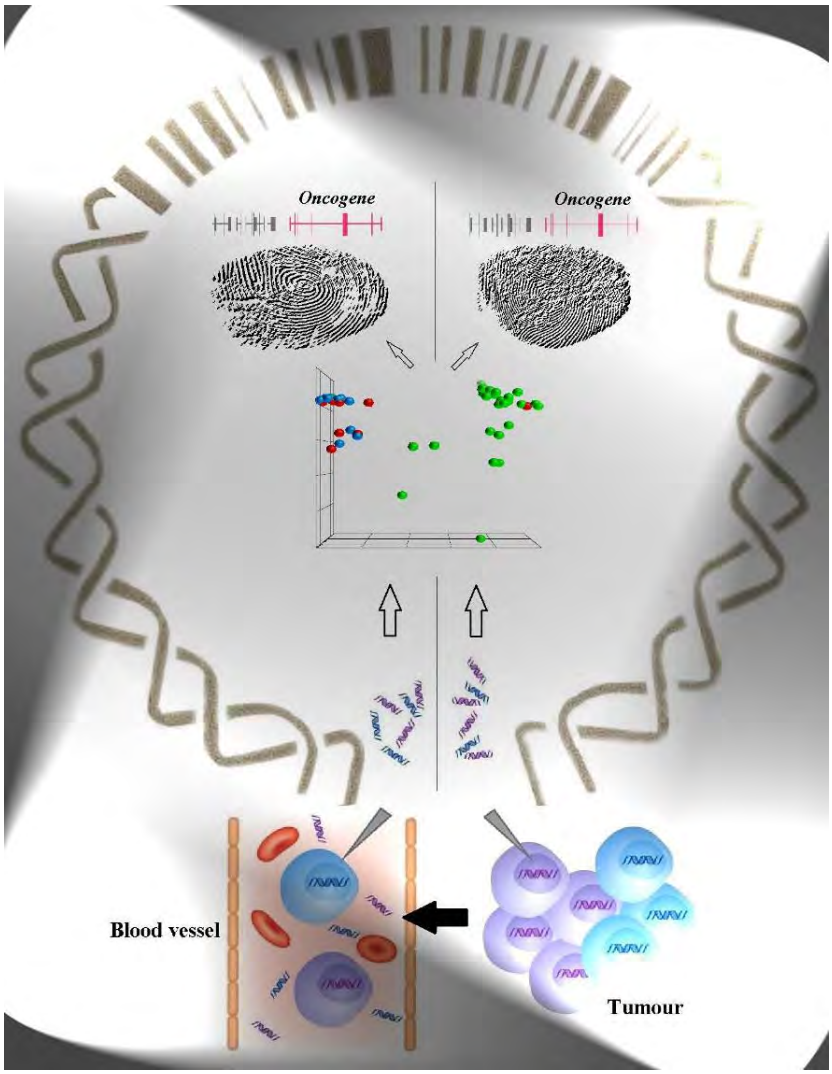
3. Principal component analysis also highlighted differences, with 3 patients' end of treatment samples appearing away from the main cluster of samples (Figure 4-7), suggesting that these samples harboured numerous markers of difference[1].
4. Binary logistic regression analysis revealed that the number of intervals harbouring commonly-amplified CNVs (Table 4-4), which significantly contributed to disease, was reduced from the baseline to the 6 week cfDNA, and also to the end of treatment cfDNA.

4.6 Future studies

Due to the fact that the clinical response status was known for just 50% of the patients in the NeoCEnT cohort at the time of writing (2 responders; 3 non-responders; 5 unknown status) - and the fact that only 10 patients' samples were studied overall - it is difficult to extrapolate whether the observed alterations correlated to a therapy response or non-response. A similar study involving a much larger number of patients could provide a better means of determining this.

[1] The variance represented by the plot was low (28%), therefore the differences observed were subtle.

Personalis EX2155



5 PAINTING A PORTRAIT OF BREAST CANCER THROUGH PRINCIPAL COMPONENT ANALYSIS

Personalis EX2155

5.1 Overview

In this chapter, an attempt is made to define a breast cancer circulating cell-free (cfDNA) 'haplotype' or genetic signature that could have utility in the diagnostic setting. Principal component analysis (PCA) is used to derive a preliminary set of markers that are then subjected to further analyses that overcome limitations of PCA and that are aimed specifically at characterising data.

5.2 Introduction

Elaborate downstream methods are required to analyse large microarray data-sets. At times, where the end goal is to look for relationships between - or patterns within - different subgroups (or indeed individual samples), large data-sets must first be filtered using statistical thresholds in order to reduce their overall volume. Such 'dimension reduction' techniques are essential to elucidate any links between CNV and tumourigenesis in microarray cancer studies. A subset of data can first be taken to represent the larger data-set - for example, during elections, where polls are taken before voting, results are used to infer the opinions of the population at large (Smith, 2002 and Van de Wiel *et al.*, 2011). However, what is the best method of choosing such a subset to represent the overall portrait?

5.2.1 Principal component analysis as a tool for analysing large bodies of data

Principal component analysis and other multivariate tools are used to analyse large volumes of data in order to tease out the differences/relationships between the logical entities being analysed (for example, a data set consisting of a large number of samples, each with their own data points/variables). It extracts the fundamental structure of the data without the need to build any model to represent it (Paschou *et al.*, 2010). This 'summary' of the data is arrived at through a process of reduction that transforms the large number of variables into a lesser number that are uncorrelated (i.e. - the 'principal' components; see Appendix 8.4 for information on how principal components are derived), whilst at the same time being capable of easy interpretation on the original data (Mazlum *et al.*, 1999 and Joliffe, 2002).

Personalis EX2155

PCA has broad applications in a wide range of areas. Examples include craniofacial recognition (Chakravarty *et al.*, 2011), analysis of water quality (Mazlum *et al.*, 1999), and to derive a set of highly confident genes (Bicciato *et al.*, 2003) or single nucleotide polymorphisms (SNPs) (Meng *et al.*, 2003 and Lin & Altman, 2004) for classification purposes. Principal component analysis has also been used in subject areas such as climatology, geology, meteorology, psychology, quality control (Joliffe, 2002), forensics and population genetics (particularly in relation to SNPs), and medical genetics (Paschou *et al.*, 2010). It can also help in the identification of subgroups within samples by visually scanning the bi-plot created to represent the data (Van de Wiel *et al.*, 2011), as utilised in Chapters 3 and 4.

In PCA experiments, two different approaches can be taken: 1) looking at relationships between variables; and 2) looking at relationships between samples. If only two variables are involved, then a simple linear correlation analysis can be employed. However, having numerous variables prevents this (Mazlum *et al.*, 1999).

5.2.1.1 Limitations of PCA

There are some limitations to PCA. For example, PCA is a linear transformation - thus any data-set that is non-linear will not be represented sufficiently after data-reduction. In addition, PCA assumes that the directions with the largest variances are of most interest (Maathuis, 2008), which might not necessarily be true. It also follows the assumption that the original data variables are correlated - if they are not, then PCA cannot reduce the data (Maathuis, 2008 and Parsons *et al.*, 2009).

5.2.2 Deriving a genetic 'signature'

A method used to minimally characterise individual samples from large data-sets was previously employed by finding haplotype-tagging SNPs (htSNPs) - i.e. a reduced list of SNP markers that captured the majority of haplotype diversity in a population (Lin & Altman, 2004). However, the original method, which was devised by Meng *et al.* (2003) (see Appendix 8.4.1.2.3), was only applied to analyse certain regions of the genome. Additionally, redundancy could still exist using this approach if some of the

Personalis EX2155

derived htSNPs mapped to regions affected by linkage disequilibrium (LD) (LaFramboise, 2009 and Paschou *et al.*, 2010), which can encompass segments ranging in size from 1-100Kb (Reich *et al.*, 2001). Due to the high correlation, measuring the value of one allele can reveal that of the other if they exhibit LD (LaFramboise, 2009).

Linkage disequilibrium can arise for different reasons - selection, favourable mutations, population mixture and migration, etc. - and, if measured around known genes, can help in the elucidation of those that are involved in disease (Reich *et al.*, 2001 and Sabatti & Risch, 2002). It is more common around genes that are 'rare' or recently evolved as these would have had less time for recombination to break the disequilibrium in question (Reich *et al.*, 2001).

Lin & Altman (2004) found that using this method could produce a list of htSNPs that achieved a 90% reconstruction precision of each observed haplotype. The process involved the generation of eigenvectors (the synonymous term 'eigenSNPs' was coined by Meng *et al.* (2003)) (see Appendix 8.3 for an explanation of eigenvectors and eigenvalues) and then reverse-mapping these to original SNPs to arrive at a minimal set that could define maximum diversity/variance. Single nucleotide polymorphisms are useful in this regard because they are the most common type of genetic variation in the human genome (Kruglyak & Nickerson, 2001).

The derivation of eigenSNPs has also been employed in population and anthropological studies. For example, the present day genetic variation in humans around the world exists and has been strengthened by the history of migration patterns (Paschou *et al.*, 2007 and 2010). What contributed to these included mutation, genetic drift, and natural selection that were each driven by an interaction with new climates, pathogens, etc. (Paschou *et al.*, 2010). This prompted Paschou *et al.* (2010) to attempt to build a scoring system to assign an individual of unknown origin to a population group based on PCA and SNPs - some success was achieved. However, their system was based on diverse population groups, between who was exhibited already-known differences/separation, both genetically and geographically. Thus, their method of assignation to a particular clade/population was made easier. Also, they only had small sample sizes from each observed group.

Personalis EX2155

The method has yet to be applied to copy number variations (CNVs) in the human genome, whether healthy or cancer. However, given the density of probes in the *Affymetrix SNP 6.0*, applying the htSNP method using both SNP and CN probes could generate a genetic signature for both healthy and cancerous samples. Hence, it could potentially provide a minimum set of diagnostic and prognostic markers.

5.2.3 Aims

The aims of this chapter were to characterise breast cancers through PCA by arriving at a concise number of genes and markers that were informative of cancer status, and also to highlight the usefulness of cfDNA as a liquid biopsy for cancer monitoring. Three groups of samples were compared: unaffected *BRCA1* mutation carriers; primary breast cancer patients on follow-up who did and did not relapse; and pre-surgical breast cancer patients receiving neoadjuvant therapy[1] (see Table 2-1 for details on these).

The objectives were:

1. To conduct comparison experiments between sample groups (using PCA) to identify genes whose alterations were reflective of the differences observed in each comparison - i.e. those which were 'drivers' of variation (5.3.1).
2. Using plasma cfDNA of healthy controls as baseline, to create distinct breast cancer cfDNA signatures using the htSNP method (5.3.2).

5.3 Results

5.3.1 Genes with known roles in breast cancer

Comparison experiments that were made using PCA are outlined in Table 5-1.

[1] Neoadjuvant chemo- or endocrine therapy (NeoCEnT) cohort.

Personalis EX2155

Group (number)	ANOVA comparison			Nature of variation
Unaffected *BRCA1* carriers (n=5)	Healthy control P	Vs	Unaffected *BRCA1* P	Specific to unaffected *BRCA1* carrier cfDNA
Patients on follow-up (n=42) - T_1N_0 (T≤20mm) (n=14) - T>20mm/node-positive (n=28)	L	Vs	T	Tumour-specific
	T_1N_0 (T≤20mm) P1	Vs	T>20mm/node-positive P1	Specific to T>20mm/node-positive cfDNA
	T_1N_0 (T≤20mm) P2	Vs	T>20mm/node-positive P2	Specific to T>20mm/node-positive cfDNA
Relapsed patients on follow-up (n=8) - T_1N_0 (T≤20mm) (n=4) - T>20mm/node-positive (n=4)	P1	Vs	P2	Alterations in relapsed cfDNA profile that could be indicative of relapse
	T	Vs	P2	Alterations in cfDNA profile, potentially tumourigenic in origin, that arose during dormancy
	T_1N_0 (T≤20mm) non-relapsed P2	Vs	Relapsed P2	Specific to relapsed cfDNA
	T>20mm/node-positive non-relapsed P2	Vs	Relapsed P2	Specific to relapsed cfDNA
Pre-surgical patients on neoadjuvant therapy (n=10)	L	Vs	T	Tumour-specific
	Baseline P	Vs	6 week P	Alterations in cfDNA profile
	Baseline P	Vs	21 week P	Alterations in cfDNA profile
	T	Vs	21 week P	Alterations in cfDNA profile, potentially tumourigenic in origin, that arose in response to therapy

Table 5-1 - PCA experiments conducted to deduce 'driver' genes of variation, including aims for each. L, lymphocyte; P, healthy control or pre-surgical plasma cfDNA; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour. NB - n=7 for 21 week P of pre-surgical group patients.

Markers having a correlation to the first principal component of >0.7 and <-0.7 (more stringent correlation coefficients were used in certain cases) from PCA were overlapped with the human genome to reveal a list of genes (Appendix Table 8-2). These genes were then cross-compared with the OMIM® disease database (Online Mendelian Inheritance in Man, http://omim.org/) using selected keywords related to 'cancer' and 'breast cancer' (PubMed Health, http://www.ncbi.nlm.nih.gov/pubmedhealth/): "carcinoma", "malignant tumor/tumour", "breast cancer", "ductal carcinoma", "lobular carcinoma", "DCIS", "LCIS", "HER2/ERBB2-positive breast cancer", "ER-positive breast cancer", ductal carcinoma *in situ*", "lobular carcinoma *in situ*"[1]. These genes are outlined in Table 5-2 (see Appendix Table 8-3 for additional information on each identified marker).

[1] Keywords "estrogen/oestrogen", "progesterone", and "growth factor" were also used.

Personalis EX2155

Comparison	SNP				CN		
	Symbol	Name	A/B allele	Prevalence % (matched lymphocyte)[1]	Symbol	Name	State, Prevalence % (matched lymphocyte)
Patients on follow-up:							
• L Vs T	*BRAF*	V-raf murine sarcoma viral oncogene homologue B1	C/G	97.5 (18)	*ESR1*	Oestrogen receptor 1	Deleted, 100 (0)
	CASP10	Apoptosis-related cysteine peptidase	C/G	82.5 (6)	*DNAJC7**	DnaJ (Hsp40) homolog, subfamily C, member 7	Deleted, 97.5 (0)
	DCC	Deleted in colorectal carcinoma	A/G	70 (4)			
	*EGFR**	Epidermal growth factor receptor	C/T	92.5 (12)	*GPC3*	Glypican 3	Deleted, 95 (0)
	EPHB2	EPH receptor B2	C/T	20 (0)			
	ESR1	Oestrogen receptor 1	C/T	82.5 (4)			
	*FYTTD1**	Forty-two-three domain containing 1	C/T	82.5 (30)			
	GRIK1	Glutamate receptor, ionotropic, kainite 1	C/T	90 (6)			
	IL1RN	Interleukin 1 receptor antagonist	G/T	97.5 (20)			
	IL1RN	Interleukin 1 receptor antagonist	G/T	97.5 (28)			
	LHCGR	Luteinising hormone, choriogonadotropin receptor	C/T	92.5 (10)			
	OPCML	Opioid binding protein, cell adhesion molecule-like	A/C	-[2]			
	OPCML	Opioid binding protein, cell adhesion molecule-like	A/G	67.5 (8)			
	PALLD	Palladin, cytoskeletal associated protein	C/T	70 (6)			
	*PDGFRA**	Platelet-derived growth factor receptor, α polypeptide	C/T	50 (14)			
	PRKCA	Protein kinase C, α	C/T	67.5 (16)			
	PTEN	Phosphatase and tensin homologue	G/T	-[3]			
	RB1	Retinoblastoma 1	C/T	80 (4)			
	TNFSF4	Tumour necrosis factor (ligand) superfamily, member 4	A/G	90 (8)			
	TUSC3	Tumour suppressor candidate 3	C/T	75 (26)			
	WWOX	WW domain-containing oxidoreductase	C/T	52.5 (2)			
	WWOX	WW domain-containing oxidoreductase	G/T	70 (4)			
	WWOX	WW domain-containing oxidoreductase	A/C	15 (2)			
• T_1N_0 (T≤20mm) P1 Vs T>20mm/node-positive P1	NA	NA	NA	NA	NA	NA	NA

[1] Includes both homo- and heterozygous changes from the baseline genotype - e.g. AA→AB or AA→BB

[2] 'No call' in 25% of tumour.

[3] 'No call' in 70% of tumour.

Personalis EX2155

• T_1N_0 (T≤20mm) P2 Vs T>20mm/node-positive P2	*BCAS3*	Breast carcinoma amplified sequence 3	C/T	71 (0)	*PKHD1*	Polycystic kidney and hepatic disease 1 (autosomal recessive)	Deleted, 44 (0)
Relapsed patients on follow-up:							
• P1 Vs P2	*MCC*	Mutated in colorectal cancers	A/G	-[1]	*ABCC4* *IGF2R* *MAML2** *PARK2*	ATP-binding cassette, sub-family C, member 4 Insulin-like growth factor 2 receptor Mastermind-like 2 (Drosophila) Parkinson disease (autosomal recessive, juvenile) 2	Amplified, 75 (100) Normal, 75 (37.5) Deleted, 62.5 (12.5) Amplified, 75 (50)
• T Vs P2	*CD209**	CD209 molecule	A/G	75 (100)	*TP63*	Tumour protein p63	Normal, 75 (62.5)
	DBC1	Deleted in bladder cancer 1	C/T	87.5 (100)			
	DEC1	Deleted in oesophageal cancer 1	C/T	75 (100)			
	GRIK1	Glutamate receptor, ionotropic, kainite 1	C/G	50 (12.5)[2]			
	GLG1	Golgi apparatus protein 1	A/G	-[3]			
	IL1RN	Interleukin 1 receptor antagonist	G/T	37.5 (50)			
	KIT	Similar to mast/stem cell growth factor receptor precursor (SCFR)	C/T	-[4]			
	OPCML	Opioid-binding protein, cell adhesion molecule-like	C/T	12.5 (12.5)			
	OPCML	Opioid-binding protein, cell adhesion molecule-like	C/T	75 (87.5)			
	PALLD	Palladin, cytoskeletal associated protein	A/G	25 (0)[5]			
	MAML2	Mastermind-like 2 (Drosophila)	A/G	87.5 (87.5)			
	*TTLL1**	Tubulin tyrosine ligase-like 1	A/G	-[6]			
	WWOX	WW domain-containing oxidoreductase	C/T	25 (37.5)			
	WWOX	WW domain-containing oxidoreductase	C/T	100 (87.5)			
• T_1N_0 (T≤20mm):	*XRCC3*	X-ray repair complementing defective repair in Chinese	C/G	100 (75)	*ABL1**	C-abl oncogene 1, receptor tyrosine kinase	Deleted, 100 (100)

[1] 'No call' in 37.5% of second plasma cfDNA samples.

[2] 'No call' in 80% of tumour DNA samples.

[3] 'No call' in 60% of tumour DNA samples.

[4] 'No call' in 60% of tumour DNA samples.

[5] 'No call' in 100% of tumour DNA samples.

[6] 'No call' in 100% of tumour DNA samples.

Personalis EX2155

Non-relapsed P2 Vs Relapsed P2		hamster cells 3			*BRCA1* *PDGFRB* *RHPN2* *SMAD4*	Breast cancer 1, early onset Platelet-derived growth factor receptor, β polypeptide Rhophilin, ρ GTPase binding protein 2 SMD family member 4	Amplified, 50 (100) Amplified, 100 (75) Amplified, 75 (50) Amplified, 50 (0)
• T>20mm/node-positive: Non-relapsed P2 Vs Relapsed P2	NA	NA	NA	NA	*PTCH1*	Patched 1	Amplified, 100 (100)
Unaffected *BRCA1* carriers:							
• Healthy control P Vs Unaffected *BRCA1* carriers P	*DIRC3* *DIRC3* *DLC* *EGFR** *PALLD* *WWOX* *WWOX*	Disrupted in renal carcinoma 3 Disrupted in renal carcinoma 3 Deleted in liver cancer 1 Epidermal growth factor receptor Palladin, cytoskeletal associated protein WW domain-containing oxidoreductase WW domain-containing oxidoreductase	A/G A/G C/T C/T A/C A/C C/T	- (87.5) 100 (20) 60 (0) - (100) -[1] - -	NA	NA	NA
Pre-surgical patients on neoadjuvant therapy:							
• L Vs T	*BCAS3* *DLC1* *DLC1* *MCC* *PALLD* *PALLD* *PARK2* *PARK2* *WWOX*	Breast carcinoma amplified sequence 3 Deleted in liver cancer 1 Deleted in liver cancer 1 Mutated in colorectal cancers Palladin, cytoskeletal associated protein Palladin, cytoskeletal associated protein Parkinson disease (autosomal recessive, juvenile) 2 Parkinson disease (autosomal recessive, juvenile) 2 WW domain-containing oxidoreductase	A/C A/G G/T A/G C/T A/G C/T C/G A/G	10 (0) 10 (0) 10 (0) 10 (0) 10 (0) 10 (0) 30 (20) 10 (0) 40 (40)	NA	NA	NA
• Baseline P Vs 6 week P	NA	NA	NA	NA	NA	NA	NA
• Baseline P Vs 21 week P	*PKHD1* *TGFBR2*	Polycystic kidney and hepatic disease 1 (autosomal recessive) Transforming growth factor, β receptor 2	A/G A/G	43 (100) 43 (60)	NA	NA	NA

[1] No call in 100% of unaffected *BRCA1* cfDNA samples.

Personalis EX2155

• T Vs 21 week P	*BCAS3* *PRKCA*	Breast carcinoma amplified sequence 3 Protein kinase C, α	A/G A/T	-[1] 14 (80)	NA	NA	NA

Table 5-2 - Genes overlapping significantly-differentiated markers (SNP or CN) through post-PCA analysis. Markers were first filtered to include those having a correlation of >0.7 or <-0.7 to the first principal component and then overlapping these with the human genome to unearth a list of genes. These were then filtered to include those only present in OMIM® and having keywords related to 'cancer' and 'breast cancer' (PubMed Health, http://www.ncbi.nlm.nih.gov/pubmedhealth/). L, lymphocyte; P, healthy control or pre-surgical plasma cfDNA; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour. Amplification=CN>2.5; deletion=CN<1.5. Appendix Table 8-3 provides additional information for these markers. NB - prevalence and state relate to the latter sample in each comparison; asterisk (*) denotes SNP/CN marker mapping to an exon; genes listed alphabetically; 'NA' denotes 'not applicable' due to no significant markers using the selected keywords; 'No call' implies a failed hybridisation and deleted locus (LaFramboise, 2009).

[1] 'No call' in 14% of 21 week cfDNA samples.

Personalis EX2155

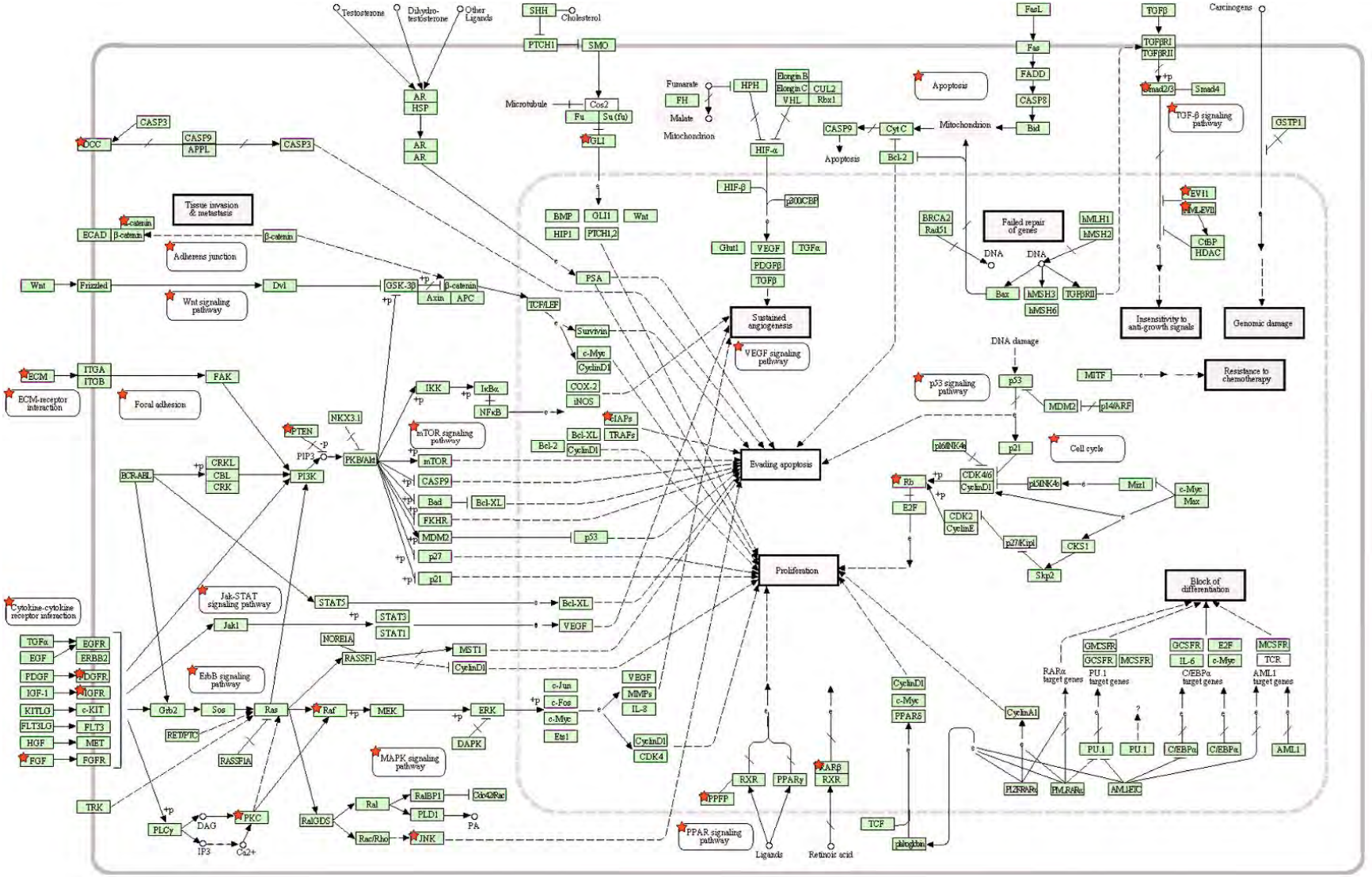
In total, there were 1,549 SNP and 476 CN markers called as significant across all comparisons made, some of which overlapped the same gene (Table 8-2). When these were filtered to include markers whose overlapping genes were known to be involved in breast cancer, 60 SNP and 15 CN markers remained (Table 5-2). The majority of markers mapped to introns, with only 9 mapping to exons.

The comparison that generated the majority of significant SNP markers (23) was the tumour DNA of patients on follow-up compared to matched normal lymphocyte DNA. The comparison that generated the most significant CN markers (5) was the T_1N_0 ($T \leq 20mm$) relapsed final cfDNA compared to non-relapsed final cfDNA.

Of all genes found in on follow-up tumour DNA that had an overlapping, significant SNP or CN marker (total 724 unique; Table 8-2), 36% are listed as being intrinsic to the cell membrane, 12% relating to the cytoskeleton, and 10% being involved in cell adhesion, according to Gene Ontology key terms (Gene Ontology Consortium, 2005). Additionally, 2.5% are included in the cancer pathway of KEGG (Kanehisa & Goto, 2000), with each type of cancer listed (numbering 14) being affected in one way or another (Figure 5-1). Numerous genes that result in the evasion of apoptosis and cellular proliferation are affected by SNP or CNV.

Personalis EX2155

Figure 5-1 - KEGG cancer pathway showing genes overlapped by significant SNP/CN markers (red stars). NB - some genes have more than 1 pathway role.



Personalis EX2155

5.3.1.1 Markers overlapping exons

Overall, there were only 9 markers identified by PCA that overlapped with exons in the human genome. In the 42 patients on follow-up that didn't relapsed, 4 genes harboured markers that were significantly different between lymphocyte and tumour DNA: *DNAJC7* was CN-altered, while *EGFR* (genotype AA to AB in 92.5% of cases), *FYTTD1* (BB to AB and AA in 60 and 22.5% of cases, respectively), and *PDGFRA* (BB to AB in 50% of cases) were each overlapped by a significant SNP marker (Table 5-2). The same SNP marker for *EGFR* was also significant in the comparison between the cfDNA of healthy controls and unaffected *BRCA1* carriers. However, this was due to the fact that it returned 'no call' in 25% of the healthy control samples.

In *DNAJC7*, the 25bp CN marker beginning at position 40,148,320 on chromosome 17 was deleted in all but one (97.5%) of the tumour DNA samples of the patients on follow-up. However, the CN marker was not significant in any cfDNA sample.

In the 8 patients on follow-up who relapsed, there were few significant changes identified in exons that could characterise relapse. However, *CD209* had a significant SNP alteration (genotypes BB) that was seen in the cfDNA samples taken just before relapse compared back to heterozygous (AB) tumour. *MAML2* was CN-altered (marker beginning at position 95,712,668 on chromosome 11), showing deletion also in these final cfDNA samples for 5 of the 8 patients who relapsed, with no deletion in any of matched first cfDNA samples. *ABL1* (marker beginning at position 133,761,278 at 9q34.12) was also deleted in these final cfDNA samples, but this case in all of the T_1N_0 ($T \leq 20mm$) relapsed second cfDNA samples. However, for each of these genes, the found variant was also present in matched normal lymphocyte DNA. The exception was *TTLL1*, whose significant overlapping SNP marker returned a 'no call' in all relapsed tumour DNA samples, suggesting failed hybridisation and a deleted locus.

Personalis EX2155

5.3.1.2 Markers overlapping introns

5.3.1.2.1 Breast cancer patients on follow-up

When comparing the tumour DNA samples from the patients on follow-up to their matched normal lymphocyte DNA, 17 individual genes of interest that were related to cancer contained a SNP or CN marker that mapped to an intron. Of these, 15 contained SNP markers, 1 a CN marker, and 1 gene contained both a SNP and CN marker). The genes affected by SNP included *BRAF*, *CASP10*, *DCC*, *EPHB2*, *ESR1* (SNP and CN markers), *GRIK1*, *IL1RN* (x2), *LHCGR*, *OPCML* (x2), *PALLD*, *PRKCA*, *PTEN*, *RB1*, *TNFSF4*, *TUSC3*, *WWOX* (x3). The SNPs occurred at a high prevalence in the tumour samples (mean 73.5%; range 15-97.5%) and low in matched lymphocyte (mean 10.9%; 0-28%). In all cases, the tumour-specific SNP genotype was also observed in numerous matched first and second cfDNA samples, which is inferential of tumour DNA circulating in the blood plasma of these patients. The two genes with overlapping significant intronic CN markers were *ESR1* and *GPC3*, which were deleted in 95 and 100% of tumour DNA, respectively, and not found in any of the matched lymphocytes.

Comparing the T_1N_0 ($T \leq 20mm$) and T>20mm/node-positive first cfDNA samples, there were no SNP or CN markers that were significant through PCA post-analysis between the 2 groups. However, when comparing the second cfDNA samples between these groups, *BCAS3* had an overlapping SNP marker, while *PKHD1* had a CN marker. The *BCAS3* SNP was prevalent in 71% of the T>20mm/node-positive second cfDNAs and not present in any of the matched lymphocyte samples, while the *PKHD1* deletion was observed in 44% of the cfDNA samples with none found in matched lymphocyte DNA samples.

5.3.1.2.2 Relapsed breast cancer patients on follow-up

Focussing on the 8 patients who relapsed, the goal was to search for markers that differed in the final (second) plasma cfDNA sample, which was the sample taken closest to the time of relapse, compared to: 1) matching first cfDNA; and 2) matching tumour DNA. SNP and CN markers were found at a high prevalence through both comparisons. Contrastingly, mostly CN markers differed between the first and second

Personalis EX2155

cfDNA samples, whereas it was mostly SNP markers that differed when comparing tumour and final plasma cfDNA.

Compared to just the first cfDNA samples, only *MCC* contained a SNP marker that differed between the two groups. However, the SNP marker was significant in the final cfDNA samples due to having returned a 'no call' in 37.5% of cases. Genes containing significant overlapping CN markers were *ABCC4* (amplified in 75% of final cfDNA and all matched lymphocyte DNA), *IGF2R* (normal in 75% of final cfDNA and mixture of normal and amplified CN in matched lymphocyte DNA). Their high prevalences in normal lymphocyte DNA reduced their importance in cancer.

Compared to their matched tumour DNA samples, many genes harbouring SNP markers of significance were found, including 5 that were previously mentioned as significant in the tumour DNA of patients on follow-up: *GRIK1*, *IL1RN*, *PALLD*, *MAML2*, and *WWOX*. Of these, only *IL1RN* featured the same SNP marker previously identified in these tumour DNA samples of patients on follow-up. However, the genotype of that marker (BB) was dissimilar to this one in cfDNA, which resembled that found in matched lymphocyte DNA, suggesting that the mutant allele of this SNP occurred in tumour. For each of the other 4 genes mentioned, the same scenario repeated itself - i.e. the identified SNP had the same genotype as that in matched lymphocyte DNA but that differed in tumour DNA.

Other genes found in this comparison (relapsed tumour DNA Vs final cfDNA) include *DBC1* and *DEC1* (87.5 and 75% prevalences, respectively). However, in both cases the significant SNP marker had the same genotype as that found in matched lymphocyte, but the estimated CN output for each locus in matched tumour was consistent with deletion/loss of heterozygosity (LOH). *KIT* and *GLG1* also harboured significant SNP markers, both due to returning 'no calls' in 60% of the tumour samples in the comparison, implying deletion. For both *KIT* and *GLG1*, the genotypes of the SNPs in the cfDNA samples resembled those found in matched lymphocytes, implying healthy cfDNA profiles. Only *TP63* was found to harbour a significantly-different CN marker in this comparison, which had normal CN in 75% of the final cfDNA samples and 62.5% of matched lymphocytes, but deleted in all of matched tumour DNA.

Personalis EX2155

Looking at just the T_1N_0 ($T \leq 20mm$) relapsed patients, when comparing the final cfDNA sample to that of the non-relapsed patients, a single significantly-different SNP marker was found overlapping *XRCC3* with a prevalence of 100%. However, this SNP had the same genotype as that appearing in the matched healthy lymphocyte DNA samples, and was thus not deemed to be a mutation. Four genes harboured intronic CN markers that highly differentiated between relapsed and non-relapsed cfDNA: *BRCA1* (amplified in half of relapsed cfDNA and all of matched lymphocyte), *PDGFRB*, *RHPN2* (amplified in 100 and 50% of relapsed cfDNA, respectively, but also showing amplification in matched lymphocyte), and *SMAD4* (deleted in all of relapsed cfDNA and amplified in all of non-relapsed cfDNA).

In the T>20mm/node-positive relapsed and non-relapsed final cfDNA comparison, only *PTCH1* had a significantly different CN marker. It was amplified in all of the relapsed cfDNA samples and deleted in 75% of the non-relapsed. However, the amplification was also observed in matched lymphocytes.

5.3.1.2.3 Unaffected BRCA1 mutation carriers

There were 4 genes that had an overlapping and significant SNP when comparing the cfDNA of the healthy controls and unaffected *BRCA1* carriers (Table 5-2): *DIRC3* (x2), *DLC1*, *PALLD*, and *WWOX* (x2). In *DIRC3*, the first SNP marker (position 218,294,420) in the unaffected *BRCA1* carriers' cfDNA had the same genotype as that found in matched lymphocyte and was therefore not regarded as a mutation. The marker seemed to have been regarded as significant due to having failed to hybridise in 1 of the healthy control cfDNA samples. The second SNP marker (position 218,301,532), however, was seen in just 20% of the matched lymphocytes and had a prevalence in the unaffected *BRCA1* carriers' cfDNA of 100% (genotype BB to AB/AA in 80/20% of cases, respectively). That of *DLC1* (deleted in liver cancer 1) (genotype AA to AB in 60% of cases) had a prevalence of 60% in the *BRCA1* carriers' cfDNA and was not observed in matched lymphocyte.

Personalis EX2155

Both *PALLD* and *WWOX* were already both found to harbour significant SNPs in the tumour DNA of patients on follow-up and also in the final cfDNA of relapsed patients. In the cfDNA of the unaffected *BRCA1* carriers, both SNPs in *WWOX* had the same genotype as that of their matched lymphocytes. However, the SNP marker contained within *PALLD* was regarded as significant because it had failed to hybridise in each of the unaffected *BRCA1* carriers' cfDNA samples, suggesting complete loss/deletion at this locus.

5.3.1.2.4 Pre-surgical breast cancer patients on neoadjuvant therapy

Between tumour and matched lymphocyte DNA, a number of significantly-different SNP markers in tumour were discovered within genes already identified in previous comparisons: *BCAS3*, *DLC1* (x2), *MCC*, *PALLD* (x2), *PARK2* (x2), and *WWOX*. However, none of the SNP markers identified matched those previously found. In addition, for each of these (except for a single SNP in *PARK2* and that found in *WWOX*), the SNP marker was found in the same pre-surgical tumour sample. Of the 2 SNPs found in multiple tumour samples, both were also identified in matched lymphocyte at high prevalences of 20 and 40%, respectively, for *PARK2* and *WWOX*, ruling them out as having significant roles in disease. The marker in *WWOX* had an equal prevalence in both tumour and lymphocyte DNA but was still regarded as significant due to individual paired samples' genotypes not matching.

Significantly-differentiated markers in this patient group were not found between the baseline and 6 week cfDNA samples. However, significant SNP markers were found for comparisons made with the 21 week cfDNA samples and: 1) baseline cfDNA; and 2) tumour DNA. In total, 2 SNPs were found to be significantly-different between the baseline and 21 week cfDNA samples. These were contained within *PKHD1* and *TGFBR2* (both at a prevalence of 43%). When compared to tumour, 2 SNPs were also found, contained within *BCAS3* and *PRKCA*. For the SNPs mapping to *PKHD1* and *TGFBR2*, genotypes resembled those found in matched lymphocyte DNA, indicating a 'healthy' profile in these 21 week cfDNA samples (and not so in the baseline cfDNA samples). In 2 samples (29%), however, the *PKHD1* SNP probes failed to hybridise,

Personalis EX2155

which suggested deletion at this locus. The same occurred in 3 samples around the *TGFBR2* SNP marker, again suggesting deletion.

The SNP marker found within *PRKCA* had a low prevalence (14%), however, the probe failed to hybridise in 5 samples (71%). A significant SNP was already found in 67.5% of the tumour DNA samples of the on follow-up patients within this gene. The SNP marker within *BCAS3* was called as significant due to its probe having failed to hybridise in 14% of cases.

5.3.2 Towards a breast cancer signature through cfDNA analysis

A breast cancer cfDNA signature was created through PCA using the cfDNA of the healthy controls as baseline/reference. This was carried out for the unaffected *BRCA1* mutation carriers, primary breast cancer patients on follow-up who did and did not relapse, and pre-surgical breast cancer patients on neoadjuvant therapy. A large number of markers remained in each comparison after filtering out markers of insignificance compared to the healthy control cfDNA (Table 5-3). The method by Meng *et al.* (2003) (see Appendix 8.4.1.2.3) was then applied to select a subset of these markers to serve as the genetic signature, as follows: all principal components were varimax-rotated and, after deducing the first few components that accounted for >85% of the variation in each comparison, the mean correlation coefficient for each marker was calculated for both the included and excluded principal components; markers to include in the genetic signature were those whose mean for the included components was greater than that of the excluded components.

Personalis EX2155

Origin of cfDNA		Significant markers for PCA[1]	Signature markers			
			Total	Redundant markers (%)	SNP (%)	CN (%)
Unaffected *BRCA1* carriers		51,749	580	98.88	55	45
Non-relapsed patients on follow-up	P1	51,866	19,606	62.2	99	1
	P2	32,876	30,673	6.7	73.5	26.5
Relapsed patients on follow-up	P1	51,258	16,936	66.96	99.7	0.3
	P2	3,462	3,399	1.82	87.56	12.44
Pre-surgical patients on neoadjuvant therapy	Baseline P	121,428	44,210	63.59	86.78	13.22
	6 week P	133,735	122,235	8.6	91.69	8.31
	21 week P	70,387	63,627	9.6	98.34	1.66

Table 5-3 - Reduction of markers for the breast cancer genetic signatures. P, plasma cfDNA of pre-surgical patients; P1 and P2, first and second plasma cfDNA samples of patients on follow-up, respectively.

When generating the signature for the cfDNA of the unaffected *BRCA1* mutation carriers, only 580 markers (98.88% reduction) were found to be capable of describing >85% of the variation (to healthy control cfDNA). When focussing on those patients who had primary breast cancer and who were on follow-up, however, markers were reduced by 62.2% and 6.7% for the first and second plasma cfDNA samples, respectively. For those patients on follow-up who relapsed, corresponding figures were 66.96% and 1.82%. In the pre-surgical breast cancer patients' cfDNA samples, there was a stark difference between the baseline and two later time-points: the baseline markers were reduced by 63.59%, while the 6 and 21 week samples that were taken during therapy were reduced by just 8.6% and 9.6%.

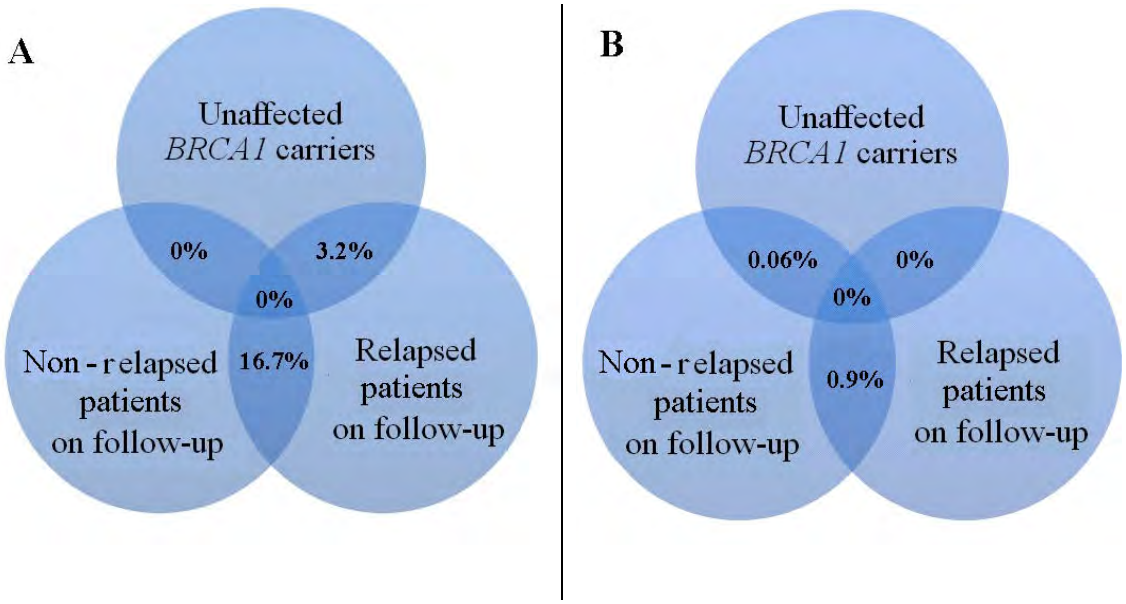
These markers did not cluster to any particular part of the genome and instead were arbitrarily distributed across all chromosomes. In each case, SNP markers contributed more to the exhibited variation than CN markers.

[1] After comparison to healthy control cfDNA through ANOVA.

Personalis EX2155

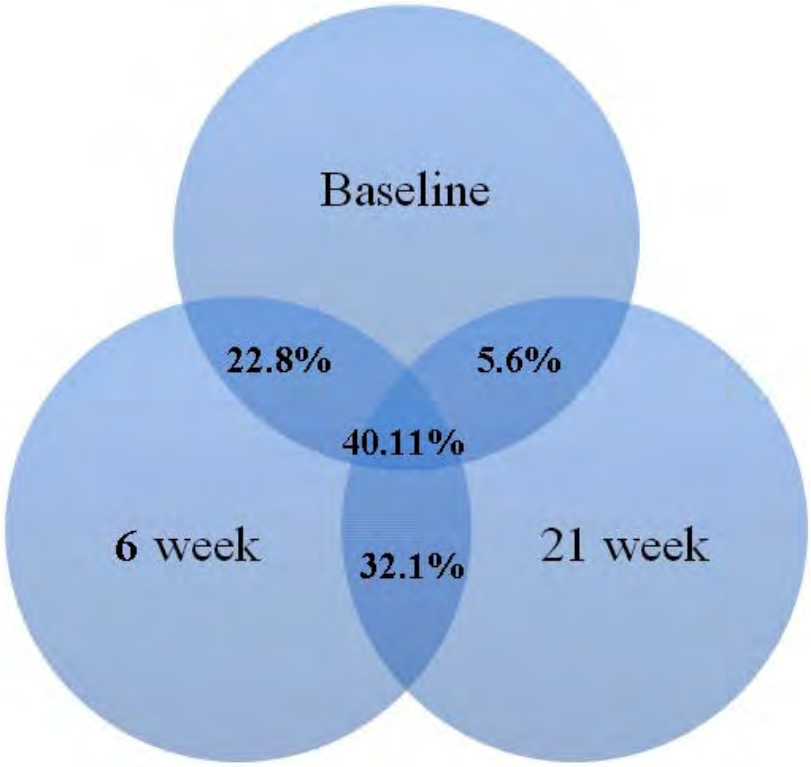
The percent overlapping markers (both SNP and CN from the numbers mentioned in Table 5-3) between each genetic signature are shown in Figure 5-2 and Figure 5-3. Looking at just the cfDNA of the unaffected *BRCA1* carriers and the first and second cfDNAs of the relapsed and non-relapsed patients on follow-up, there were no overlapping markers between these (Figure 5-2A). However, between the first cfDNA sample of the primary and relapsed breast cancer patients, 16.7% markers overlapped. For the same comparison when the second blood plasma sample was the focus, there was just 0.9% overlap (Figure 5-2B). In the pre-surgical patients' samples, there was only a 22.8% overlap between the baseline and 6 week cfDNA sample, whilst the figure was much less as treatment progressed, with overlap between the baseline and final (21 week) sample being reduced to just 5.6% (Figure 5-3).

Figure 5-2 - Percent overlapping markers for the derived cfDNA genetic signatures between the unaffected *BRCA1* carriers and the first (A) and second (B) samples of the relapsed and non-relapsed patients on follow-up.



Personalis EX2155

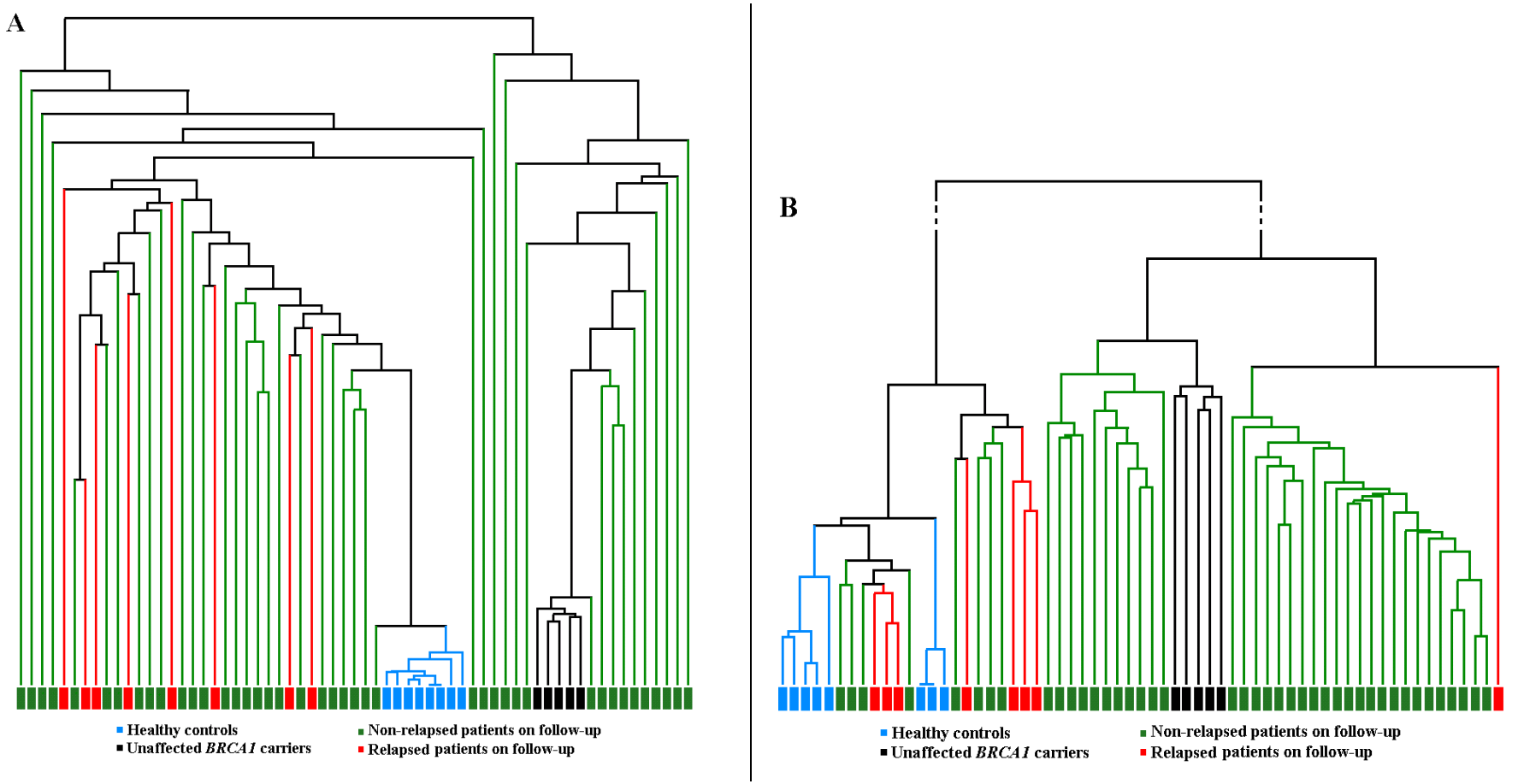
Figure 5-3 - Percent overlapping markers between each time-point in the pre-surgical patients' plasma cfDNA samples.



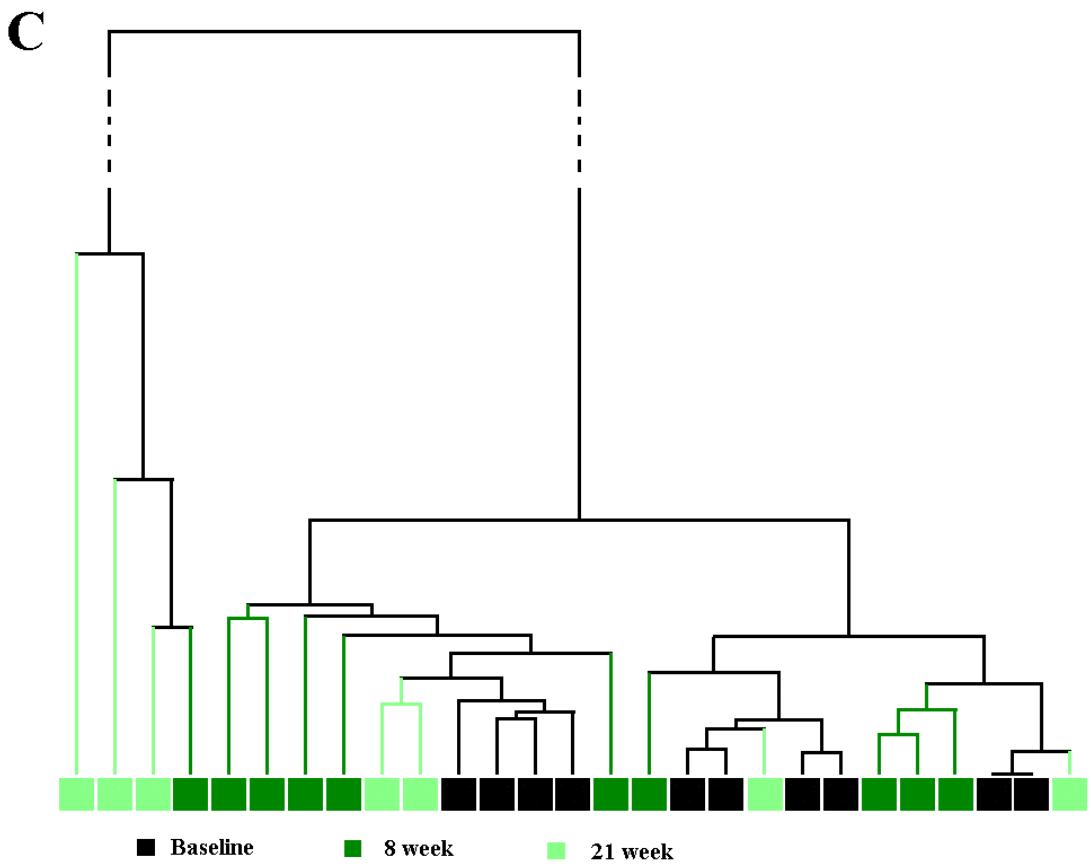
To test the ability of this method to distinguish the different types of cfDNA, all derived genetic signature markers were combined and then used to hierarchically cluster each sample (Figure 5-4A and B). Both dendrograms were recreated to include the markers comprising the pre-surgical breast cancer patients' cfDNA signature but the general tree structure was maintained, with these samples grouping amongst those of the patients on follow-up. These pre-surgical patients' cfDNA samples were then clustered together to determine the method's ability at distinguishing these (Figure 5-4C). Although three main groupings were identified, these did not relate to sample time-point.

Personalis EX2155

Figure 5-4 - Hierarchical clustering using signature markers reveals distinct clusters that distinguish the plasma cfDNA of unaffected *BRCA1* carriers from the first (A) and second (B) plasma cfDNA samples of the relapsed and non-relapsed breast cancer patients on follow-up (also shown are the plasma cfDNA samples of the healthy controls). The cfDNA samples of pre-surgical patients on neoadjuvant therapy are also shown (C).



Personalis EX2155



5.4 Discussion

The aims of this chapter were to characterise breast cancers through PCA by arriving at a concise number of genes and markers that were informative of cancer status, and also to highlight the usefulness of cfDNA as a liquid biopsy for cancer monitoring. Three groups of samples were compared: unaffected *BRCA1* mutation carriers; primary breast cancer patients on follow-up who did and did not relapse; and pre-surgical breast cancer patients receiving neoadjuvant therapy[1] (see Table 2-1 for details on these).

There were many genes that had overlapping significant SNP and CN markers. By focussing on the entire gene (as opposed to the individual marker) as the unit of

[1] Neoadjuvant chemo- or endocrine therapy (NeoCEnT) cohort.

Personalis EX2155

association, a broader picture is portrayed of what is affected at the biological level (Hicks *et al.*, 2012).

5.4.1 Exons were infrequently affected by variation

Overall, exons within genes associated with cancer were infrequently affected by significant variation. *DNAJC7* harboured an exonic CN deletion in all but 1 of the tumour DNA samples of the patients on follow-up. It is a co-chaperone protein that binds heat-shock protein (HSP) 70 and 90 and has been shown to reduce the activity of the ERα and ERβ receptors (Schülke *et al.*, 2010). Conversely, significant SNPs were found in exons of *EGFR*, *FYTTD1*, and *PDGFRA*, also in these tumour samples. EGFR has a well-documented role in breast cancer (Bronte *et al.*, 2011). FYTTD1 (UIF) is a protein that is involved in the relocation of mRNAs to the nuclear pore. Interestingly, for this to occur, another chaperone protein (FACT) is required (Hautbergue *et al.*, 2009). Meanwhile, *PDGFRA* encodes a receptor for the platelet-derive growth factors, which act as mitogens in cells of mesenchymal origin (Pruitt *et al.*, 2005). Mutations within exons of this gene have recently been reported in tumours of the gastrointestinal tract (Huss *et al.*, in press). The prevalences of each SNP were high in tumour DNA (92.5, 82.5, and 50% for *EGFR*, *FYTTD1*, and *PDGFRA* respectively) and low in matched lymphocyte DNA, except for *FYTTD1* (12, 30, and 14% respectively).

In the 8 patients who relapsed, *CD209* and *TTLL1* harboured significant SNPs in the cfDNA samples taken just before relapse (when compared back to tumour DNA taken during surgery), whilst *MAML2* and *ABL1* harboured significant CN markers. CD209 (also known as DC-SIGN), is expressed on the surface of dendrites and macrophages and was shown to interact with a protein expressed on the surface of colorectal carcinoma cells (Nonaka *et al.*, 2011).

TTLL1, however, encodes a tubulin protein and was previously identified as harbouring a SNP associated with prostate cancer (Kim *et al.*, 2010), although it was a different SNP from the one identified in this study. The *TTLL1* SNP in this study was regarded as significant due to the fact that it had returned a 'no call' in 100% of the tumour samples, suggesting that the SNP was deleted in the primary tumours. However, in

Personalis EX2155

75% of the cfDNA samples, genotype calls matched those found in normal lymphocyte DNA, indicating a healthy cfDNA profile in these final cfDNA samples of relapsed patients. The SNP genotype identified in *CD209* also had a genotype resembling that in the matched lymphocytes (100% of cases), thus also implying a healthy cfDNA profile. Thus, these two genes appear to not be significant in the cfDNA of relapsed patients on follow-up.

MAML2 has a role as a NOTCH signalling coactivator (Köchert *et al.*, 2010) - it was deleted in 62.5% of the final plasma cfDNA samples in the relapsed patients, a deletion not observed in any of the matched first cfDNA samples. Conversely, ABL1 was implicated in ER signalling and was shown to physically interact with ERα, whereby it could cause resistance to the ER antagonist fulvestrant (Zhao *et al.*, 2011). The particular locus in question was deleted in all of the T_1N_0 ($T \leq 20mm$) final plasma cfDNA samples and also in matched lymphocytes, indicating a germline mutation. However, it was amplified in all tumour DNA samples in the patients on follow-up, perhaps indicating one mechanism through which ERα can become resistant to therapy.

The scarcity of variation found in exons observed in this study could point to the hypothesis that variation in coding regions suffers deselection in populations of tumour cells (Schuster-Böckler *et al.*, 2010) - i.e. unless the variation conferred an advantage to the cancer, it would not persist. Therefore, it is possible that the variations mentioned here conferred such advantages, which is why they had been detected through analysis in this study.

5.4.2 Principal component analysis revealed numerous existing and novel gene targets of oncogenic potential

Although many of the significant markers mapped to intronic regions (Table 5-2), variation within introns can still be pathogenic if they interfere with gene splicing or affect promoter regions (Conrad *et al.*, 2009), and can have oncogenic potential (Haimovich, 2011). Through PCA, numerous genes that harboured already-documented mutations were found, including both oncogenic and tumour suppressor genes. However, genes of novel interest were also found.

Personalis EX2155

5.4.2.1 Patients on follow-up

5.4.2.1.1 Lymphocyte Vs Tumour DNA

There was a total of 15 genes (not including *ESR1*, which had an overlapping SNP and CN marker) that had an overlapping SNP that differed significantly between tumour and matched lymphocyte DNA. *BRAF* is a known oncogene and has been linked to several cancers. Its translated protein is a kinase that is involved in cell division, differentiation, and secretion through the MAPK and ERK signalling pathways (Garnett & Marais, 2004). In breast cancer, mutations are rare within this gene, whereas in malignant melanoma its mutation rate is 70% (Davies *et al.*, 2002). In this study, the SNP was observed in 97.5% of tumour and 18% of matched lymphocyte DNA.

Of further interest, DCC is a receptor that, when bound to NTN1 (Netrin-1), was found to activate the MAPK pathway through MAPK13 (Forcet *et al.*, 2002). *DCC* is also known to function as a tumour suppressor and to frequently undergo LOH in colorectal and oesophageal carcinomas (Arakawa, 2004). The SNP was observed in 70% of tumour and just 4% of matched lymphocyte. *PTEN* and *RB1* are other documented tumour suppressor genes, with *RB1* being regarded as the first tumour suppressor gene found and having an involvement in the maintenance of chromatin (Pruitt *et al.*, 2005). *PTEN* is mutated or lost in many tumours and is associated with resistance to EGFR inhibitors (Li *et al.*, 2012). An *in vitro* endometrial cancer study by Li *et al.* (2011) found that loss of *PTEN* resulted in cellular proliferation and a significant increase in EGFR expression (of importance, in this study, *EGFR* harboured an exonic SNP in 92.5% of tumour DNA samples). The SNP within *PTEN* was called due to a 'no call' in 70% of the tumours, implying a deleted locus (LaFramboise, 2009), while that affecting *RB1* was seen in 80% (and just 4% of matched lymphocytes).

TUSC3 (75% of tumours) is expressed in tissues of an epithelial origin and its gene is a candidate tumour suppressor that is found within a homozygously-deleted region in prostate cancer (Garshasbi *et al.*, 2008). It was also found to be deleted in squamous cell carcinomas of the larynx and pharynx (Guervós *et al.*, 2007). However, it had a

Personalis EX2155

high prevalence in matched lymphocyte of 26%. TNFSF4 (90% of tumours affected; 8% of matched lymphocytes), on the other hand, belongs to the tumour necrosis factor (ligand) superfamily, which are proteins involved in the upregulation of cell-adhesion molecules, antigens, and costimulatory proteins (Gruss & Dower, 1995). Finally, ESR1 (oestrogen receptor 1-α) has a well-documented role in breast cancer: ESR1 binds 17β-estradiol, a hormone that is the driver in the majority of breast cancers and one that can dictate the course of treatment to apply in the post-surgery/adjuvant clinical setting (Coombes *et al.*, 2004 and Chlebowski *et al.*, 2009). Numerous drugs have already been developed that target this receptor either directly or indirectly, including tamoxifen, fulvestrant, anastrozole, and raloxifene (CRUK, 2011c). The prevalence of the SNP in tumour in this study was high (82.5%; 8% in matched lymphocyte DNA).

Other genes with overlapping SNPs in these tumour DNA samples that merit a mention include: *CASP10*, *LHCGR*, *PRKCA*, and *WWOX*. CASP10 is a cysteine-aspartic acid protease, a family of proteins that are involved in the execution phase of apoptosis, resulting in the cleaving of vital proteins and destruction of the nucleus (Fischer *et al.*, 2003). Mühlethaler-Mottet *et al.* (2011) deduced that CASP10 plays a critical role in this apoptopic process and noted a transcriptional silencing of its gene in tumour. In the tumour DNA samples of this study, the SNP appeared in 82.5% of cases and just 6% of lymphocytes. LHCGR, however, is a receptor for luteinising hormone (LH) - inactivating mutations within which are associated with resistance to - and higher levels in the blood of - LH (Che *et al.*, 2011). Conversely, Li *et al.* (2011) found that decreased levels of IGF1 (insulin-like growth factor 1) resulted in subsequent decreases of LH and oestrogen - the role of the IGFs has already been reported in breast cancer (Tomlinson *et al.*, in press). In addition, a LHRH (luteinising hormone releasing hormone) analogue is used as a treatment in breast cancer and can reduce the risk of recurrence (Hackshaw, 2009 and Cruz Jurado *et al.*, 2011). The SNP was present in 92.5% and 10% of tumours and matched lymphocytes, respectively, in this study.

PRKCA is a protein kinase that can function as receptor for phorbol esters, a class of tumour promoters, and its expression is elevated in prostate cancer (Rumsby *et al.*, 2011). Higher expression was also linked with cervical cancer (Yu *et al.*, 2011) and metastasis in non-small cell lung cancer (Arora *et al.*, 2011). The SNP prevailed in

Personalis EX2155

67.5% of the tumour DNA samples, but was high in lymphocyte (16%). *WWOX*, however, harboured 3 different SNPs in tumour (at rates of 52.5, 70, and 15%; 2, 4, and 2%, respectively, in matched lymphocyte) - its protein is an oxidoreductase that is also regarded as a tumour suppressor (Aqeilan *et al.*, 2007). Wang *et al.* (2011) found that low expression of WWOX in breast cancer was associated with poor disease-free survival, whereas Hezova *et al.* (2007) found that improved disease-free survival was associated with higher expression of WWOX in glioma. It was also found that WWOX was co-expressed with ERBB4, and that this was again associated with better outcome in cancer (Aqeilan *et al.*, 2007). Thus, mutations affecting this gene could be associated with poor survival.

There were two intronic CN-affected genes in these tumour samples: *GPC3* and *ESR1*. The affected region in *ESR1* was deleted in 100% of the tumour samples and not found in any of the matching lymphocytes. To this author's knowledge, this is the first time that a deletion at this prevalence in a cohort of patients within the *ESR1* gene has been reported. Previous studies have focussed on genome-wide differences in CN patterns between ER-positive and -negative breast cancers (Fang *et al.*, 2011 and Thompson *et al.*, 2011) but never on *ESR1* itself. Considering that this deletion was observed in all tumours, it was correlated to neither tumour stage, therapy received, nor any other factor, highlighting its potential importance in breast cancer in general. *GPC3* is a gene in which deletions and other mutations have been reported previously (Sakazume *et al.*, 2007 and Weichert *et al.*, 2011). Its association with breast cancer has not been studied, although it was found to not be involved in Wilms' tumourigenesis (Gillan *et al.*, 2003). In this study, it was deleted in 95% of the tumour samples and none of the matched lymphocytes. Both of these deleted loci - and also that which affected the exon of *DNAJC7* in the on follow-up tumour samples - could be reflective of larger deletions within these genes that could alter their level of expression or function.

5.4.2.1.2 T1N0 (T≤20mm) Vs T>20mm/node-positive (first and second cfDNA)

Only the comparison between the T1N0 (T≤20mm) and T>20mm/node-positive patients' second cfDNA samples yielded markers that differed significantly, with *BCAS3* (breast carcinoma amplified sequence 3) harbouring a SNP that did not occur in

Personalis EX2155

matched normal lymphocyte. *BCAS3* is frequently amplified in breast cancer (Bärlund *et al.*, 2002) and is thought to function as a transcriptional coactivator and is also regulated by ESR1 (Gururaj *et al.*, 2007). It had earlier been shown by the same group that interplay existed between the *BCAS3* gene, ESR1, and MTA1, with MTA1 functioning as a transcriptional activator of *BCAS3* in a mechanism that required ESR1 (Gururaj *et al.*, 2006). *MTA1* was found in this study to belong to the common amplified CNV at 14q32.33 (Table 3-6), which was amplified in one-third of the tumours studied in the patients on follow-up (including relapsed and non-relapsed). The SNP identified was not present in matched lymphocyte and could be a mutation specific to cancer and have importance in disease-progression in these T>20mm/node-positive patients.

PKHD1, which was deleted in the cfDNA samples, encodes a membrane-associated protein that localises to epithelial basal bodies - mutations in its gene result in polycystic kidney and hepatic disease (Zhang *et al.*, 2004).

5.4.2.2 Relapsed patients' final cfDNA samples

5.4.2.2.1 First Vs Final/second cfDNA

Compared to just the first cfDNA samples, only *MCC* contained a SNP marker that differed between the first and final/second cfDNA samples of these relapsed patients. Loss of function of *MCC* through the methylation of its promoter was seen in colorectal cancer, and its translated protein is believed to be involved in the DNA damage-response due to its relocation to the nucleus and subsequent phosphorylation (Pangon *et al.*, 2010). The SNP marker was significant in the final cfDNA samples due to having returned a 'no call' in 37.5% of cases, implying a deleted locus. Deletion was not observed in matched lymphocyte DNA. Thus, a deletion could result in one less mechanism through which DNA is able to be repaired, leading to genomic instability.

The genes containing significant CN markers were *ABCC4*, *IGF2R*, and *PARK2*. ABCC4 is an efflux pump of endogenous and xenobiotic substances (Ban *et al.*, 2010 and Tagami *et al.*, 2010). Interestingly, it was found that small interfering RNA

Personalis EX2155

knockdown of this gene product led to enhanced cellular migration (Tagami *et al.*, 2010). In the samples in this study, the gene was amplified in 75% of the final plasma cfDNA samples and all of the matched healthy lymphocytes. Thus it can be assumed that these samples that harboured the amplification had a 'healthy' cfDNA profile, as this was also present in the germline. Conversely, whilst the precise role of *PARK2* (amplified in 75% of cfDNA and 50% of matched lymphocyte DNA) is unknown, what is known is that mutations in *PARK2* cause Parkinson Disease (Pruitt *et al.*, 2005). Indeed, CNV has long-been associated with neurological disorders (Sebat *et al.*, 2004 and 2007; Kirov *et al.*, 2009, Tam *et al.*, 2009; and Williams *et al.*, 2010).

Finally, IGF2R is a receptor of the insulin-like growth factor family, the role in breast cancer of which has already been discussed in relation to therapy resistance (Tomlinson *et al.*, in press; see section 4.2.2). This marker was found to have amplified CN in all of the first plasma cfDNA samples, normal CN in 75% of the final, and deleted in 80% of tumour. The matched healthy lymphocytes ranged from normal to amplified CN, thus implying that both the first and final cfDNA samples were 'healthy'.

5.4.2.2.2 Tumour DNA Vs Final cfDNA

When the final cfDNA samples of the relapsed patients were compared to their matched tumour DNA, 5 of the genes that were found were already found to harbour overlapping significant SNP markers in the tumour DNA samples of the patients on follow-up (*GRIK1*, *IL1RN*, *PALLD*, *MAML2*, and *WWOX*), whilst other genes included *DBC1* and *DEC1*. *DBC1* (deleted in bladder cancer 1) is located in an area that undergoes LOH in some bladder cancers (Nishiyama *et al.*, 1999) and it has been shown to modulate the expression of ERα, promoting breast cancer survival, and also to repress the function of ERβ receptor, which has tumour suppressive functions (Saha Roy & Vadlamudi, in press) and normally protects against the mitogenic effects of oestrogen (Koyama *et al.*, 2010). *DEC1*, which belongs to a region commonly deleted in oesophageal cancer (Nishiwaki *et al.*, 2000), was shown to interact with an isoform of TP63 to aid the suppression of growth and differentiation of keratinocytes (Qian *et al.*, 2011). TP63 exhibited CN deletion in all relapsed tumour DNA samples in this study. It was also revealed that lower expression of DEC1 resulted in cells that were more motile than

Personalis EX2155

those with normal/higher levels of expression (Kunimoto *et al.*, 2011). Both *DBC1* and *DEC1* showed deletion/LOH across the locus where the significant SNPs were identified, thus, the primary tumours of these relapsed patients would have had aberrant oestrogen signalling (due to loss of DBC1) and an ability to be more 'fluid' than others (due to loss of DEC1) from the outset.

KIT and *GLG1* also harboured significant SNP markers, but in this case due to 'no calls' made in matched tumour, implying deletion. *KIT* is a proto-oncogene that encodes a mast cell growth factor receptor and, although its role in breast cancer has not been extensively studied, higher expression levels of it were found in *BRCA1* breast tumours in a recent study by Smart *et al.*, (2011). *GLG1* was previously found to harbour a significant SNP in a separate genome-wide association study involving breast cancer (Kibriya *et al.*, 2009).

TP63 was shown to have normal CN in 75% of final cfDNA and 62.5% of matched lymphocyte, but deleted in all of matched tumour. TP63 is a transcription factor belonging to the P53 family that is involved in the development of epithelial tissues. It also regulates cell-adhesion and plays a role in epithelial-to-mesenchymal transition (EMT). Graziano & De Laurenzi (2011) concluded that TP63-negative tumours would metastasise earlier and more frequently, thus having a poorer outcome. A deletion within this gene in the primary tumour could therefore predispose the patient to later recurrence of the disease and could be useful as a prognostic marker for relapse.

5.4.2.2.3 Non-relapsed T1N0 (T≤20mm) Vs Relapsed T1N0 (T≤20mm) final cfDNA

In the T1N0 (T≤20mm) relapsed patients, when comparing the final cfDNA sample to that of the non-relapsed patients, *XRCC3*'s significantly-different SNP marker had the same genotype as that found in matched lymphocyte DNA, thus implying a healthy profile.

Four genes harboured CN markers that highly differentiated between relapsed and non-relapsed cfDNA: *BRCA1*, *PDGFRB*, *RHPN2*, and *SMAD4*. *BRCA1* requires little

Personalis EX2155

introduction - it, along with *BRCA2*, accounts for 40% of hereditary breast cancer cases and its protein is involved in the repair of DNA double strand breaks, checkpoint control, and spindle and transcriptional regulation (Shuen & Foulkes, 2011). The gene harboured amplification in half of the relapsed cfDNA samples but also all of the matched lymphocytes. However, the same amplification also appeared in all of the non-relapsed cfDNA and tumour samples. This result could allude to an error with the microarray technology, possibly relating to the signal intensity of the probe in question. For example, if the intensity was erroneously-interpreted above the baseline, then it would have been regarded as amplification, whereas in reality it could have been normal CN. This was made more apparent by observing the estimated CN in the tumour samples across this locus, which ranged from 8-55.

PDGFRB (amplified in 100% of the relapsed cfDNA samples) encodes the β-receptor for the platelet-derived growth factors - a mutation in an exon of the α-receptor was already found in this study in the tumour DNA samples of the patients on follow-up. Conversely, *RHPN2* (amplified in 50%) is a gene that has been found to harbour SNPs in relation to colorectal cancer in numerous genome-wide association studies (Abulí *et al.*, 2010; Kupfer *et al.*, 2010; Von Holst *et al.*, 2010; and Xiong *et al.*, 2010) - it encodes a protein (rhophilin-2) that belongs to the family of ras-homologous-GTPase binding proteins (Pruitt *et al.*, 2005). However, both of the amplifications within these genes were also found in matched normal lymphocytes (75 and 50% prevalences).

Finally, SMAD4 is a protein that is phosphorylated in response to TGF-β signalling and binds to an 8bp DNA sequence (GTCTAGAC) called the SMAD binding element (Pruitt *et al.*, 2005). The marker identified within the *SMAD4* gene was deleted in all of the relapsed cfDNA samples and amplified in all of the non-relapsed. In matched normal lymphocyte, however, the same deletion was observed - thus, it was most likely a germline deletion.

5.4.2.2.4 Non-relapsed T>20mm/node-positive Vs Relapsed T>20mm/node-positive final cfDNA

Personalis EX2155

Only *PTCH1* was significantly CN-affected between the T>20mm/node-positive relapsed and non-relapsed final cfDNA samples. Mutations within this tumour suppressor gene have previously been found in relation to breast cancer (Sinha *et al.*, 2008), and it was concluded by Santos *et al.* (2011) that mutations would contribute to the development of tumours. It was amplified in all of the relapsed cfDNA samples and deleted in 75% of the non-relapsed. However, the amplification was also observed in matched lymphocytes.

5.4.2.3 Healthy control Vs Unaffected *BRCA1* carriers' cfDNA

DIRC3 (disrupted in renal carcinoma), which spans a breakpoint on chromosome 2 that was found in renal cancer (Bodmer *et al.*, 2003), and *DLC1* (deleted in liver cancer 1), a gene whose loss was linked to numerous cancers, including that of the breast (Heering *et al.*, 2009 and Feng *et al.*, 2011), both harboured SNP markers whose values differed from healthy control cfDNA. *PALLD* also contained a SNP marker that failed to hybridise in all of the cfDNA samples of the unaffected *BRCA1* carriers, strongly suggesting deletion. Loss of expression of *PALLD*, on the other hand, was deduced as being an important part of EMT (through helping E-cadherin localise to adherens junctions) and as a potential early step in the metastatic spread of tumour cells in colon cancer (Tay *et al.*, 2010). PALLD is a protein component of actin microfilaments in the cytoskeleton and is involved in cell adhesion and contractility (Tay *et al.*, 2010). It is possible that loss of this gene's function could severely disrupt the formation of the cytoskeleton, resulting in cells becoming more 'plastic'.

Could these loci be signs of genomic instability and act as a useful target for early screening of women carrying *BRCA1* mutations?

5.4.2.4 Pre-surgical patients' cfDNA

In the pre-surgical breast cancer patients' samples, significant SNP markers were only found for comparisons made between the 21 week cfDNA samples and baseline cfDNA and tumour DNA. This is interesting because the 21 week cfDNA sample was the end of treatment sample taken from these patients after having received neoadjuvant

Personalis EX2155

therapy. Thus, differences between these and their beginning of treatment/baseline samples were a possible reflection of responses to therapy.

The SNP markers identified within *PKHD1* and *TGFBR2* suggested deletion at these loci due to returning 'no call' in 29 and 43% of samples, respectively. Mutations in *PKHD1* resulted in polycystic kidney and hepatic disease (Zhang *et al.*, 2004), whereas a mutation within *TGFBR2* could indirectly affect the expression of this receptor on cell surfaces and the regulation of cell growth - it encodes a transmembrane protein that binds TGF-β, which indirectly results in the regulation of the transcription of genes involved in cell proliferation (Park *et al.*, 2012).

5.4.3 ER receptor activity is altered in tumour and cfDNA

Oestrogen receptor signalling is complex and its role in breast cancer metastasis is not yet understood - shedding light on this role will help the development of novel therapies (Saha Roy & Vadlamudi, in press). In this study, significant SNP and CN variants that mapped to genes whose functions centered on the activity of the oestrogen receptor were found in both cfDNA and tumour DNA. This is of high interest, considering that the majority (~70%) of breast cancers are ER-positive (Koyama *et al.*, 2010; Hicks *et al.*, 2012, and Saha Roy & Vadlamudi, in press) and 65-70% of primary tumours retain an ER-positive status at distant metastatic sites (Saha Roy & Vadlamudi, in press).

In tumour, of the exonically-affected genes, one stood out as warranting further attention - i.e. *DNAJC7*. An exon within this gene harboured a deletion in 97.5% of the breast tumours analysed in the on follow-up group. This protein is a co-chaperone that binds HSP 70 and 90 and it has been shown to reduce the activity of the ERα (encoded by *ESR1*) and ERβ (encoded by *ESR2*) (Schülke *et al.*, 2010). Thus, a deletion in an exon of this gene could result in a non-functional, translated protein that could result in a heightened activity of the ER, thereby providing one of the necessary steps through which the ER becomes oversensitised in certain breast cancers. To further the case, *ESR1* was itself both SNP- and CN-affected in on follow-up tumour DNA, with prevalences of 82.5% and 100%, respectively. Although intronic, both variations could be alluding to more widespread variation within the *ESR1* gene that could have strong

Personalis EX2155

implications in cancer development and progression. However, it is important to note that not all of these tumours were ER-positive. Therefore, the alterations mentioned, while having high prevalences, are not correlated with receptor status.

Arguably of most interest was *ABL1*, which was deleted in 75% of the relapsed final/second plasma cfDNA samples (T_1N_0 ($T \leq 20mm$) group). Its related protein had been shown to interact with the ER and was associated with resistance to the ER antagonist fulvestrant (Zhao *et al.*, 2011). If a patient acquired resistance to this SERM, through deregulation of *ABL1*, it could affect other SERMs and thus position the patient in a state of vulnerability. Although deleted in these cfDNA samples, it was amplified in each of the on follow-up tumour samples, perhaps giving a partial insight into the process through which the ER becomes resistant to therapy. Moreover, it was the terminal exon of this gene that was amplified.

Tied indirectly to the ER is the SNP overlapping *LHCGR*, which had a prevalence of 92.5% in the tumour DNA samples of non-relapsed patients on follow-up. A previous study found that a decrease in IGF1 resulted in a decrease in LH and oestrogen (Li *et al.*, 2011), whereas another study reported a link between IGFR activation and that of the ER (Tomlinson *et al.*, in press). Taken together, this reflects interplay between the growth factors and hormone receptors in breast cancer, where alterations at the genomic and/or proteomic level could result in indirect effects on other genes/proteins.

When honing in on just the relapsed tumour samples, a SNP was identified within *DBC1*. Although the precise function of DBC1 is unknown, it was observed in breast cancer cells to bind ERα and promote the survival of the cancer, while at the same time repressing the function of the β receptor, which is normally involved in the mitigation of the mitogenic effects of oestrogen (Koyama *et al.*, 2010). Thus, *DBC1* provides a double advantage for tumours to develop and progress through oestrogen signalling. The SNP genotype within the gene changed from BB (all lymphocyte) to AB (all matched tumour). Additionally, Gaussian-estimated CN across this locus revealed deletion in all relapsed tumour DNA samples but none of the healthy lymphocytes. Considering that the SNP and deletion were only prevalent in the relapsed tumour

Personalis EX2155

samples, a possible pre-disposition to recurrence within these tumour samples could have been elucidated.

In cfDNA, *BCAS3* was the only gene known to be related to oestrogen signalling that harboured significant SNPs, with a prevalence of 71% in the final cfDNA of the T>20mm/node-positive patients (genotype AA→AB and absent from matched lymphocyte). A different SNP marker within the same gene was significant when comparing the pre-surgical 21 week cfDNA to matched tumour DNA (although in this case the SNP marker's 'no call' implied that the locus was deleted). *BCAS3* is transcriptionally activated by MTA1 in a mechanism that requires ER (Gururaj *et al.*, 2006 and 2007). *MTA1* was already found in this study to belong to a commonly-amplified CNV in tumour (Table 3-6 and Table 4-4) and cfDNA. Thus, both *BCAS3* and *MTA1* could be useful biomarkers in cfDNA.

5.4.4 SNP is more conducive to variation than CNV

The distribution between SNP and CNV in each breast cancer cfDNA signature created for the unaffected *BRCA1* carriers, relapsed and non-relapsed patients on follow-up after surgery, and pre-surgical patients on neoadjuvant therapy (Table 5-3) indicates that SNPs are more conducive to variation in cfDNA than CNVs. The most even distribution was observed in the unaffected *BRCA1* carriers, but even here SNPs dominated, with 55% of the variation. The most extreme example was the first plasma cfDNA of the relapsed patients, where SNPs contributed towards 99.7% of the variation.

These findings could in part be due to the nature of SNP and CNV. For example, SNPs affect just a single nucleotide, whereas CNV can affect from kilobases up to megabases (Scherer *et al.*, 2007 and Grayson & Aune, 2011), with a minimum size generally accepted as being 1Kb (LaFramboise, 2009). Thus, despite the fact that CNV is common in the healthy genome (Redon *et al.*, 2006), it should be expected that SNP variation in any genome - cancerous or otherwise - will be more common than CNV. However, observing the distribution between these two types of variation, it can be inferred that genome-wide changes occurred between the first and second/final plasma

Personalis EX2155

cfDNA samples of these patients on follow-up who relapsed and did not relapse, owing to the fact that - in both cases - the proportion of CNVs went from 0.3 to 12.44% and 1 to 26.5%, respectively.

5.4.5 Principal component analysis creates breast cancer cfDNA genetic signatures

The *Affymetrix SNP 6.0* microarray genotypes 906,600 SNPs and has 945,826 probes for detecting CNV across the human genome. Thus, it is capable of building a genome-wide 'profile' of a sample based on SNPs and CNVs. Particularly, using the htSNP method described by Meng *et al.* (2003), PCA was capable of identifying sets of SNP and CN markers for the purpose of creating a cfDNA genetic signature of unaffected *BRCA1* mutation carriers, breast cancer patients on follow-up who did and did not relapse, and pre-surgical breast cancer patients on neoadjuvant therapy (Table 5-3).

The derived markers for the relapsed patient group's cfDNA are important because the final sample taken from each patient was the one taken before relapse. It had previously been shown by Gangnus *et al.* (2004) that different genetic changes can occur with progressive micrometastases from solid tumours, thus any changes in the blood plasma of these patients could be reflective of a tumour metastasising. For the first cfDNA sample, 66.96% of the markers were redundant and eliminated through the htSNP method (Table 5-3). This left 16,936 markers that comprised the genetic signature and that were capable of describing >85% of the variation amongst these samples. However, for the final cfDNA samples, just 3,399 markers were capable of achieving the same. Thus, the final plasma cfDNA samples were a more homogenous group than their matched first samples.

The reverse was seen for the breast cancer patients on follow-up who did not relapse. The final cfDNA genetic signature required more markers than that of the first cfDNA (22,532 compared to 19,606) (Table 5-3). Thus, the final cfDNA samples were more heterogeneous than the first.

Personalis EX2155

A much smaller number of markers (580) were derived for the unaffected *BRCA1* carrier cfDNA signature. This suggests that the cfDNA samples of these women are a homogenous group and do not differ greatly from those of healthy controls (which were used as the baseline comparison).

5.4.5.1 Genetic signatures are highly defined and narrow

Having too few or many markers when screening patients can affect sensitivity and specificity. Additionally, if different types and stages of cancer were to be distinguished by the same test, the markers employed would have to uniquely identify each stage. In this study, the narrowness of each derived signature was evidenced by the percentage of overlapping markers found between each. There was no marker common to all 3 sample groups when focussing on neither the first (Figure 5-2A) nor second (Figure 5-2B) plasma cfDNA samples. This suggests that these sample groups were of different natures/origins, as the variation amongst them was described by independent sets of markers, and is a useful finding for classification purposes.

Also, the relapsed patients' samples seemed to segregate more when going from the first to the final cfDNA samples: For the first cfDNA samples, there were 16.7% and 3.2% overlapping markers with the non-relapsed patients on follow-up and the unaffected *BRCA1* carriers, respectively. However, for the final cfDNA samples, the corresponding amount of overlap was 0.9% and 0%. Thus, the relapsed cfDNA genetic signature became more dissimilar from the non-relapsed on follow-up patients and unique from the unaffected *BRCA1* carriers. These could be the changes that are involved in relapse.

Classifying these samples through hierarchical clustering using the derived sets of markers revealed that they are useful for distinguishing between these groups when focussing on the first cfDNA samples (Figure 5-4A). In this case, however, the relapsed patients' samples did not cluster on their own, instead being arbitrarily distributed in the dendrogram structure amongst those of the other patients on follow-up. Also, in this instance, the healthy control and unaffected *BRCA1* carriers' samples clustered tightly and separately, highlighting their respective homogeneity.

Personalis EX2155

When focussing on the final cfDNA samples, the relapsed samples clustered more tightly together, again indicating an increase in their homogeneity and segregation from the other samples (Figure 5-4B). Although the healthy control cfDNA cluster featured 3 of the relapsed samples in this instance, there were 5 further branches off this cluster before these 3 relapsed samples were reached. Interestingly, amongst these relapsed samples were grouped 8 samples from the primary breast cancer patients who had not relapsed. This could indicate that these 8 women were at risk of recurrence.

There was also one relapsed sample that did not cluster with the others - this was taken from a patient who had a second primary, and who carried a *BRCA2* mutation, but was grouped with the relapsed cohort for simplicity.

The height of the lines in each of these dendrograms between each of the on follow-up cfDNA samples highlights a general heterogeneity between these samples, whereas the short lines for the healthy controls and unaffected *BRCA1* carriers highlights more homogenous groups of samples. The line-heights are also reflective of the method employed - i.e. complete linkage - which looks for greatest distance between samples. Single linkage, conversely, searches for the shortest distance.

5.5 Conclusion

The overall picture that has been painted is that tumour DNA is afflicted with numerous SNP and CN mutations that affect key genes in cancer, many related to the oestrogen receptor and its signalling pathways. The high prevalences of these markers point to mutation 'hot spots' in the breast cancer genome. Were they at low prevalences, one could put it down to isolated mutational events, but their commonality implies oncogenic potential. Moreover, through the analysis of cfDNA, these mutations could also be detected.

The diagnostic and prognostic potential of cfDNA through the creation of a genetic signature has also been shown. The derived markers could be incorporated in tandem

Personalis EX2155

on a single custom assay (similar to a microarray) and then used to screen patients in the future at low cost and with short processing time.

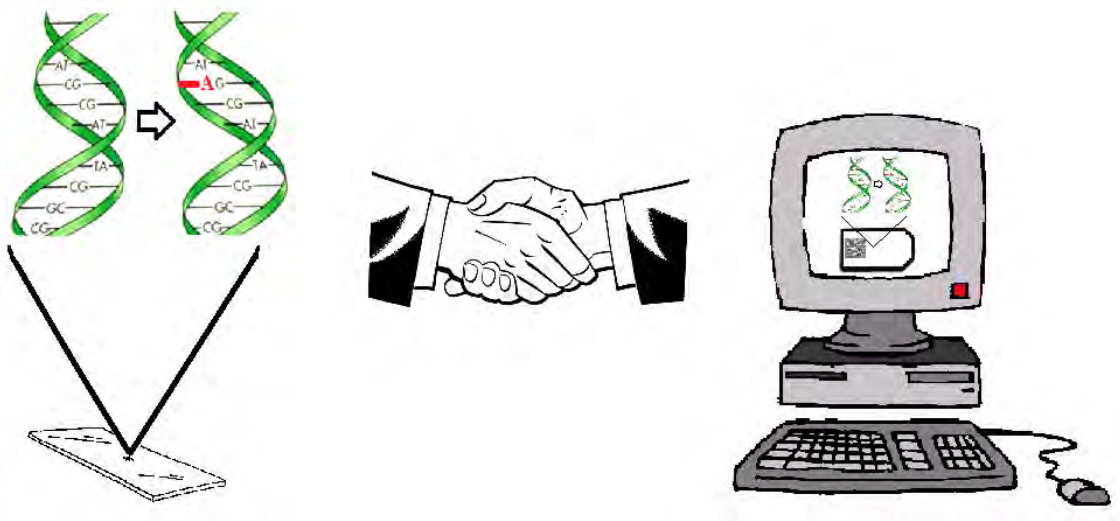
Also, a strong indication has been made that the primary tumour of patients can give clues as to their metastatic potential. The primary tumours of the relapsed patients exhibited 3 deletions that could be key in deciding which patients were more likely to relapse: complete loss of a region within *TP63* and LOH within *DBC1* and *DEC1*. TP63-negative tumours would metastasise more frequently (Graziano & De Laurenzi, 2011), while DEC1 provided increased motility of cells (Kunimoto *et al.*, 2011). Concurrently, DBC1 would be encouraging the proliferative effects of oestrogen by heightening the activity of ERα and diminishing the protective effects of ERβ (Koyama *et al.*, 2010).

5.6 Future studies

Future studies are essential in order to define the role that variation within certain genes related to oestrogen signalling has in breast cancer development and progression. Such SNP and CNVs found in this study could be useful as diagnostic and prognostic biomarkers, including that of disease recurrence. The deletions within *DNAJC7* and *DBC1*, for example, could result in an increase in ER activity and promote cell survival, whereas the amplification of the terminal exon of *ABL1* could be a necessary step through which a tumour must pass en-route to resistance to therapy. In cfDNA, *BCAS3* and *MTA1* were identified as useful biomarkers, but the results are not clear-cut. Focussing on ERβ as a therapeutic target in order to heighten and foster its tumour suppressive functions is also encouraged.

Future work is also encouraged to determine the role that each gene found to harbour a significant SNP or CN markers plays in breast cancer (Table 8-2). In this study, only those genes already-known to be involved in cancer and its known molecular pathways were the focus. It could be that many in the complete list of genes are putative oncogenes with critical roles in breast cancer.

Personalis EX2155



6 MICROARRAY DATA VALIDATION

Personalis EX2155

6.1 Overview

The focus of this chapter is to validate the results produced in previous chapters through the analysis of both test and independent sample-sets using qPCR. The utility and reproducibility of this microarray technology is also tested by making comparisons of the data to other studies of a similar nature.

6.2 Introduction

There are large differences between microarray analysis program and their associated algorithms. For example, a 91% agreement was found between Birdsuite, a program that also utilises the Birdseed algorithm (Broad Institute, http://www.broadinstitute.org/mpg/birdsuite/index.html/), and quantitative PCR (qPCR), whereas only 76% agreement was found between qPCR and Partek® Genomics Suite™ (Partek Inc., http://www.partek.com/) (Grayson & Aune, 2011). The highest agreement between each was found when the copy number (CN) state was 2 (i.e. - normal). Therefore, a reference or standardised method of data analysis and validation are needed. The current preferred method for microarray validation is qPCR (Grayson & Aune, 2011). However, it is unfeasible from a financial and time point of view to use qPCR to validate an entire genome's worth of microarray data.

6.3 Aims

The aim of this chapter was to analyse key alterations identified by the *Affymetrix SNP 6.0* microarray as data validation. This was achieved through focussing on different variants - single nucleotide polymorphisms (SNPs) and copy number variations (CNVs) - in the genome and comparing their values to those derived through independent methods (such as real time qPCR) or other studies.

Personalis EX2155

6.4 Results

6.4.1 Examples of known genomic regions of variance

The *Affymetrix SNP 6.0* microarray genomic profiling was capable of identifying numerous known regions of variance. For example, *CCL3L1* at 17q12, which was one of the common amplified CNVs in the pre-surgical breast cancer patients (Table 4-4), is known to have variable CN in the general population, ranging from 1-10 (Field *et al.*, 2009). The microarray also reported variation at 5q13.2 and 10q11.22. 5q13.2 harbours a 500Kb inverted duplication and is prone to rearrangements and deletions (Huang *et al.*, 2010 and Wang & Dreyfuss, 2001). 10q11.22 harbours a segmental duplication that affects a protein tyrosine phosphatase, designated *PTPN20A* and *PTPN20B*, with 'B' resulting from a duplication of 'A' (a third copy, *PTPN20C*, is only partially-duplicated) (Pruitt *et al.*, 2005). Other detected regions included 7q11.23, which has an associated duplication syndrome (Mervis, 2008), 15q25.2, which shows trisomy in autism (Bonati *et al.*, 2005), and 16p12.3, which is also known to harbour duplication (Pruitt *et al.*, 2005).

6.4.2 *ERBB2* gene amplification

The ERBB2 receptor is overexpressed in approximately 20% of breast cancers, mainly as a result of gene amplification (Mukohara, in press), and can dictate the course of patient treatment (Awada *et al.*, in press and Bayoudh *et al.*, 2012). The ERBB2 status was known for the primary tumour of the 50 patients on follow-up analysed in this study (Table 2-1). Thus, it was possible to compare the CN output for *ERBB2* by *Affymetrix SNP 6.0* and real-time qPCR against immunohistochemistry (IHC)[1]. The qPCR assay targeted exon 16. However, as the CN probe density for this exon was low, caution must be exercised when interpreting this data.

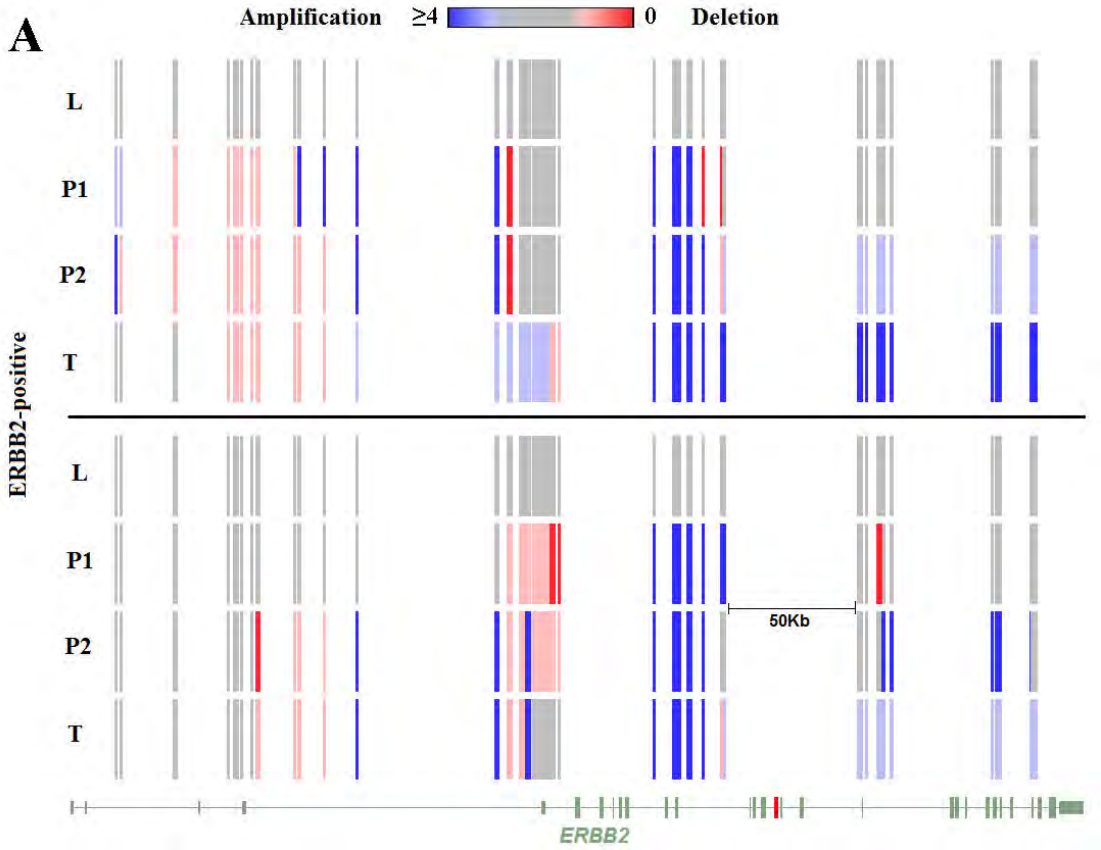
Figure 6-1A and B show amplification and deletion for the *ERBB2* gene interval in ERBB2-positive and -negative patients, respectively, by both qPCR and IHC. In the

[1] Quantification by IHC is often subjective and prone to inter-observational bias (Taylor & Levenson, 2006).

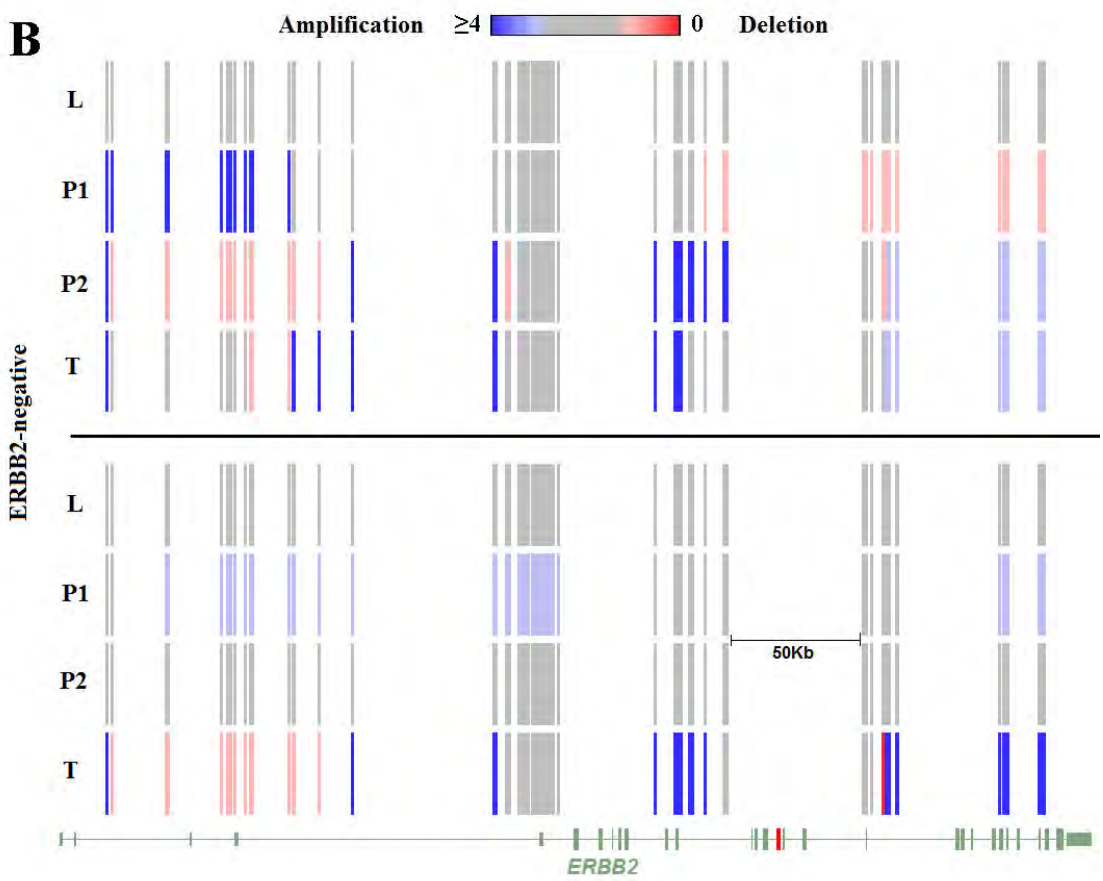
Personalis EX2155

two ERBB2-positive tumours, amplification was observed surrounding the targeted exon, while in the ERBB2-negative patients, no amplification was seen. In one ERBB2-negative patient, however, amplification of the region was picked-up in the second cfDNA sample, which might suggest an acquired genomic change.

Figure 6-1 - *Affymetrix SNP 6.0* CN state for the *ERBB2* gene interval in patients on follow-up. Two ERBB2-positive (A) and -negative (B) patients by real-time qPCR and IHC are shown. Each vertical bar represents one CN marker (or consecutive markers) and values are as inferred by the colour-scale (at top). The exons and introns of the *ERBB2* gene are indicated at the bottom of the diagram, with exon 16 marked in red. L, lymphocyte; P1 and P2, first and second plasma cfDNA, respectively; T, tumour.



Personalis EX2155



6.4.3 Analysis of commonly-amplified CNVs by qPCR

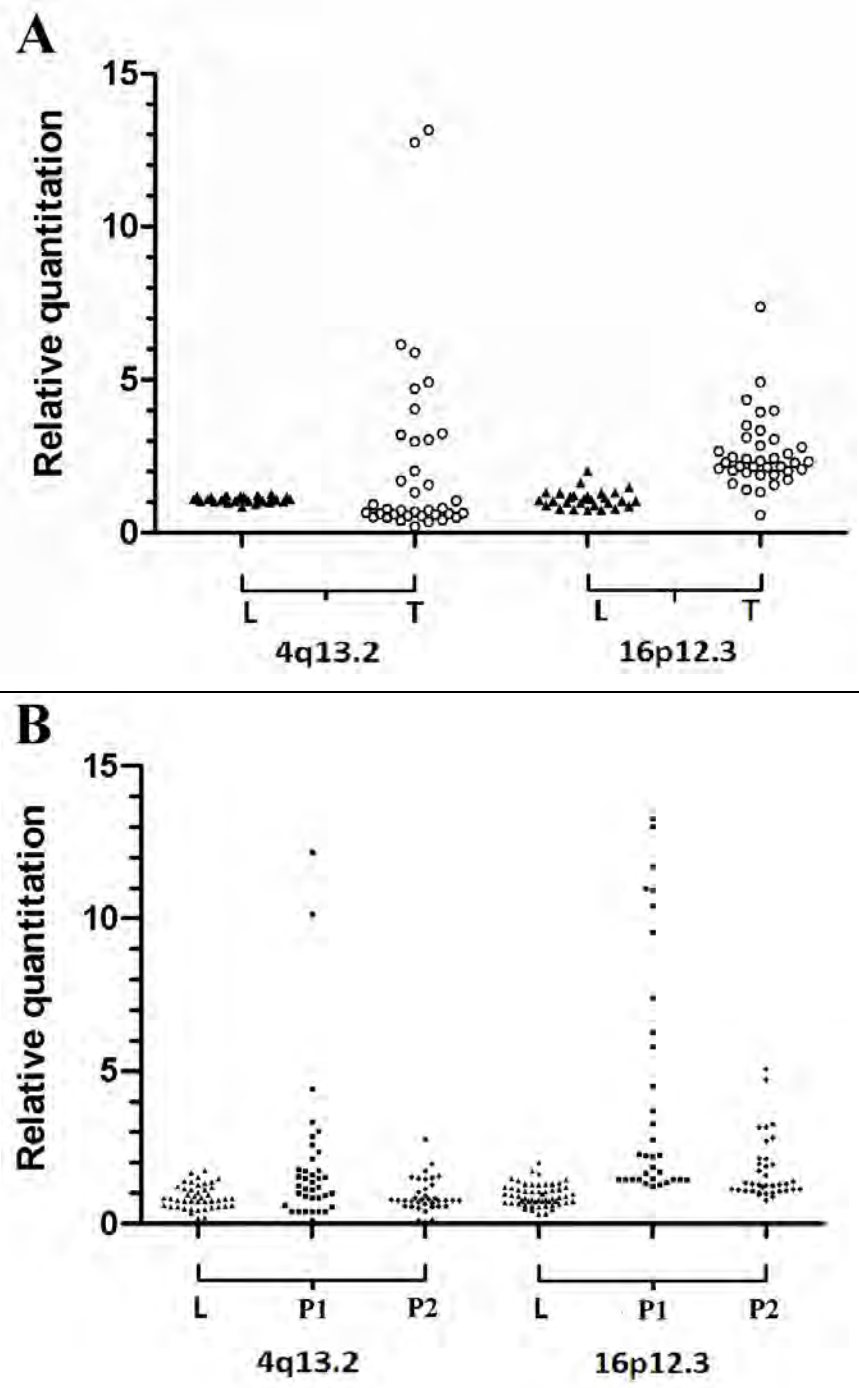
Five of the common CNVs (Table 3-6 and Table 4-4) were excluded from further analysis due to having a low marker density: 5q13.2, 7p11.2, 7q11.21, 9q12, and 14q11.1. Two assays were initially designed to map to the CNVs at 4q13.2 and 16p12.3, and tested by real-time qPCR. Amplification (relative quantitation>1.5) at 4q13.2 (38%) and 16p12.3 (94%) was detected in an independent cohort of patients that included 37 unamplified (non-whole genome amplification (WGA)) formalin-fixed paraffin-embedded (FFPE) tumour DNA samples, with little amplification detected in normal lymphocyte DNA samples from the patients on follow-up (Figure 6-2A).

Separately, matched, unamplified, first and second plasma cfDNA from 40 of the patients on follow-up also showed amplification (45 and 20%, at 4q13.2, and 67.5 and

Personalis EX2155

37.5%, at 16p12.3, for first and second cfDNA, respectively) that was scarce in the matched normal lymphocyte DNA (Figure 6-2B).

Figure 6-2 - Real-time qPCR scatter-plots for common CNVs at 4q13.2 and 16p12.3. Normal diploid CN was seen in lymphocyte DNA, while some tumours (independent cohort) exhibited amplification (A). Amplification was also observed in some cfDNA samples (B). L, lymphocyte; P1 and P2, first and second plasma cfDNA, respectively; T, tumour.



Personalis EX2155

6.4.4 Comparison to other studies

The CN of the *LMTK3* gene, which was identified as a potential oncogene by a collaborator (Giamas *et al.*, 2011), was measured in order to highlight the prospective use of the microarray data to confirm existing - or elucidate new - results. Comparison of CNV regions identified in this study with those of two others of similar natures (Naylor *et al.*, 2005 and Bae *et al.*, 2010) was also performed in order to highlight reproducibility in independent data-sets using separate methods (aCGH).

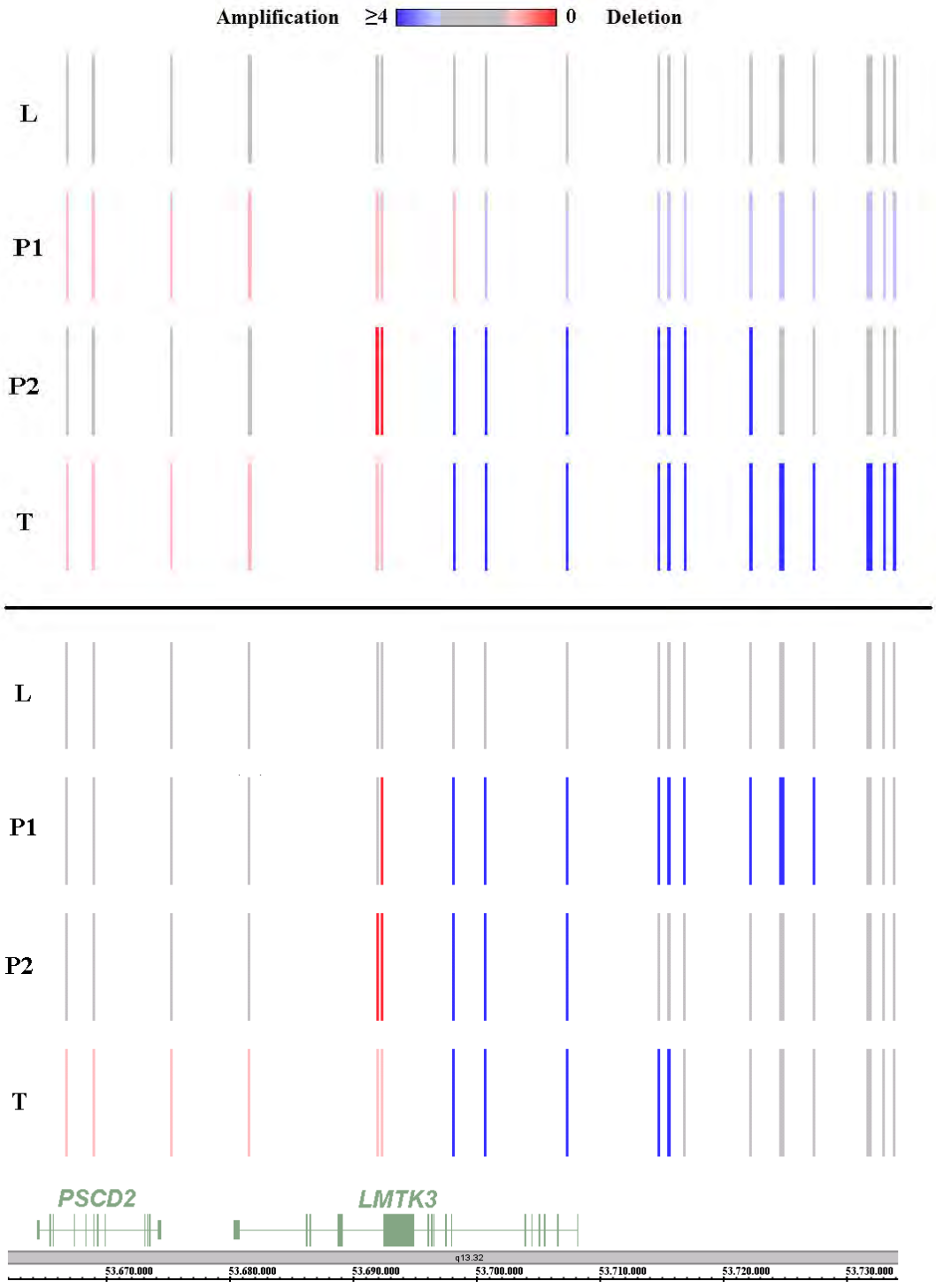
6.4.4.1 *LMTK3* as a potential oncogene

The *LMTK3* gene was recently suggested to be a potential novel breast cancer oncogene with a significant role in the ERα response to therapy (Giamas *et al.*, 2011). Using Gaussian-smoothed CN values of >2.5 and <1.5 as thresholds for amplification and deletion, respectively, there was no amplification or deletion identified in any of the normal lymphocyte DNA samples of the 50 patients on follow-up, or in the lymphocyte DNA and cfDNA of healthy controls. However, the 3' (distal) end of the gene was amplified in 32.5% of the primary tumour DNA samples (total=40) and the entire gene was amplified in 12.5%. Moreover, this tumour-specific amplification was also present in 20% of the first cfDNA samples and 25% of the second (taken an average of 6 and 9 years after surgery, respectively) (Figure 6-3). Comparing primary tumour and matched cfDNA samples, the results agreed for 60% of first cfDNA and 52.5% of the second. A small percentage of the cfDNA samples (4% of the first and 2% of the second) also showed deletion at the 5' end of the gene (Table 6-1).

Twenty-five of the patients on follow-up were receiving tamoxifen and 6 went on to relapse. By first filtering-out markers of difference between the first and final cfDNA samples of these 25 patients (which was the sample taken closest to the time of relapse in the 6 patients who relapsed) - and then applying Gaussian-smoothing on the reduced number of markers - significant amplification ($p<0.0001$) in *ESR1* and *LMTK3* was revealed in the final cfDNA samples that was not seen in the first (*manuscript in preparation*).

Personalis EX2155

Figure 6-3 - *LMTK3* gene interval for two on follow-up patients. Each vertical bar represents one CN marker (or consecutive markers) and values are as inferred by the colour-scale (at top). The exons and introns of the *LMTK3* gene are indicated at the bottom of the diagram. L, lymphocyte; P1 and P2, first and second plasma cfDNA, respectively; T, tumour.



Personalis EX2155

	Patient ID	L	T	P1	P2
T1N0 (T≤20mm)	**1**	-	-	-	-
	2	-	NA	-	5' 3' up
	5^{R}	-	3' up	-	-
	6	-	-	-	-
	7	-	5' 3' up	-	5' down
	8	-	NA	3' up	-
	10^{R}	-	NA	5' up	-
	11	-	3' up	-	-
	13	-	-	-	-
	14^{R}	-	5' 3' up	5' down	-
	15^{R}	-	3' up	-	-
	16	-	-	-	-
	17	-	NA	-	-
	18	-	-	-	-
	19	-	3' up	-	-
	20	-	-	-	-
	22	-	NA	-	-
	26	-	-	3' up	-
T>20mm/node-positive	**9**	-	-	5' down	5' 3' up
	23	-	-	-	5' 3' up
	24	-	NA	-	-
	25	-	-	3' up	-
	27	-	-	-	3' up
	30^{R}	-	3' up	-	-
	33^{R}	-	NA	-	-
	34	-	5' 3' up	-	-
	35	-	3' up	-	-
	36	-	-	-	-
	37	-	NA	-	5' 3' up
	38	-	-	-	-
	39	-	-	-	-
	40	-	-	-	-
	41	-	NA	-	-
	42^{R}	-	NA	-	-
	43	-	-	-	-
	44	-	-	-	-
	45	-	-	5' 3' up	3' up
	47	-	-	-	-
	48	-	-	-	-
	49	-	3' up	3' up	-
	51	-	-	-	3' up
	52	-	-	-	-
	53	-	-	-	-
	54	-	-	3' up	-
	55	-	-	-	3' up
	56^{R}	-	-	-	-
	59	-	5' 3' up	3' up	-
	61	-	3' up	-	-
	62	-	-	-	3' up
	64	-	5' 3' up	-	3' up

Table 6-1 - Copy number of *LMTK3* gene interval in the patients on follow-up. 'Up' denotes CN>2.5; 'down' denotes CN<1.5. L, lymphocyte; P1 and P2, first and second plasma cfDNA, respectively; T, tumour; NA, no sample available; Superscript 'R' denotes relapsed patient.

Personalis EX2155

6.4.4.2 Comparison to aCGH data of breast primary tumour and blood

The *Affymetrix SNP 6.0* data were also compared with the findings of a separate study by Bae *et al.* (2010). Using a threshold of CN>6.0 for amplification and <1.0 for deletion, good correlation was found: 60.9% of the cytobands identified by Bae's group as showing amplification in tumour (total=46) exhibited amplification in the tumour DNA samples of the patients on follow-up in this study, while the figure for deletions (total=11) was 45.5% (Table 6-2).

In the on follow-up group's plasma cfDNA samples, the correlation with cytobands identified in blood by Bae's group was 52.9% for amplifications (total=70) and 100% for deletions (total=16) in first plasma cfDNA, while corresponding figures for second plasma cfDNA were and 78.6 and 100% (Table 6-3).

Personalis EX2155

Amplification		Deletion	
Cytoband (frequency) n=30	**On follow-up tumour n=40**	**Cytoband (frequency) n=30**	**On follow-up tumour n=40**
1q23.1 (47)	2.5	3p21.31 (17)	7.5
2p11.2 (37)	25	4q35.2 (20)	5
3q21.3 (50)	2.5	6p25.3 (17)	5
3q27.1 (33)	2.5	22q11.1 (17)	30
7q11.23 (40)	100	22q11.21 (37)	87.5
8q24.3 (67)	2.5		
9p12 (37)	55		
10q22.3 (37)	22.5		
11p15.5 (60)	2.5		
11p15.4 (40)	5		
11p11.2 (40)	2.5		
11q12.3 (43)	2.5		
11q23.1 (33)	5		
12p13.31 (60)	2.5		
12q21.2 (33)	5		
15q11.2 (37)	57.5		
15q12 (33)	2.5		
15q25.2 (33)	85		
16p11.2 (37)	50		
17p13.3 (47)	2.5		
17p11.2 (43)	7.5		
17q12 (43)	2.5		
17q21.2 (37)	2.5		
17q25.3 (60)	2.5		
19p13.3 (53)	90		
19p13.11 (50)	2.5		
Xp11.22 (47)	2.5		
Xq23 (37)	2.5		

Table 6-2 - Frequencies of amplification/deletion (%) for cytobands identified by Bae *et al.*, (2010) in tumour. Amplification and deletion was based on a threshold of CN>6.0 and <1.0, respectively, by Gaussian-smoothed signal.

Personalis EX2155

Amplification			Deletion		
Cytoband (frequency) n=30	**On follow-up P1 n=50**	**On follow-up P2 n=50**	**Cytoband (frequency) n=30**	**On follow-up P1 n=50**	**On follow-up P2 n=50**
1p36.13 (40)	4	4	1q44 (17)	38	38
1p36.12 (43)	2	2	2p25.3 (17)	52	36
1p35.1 (43)	-	2	2q37.3 (23)	60	40
2p13.1 (47)	-	6	3p26.3 (17)	28	38
2p11.2 (37)	28	46	5q13.2 (20)	52	40
3p22.1 (33)	-	2	6p25.3 (20)	50	60
3q21.3 (50)	-	4	7q22.1 (20)	32	42
4p16.3 (50)	-	2	10q11.22 (27)	92	96
5p15.33 (57)	-	4	11q25 (20)	6	6
5p13.3 (57)	2	8	13q34 (20)	78	78
5q33.1 (40)	2	-	14q32.33 (27)	42	32
7q11.23 (40)	42	76	16q21 (17)	18	18
7q36.3 (33)	-	4	18p11.32 (27)	88	80
8p21.2 (37)	2	4	21q21.1 (17)	32	34
8q24.3 (67)	2	2	22q11.1 (23)	22	6
9p12 (37)	50	68	22q11.21 (47)	62	2
10p15.3 (37)	2	2			
10q22.3 (37)	10	20			
10q26.3 (37)	4	8			
11p15.4 (40)	8	14			
11p11.2 (40)	4	8			
11q12.3 (43)	2	-			
11q23.1 (33)	4	2			
11q23.3 (37)	2	-			
12p13.33 (40)	-	6			
12p13.31 (60)	2	6			
12q13.13 (33)	-	6			
12q21.2 (33)	2	4			
14q32.12 (43)	-	2			
15q11.2 (37)	16	32			
15q12 (33)	-	6			
15q25.2 (33)	42	60			
16p13.3 (53)	-	2			
16p11.2 (37)	22	60			
16q22.1 (40)	2	4			
16q24.1 (43)	-	-			
16q24.3 (37)	10	6			
17p13.3 (47)	-	2			
17p13.2 (33)	-	4			
17p12 (37)	2	2			
17p11.2 (57)	2	8			
17q11.2 (57)	-	4			
17q12 (43)	4	8			
17q21.2 (37)	-	4			
17q21.32 (43)	2	4			
17q25.3 (60)	-	6			
18q23 (33)	-	2			
19p13.3 (53)	58	82			
19p13.11 (50)	2	8			
19q13.2 (37)	4	4			
19q13.33 (37)	-	2			
19q13.43 (37)	-	2			
20p12.2 (33)		6			
20q13.12 (33)	2	-			
20q13.33 (53)	-	2			
22q12.2 (63)	-	2			
22q13.31 (50)	2	4			
22q13.33 (50)	2	-			
Xp11.23 (33)	-	8			
Xp11.22 (47)	4	8			
Xq23 (37)	4	12			

Table 6-3 - Frequencies of amplification/deletion (%) for cytobands identified by Bae *et al.*, (2010) in blood. Amplification and deletion was based on a threshold of CN>6.0 and <1.0, respectively, by Gaussian-smoothed signal. P1 and P2, first and second plasma cfDNA, respectively.

Personalis EX2155

6.4.4.3 Comparison to aCGH data of breast primary tumour and cancer cell-line

Comparisons were also made to a separate aCGH study (Naylor *et al.*, 2005). Numerous amplified and deleted genes were common (Table 6-4). In addition, although Naylor *et al.* (2005) identified 8p23.1 as a region of hemizygous loss, it was identified as a region of common amplification in this study (Table 3-6 and Table 4-4).

Amplification				Deletion			
	(Naylor *et al.*, 2005)				(Naylor *et al.*, 2005)		
Gene	**Primary tumour n=47**	**Cell line n=18**	**Current study**	**Genes**	**Primary tumour n=47**	**Cell line n=18**	**Current study**
BRCA1	34	44	Amplification in T1N0 (T≤20mm) relapsed final plasma cfDNA (Table 5-2)	*RB1*	26	50	SNP in on follow-up tumour (Table 5-2)
EXT1	32	78	SNP in on follow-up tumour (Table 8-2)	*CYLD*	26	61	-
TSC1	32	44	-	*MAP2K4*	23	28	-
EP300	30	39	-	*PTCH1*	19	33	Deletion in T<20mm/node-positive non-relapsed final plasma cfDNA (Table 5-2)
TSC2	30	22	-	*CDKN2A*	17	39	SNP in pre-surgical/NeoCEnT tumour (Table 8-2)
NF2	28	39	-	*APC*	14	22	-
SMARCB1	28	56	-	*MLH1*	13	44	-
CDH1	21	61	-[1]	*PTEN*	11	39	Deletion in on follow-up tumour (Table 5-2)
MEN1	19	39	-	*BRCA2*	9	28	-
STK11	13	18	Commonly amplified CNV at 19p13.3 (Table 3-6 and Table 4-4)	*MSH2*	9	17	-

Table 6-4 - Top 10 amplified and deleted tumour suppressor genes reported by Naylor *et al.* (2005) and their correlation to findings in this study from on follow-up tumour. For Naylor *et al.* (2005) data, frequencies of amplification/deletion (%) are shown.

[1] *CDH12* and *13* harboured SNPs in tumour (Table 8-2).

Personalis EX2155

6.5 Discussion

The aim of this chapter was to analyse key alterations identified by the *Affymetrix SNP 6.0* microarray as data validation. This was achieved through focussing on different variants - SNPs and CNVs - in the genome and comparing their values to those derived through independent methods (such as real time qPCR) or other studies.

6.5.1 Real-time qPCR validates microarray results

Through real-time qPCR, the microarray results were validated by focussing on two common CNVs that were amplified in a high proportion of tumour DNA and cfDNA samples by *Affymetrix SNP 6.0*. In addition, these amplifications were confirmed in an independent tumour series (Figure 6-2A), alluding to potential role for these two CNVs in breast cancer. Little amplification was detected in normal lymphocyte DNA, while some was detected in matched first and second plasma cfDNA (Figure 6-2B). In these experiments, unamplified DNA was used for qPCR, whereas WGA DNA was used in the microarray, which confirms the efficiency and unbiased nature of the WGA process.

6.5.2 LMTK3 is involved in resistance to therapy

Giamas *et al.* (2011) deduced that LMTK3 helped to 'switch on' the cellular response to oestrogen, thereby counteracting the effect of the ER antagonist tamoxifen. This was shown through an *in vivo* mouse xenograft model where tumour growth inhibition by LMTK3-siRNA and tamoxifen therapy was observed. So, LMTK3 knockdown allowed the tumour to respond to the therapy. Overall this suggests that the oncogenic activation of LMTK3 in breast cancer has two important roles: 1) it promotes tumour growth; and 2) it is one possible route to endocrine resistance.

In this study, amplification observed in the *LMTK3* gene interval was of interest and the results overall showed that many patients had specific CN changes within this gene that were detectable in cfDNA. Moreover, some of these changes mirrored those observed in the matched primary tumour (Figure 6-3) and persisted over a number of years, despite some of the patients having received adjuvant therapy. These data support previous findings in this study that revealed the persistence of tumour-associated CNVs

Personalis EX2155

in the first and second plasma cfDNA samples years after surgery and treatment, despite the fact that patients had no clinically-evident recurrent disease (Shaw *et al.*, 2012).

Significant amplification ($p<0.0001$) in *ESR1* and *LMTK3* was detected in the cfDNA samples of the 6 patients who relapsed while on tamoxifen therapy, and may suggest a mechanism leading to tamoxifen resistance and endocrine therapy in general. The amplification observed was most likely subtle in origin as it was only detected when applying Gaussian-smoothing on filtered-data and did not appear in Table 6-1, which was based on unfiltered data.

The response to tamoxifen involves many genes and pathways, and approximately 50% of ER-positive cancer patients fail to respond to endocrine therapy, while some who do initially respond later develop resistance (Hicks *et al.*, 2012). Tumour cells that persist in a dormant state in micrometastatic deposits may be under selective pressure to harbour mutations that permit them to 'resist' the therapy being applied and to later relapse. Amplifications in *ESR1* and *LMTK3* are 2 possible milestones on this path to relapse.

In relation to this, a previous study identified a signature of genes exhibiting significant differences between tamoxifen-treated and non-treated patients, including SNPs in *ABCC4*, *ESR1*, *GRIK1*, and *RB1* (Hicks *et al.*, 2012). They also found that *ESR1*, *MCC*, and *WWOX* were up-regulated, whilst *GRIK1*, *RB1*, and two *XRCC* genes were down-regulated, when just looking at the treated patients. Each of these genes was also identified in this thesis through principal component analysis as potential cancer biomarkers (Table 5-2).

6.5.3 Microarray can serve as a retrospective and prospective tool

It was possible to interrogate the microarray data to retrospectively confirm results of previously-published studies, including those of Naylor *et al.* (2005), Bae *et al.* (2010), and Giamas *et al.* (2011). Conversely, the data also served as a prospective tool for building new hypotheses. For example, the successful analysis of the *LMTK3* gene has

Personalis EX2155

formed a new collaboration with Professor Justin Stebbing at Imperial College London and has provided data for a forthcoming publication.

The *Affymetrix SNP 6.0* data is continually being mined as research collaborators search for clues and evidence that back-up their own hypotheses or independent data sets. As a further example, two genes previously identified as having roles in ovarian cancer - *OPCML* (Sellar *et al.*, 2003) and *WWOX* (Paige *et al.*, 2001 and Gourley *et al.*, 2005) - have been analysed for Professor Hani Gabra at Imperial College London as potentially having prognostic and predictive roles in breast cancer.

6.6 Conclusion

To conclude, it has been demonstrated that *Affymetrix SNP 6.0* microarray technology is capable of describing true SNP and CNV in both cancer and healthy tissue, and that it serves as a useful platform for comparison to other approaches. Moreover, it has been confirmed that variation found in cfDNA detected by microarray can be validated by real-time qPCR for a number of candidate markers.

Personalis EX2155



7 CONCLUSIONS AND FUTURE DIRECTION

Personalis EX2155

The aim of this thesis was to test these two related hypotheses using bioinformatics to analyse/interrogate *Affymetrix SNP 6.0* microarray data and compare cfDNA extracted from blood plasma with matched lymphocyte and tumour DNA (where available).

7.1 Illuminating the breast cancer genome

Through various analytical approaches, snapshots of genome-wide breast cancer variation have been illuminated in this thesis. This variation included: 1) the 23 common and highly-amplified CNVs in cfDNA and tumour of the patients on follow-up (Table 3-6) (21 in the NeoCEnT cohort (Table 4-4)); 2) the suite of derived markers that acted as genetic signatures for the cfDNA of unaffected *BRCA1* mutation carriers and primary breast cancer patients who did and did not relapse (Table 5-3); and 3) the complete list of genes that were significantly altered by SNP and/or CNV based on PCA comparisons (Table 5-2 and Table 8-2).

7.1.1 Common CNVs and the cfDNA genetic signatures may have utility in the clinical setting

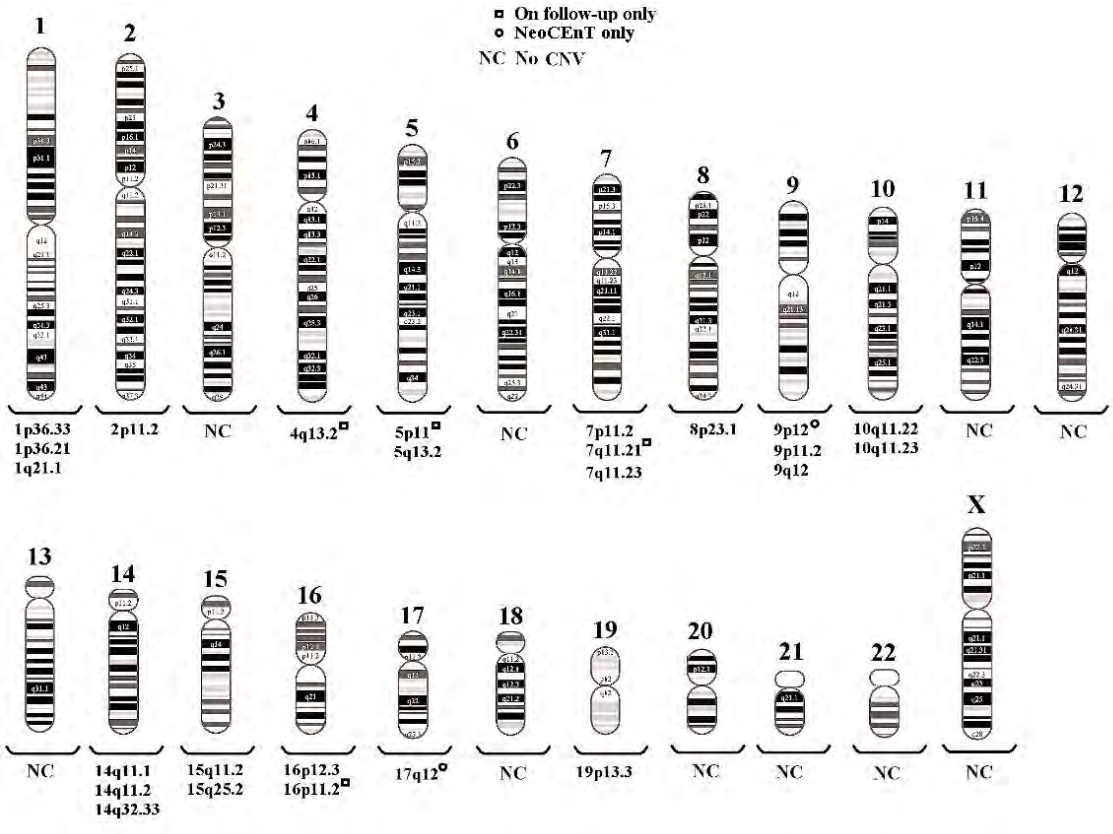
The combination of the common amplified CNVs and the cfDNA genetic signature may have great potential in the clinical setting. The markers comprising the signatures were shown to segregate cfDNA of healthy controls, unaffected *BRCA1* mutation carriers, and primary breast cancer patients on follow-up who did and did not relapse, whilst the common CNVs could serve as a baseline (or 'safeguard') for the presence or absence of cancer.

The common amplifications were screened to include only those that had a copy number greater than 6. They exhibited high prevalences in both cfDNA and tumour, with 8 of these having prevalences of >90% in the tumour DNA of on follow-up patients (Table 3-8), indicating their importance in breast cancer. Also, by focussing on these CNVs, it was shown that tumour-specific alterations persisted in the plasma cfDNA of breast cancer patients up to 12 years after surgery to remove the primary tumour (Figure 3-14). These CNVs have potential as diagnostic markers for cancer through the analysis of cfDNA (Figure 7-1).

Personalis EX2155

The markers comprising the genetic signatures were derived through PCA using the method by Meng *et al.* (2003) and were capable of segregating the different cfDNA sample groups through hierarchical clustering (Figure 5-4), thus highlighting their utility for classifying any sourced cfDNA sample into a particular grouping and showing their diagnostic and prognostic potential (Figure 7-2). Principal component analysis was also used to arrive at lists of genes within which were contained SNPs or CNVs that different significantly between different sets of samples. Many of the affected genes had known roles in cancer and related to the oestrogen receptor and its signalling pathways. Also, the high prevalences of the variants mapping to these genes pointed to mutation 'hot spots' in the breast cancer genome.

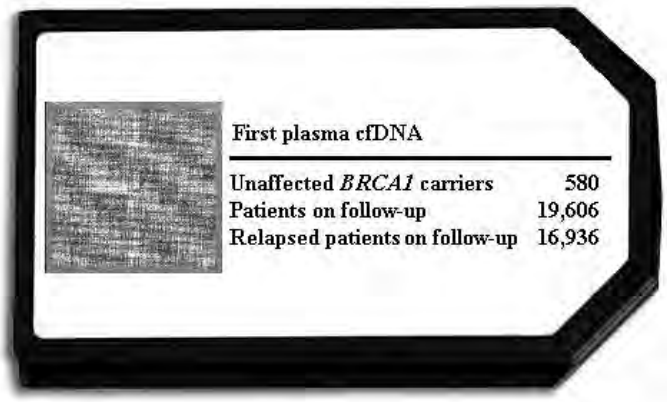
Figure 7-1 - Common and amplified CNVs to be screened for through plasma cfDNA as a diagnostic and prognostic tool for cancer. Each CNV is listed under its respective chromosome.



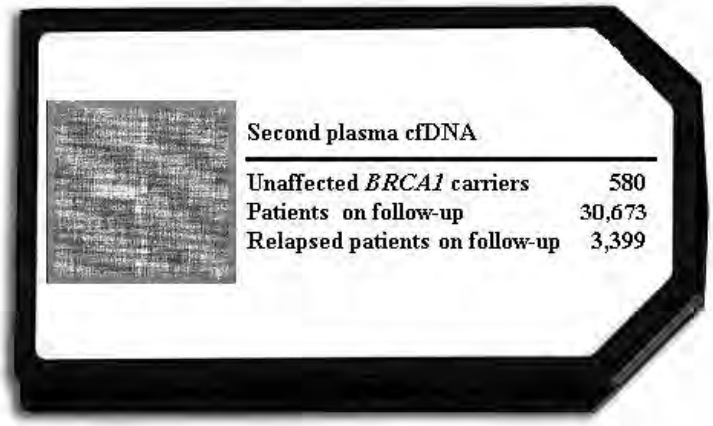
Personalis EX2155

Figure 7-2 - Number of markers previously shown (Figure 5-4) to be capable of distinguishing the cfDNA of healthy controls, unaffected *BRCA1* mutation carriers, and patients on follow-up who did and did not relapse for the first (A) and second (B) plasma cfDNA. The second plasma cfDNA markers have more utility at distinguishing relapsed patients due to these samples having been taken closer to time of relapse than their first. Also shown is the number of markers derived for the time-point plasma cfDNA samples of the pre-surgical/NeoCEnT breast cancer patients on neoadjuvant therapy (C).

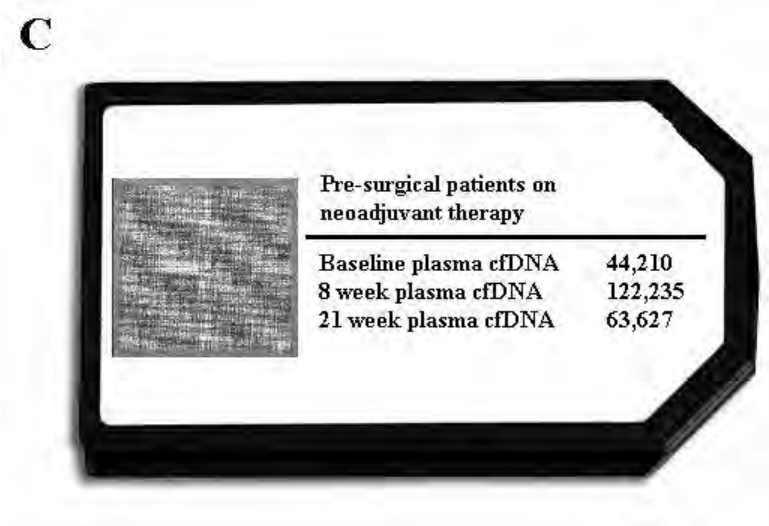
A



B



Personalis EX2155



7.1.2 Genes significantly-different through PCA can be used as panel of cancer biomarkers

Alternatively, one could solely focus on the list of genes that were significantly altered in the primary tumour DNA samples (Table 5-2) and to investigate each of these as potential biomarkers. The most important cost elements in cancer are that of screening and treating advanced disease, thus it is important to find a way of diagnosing breast cancer at the earliest possible stage when therapy is most effective (Vainio *et al.*, 2002). If a cheaper, non-invasive, and more sensitive and specific method of diagnosis existed that could detect breast cancer in its earliest stage, it would result in a reduction of breast cancer-related mortality. It would also permit governments to replace existing screening programmes, which can be both costly and imperfect due to returning false-negatives or -positives (Armstrong *et al.*, 2007), whilst also offering it to wider age groups.

Currently, Applied Biosystems™ Ion AmpliSeq™ Cancer Panel (Applied Biosystems™, http://www.appliedbiosystems.com/) searches for 739 mutations in 46 genes known to be involved in cancer. Using data from this study, the spectrum is widened to include 761 variants (571 SNP and 190 CNV) (Table 8-2) based on data from the tumour DNA of the patients on follow-up. Many variants were also identified in the cfDNA circulating in blood plasma, some of which have been shown to be

Personalis EX2155

tumour-specific. However, the majority of these variants are yet to be tested. More targets could be included in diagnostic and prognostic tests, but it becomes a battle between sensitivity, specificity, and cost.

7.2 Missing heritability?

Related to the variants unearthed in this study is the concept of 'missing heritability' in disease - i.e. the amount of variance associated with a disease that cannot be explained solely by heritable effects (Manolio *et al.*, 2009). Age-related macular degeneration, for example, has 5 identified loci that only account for 50% of the explained heritability (Maller *et al.*, 2006), while corresponding figures for systemic lupus erythematosus are that 6 loci account for 15% of the heritability (International Consortium for Systemic Lupus Erythematosus Genetics *et al.*, 2008). A possible explanation for this missing heritability are *de novo* variants, such as much of the CNV found in this study, which accounted for more than two-fifths of the variation found in some cases (see section 3.3.4.1) - these were CNVs that did not overlap with any listed in the Toronto database of genomic variants (Iafrate *et al.*, 2004).

In addition, it was noted in two other studies that more than 80% of disease-associated variants fell outside protein coding regions (Hindorff *et al.*, 2009 and Manolio *et al.*, 2009), while in another, CNV affecting dosage-sensitive genes appeared to be negatively-selected by evolution (Schuster-Böckler *et al.*, 2010). These findings are backed-up by data in this thesis where, through PCA analysis, the majority of significantly-different markers mapped to introns. It is possible that these have effects on disease phenotype and may also account for missing heritability in disease.

7.3 Unresolved questions

One question that remains pertains to the origin of circulating cfDNA. A potential origin is circulating tumour cells (CTCs). To elucidate this, CTCs would have to be extracted from blood and then compared to cfDNA extracted from the same patient. There are different methods for isolating CTCs from blood, each with positives and

Personalis EX2155

negatives (Cristofanilli *et al.*, 2004; Paterlini-Brechot & Benali, 2007; and Farace *et al.*, 2011), and it has already been strongly supported in this thesis that a representative human genome can be extracted as cfDNA. It is possible that the cfDNA profile changes with the evolution of breast cancer in a person, long before cells become malignant, and that minute alterations detected in cfDNA represent the first signs of genomic instability in pre-cancerous cells.

7.4 Conclusion and future direction

To conclude, although the numbers of patients and controls analysed in this study were small, the results presented in this thesis have demonstrated that *Affymetrix SNP 6.0* analysis of cfDNA can distinguish between primary breast cancer patients and healthy controls, and between pre-surgical breast cancer patients and those who already had surgery and treatment. High level CNV amplification was observed in matched tumour and cfDNA at numerous loci on different chromosome arms. Moreover, these tumour-specific CNVs were detectable in cfDNA up to 12 years after initial diagnosis of cancer in patients on follow-up. This finding strongly infers breast cancer dormancy in the majority of the patients studied in this thesis. A similar study involving a larger and independent cohort of patients is necessary to further support this.

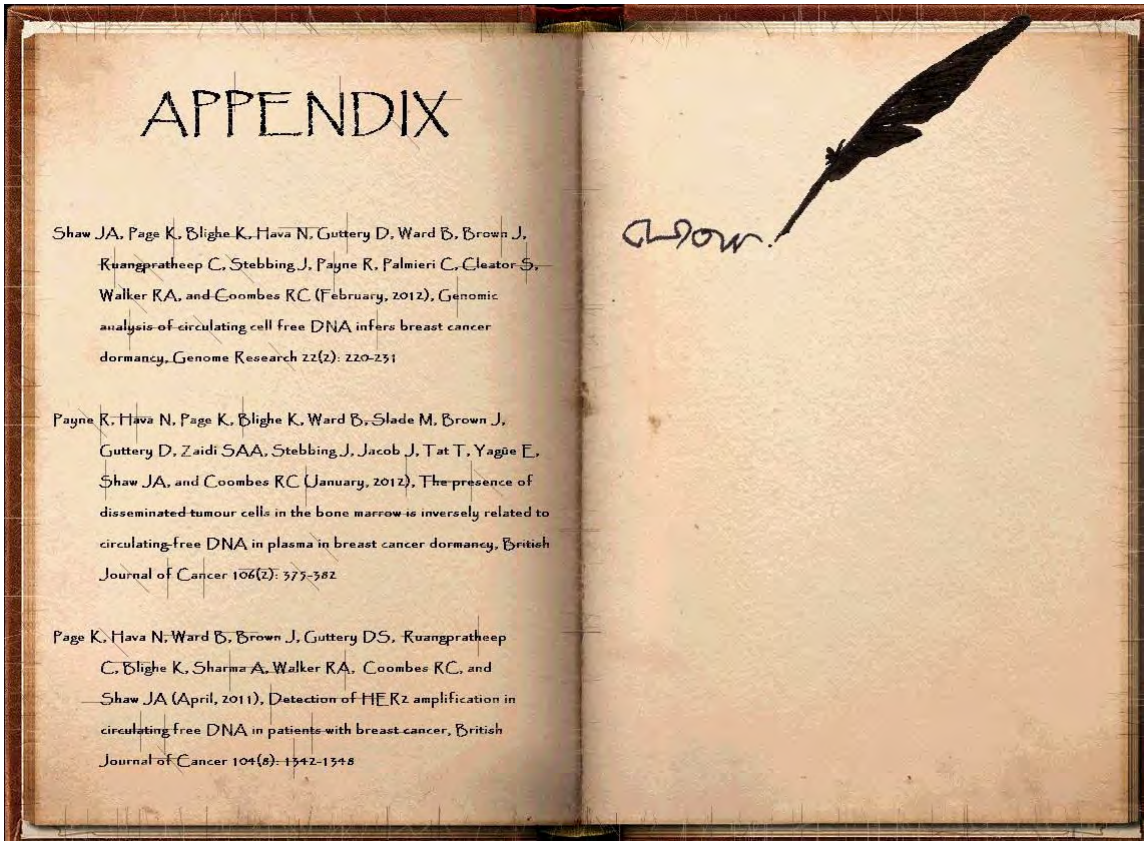
Hence, there is strong evidence from the data presented in this thesis that the analysis of circulating cfDNA by microarray technology can be a useful substitute to tumour biopsy. The results, if validated in other series of patients and samples, would represent the first step on a long road that could offer real, translational potential and alter the diagnostic and prognostic landscape of breast cancer. Therefore, one key goal of research in the future must be to develop methods of achieving this based on key SNP/CNV targets. Future studies will therefore focus on developing high throughput approaches to target common CNVs for screening and monitoring. A targeted chip could be developed to be loaded with a cfDNA sample extracted from the blood of the individual being tested. In this scenario, the probe sequences included on the chip could be representative of those SNPs and CNVs that were identified as having significant involvement in breast cancer. For example, the set of markers identified through PCA

Personalis EX2155

that provided a genetic signature of the cfDNA of healthy controls, unaffected *BRCA1* mutation carriers, and primary breast cancer patients who did/did not relapse (Table 5-3) could all be included on the chip to assist in screening and also in the adjuvant setting, potentially to infer those at risk of relapse. The chip could also detect the commonly-amplified CNVs (Table 3-6 and Table 4-4) using a suite of CN probes that swept these particular intervals for high level amplification.

Ultimately, the method for detecting genomic variation that offers the highest resolution is next generation sequencing (Van de Wiel *et al.*, 2011 and Manolio *et al.*, 2009). However, sequencing technologies are still in their infancy and are not yet ready for use in diagnostics (Meldrum *et al.*, 2011). Such methods also generate tremendous amounts of data - much more than microarrays - that equally require intensive interpretation. For example, solely from pilot data, the 1,000 Genomes Project (The 1,000 Genomes Project Consortium, http://www.1000genomes.org/) has already discovered around 15 million SNPs and 1 million insertions and deletions, but to what end? Whole genome sequencing may merely provide another more expensive way of determining potential targets that then have to be examined, as usual, in the laboratory. However, targeted sequencing could be a way forward if it can be developed to cover key somatic mutations and disease-causing SNPs/CNVs.

Personalis EX2155



8 APPENDIX

Personalis EX2155

8.1 23 commonly-amplified intervals - additional information

		Healthy control		Pre-surgical		On follow-up		
		L	P	P	T	P1	P2	T
		n=8	n=8	n=15	n=15	n=50	n=50	n=40
1p36.33	**Size**	84-762 [343]	106-482 [294]	151-274 [212]	207-624 [335]	72-531 [270]	84-500 [266]	146-298 [244]
	Probes	51-193 [81]	50-126 [77]	59-62 [61]	51-174 [94]	50-147 [75]	51-147 [72]	50-71 [58]
1p36.21	**Size**	49-711 [116]	39-928 [149]	45-636 [109]	42-634 [220]	40-1022 [172]	37-952 [113]	42-861 [178]
	Probes	50-176 [77]	50-129 [77]	50-124 [65]	60-108 [77]	50-248 [76]	50-212 [68]	51-77 [60]
1q21.1	**Size**	47-543 [156]	50-454 [115]	57-701 [133]	86-621 [338]	48-1054 [200]	46-679 [164]	88-607 [175]
	Probes	51-167 [79]	50-84 [57]	50-74 [61]	53-195 [116]	50-128 [66]	50-133 [65]	50-86 [57]
2p11.2	**Size**	62-1438 [254]	58-1202 [170]	45-1218 [153]	69-451 [187]	49-1317 [220]	32-1207 [167]	55-1327 [232]
	Probes	51-147 [85]	50-161 [74]	50-146 [72]	50-166 [96]	50-149 [76]	50-180 [74]	50-131 [90]
4q13.2	**Size**	43-543 [166]	50-745 [163]	45-555 [157]	63-835 [400]	49-652 [163]	37-626 [163]	61-595 [386]
	Probes	50-165 [76]	52-220 [74]	50-123 [64]	54-363 [127]	50-175 [70]	50-150 [67]	52-86 [65]
5p11	**Size**	3352-3645 [3454]	142-3379 [2276]	-	216-3475 [2348]	196-3646 [2785]	103-3675 [2215]	268-320 [291]
	Probes	58-106 [75]	51-72 [61]	-	51-100 [69]	53-106 [70]	50-115 [67]	60-65 [63]
5q13.2	**Size**	42-1660 [314]	42-212 [101]	63-191 [106]	44-301 [141]	45-600 [108]	34-1936 [121]	52-145 [103]
	Probes	51-78 [59]	50-114 [65]	50-122 [69]	50-118 [75]	50-117 [66]	50-117 [64]	51-72 [57]
7p11.2	**Size**	40-113 [87]	38-98 [72]	27-172 [77]	26-129 [90]	24-543 [99]	23-438 [87]	26-128 [73]
	Probes	51-137 [82]	60-96 [82]	50-90 [61]	55-110 [79]	50-189 [71]	50-188 [68]	51-107 [67]
7q11.21	**Size**	57-520 [241]	75-417 [180]	31-180 [88]	108-597 [262]	61-513 [194]	48-634 [180]	55-292 [150]
	Probes	51-212 [77]	50-137 [78]	50-134 [67]	55-133 [98]	51-160 [70]	50-128 [63]	50-74 [58]
7q11.23	**Size**	69-366 [182]	68-186 [122]	59-215 [114]	68-222 [157]	35-1085 [171]	36-1226 [142]	44-1232 [261]
	Probes	57-174 [90]	50-95 [66]	51-78 [63]	50-148 [92]	50-226 [71]	50-147 [67]	50-69 [56]
8p23.1	**Size**	25-475 [87]	25-356 [88]	22-365 [82]	30-464 [123]	21-714 [103]	22-949 [100]	22-361 [86]
	Probes	50-93 [63]	50-190 [68]	50-122 [63]	50-196 88]	50-146 [66]	50-205 [66]	50-70 [57]
9p11.2	**Size**	-	1522	663-21784 [4018]	1503	835-21570 [7978]	983-22860 [13125]	641-1503 [976]
	Probes	-	63	58-108 [75]	52	65-86 [75]	56-136 [82]	51-63 [55]

Personalis EX2155

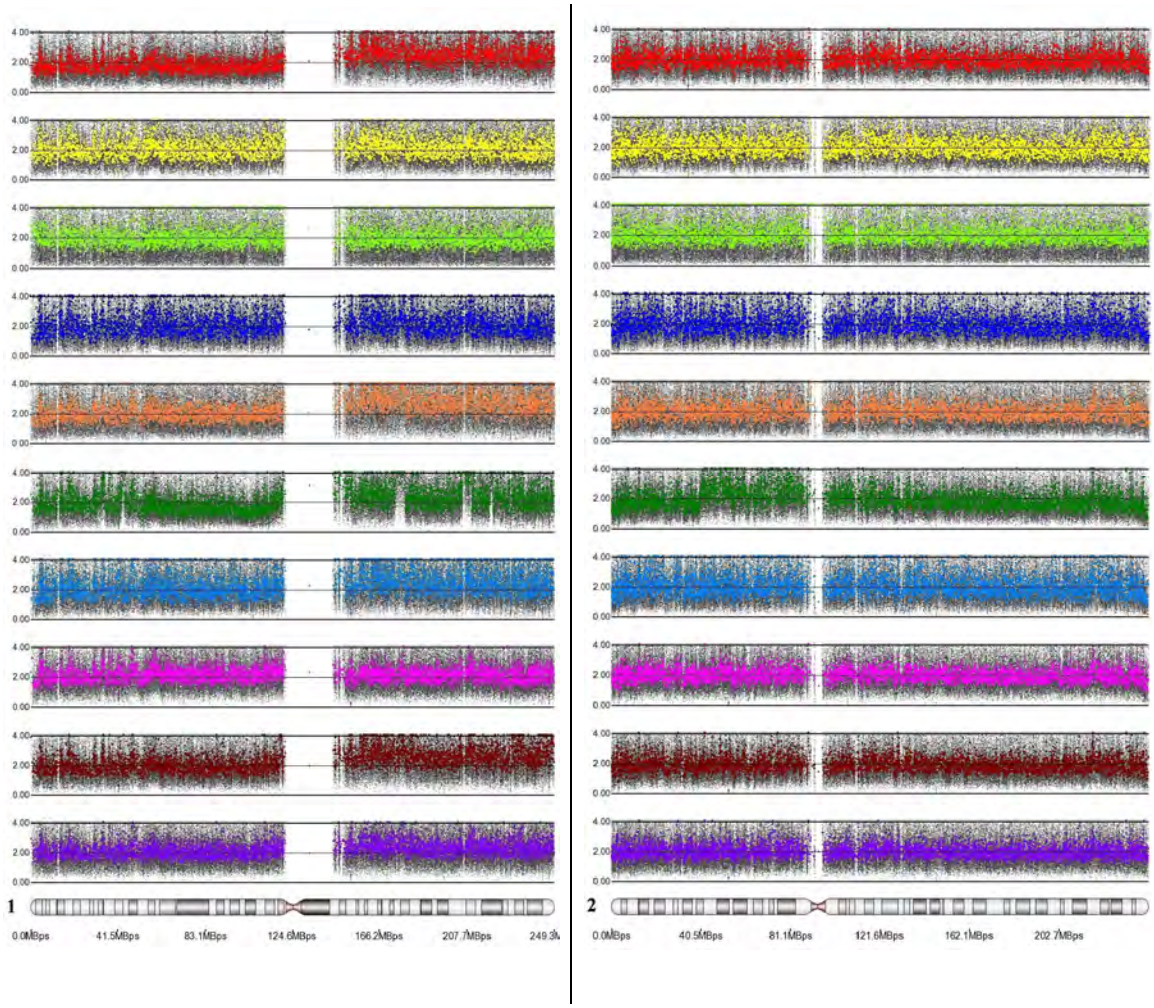
9q12	**Size**	-	1363	654-1142 [970]	1099	76-2416 [1230]	508-1358 [969]	445-573 [509]
	Probes	-	70	54-80 [63]	65	51-71 [59]	50-81 [59]	56-60 [58]
10q11.22	**Size**	27-543 [105]	33-826 [97]	32-332 [80]	55-1073 [210]	27-865 [110]	26-832 [115]	31-801 [146]
	Probes	50-190 [79]	50-120 [72]	50-144 [68]	51-265 [108]	50-170 [74]	50-157 [70]	50-75 [60]
10q11.23	**Size**	51-133 [84]	42-265 [80]	27-115 [71]	41-131 [84]	21-272 [84]	19-469 [89]	25-281 [98]
	Probes	50-77 [60]	51-109 [69]	51-124 [63]	50-85 [60]	50-118 [67]	50-134 [65]	52-84 [61]
14q11.1	**Size**	314-1087 [808]	811-1321 [1081]	1004-1005 [1029]	630-970 [800]	563-1238 [1013]	563-1248 [960]	-
	Probes	52-142 [84]	52-130 [84]	64-82 [73]	64-136 [100]	50-143 [76]	51-83 [65]	-
14q11.2	**Size**	39-231 [101]	26-222 [94]	41-139 [74]	40-215 [102]	23-366 [94]	23-313 [91]	45-150 [82]
	Probes	50-146 [78]	50-149 [68]	50-110 [61]	50-148 [79]	50-321 [72]	50-286 [71]	50-80 [60]
14q32.33	**Size**	55-478 [208]	62-460 [200]	61-315 [189]	74-441 [224]	56-572 [200]	55-695 [188]	52-447 [216]
	Probes	51-174 [76]	51-171 [80]	51-144 [71]	50-164 [83]	50-240 [74]	50-268 [72]	50-149 [77]
15q11.2	**Size**	46-399 [123]	49-579 [135]	51-474 [147]	56-446 [158]	41-545 [144]	39-942 [144]	45-257 [144]
	Probes	50-194 [74]	50-139 [73]	50-96 [65]	50-133 [68]	50-212 [76]	50-174 [73]	50-98 [64]
15q25.2	**Size**	57-778 [191]	47-596 [141]	66-324 [119]	92-796 [229]	35-715 [145]	37-695 [136]	83-133 [110]
	Probes	50-139 [79]	50-91 [63]	50-98 [64]	53-107 [80]	50-141 [67]	50-161 [64]	50-69 [57]
16p12.3	**Size**	48-686 [118]	45-674 [120]	52-215 [87]	45-731 [215]	43-908 [126]	49-705 [102]	54-144 [86]
	Probes	50-125 [71]	53-128 [75]	50-194 [64]	53-246 [127]	50-262 [75]	50-230 [71]	50-95 [60]
16p11.2	**Size**	71-326 [149]	68-716 [210]	61-557 [179]	68-518 [218]	51-816 [217]	39-698 [200]	75-272 [189]
	Probes	50-84 [59]	50-86 [60]	50-92 [62]	51-138 [76]	50-142 [64]	50-115 [64]	50-69 [60]
19p13.3	**Size**	95-748 [333]	100-502 [232]	61-371 [175]	99-582 [299]	52-1158 [240]	52-903 [230]	63-297 [163]
	Probes	50-183 [99]	51-137 [71]	50-93 [64]	53-148 [85]	50-274 [80]	50-202 [74]	50-121 [63]

Table 8-1 - CNV size range (Kb) and marker density for 23 cytobands that exhibited a commonly-amplified region (Table 3-6). Amplifications and deletions for the entire cytoband containing the CNV were taken into account; mean shown in square-brackets; single values indicate the presence of just one CNV; absent values indicate no CNV. L, lymphocyte; P, plasma cfDNA of healthy controls or pre-surgical breast cancer patients; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour.

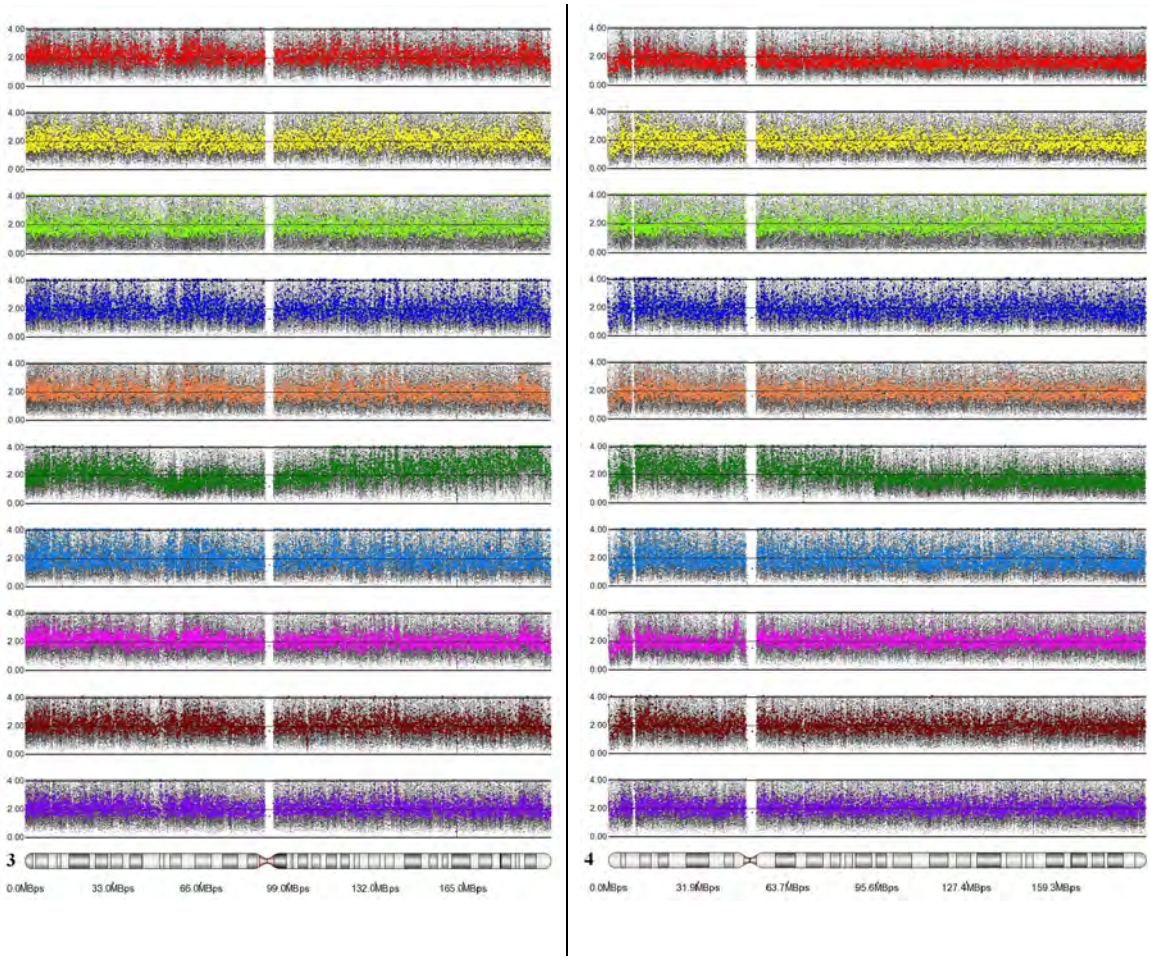
Personalis EX2155

8.2 Genomic aberrations

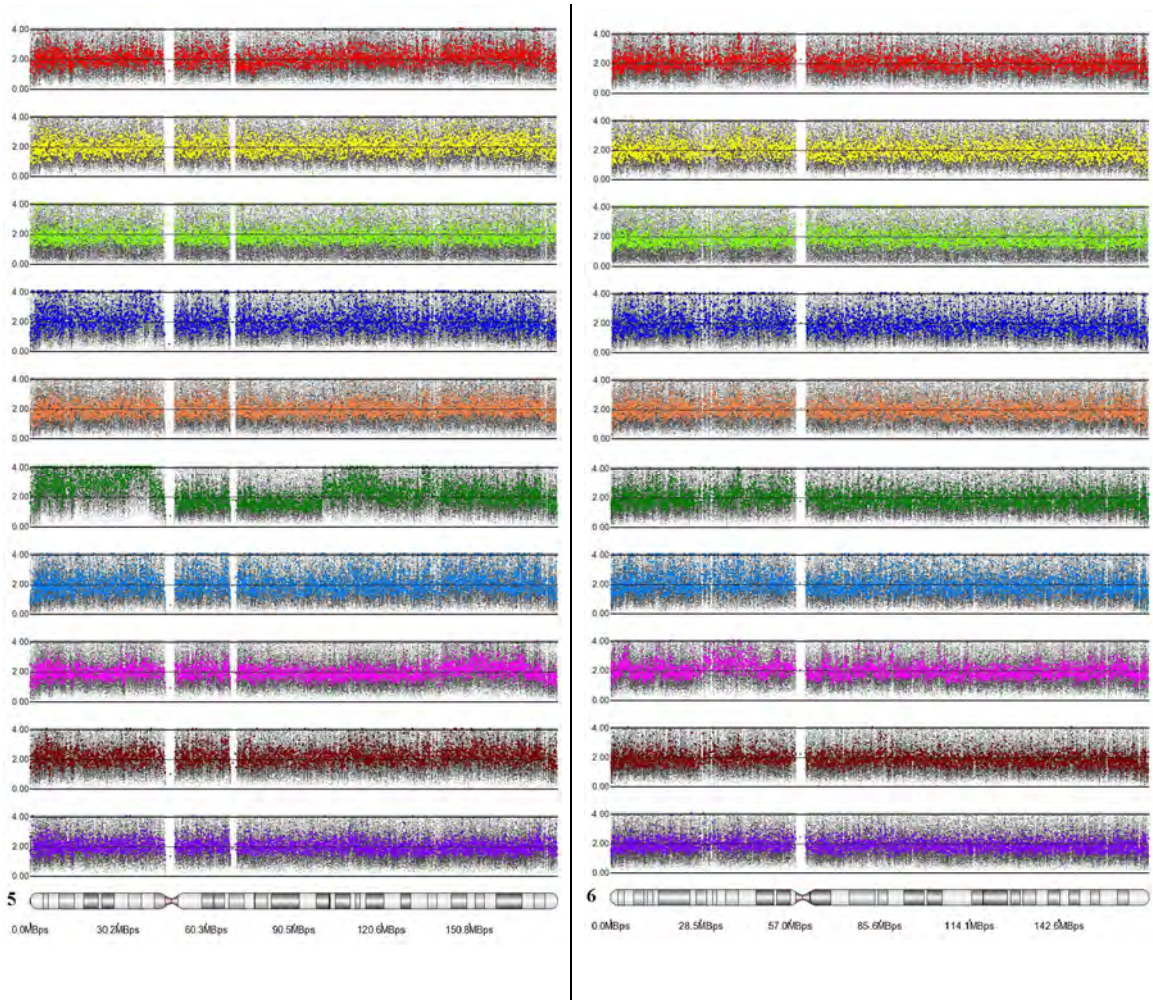
Figure 8-1 - Gaussian-smoothed (smoothed points shown as coloured dots), CN output of NeoCEnT tumour samples. Samples are colour-coded as follows: red, N1; yellow, N2; light-green, N3; dark-blue, N6; beige, N7; dark-green, N8; light-blue, N10; pink, N11; brown, N12; purple, N14. CN scale fixed to 0-4; chromosomes not to scale.



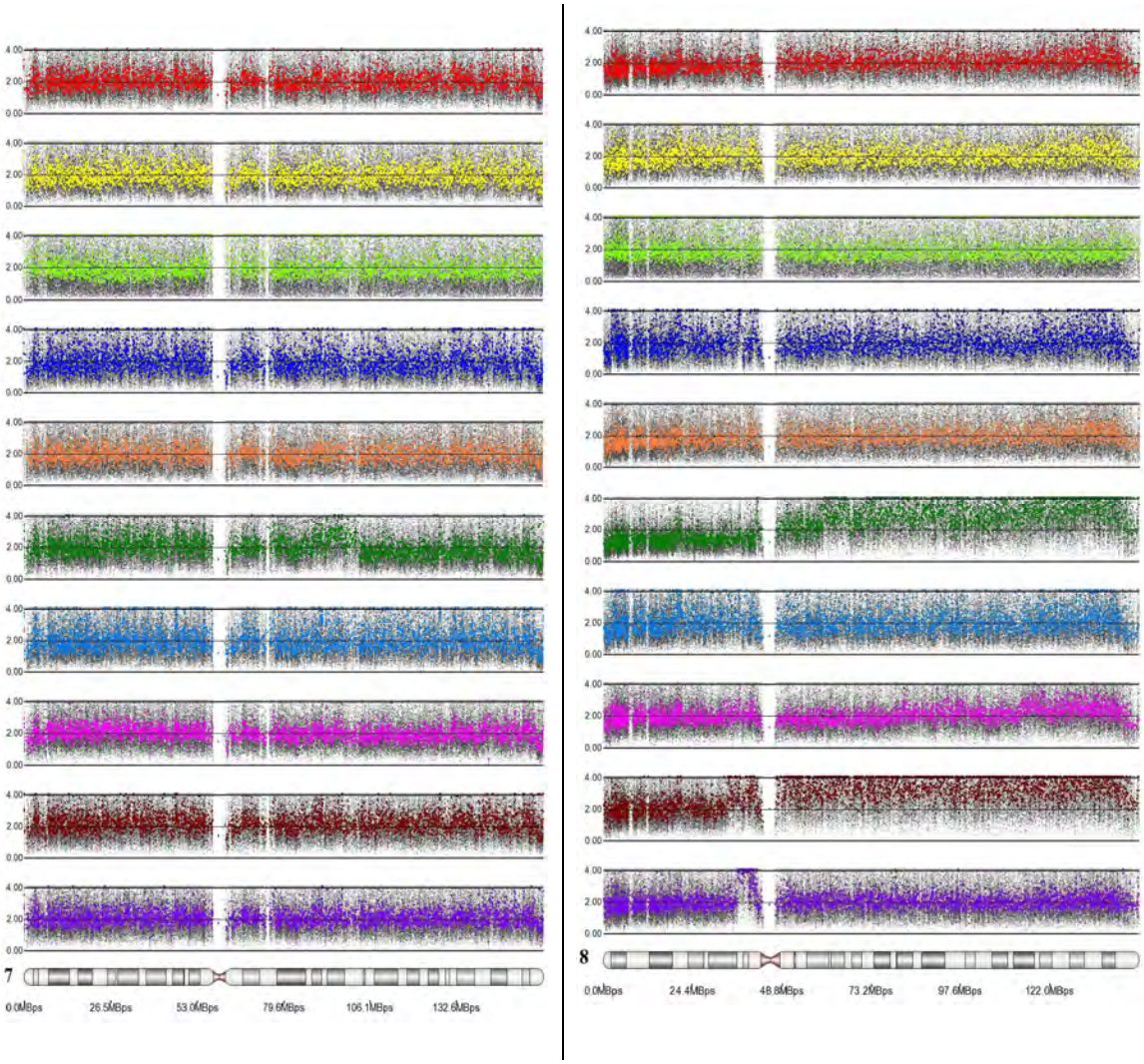
Personalis EX2155



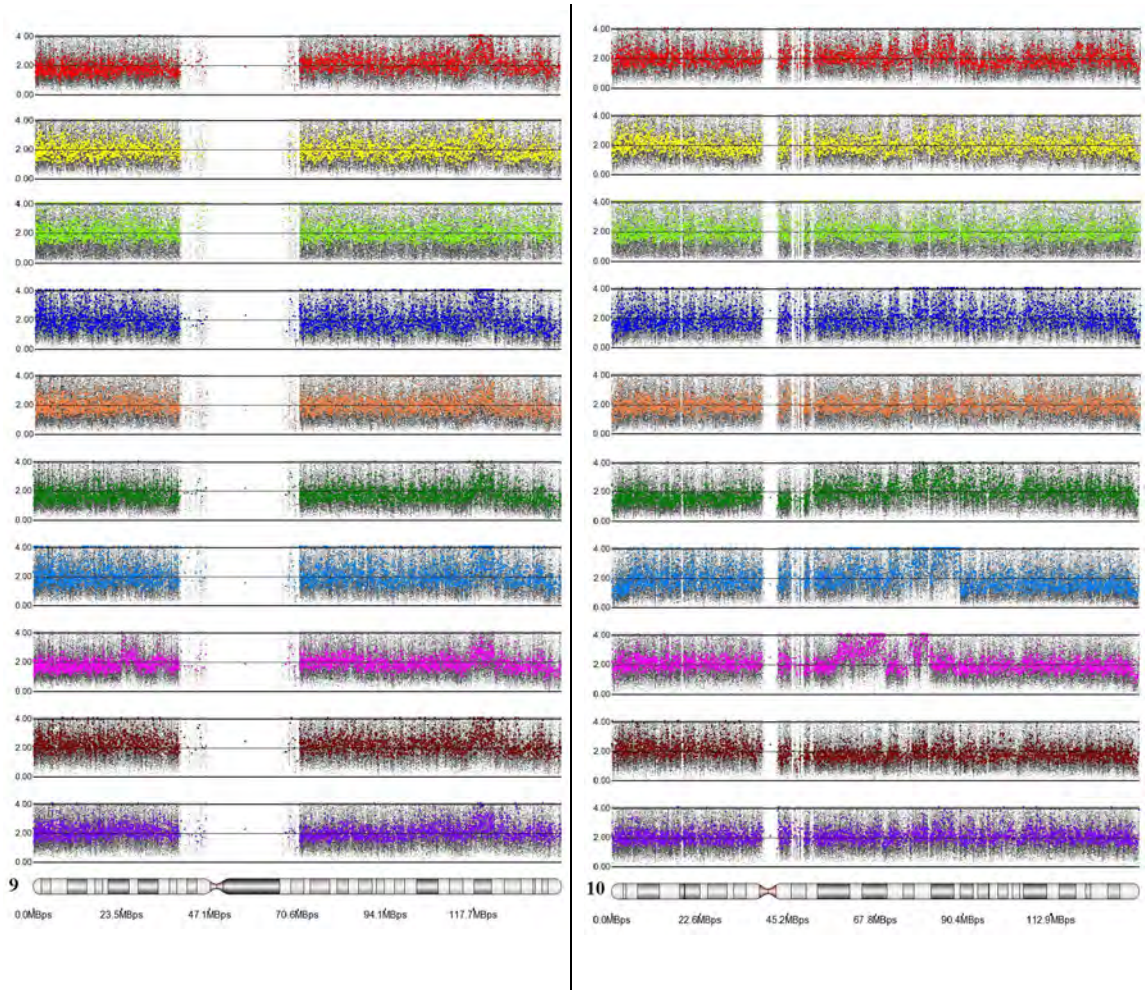
Personalis EX2155



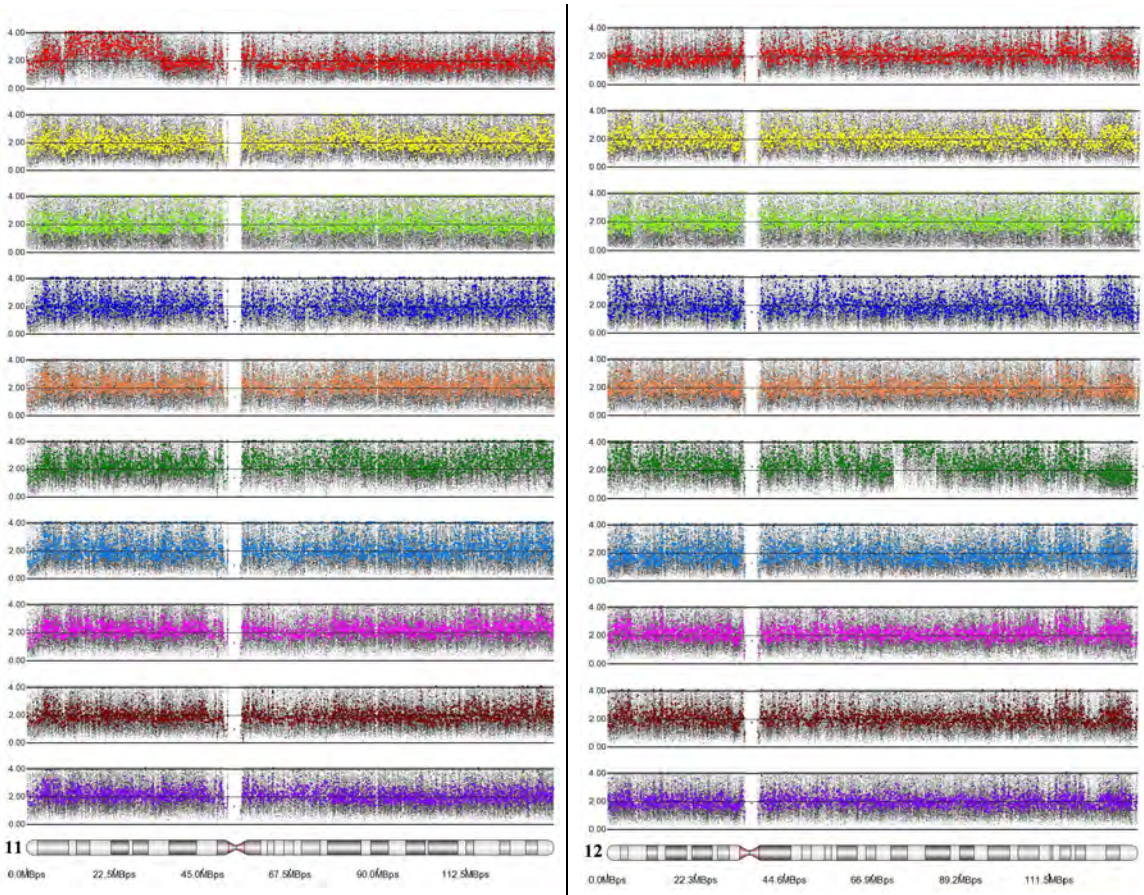
Personalis EX2155



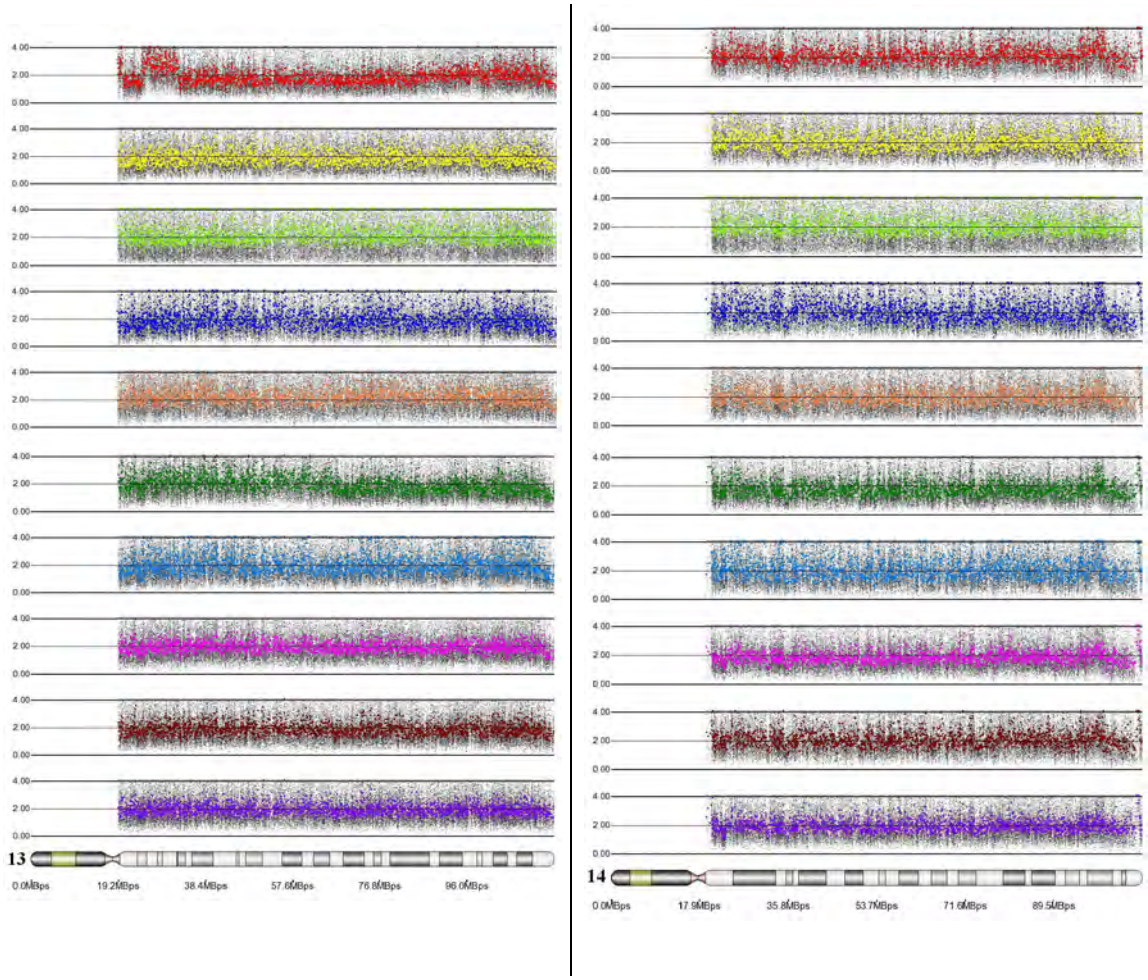
Personalis EX2155



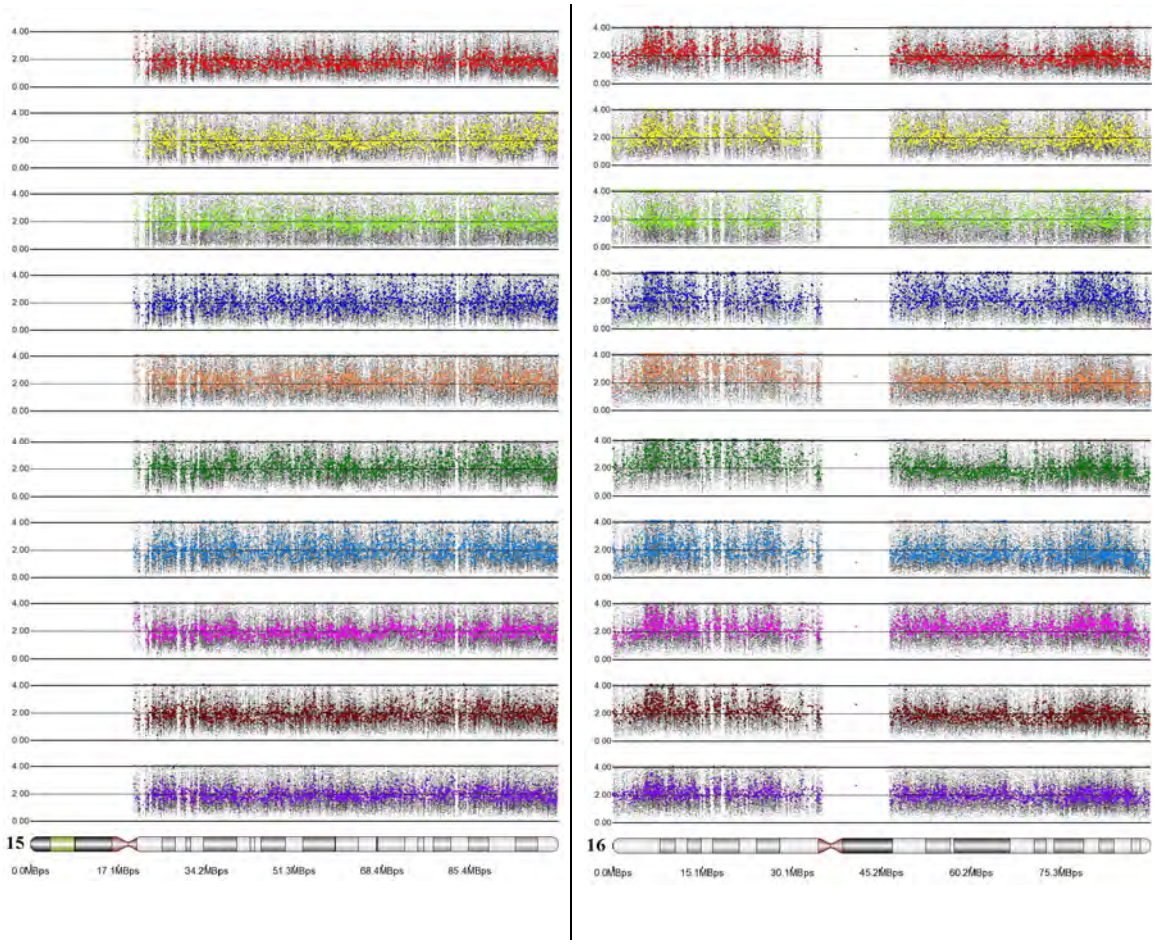
Personalis EX2155



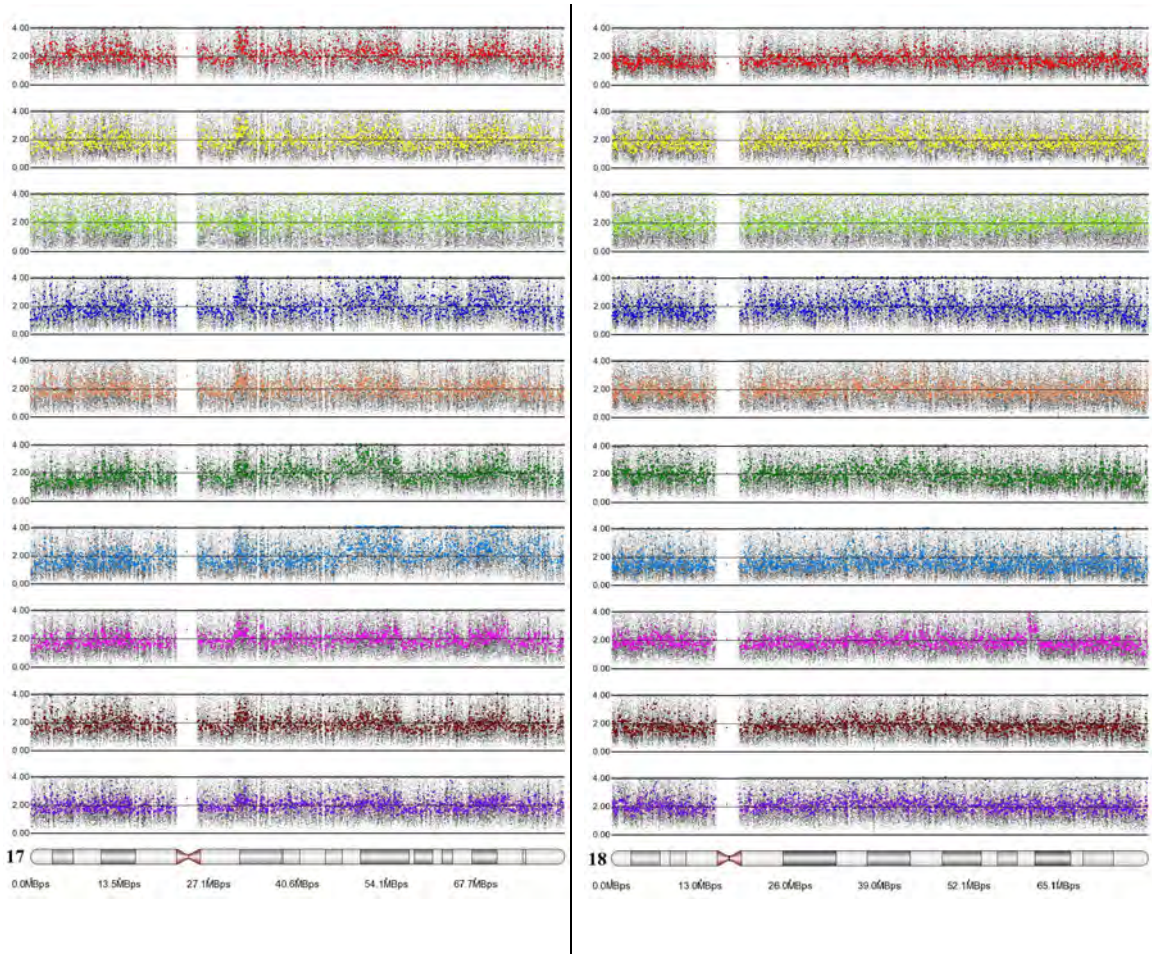
Personalis EX2155



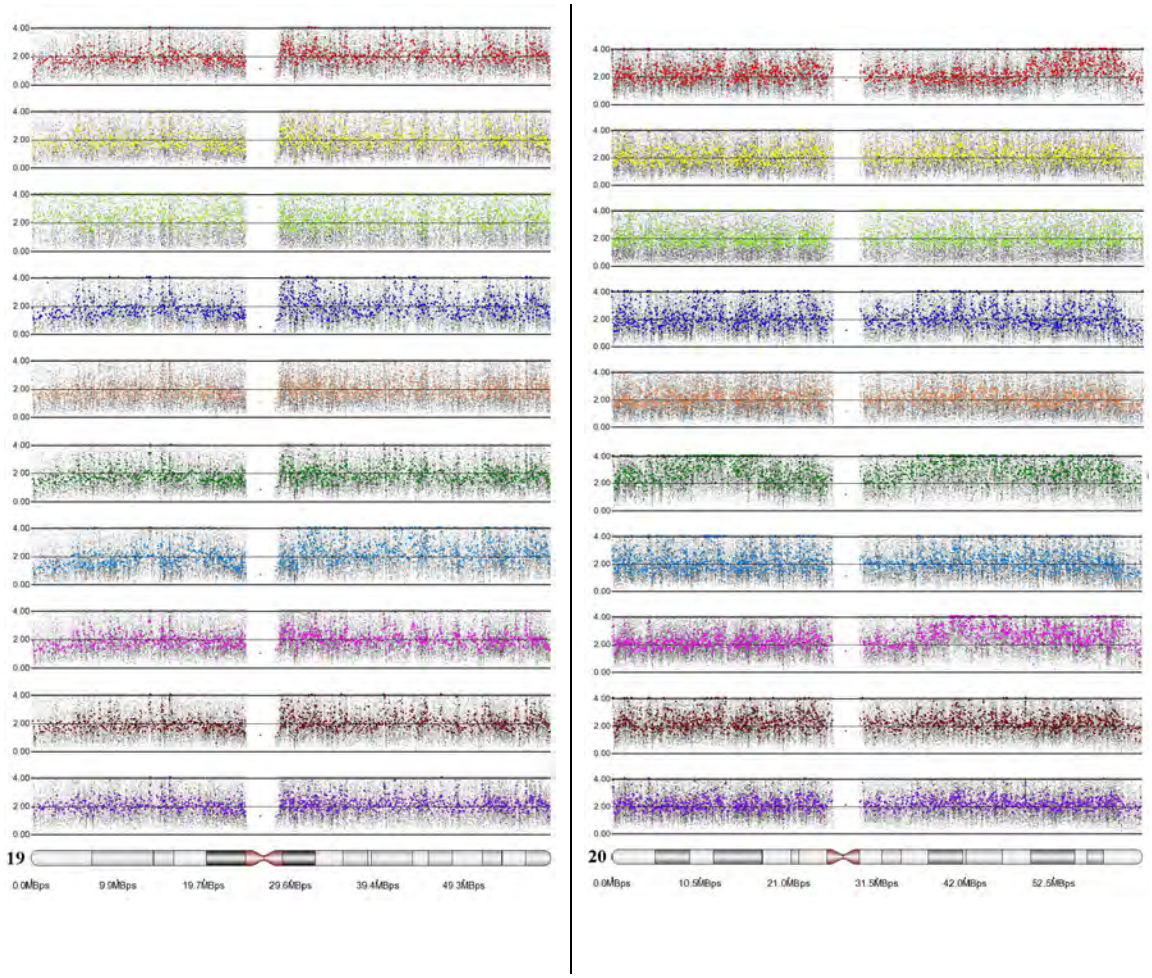
Personalis EX2155



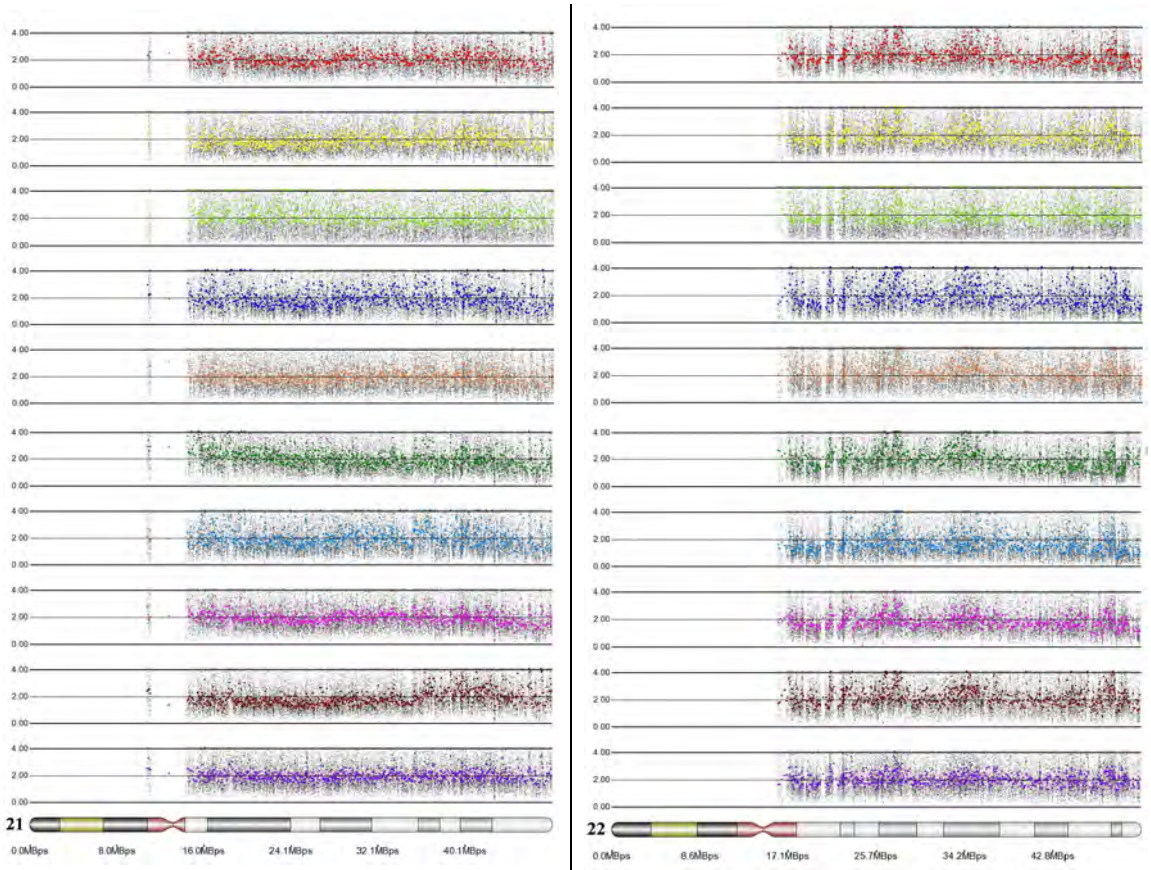
Personalis EX2155



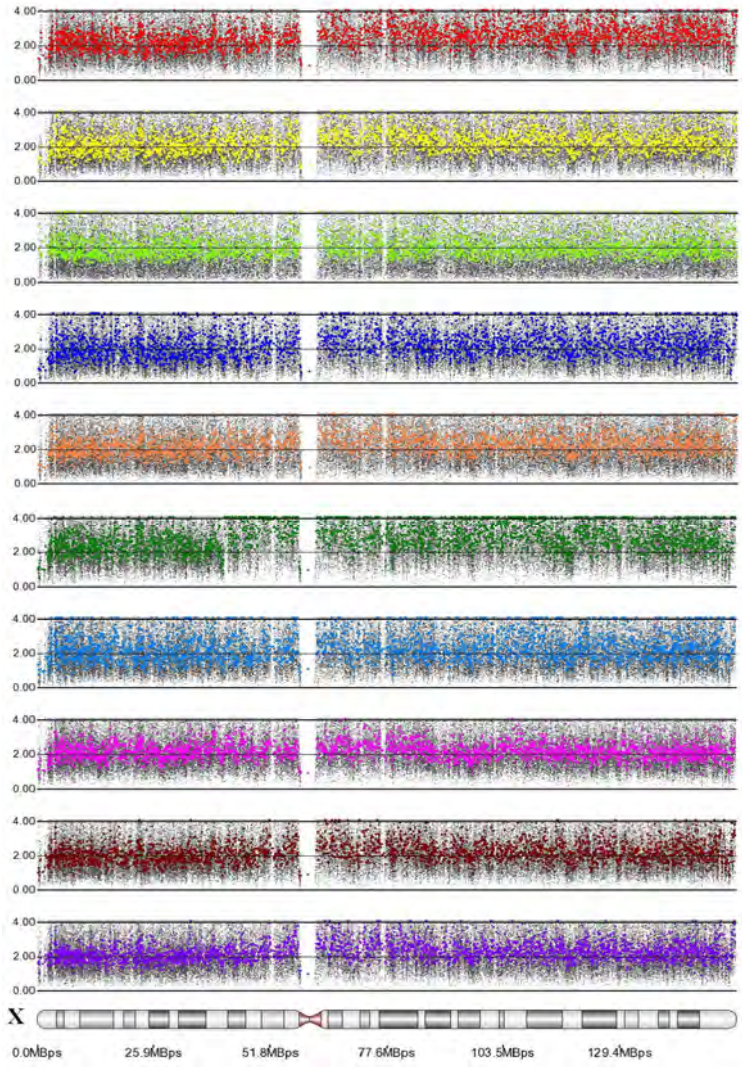
Personalis EX2155



Personalis EX2155



Personalis EX2155



8.3 Eigenvectors and eigenvalues

Common statistical measures include the mean ($\overline{X}$) and standard deviation (σ). However, variance (var), which is merely the square of the standard deviation (σ^2), and covariance (cov) can also be used. To derive the eigenvectors and eigenvalues for a set of data, the covariance matrix must first be derived.

The standard deviation and variance are represented as follows:

Personalis EX2155

$$\sigma = \sqrt{\frac{\sum_{i=i}^{n}\left(X_i - \overline{X}\right)^2}{(n-1)}}$$

$$\text{var} = \sigma^2 = \frac{\sum_{i=i}^{n}\left(X_i - \overline{X}\right)^2}{(n-1)}$$

Where:

n=Number of values in the data-set

$\overline{X}$ =Mean of the data-set

X_i=ith element of the data-set

These measure the spread of the data. However, whereas the variance is one-dimensional, covariance is used on two-dimensional data (for example, measuring the relationships between the height of a person and their body-weight, or between hours studied and exam results). Covariance retains the same formula as variance, except for a minor difference:

$$\text{cov}(XY) = \frac{\sum_{i=1}^{n}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{(n-1)}$$

Where:

n=Number of values in the data-set

$\overline{X}$ & $\overline{Y}$ =Means of data-sets X and Y

X_i & Y_i=ith elements of data-sets X and Y

Personalis EX2155

This can be a positive or negative value: positive indicates that both variables being analysed increase in tandem (for example, as hours studied increase, so do exam results); negative indicates that as one increases, the other decreases (Smith, 2002).

For n-dimensional data-sets, a covariance matrix can be constructed that represents the covariance between each possible combination of data-sets. For example, a covariance matrix for a three-dimensional data-set (X, Y, Z) would look like the following:

$$C = \begin{bmatrix} \text{cov}_{XX} & \text{cov}_{XY} & \text{cov}_{XZ} \\ \text{cov}_{YX} & \text{cov}_{YY} & \text{cov}_{YZ} \\ \text{cov}_{ZX} & \text{cov}_{ZY} & \text{cov}_{ZZ} \end{bmatrix}$$

The matrix is symmetrical across the main diagonal. Also, the covariance values on this diagonal are equivalent to simply finding the covariance of each particular data-set to itself (Smith, 2002)

8.3.1 Derive the eigenvectors and eigenvalues

If we multiplied two matrices together (one a transformation matrix and the other a vector) the result would be a matrix that is either an integer-multiple of the original vector or not. Integer-multiples are eigenvectors, with the scaling value being the eigenvalue (Smith, 2002). For example, observe the following:

$$A\begin{bmatrix} 6 & 4 \\ 3 & 7 \end{bmatrix} B\begin{bmatrix} 2 \\ 2 \end{bmatrix} = \begin{bmatrix} 20 \\ 20 \end{bmatrix} = 10\begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

Where:

A=Transformation matrix

B=Vector

10=Eigenvalue

$\begin{bmatrix} 2 \\ 2 \end{bmatrix}$=Eigenvector

Personalis EX2155

Eigenvectors are orthogonal (i.e. - perpendicular) to each other. For any nXn matrix, only n possible eigenvectors can be found, and they are scaled to have a length of 1 (Smith, 2002). Visually, the orthogonal nature of eigenvectors means that at most 3 eigenvectors can be displayed on a three-dimensional plot.

8.4 Principal component analysis

The process of reducing data using principal component analysis (PCA) involves the following steps:

1. Subtract mean
2. Build covariance matrix
3. Calculate eigenvectors and eigenvalues
4. Transform original data
5. End-analysis

(Smith, 2002)

The first step in PCA is to subtract the mean of each data-set from each data-point in the set. For example, if we had data-sets X, Y, and Z, then from each value in X, Y, and Z, we would subtract the mean of each data-set, respectively. This would result in the data-sets having means of 0. The covariance matrix is then built using the mean-subtracted data-sets[1]. Thirdly, eigenvectors and eigenvalues for the built covariance matrix are determined (see Appendix 8.3) - eigenvectors characterise the data using straight, orthogonal lines and each is scaled with an eigenvalue. These are then ordered by their eigenvalue (highest to lowest), indicating the level of significance to the data-set.

[1] If the variables were measured in different units, then a correlation matrix should be employed (Karpuzcu & Şeneş, 1987), whereas if the variables were in the same units and are mean-centered, then a covariance matrix should be employed (Chatfield & Collins, 1980 and Partek Inc., 2009).

Personalis EX2155

Transformation of the original data then occurs by multiplying a matrix containing the derived eigenvectors by one containing the mean-subtracted original data. The result is a matrix with the same dimensions as that of the mean-subtracted data but whose values have been transformed. The original data can then be said to be expressed with regard to the patterns within it (Smith, 2002). Finally, once the data has been transformed, end-analysis can be performed, which can involve the selection of significant data-points and eigenvectors - it can also involve viewing the data on bi-plots.

The transformed data contains the principal components and each is assigned a percentage that corresponds to the amount of total variance in the data towards which the component contributes. The level of correlation of the original variables to the new components is represented by numbers of no unitary value that are scaled between -1 and +1. Eigenvalues (e) and eigenvectors (a) are also assigned to the components: eigenvalues indicate the amount of variance a component contains (similar to the percentage of variance), whereas the eigenvector is the direction cosine of the axis of the principal component (Partek Inc., 2009), such that[1]:

$$\sum_{v=1}^{n} z_v = a_v x$$

Where:

n=Number of original variables

Z_v=v^{th} principal component

a_v=v^{th} eigenvector

x=vector of the original variables

As mentioned, it also holds true that the variance of a principal component is equivalent to its corresponding eigenvalue[2]:

[1] Adapted from Joliffe (2002).

[2] Adapted from Joliffe (2002).

Personalis EX2155

$$\text{var}(z_v) = e_v$$

Where:

Z_v=v^{th} principal component

e_v=v^{th} eigenvalue

If there are variables having high correlation to one or more of the derived components, then most of the variation will be accounted for by these components and such variables will be the ones that are accounting for the differences amongst samples. The last few, in such a case, will account for little variation as they will define constant or near-constant linear relationships amongst the variables (Joliffe, 2002) (i.e. - variables whose values were common across samples).

8.4.1 End analysis

8.4.1.1 Selecting principal components

There is no standard way of choosing how many components to include or exclude in end-analysis, which is probably related to the broad spectrum of analyses on which PCA is carried-out. The nature of the study is also important: if you wanted to determine the variables that defined differences between samples, then you would observe the first few principal components (or even just the first); if you wanted to determine the variables that were common across samples, then you would observe the last few. If it is the former, then choosing a certain number of components whose combined percentage of variance accumulates to a pre-determined level (generally ≥70%) is employed (Chatfield & Collins, 1980 and Joliffe, 2002). They can also be chosen using 'Kaiser's rule', which states that all principal components with an eigenvalue greater than 1 should be retained (Kaiser, 1960), or by invoking and analysing a scree plot (Cattell, 1966), which shows a line-graph of the eigenvalues of each principal component.

Personalis EX2155

8.4.1.2 Selecting variables

Choosing which variables (corresponding to SNP and CN markers in this study) to retain can pose some difficulty due to the fact that there is no standard way of selecting a subset. The ultimate aim is to find a subset that best represents the internal variation within the data (Joliffe, 2002). Thereafter, in the case of this study, the chosen subset of markers could be mapped to the genome to elucidate genes and other genomic regions that are related to a particular disease group or tissue.

Three methods for arriving at a subset of variables/markers for end-analysis are now outlined.

8.4.1.2.1 Remove and compare

First, remove a pre-determined number of markers from the full list of markers and re-create the PCA bi-plot to see if the original structure and variance is still visually/graphically represented. If the same structure is present, then remove further markers and repeat until a manageable subset of markers is present that still visually/graphically portray the variance in the data. In essence, this method involves comparing the PCs derived from subsets of markers to those of the full set, but key markers could be lost in the process.

8.4.1.2.2 High correlation

Only choose markers that have high correlations to each of the generated eigenvectors of the chosen principal components and then search for overlap between them (Joliffe, 2002). There is no standard cut-off value for the correlation coefficients, but Mahloch (1974) indicated that a value larger than 0.6 is sufficient in PCA, while a coefficient larger than 0.8 is regarded as good.

Additionally, one must distinguish between positive and negative correlation coefficients: a positive correlation coefficient indicates that as the value of a marker increases in one set of samples, its value in the other set decreases; negative correlation indicates the same except that the first marker is decreased while the other increased. It

Personalis EX2155

follows that a correlation coefficient close to 0 indicates that a marker is not significantly-contributing to variance and is common across samples (Joliffe, 2002 and Madsen, 2007).

8.4.1.2.3 Rotation of principal components

A different method of selecting which markers to include in end-analysis is to first perform an orthogonal rotation of the derived principal components. There are a number of ways to do this, including varimax, quartimax, equamax, and promax rotation. The rotation of principal components is performed to increase the accuracy of the relationship/correlation of the original markers to the newly-derived eigenvectors (principal components). As a result, this also serves to maximise the differences between each eigenvector (Mazlum *et al.*, 1999 and Meng *et al.*, 2003).

After rotation, a method according to Meng *et al.* (2003) for the selection of original variables/markers to include in end-analysis is preceded as follows (using absolute values): for the principal components that were included in end-analysis, obtain the mean correlation coefficient of each marker to these selected components. Then, for the components that were not included in end-analysis, derive the same value. If the mean of a marker for the included components was greater than that its mean for the excluded components, then that marker would be included in the end-analysis. This method is used with the aim of characterising samples with as minimum a number of variables as possible.

Personalis EX2155

8.5 Variants with roles in breast cancer - additional information

Comparison	Genes
Patients on follow-up:	
• L Vs T	**SNP (571)** *A2BP1 A2ML1 ABCA13 ACCN1 ACPP ADAM12 ADAM23 ADAMTS17 ADAMTS6 ADCY2 AFF3 AGAP1 AGBL1 AGTPBP1 AIF1L AIG1 AKAP13* *AKAP6 AKR1B10 ALCAM ALPK2 ALS2CR12 ANK2 ANKRD22 ANKRD42 ANKS1B ANO4 ANXA4 APBA2 APOB APOLD1 AQPEP ARHGAP22* *ARHGAP24 ARL15 ASTN2 ATP6V1B2 ATRNL1 AUTS2 BACH2 BAT3 BCKDHB BCOR BICD1 BMP2K BMPR1B BMX BRAF BTBD9 BUB3* *C10orf11 C13orf18 C14orf143 C15orf41 C16orf54 C1orf173 C22orf27 C2orf86 C3orf50 C4orf19 C5orf25 C6orf186 C6orf191 C6orf97 C9orf46 CA10 CA12* *CABIN1 CACNA2D1 CACNA2D3 CADM2 CAMK4 CAPN9 CASP10 CATSPERB CCDC149 CCDC3 CCDC38 CCDC39 CCNJ CD59 CDC14A CDH12 CDH13* *CDK14 CDK5RAP2 CENPF CHAT CHCHD3 CHD1L CHD9 CHN2 CIITA CLCN1 CLCN3 CLCN4 CLIC5 CNBD1 CNTN1 CNTN4 CNTN5* *CNTNAP2 CNTNAP5 COL22A1 COL24A1 COL25A1 COL28A1 COL4A5 CORIN CP CPNE4 CPS1 CPXM2 CR2 CSGALNACT1 CSMD1 CSMD2* *CSTL1 CTNNA2 CTNND1 CTNND2 DAB2 DACH2 DCAF8 DCC DCDC2 DDO DDR2 DDX12 DDX19B DDX51 DENND4C DFNA5 DGKB* *DGKI DIAPH3 DIRC3 DIS3L2 DLEU7 DLG2 DLG5 DNAH11 DNAH14 DNAH6 DNAJC1 DNM3 DOCK5 DOCK8 DOCK9 DPP10 DPP6* *DTNBP1 DUSP16 DYNC1I1 DYNC2H1 DZIP1L E2F7 EBF1 EDA EDARADD ENDOD1 ENDOU ENOX1 EPB41L4A EPHA5 EPHB1 EPHB2* *ERBB2IP ERC1 ERC2 ERG ERN1 ERP44 ESR1 EVC EXOC4 EXOC6B EXT1 EYA2 EYS F2RL2 FAHD2A FAM102B* *FAM135B FAM174A FAM178B FAM185A FAM190A FAM19A4 FAM19A5 FANCC FARP1 FARS2 FAT3 FCER1A FER FGF12 FGF14 FHIT FILIP1* *FLJ35390 FLRT2 FMN1 FMO1 FOXN3 FRAS1 FRK FUT2 FYTTD1 GABRB2 GALNT7 GALNTL6 GAS2 GBA3 GDF2 GLI3 GLIS3* *GPC6 GPR116 GPR158 GPR19 GPR64 GPR98 GPSM2 GRIA1 GRID2 GRIK1 GRIK2 GRIN2A GRM1 GRM8 GSG1 HBG2 HDAC4* *HDAC8 HDAC9 HECW2 HEPH HFM1 HRNBP3 HS3ST4 HSD17B12 HSD17B4 ICOS IGF1R IGF2BP2 IGSF1 IGSF6 IKBKAP IL18R1* *IL1RAPL1 IL1RL2 IL1RN IL22RA2 IQGAP2 ITGB6 ITPR1 ITPR2 JAM3 KANK1 KBTBD11 KCNAB1 KCNB2 KCND2 KCND3 KCNH1 KCNH5* *KCNH7 KCNIP4 KCNK10 KCNN3 KCNQ5 KDM4C KIAA0922 KIAA1009 KIAA1267 KIAA1486 KIAA1797 KIAA2022 KIF21A KIRREL KLHL32 KNG1 KRR1* *KRT6B L3MBTL4 LAMA2 LANCL3 LBP LDB3 LDLR LEMD3 LGR5 LHCGR LOC100132247 LOC202181 LOC283867 LOC613037 LOC647946* *LOC91948 LPHN3 LPP LPPR1 LRBA LRFN5 LRMP LRP1B LRP2 LRRC1 LRRC16A LRRFIP1 LTBP1 MAMDC2 MAML3 MAMLD1* *MAP3K4 MAP7D2 MAPK4 MAPK9 MCM10 MDGA2 MECOM MED6 MEF2C MEGF10 METTL9 MGAT4C MMP16 MOBKL2B MPPED2 MREG MRPS22* *MRPS28 MSI2 MTUS2 MYH9 MYLK4 MYO10 MYO1D MYOF MYRIP MYST4 MYT1L NBEAL2 NCAM2 NCK2 NCOA7 NDST1 NDST3* *NEDD9 NEK11 NELL1 NELL2 NFS1 NHSL1 NKAIN2 NMUR2 NOTCH2NL NPAS3 NPSR1 NR3C2 NRG1 NRG2 NRG3 NRXN3 NSMAF* *NTRK3 NUAK1 NUDT7 ODZ2 ODZ3 ODZ4 OPCML ORC4L OSBPL10 OTUD7A OVGP1 PAK3 PALLD PAMR1 PARVB PAX7 PAX8* *PDE4B PDE4DIP PDGFRA PDSS2 PFKFB1 PGM5 PHLDB2 PID1 PION PKD2L1 PLCB1 PLCE1 PLCXD2 PLD5 PLEKHG1 PLXNA4* *PPP1R1C PRDM5 PRKCA PRKCQ PRKG2 PRUNE2 PSD3 PTEN PTN PTPRD PTPRG PTPRK PTPRM PTPRO PTPRR PTPRT PWRN1* *PWRN2 PXDN RAB30 RAB3GAP2 RABGAP1L RAD51L1 RALGAPA2 RAPGEF4 RAPGEF5 RARB RB1 RBM20 RBP1 RELN RFC3 RGSL1 RNF220* *RNF8 RPH3A RPRD1A RPTOR RTKN2 RUNDC3B RYBP RYR1 RYR3 S100A4 SAMD12 SCARA3 SCFD1 SCFD2 SCHIP1 SCNN1A SCTR* *SDK1 SEL1L3 SEMA3C SEMA5A SERPINA12 SFI1 SGIP1 SGMS1 SH2D4A SH3KBP1 SH3RF2 SH3TC1 SH3TC2 SHANK2 SHISA9 SKP1* *SLC15A1 SLC18A1 SLC22A10 SLC24A4 SLC28A3 SLC35F3 SLC36A2 SLC41A2 SLC4A8 SLC6A11 SLC9A4 SLC9A9 SLCO1B3 SMAD3 SMC3 SMOC1 SNX29* *SORBS2 SORCS1 SOX5 SPATA13 SPATC1 SPDYE5 SPHKAP SPTBN1 SRBD1 SRFBP1 SRPK2 ST3GAL1 ST8SIA5 STEAP2 STK10 STK32B STOX1* *STX11 STX3 STX8 SYTL2 TBC1D22B TCF12 TCP11L2 TEAD1 TFP1 THOC2 THSD7B TIPARP TMCO4 TMEM132B TMEM132C TMEM132D* *TMEM164 TMEM217 TMLHE TNFSF4 TNN TNR TPD52L2 TRIM16 TRIM16L TRIM21 TRIO TRPC4 TRPM3 TSHZ3 TSPAN7 TSSC1 TTC27* *TTC37 TTK TULP4 TUSC3 TXNDC3 UBR5 UCP1 UGT2B10 UGT8 ULK4 UNC5C UNC80 USP11 USP13 USP49 USP54 USP6NL* *UTRN VIM VPS13A VWC2 VWC2L WBSCR17 WDHD1 WDSUB1 WWOX XIAP ZBBX ZC4H2 ZCCHC16 ZDHHC5 ZFAND3 ZNF185* *ZNF385B ZNF385D ZNF595 ZNF718 ZZEF1* **CN (190)** *ADAMTS3 AGBL4 AKAP2 AMOTL1 ANAPC10 ANGPTL2 ARHGEF3 ATP13A4 ATP2B2 ATRNL1 BDNF BDNFOS BMP7 BTBD9 C10orf47 C12orf23* *C14orf143 C3orf59 CA10 CACNA1S CAMTA1 CAMTA2 CCBE1 CDKAL1 CELF2 CEP110 CEP170 CEP350 CHN2 CHRDL1 CLCN5 CMTM7 CNTN6* *CNTNAP2 COL22A1 COQ9 CPA4 CSMD1 CTNND2 CXXC4 DAAM1 DACH2 DLG2 DMD DNAJC7 DNM3 DOCK8 DPP4 DYNC1I1 EIF2C3* *EPB41L5 ESR1 ESR2 EXOC4 EYS FAM135B FAM155B FAM177A1 FAT3 FBXO34 FHIT FHOD3 FILIP1 FMN2 FRMD5 FRMPD4* *GABRA3 GABRB1 GLP2R GORASP1 GPC3 GRM5 GRM8 GYPE HDX HEATR7B2 HIPK2 HIST2H2BF HPSE2 IQGAP1 ITPKB KAZ KCNH7* *KCNQ1 KDM4C KIAA0753 KIAA1210 KIF5C KLC4 KRT35 LARGE LDB2 LGMN LILRA6 LINGO2 LUZP2 MAST4 MDGA2 MGAM MLL* *MPHOSPH9 MYO16 MYO18A NAV2 NCAM1 NHSL2 NRXN3 OPHN1 OSBPL10 PAG1 PALM2 PARP15 PCCA PCDHGA1 PCDHGA10 PCDHGA11* *PCDHGA12 PCDHGA2 PCDHGA3 PCDHGA4 PCDHGA5 PCDHGA6 PCDHGA7 PCDHGA8 PCDHGA9 PCDHGB1 PCDHGB2 PCDHGB3 PCDHGB5 PCDHGB6 PCDHGB7 PCDHGC3* *PCDHGC4 PCDHGC5 PDE3B PDE7B PDHX PITPNC1 PLCXD3 PPFIA2 PTPRA RABL3 RAD51L1 RALGPS1 RAPGEF4 RICH2 ROBO1 RTN4 SASH3* *SCFD2 SDCBP SELP SETBP1 SHROOM3 SLC16A12 SLC16A2 SLCO6A1 SND1 SNTB2 SORBS2 SPHKAP SPTB STRADB SYT15 TAF1 TASP1* *TBC1D15 TBL1X TBX22 TESK2 TEX9 TG THEM4 TM4SF4 TMC2 TMEM132D TRIO TRPC7 TYW1 TYW1B UTRN XIRP2* *ZCCHC11 ZCCHC16 ZFAND3 ZGPAT ZNF208 ZNF341 ZNF782 ZNF804B*
• T1N0 (T≤20mm) P1	**SNP (12)** *ALK CACNA2D3 CDH7 ENOX1 FLT3 GRID2 IL6ST LRIG1 NLRP5 TMCC3 UGGT2 VGLL3*

Personalis EX2155

Vs T>20mm/node-positive P1	**CN (10)** *DGCR8* *FAM90A10* *FAM90A13* *FAM90A5* *FAM90A7* *FAM90A8* *FAM90A9* *LOC100132396* *ZNF705A* *ZNF705G*
• T1N0 (T≤20mm) P2 Vs T>20mm/node-positive P2	**SNP (28)** *ADAMTSL1* *APPL2* *ARPP21* *BCAS3* *BMP2K* *CACNA2D3* *CHRNA7* *CTNNA3* *CYP19A1* *DNAH8* *ENOX1* *EPB41L4A* *GALNTL6* *GARNL3* *GSG1* *HEPH* *KIFAP3* *LRRC4C* *NIPAL2* *NRG1* *PID1* *PTPRG* *RELN* *RGL1* *TPTE2P1* *TRIO* *ZBTB20* *ZNF385B* **CN (10)** *CNDP1* *GALNTL6* *GPR149* *LHX1* *MGAT4C* *NRG1* *PKHD1* *SLC39A11* *WARS2* *ZHX2*
Relapsed patients on follow-up:	
• P1 Vs P2	**SNP (24)** *ABCA4* *ACOX1* *BTBD9* *C20orf46* *CA8* *CSMD3* *CTNNA2* *DCX* *DGKB* *EIF2AK4* *KCTD16* *KIF6* *MBTPS1* *MCC* *MGMT* *NPAS3* *NTM* *PANK1* *PRR12* *PTN* *PTPRD* *RYR3* *SORCS1* *TNIP1* **CN (23)** *ABCC4* *ASTN2* *BTBD9* *C9orf135* *CD36* *CLIC5* *CTNNA2* *GPR39* *GTF3A* *IGF2R* *LOC100132891* *MAML2* *MAPK10* *MIPOL1* *NOP14* *PARK2* *RNASE11* *SLC7A14* *SLC9A9* *TACR3* *UBP1* *ULK4* *UVRAG*
• T Vs P2	**SNP (348)** *A2BP1* *ABCB1* *ACCN1* *ADAMTSL1* *ADIPOQ* *AFF1* *AKT3* *ALK* *ANKRD20A3* *ANKRD20B* *APBA2* *ARHGAP15* *ARHGAP22* *ARHGAP25* *ARHGAP6* *ARHGEF12* *ARHGEF6* *ASB12* *ASTN2* *ATP8A2* *B3GALT5* *BACH2* *BAGE2* *BBS9* *BCAS1* *BCLAF1* *BICC1* *BTBD9* *BTC* *C14orf143* *C1orf161* *C6orf103* *C7orf23* *C9* *C9orf5* *CALB1* *CAMK1G* *CCDC102B* *CCDC146* *CCND3* *CD1C* *CD209* *CDC14A* *CDC5L* *CDH13* *CHCHD6* *CHD1L* *CHRM2* *CHRM3* *CLCN1* *CLRN1* *CNNM2* *CNTNAP3* *COL14A1* *COL4A2* *COQ7* *CPNE5* *CPXM2* *CRISPLD1* *CSMD1* *CSTL1* *CTNNA2* *CTNNA3* *CTNND2* *CUBN* *CXorf38* *DARS* *DBC1* *DDR2* *DDX19B* *DDX3X* *DEC1* *DEFB135* *DGKH* *DGKI* *DIAPH2* *DNAH5* *DNAH6* *DNAH8* *DNER* *DOCK2* *DPT* *DPYD* *DPYSL5* *DRD2* *DYNC2H1* *EBF1* *EBF2* *ENOX2* *EPB41L2* *ERBB2IP* *ERC2* *EXOC4* *F2RL2* *F5* *F8* *FAM174A* *FAM189A1* *FAM190A* *FAM196B* *FAM19A4* *FAM38B* *FGGY* *FHIT* *FILIP1* *FLJ12825* *FMN2* *FNIP2* *FOXN3* *FRMD4A* *FRMD6* *FSIP1* *FSTL5* *GAGE12B* *GAGE12C* *GAGE12D* *GAGE12E* *GAGE12G* *GAGE12H* *GAGE12I* *GAGE12J* *GAGE13* *GAGE2A* *GAGE2B* *GAGE2C* *GAGE2D* *GAGE2E* *GAGE4* *GAGE5* *GAGE6* *GAGE7* *GAGE8* *GAN* *GFI1B* *GLG1* *GNL3L* *GOLGA6L9* *GPR26* *GRIA1* *GRIK1* *GRIK3* *GRIN2A* *GRIP1* *GRM1* *GRM7* *GSK3B* *HADHA* *HECW2* *HERC4* *HHAT* *HNMT* *HS3ST4* *HSDL2* *HSP90AA1* *HUNK* *HYDIN* *IGSF1* *IKBKAP* *IL1RN* *IMMP1L* *IQGAP2* *ITGA10* *ITGB8* *ITPR1* *JAK1* *JAM3* *JPH1* *KALRN* *KCNH1* *KCNH5* *KCNN2* *KEL* *KHDRBS2* *KIAA1244* *KIAA1267* *KIRREL* *KIT* *KLHL18* *KRT6B* *LCE2C* *LIFR* *LINGO2* *LMO2* *LMO7* *LOC100170939* *LOC441666* *LOC442459* *LOC595101* *LOC728989* *LRBA* *LRFN2* *LRFN5* *LRMP* *LRRC59* *LRRTM3* *MACROD2* *MAGI1* *MAML2* *MAML3* *MAP1B* *MARCH1* *MAST4* *MBTD1* *MECOM* *MED6* *MTMR8* *MYBPC1* *MYO1D* *MYO5B* *NAV2* *NCALD* *NCKAP5* *NEGR1* *NELL1* *NHS* *NLRP11* *NLRP14* *NOTCH2NL* *NR3C2* *NRCAM* *NRG1* *NRG3* *NTNG1* *NUAK1* *NUP50* *OAS2* *ODZ1* *ODZ3* *ODZ4* *OPCML* *OSBPL10* *OSBPL1A* *PALLD* *PAPPA* *PARD3B* *PARP8* *PARVA* *PCDH19* *PDE11A* *PDE4DIP* *PDSS2* *PDXK* *PEX5L* *PGM5* *PGPEP1L* *PHACTR1* *PKD2L1* *PLA2R1* *PLCXD3* *PLXNA4* *PLXNC1* *POLH* *PPM1H* *PPP1R14C* *PPP4R4* *PRAMEF11* *PRAMEF15* *PRAMEF4* *PRAMEF5* *PRAMEF6* *PRAMEF9* *PREX2* *PRIM2* *PRKG1* *PRKRIP1* *PRRX1* *PSG8* *PTPRD* *PTPRK* *PTPRR* *QKI* *RABGAP1L* *RALGPS1* *RAP1B* *RBM28* *RECQL5* *REG4* *RFX4* *RGAG1* *RGL1* *RIT2* *RNF220* *RPS6KA5* *RUNDC3B* *RYBP* *SAMD12* *SCFD2* *SELP* *SESTD1* *SGCZ* *SGMS1* *SHROOM4* *SKP1* *SLC2A13* *SLC35F4* *SLC4A4* *SLC6A11* *SLC8A3* *SLC9A9* *SLN* *SMG1* *SMYD3* *SNX29* *SORCS3* *SOX5* *SPATA13* *SPDYE5* *SPIN1* *SPOCK1* *SPTLC2* *ST18* *ST8SIA1* *STK32B* *STK38L* *STK39* *SUMF1* *TADA1* *TBL1X* *TBX15* *TCP11L2* *TDRD5* *TEKT3* *TG* *THADA* *THSD7B* *TMEM132C* *TMEM132D* *TMEM84* *TPRG1* *TPX2* *TRPM8* *TTLL1* *TTLL4* *TXNRD3* *USH2A* *USP49* *VPS41* *VSNL1* *WDHD1* *WDSUB1* *WWC2* *WWOX* *ZBTB16* *ZNF185* *ZNF426* *ZNF462* *ZNF761* *ZNF775* **CN (96)** *AFF3* *AKAP6* *AKAP9* *ALK* *ANGEL2* *ARHGAP12* *ATP10B* *BAGE2* *BBX* *BCAS1* *BICC1* *C1orf129* *C3orf50* *CA5BP* *CBWD1* *CBWD2* *CBWD3* *CBWD5* *CBWD6* *CD300E* *CEP170* *CLSTN2* *COL17A1* *COL4A2* *CSMD2* *CTNNA2* *DACH2* *DCX* *DDI2* *DLAT* *DMD* *DPP6* *DZIP3* *EXOC4* *FCHO2* *FGF13* *FRMD3* *FSIP1* *GABRG3* *GALNTL2* *GPC5* *GPR158* *GRIN3A* *GTF3A* *GYPB* *HDX* *IQCG* *KCNIP1* *KCNT2* *KIAA0564* *LHCGR* *LIFR* *LOC139201* *MACROD2* *MANIA2* *MED12* *MED12L* *MSRB3* *NCAM1* *NOS1* *NR1I3* *NRG3* *OTUD3* *OXCT1* *PDZRN3* *PI4KA* *PIGK* *PITPNC1* *PRIM2* *PRKCB* *PTPRO* *REPS2* *RHOG* *SCFD2* *SCML2* *SHROOM2* *SLC16A2* *SLC8A2* *SPAG16* *SPDYE5* *TBL1X* *TEKT5* *THEM4* *TMEM164* *TMEM195* *TMEM59* *TP63* *TSHZ2* *UQCC* *USH2A* *UVRAG* *WBSCR17* *ZFAT* *ZNF419* *ZNF521* *ZSCAN4*

Personalis EX2155

• T1N0 (T≤20mm): Non-relapsed P2 Vs Relapsed P2	**SNP (115)** *ABCA7 ABCB7 ACTN2 ADAMTSL1 ADRBK2 AHI1 AIM1 ATP10A ATP1B1 BACE2 C11orf45 C18orf1 CA5B CADPS2 CD9 CDH13 CIT* *CNTN5 CORO2A CSMD2 CTNNA2 DAPK1 DFNA5 DLG2 DMD DPP10 DYNC2H1 EIF4G3 EPS15L1 ERGIC1 EXOC6 FAM110B FAM66B* *FAM66C FAM66D FAM66E FER FER1L6 FHIT GLRA4 GMPPA GPC5 GRM7 HDAC9 INPP4B KCNJ5 KCNK1 KCNQ3 KCNT1 KDM5B* *KIRREL3 KLHL29 LTBP1 LTBP2 MAD1L1 MAMDC2 MAN2A1 MAPK4 MCPH1 MEFV MMP16 MTUS1 MYH7B MYO1D MYO5B NAV2 NPAS3* *NR6A1 NSMCE2 NTRK2 NUBP1 PAPD7 PAX7 PCDH15 PCDHA1 PCDHA3 PCDHA4 PCDHA6 PCDHA7 PDE4D PDLIM1 PHACTR3 PTPRD* *RABGAP1L RAI14 RALA RALGAPA2 RASGRF1 RGS6 RIN2 RNF157 RORA RRN3P2 RSPH1 SCFD2 SEZ6L SLC35A4 SLC39A11 SLC44A1 SLIT3* *SSPO ST5 ST6GALNAC4 SUPT3H SWAP70 SYN3 SYT6 TEAD1 TMEM63C TNR UBL4A USP34 UTRN XRCC3 ZC3H18* **CN (49)** *ABL1 ABL2 AGAP1 ANAPC16 ANKS1B ARRB1 BICD1 BOKAS BRCA1 BTBD1 C1orf25 C4orf21 C5orf36 CARD14 CD2AP CDH22 COG4* *CTDSPL2 EPHB6 ERC2 ETS2 FKBP9 GABRR2 GYS2 IL17D INPP5B KIF27 LOC100101938 LOC400891 MAGI2 MAPK4 MGMT* *MTHFD2L NINJ1 NRCAM PDGFRB RBL1 RHPN2 RIT2 SGSM1 SLC22A25 SMAD4 SNX9 SPHKAP TARBP1 TARP TEC TRIM50 ZNF484*
• T>20mm/node-positive: Non-relapsed P2 Vs Relapsed P2	**SNP (40)** *ADA AFF3 ARFIP1 C10orf119 C12orf42 C18orf1 CAPN6 DCC EDDM3A ENOX1 FAM19A4 FAM98A FRMD4A GARNL3 HMSD INTS7* *KCNK13 KIAA0355 KIAA1211 KIF5C KIRREL3 KLHL18 LPCAT2 MAST4 MBD4 MTMR1 OCA2 PARVB PRKCB RANBP17 RBM33 REPS1 RIT2* *SORCS1 TMC2 TMEM232 UVRAG XPA ZEB2 ZNRF1* **CN (22)** *ACMSD BIRC6 CASS4 CLSTN2 CYP4Z1 EPB41L4A EPHB1 GARNL3 GPR112 HNF4A IL1RAPL2 KLHL18 LAMA3 PLD5 PTCH1 REPS1 RGS5* *ROCK1 S100A7 S100A7A SDK2 TSSC1*
Healthy controls: • P Vs Unaffected *BRCA1* carriers P	**SNP (26)** *CACNA1A CAND1 CCNY DIRC3 DLC1 EGFR FAM98A GDPD4 GLRA2 MAGI2 NHS OPHN1 PALLD RGS6 ROPN1 RYR3 SDC2* *SHROOM3 SMOC1 SRPK2 TG TSPAN7 WWOX ZMAT1* **CN (12)** *ADAMTS3 COX4NB DISP1 DMD F8 GABRA3 GRIK2 GUCY2F KIAA1210 KIAA2022 KLHL20 RIT2*
Pre-surgical (NeoCEnT) patients: • L Vs T	**SNP (228)** *A2BP1 ABCA12 ABCB5 ABCC1 ACBD6 ACOXL ADAM12 ADAM32 ADAMTS12 ADAMTS2 ADAMTSL1 ADAMTSL4 ADD2 ADRA1A AIF1L ANKRD32 APP* *ATG4B AUTS2 BAT2L1 BCAS3 BNC2 C13orf39 C14orf143 C1orf105 C21orf131 C21orf29 C21orf34 C2orf34 C3orf67 C4orf21 CABLES1 CCDC13 CCL14* *CCL15 CD3EAP CDH11 CDH23 CDH4 CDKN2A CEACAM8 CGRRF1 CLIC5 CNTN6 COL13A1 COL24A1 CPPED1 CSMD2 CTDSP2 CTNNA2 DAB1* *DHX33 DIAPH2 DLC1 DLG1 DMRT1 DNAH14 DNAJA1 DNM1L DPF3 DUSP19 EDIL3 EGFLAM ENKUR EPHB1 ERBB4 ERCC1 ERICH1* *FAM184A FAM19A2 FAM83D FANCI FARS2 FBLN2 FGF13 FLG FLJ43663 FLT4 FOXN3 FRMD4B GABRB3 GLI3 GPC6 GRHL2 GRIA3* *GTDC1 GUCY1A2 HDDC2 HERC2 HMCN1 HPS5 HS3ST5 IGSF21 IL1RAPL1 IL1RAPL2 IMMP2L IMPA2 IPCEF1 IQCH ITGB8 ITIH2 ITPR1* *IYD JAKMIP1 KCNIP1 KCNT2 KIF26B KIRREL3 LAD1 LAMC1 LASS6 LIFR LIMD1 LOC116437 LRP1B LRRC8D LZTFL1 MACROD2 MADD* *MAP2K5 MAP2K6 MAST4 MCC MCF2L2 MCTP1 MEGF11 MID1 MLPH MMP26 MSH4 MSI2 MSRA MYO10 MYT1L NCAM1 NCOA1* *NEDD4L NELL2 NFATC2 NR5A2 NRK NSD1 NSMCE2 OXNAD1 PALLD PARK2 PCDHB7 PCYT1B PDE4D PDZRN3 PHF12 PIP4K2A* *PLA2G4A PLCB1 PLD1 PLS3 PLXDC2 PLXNA4 POGZ POLA1 PPARA PPIL3 PPME1 PPP1R16B PRKCB PRKCH PSD3 PTH1R PTH2R* *PTN PTPN14 PTPRK PTPRM RBMS3 RERE RERG RFC2 RGNEF RGSL1 RIMS1 RIMS2 RTN4 RYBP RYR1 SCAPER SDC1* *SDK1 SIDT1 SIGLEC12 SIK3 SLC12A8 SLC36A1 SLC43A2 SMC2 SNX30 SPAG6 SPATA16 SPTLC1 STARD9 STX8 SUSD1 TAF1B TAOK1* *TBC1D5 TBL1X TIPARP TMEM117 TMEM132C TMEM39B TNKS TRIO TRPM3 TRPM6 TTC38 TUB U2AF1 USP7 UTP6 VAV2 VGLL4* *VIM WAC WARS2 WWOX XKR6 ZCCHC16 ZFAND3 ZNF385D* **CN (32)** *AAK1 AGK APBA1 CYP2C18 DARS DCHS1 DPP10 DUSP19 FBLN2 FBXW11 FCER2 FHIT GLCE GPR114 GRAMD3 HELZ HMP19* *HNMT KYNU LRRFIP1 MRPS22 MSH4 MTG1 OBFC1 OS9 RGS12 SCHIP1 SGCZ SNTG2 SNX13 TMEM144 TTBK2*

Personalis EX2155

• Baseline P Vs 6 week P	**SNP (3)**	*ALOX5AP*	*GCOM1*	*MCF2*														
	CN (4)	*ALOX5AP*	*GCOM1*	*MCF2*	*RBM27*													
• Baseline P Vs 21 week P	**SNP (12)**	*BNC2*	*CTNNA2*	*DMD*	*DOCK8*	*FHIT*	*KCNAB1*	*NELL1*	*PKHD1*	*PLEKHG1*	*PTPRD*	*TGFBR2*	*ZZEF1*					
	CN (13)	*BNC2*	*CTNNA2*	*DMD*	*DTNB*	*FAM3C*	*NKAIN2*	*PARD3B*	*PCDH9*	*PTPRD*	*SPATA13*	*SPEF2*	*TACR3*	*ZZEF1*				
• T Vs 21 week P	**SNP (142)**	*ACAN*	*ACSS1*	*ADARB2*	*ADK*	*AIF1L*	*AKAP6*	*ALK*	*ANGPT2*	*ANK1*	*ANKS1B*	*ANO1*	*AQP9*	*ARHGEF11*	*ASXL3*	*BCAS3*	*C20orf185*	*C2CD2*
		C2orf86	*C6orf203*	*C7orf50*	*CADM3*	*CASR*	*CCDC46*	*CCDC68*	*CDC14A*	*CDH12*	*CDH18*	*CDYL2*	*CEACAM5*	*CHRNA9*	*CNTNAP2*	*COL18A1*	*CTBP1*	*DCLK1*
		DDX3Y	*DHX29*	*DNAH12*	*DOCK3*	*DOK5*	*DPYSL2*	*EBF3*	*EML6*	*EMP2*	*ERBB4*	*ERLIN1*	*FAM114A1*	*FAM190A*	*FAM36A*	*FARP1*	*FARS2*	*GFRAL*
		GTDC1	*GTPBP5*	*HDAC9*	*HECW1*	*HERC1*	*HIVEP2*	*HNF4G*	*HP1BP3*	*IL1RAPL1*	*ITGA8*	*JAKMIP1*	*JAKMIP3*	*KCNE4*	*KCNH4*	*KCNJ12*	*KDM4C*	*KDM5B*
		KIFC3	*KSR2*	*LAS1L*	*LPP*	*LRRC16A*	*LRRFIP1*	*MACROD2*	*MAGI2*	*MAOA*	*MARK1*	*MCF2*	*MCPH1*	*MPP2*	*MTMR3*	*MYBPHL*	*MYO18B*	*NID1*
		NISCH	*NOS3*	*NSMCE2*	*NTN4*	*ODZ4*	*PARD3B*	*PIAS4*	*PLA2G3*	*PLA2R1*	*PLCH1*	*PLEKHG4B*	*PLEKHH2*	*POLR1D*	*PPP1R12A*	*PRKACB*	*PRKCA*	*PRKCH*
		PRKG1	*PTPRT*	*PYGL*	*RAD51L1*	*RASGRF2*	*RCL1*	*RER1*	*RIMS1*	*RNF128*	*RNF152*	*ROBO1*	*RORA*	*RYR2*	*SCAPER*	*SEMA4F*	*SLC2A13*	
		SLC37A2	*SORCS2*	*SORCS3*	*SPOCK1*	*SPTBN1*	*ST18*	*ST6GALNAC3*	*STK17B*	*TBC1D22A*	*TGFA*	*TLL1*	*TMEFF2*	*TMEM117*	*TNRC18*	*TSPO*	*TTC39B*	*UBR3*
		USP9Y	*UTY*	*UVRAG*	*WNK2*	*WWTR1*	*ZMAT1*	*ZNF775*										
	CN (15)	*BAHCC1*	*COL24A1*	*DNAH9*	*DYSF*	*FRAS1*	*GLCE*	*IMMP2L*	*MED12L*	*MYST3*	*PGCP*	*PPPDE1*	*RYR2*	*TNIK*	*TTN*	*ZNF395*		

Table 8-2 - Complete listing of genes harbouring significantly-differentiated markers (SNP or CN) for post-PCA comparisons carried out in Table 5-2. Number of genes for each comparison in parentheses. L, lymphocyte; P, healthy control or pre-surgical plasma cfDNA; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour.

Personalis EX2155

Comparison	SNP					Copy number					
	Gene	Probe	Chromosome	Position (bp)	Cytoband	Gene	Probe	Chromosome	Start (bp)	End (bp)	Cytoband
Non-relapsed patients on follow-up:											
• L Vs T	*BRAF*	SNP_A-2243994	7	140505853	q34	*ESR1*	CN_1174833	6	152170874	152170899	q25.1
	CASP10	SNP_A-1876453	2	202061050	q33.1	*DNAJC7**	CN_750041	17	40148320	40148345	q21.2
	DCC	SNP_A-8312888	18	50181120	q21.2	*GPC3*	CN_938512	X	133107737	133107762	q26.2
	*EGFR**	SNP_A-1894491	7	55268916	p11.2						
	EPHB2	SNP_A-8401568	1	23060250	p36.12						
	ESR1	SNP_A-2241225	6	152082646	q25.1						
	*FYTTD1**	SNP_A-8607393	3	197495334	q29						
	GRIK1	SNP_A-4263006	21	30888877	q21.3						
	IL1RN	SNP_A-2272677	2	113877495	q13						
	IL1RN	SNP_A-2175766	2	113877413	q13						
	LHCGR	SNP_A-4286786	2	48937154	p16.3						
	OPCML	SNP_A-8679702	11	132290674	q25						
	OPCML	SNP_A-8644627	11	132492654	q25						
	PALLD	SNP_A-2230481	4	169784122	q32.3						
	*PDGFRA**	SNP_A-205677	4	55162271	q12						
	PRKCA	SNP_A-8328793	17	64535879	q24.2						
	PTEN	SNP_A-2003660	10	89686509	q23.31						
	RB1	SNP_A-8332880	13	48942814	q14.2						
	TNFSF4	SNP_A-2157403	1	173141691	q25.1						
	TUSC3	SNP_A-4266816	8	15423450	p22						
	WWOX	SNP_A-8380063	16	78707178	q23.1						
	WWOX	SNP_A-2123845	16	78717672	q23.1						
	WWOX	SNP_A-8303843	16	79199218	q23.1						
• T1N0 (T≤20mm) P1 Vs T>20mm/node-positive P1	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
• T1N0 (T≤20mm) P2 Vs T>20mm/node-positive P2	*BCAS3*	SNP_A-2058004	17	59226527	q23.2	*PKHD1*	CN_1177822	6	51756420	51756445	p12.3
Relapsed patients on follow-up:											
• P1 Vs P2	*MCC*	SNP_A-4211179	5	112579419	q22.2	*ABCC4*	CN_635710	13	95770808	95770833	q32.1
						IGF2R	CN_1186010	6	160465104	160465129	q25.3
						*MAML2**	CN_560357	11	95712668	95712693	q21
						PARK2	CN_1161842	6	162256120	162256145	q26

Personalis EX2155

• T Vs P2	*CD209**	SNP_A-8583612	19	7807550	p13.2	*TP63*	CN_1007723	3	189401640	189401665	q28
	DBC1	SNP_A-2313075	9	122067787	q33.1						
	DEC1	SNP_A-1797781	9	117966943	q33.1						
	GRIK1	SNP_A-8339473	21	31230702	q21.3						
	GLG1	SNP_A-2149277	16	74574830	q23.1						
	IL1RN	SNP_A-2272677	2	113877495	q13						
	KIT	SNP_A-4276826	4	55553046	q12						
	OPCML	SNP_A-8412788	11	132304552	q25						
	OPCML	SNP_A-8508943	11	133389115	q25						
	PALLD	SNP_A-2277928	4	169530258	q32.3						
	MAML2	SNP_A-8338242	11	95771210	q21						
	*TTLL1**	SNP_A-8706255	22	43465814	q13.2						
	WWOX	SNP_A-4244914	16	78417710	q23.1						
	WWOX	SNP_A-2131861	16	78679045	q23.1						
• T_1N_0: Non-relapsed P2 Vs Relapsed P2	*XRCC3*	SNP_A-8703759	14	104172253	q32.33	*ABL1**	CN_1306000	9	133761278	133761303	q34.12
						BRCA1	CN_756593	17	41248293	41248318	q21.31
						PDGFRB	CN_1093773	5	149521022	149521047	q32
						RHPN2	CN_787341	19	33500261	33500286	q13.11
						SMAD4	CN_768778	18	48572583	48572608	q21.2
• T>20mm/node-positive: Non-relapsed P2 Vs Relapsed P2	NA	NA	NA	NA	NA	*PTCH1*	CN_1332424	9	98241762	98241787	q22.32
Healthy controls:											
• P Vs Unaffected *BRCA1* carriers P	*DIRC3*	SNP_A-4208690	2	218294420	q35	NA	NA	NA	NA	NA	NA
	DIRC3	SNP_A-2100871	2	218301532	q35						
	DLC1	SNP_A-4258553	8	13280398	p22						
	*EGFR**	SNP_A-1894491	7	55268916	p11.2						
	PALLD	SNP_A-8304889	4	169585953	q32.3						
	WWOX	SNP_A-8400779	16	78499157	q23.1						
	WWOX	SNP_A-8389764	16	78500945	q23.1						
Pre-surgical patients on neoadjuvant therapy:											
• L Vs T	*BCAS3*	SNP_A-4298084	17	59362798	q23.2	NA	NA	NA	NA	NA	NA
	DLC1	SNP_A-2198754	8	13005559	p22						
	DLC1	SNP_A-8330309	8	13185233	p22						
	MCC	SNP_A-4275195	5	112549224	q22.2						

Personalis EX2155

	PALLD	SNP_A-2252644	4	169558119	q32.3							
	PALLD	SNP_A-8325894	4	169593407	q32.3							
	PARK2	SNP_A-1843669	6	162779001	q26							
	PARK2	SNP_A-8545610	6	162981381	q26							
	WWOX	SNP_A-1816212	16	78406523	q23.1							
• Baseline P Vs 6 week P	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
• Baseline P Vs 21 week P	*PKHD1*	SNP_A-1986034	6	51514608	p12.3	NA	NA	NA	NA	NA	NA	NA
	TGFBR2	SNP_A-1793596	3	30669358	p24.1							
• T Vs 21 week P	*BCAS3*	SNP_A-4201037	17	59070563	q23.2	NA	NA	NA	NA	NA	NA	NA
	PRKCA	SNP_A-8382030	17	64454921	q24.2							

Table 8-3 - Overlapping gene, marker ID, chromosome, genomic position, and cytoband for each significant marker outlined in Table 5-2. NB - asterisk (*) denotes SNP/copy number probe maps to an exon. L, lymphocyte; P, healthy control or pre-surgical plasma cfDNA; P1 and P2, first and second plasma cfDNA of patients on follow-up, respectively; T, tumour. NB - asterisk (*) denotes SNP/CN marker mapping to an exon; genes listed alphabetically; 'NA' denotes 'not applicable' due to no significant markers using the selected keywords.

Personalis EX2155

8.6 References

Aalfs CM, Hoovers JM, Nieste-Otter MA, Mannens MM, Hennekam RC, and Leschot NJ (December, 1995), *Further delineation of the partial proximal trisomy 10q syndrome*, Journal of Medical Genetics 32(12): 968-971

Abulí A, Bessa X, González JR, Ruiz-Ponte C, Cáceres A, Muñoz J, Gonzalo V, Balaguer F, Fernández-Rozadilla C, González D, de Castro L, Clofent J, Bujanda L, Cubiella J, Reñé JM, Morillas JD, Lanas A, Rigau J, García AM, Latorre M, Saló J, Fernández Bañares F, Argüello L, Peña E, Vilella A, Riestra S, Carreño R, Paya A, Alenda C, Xicola RM, Doyle BJ, Jover R, Llor X, Carracedo A, Castells A, Castellví-Bel S, Andreu M, and Gastrointestinal Oncology Group of the Spanish Gastroenterological Association (September, 2010), *Susceptibility genetic variants associated with colorectal cancer risk correlate with cancer phenotype*, Gastroenterology 139(3): 788-796

Adorno M, Cordenonsi M, Montagner M, Dupont S, Wong C, Hann B, Solari A, Bobisse S, Rondina MB, Guzzardo V, Parenti AR, Rosato A, Bicciato S, Balmain A, and Piccolo S (April, 2009), *A Mutant-p53/Smad complex opposes p63 to empower TGFbeta-induced metastasis*, Cell 137(1): 87-98

Affymetrix Inc. (2008a), *Genome-Wide Human SNP Nsp/Sty 6.0 User Guide* (revision 4), Affymetrix Inc., 3420 Central Expressway, Santa Clara, CA 95051, USA, available at http://www.affymetrix.com/, last accessed 27th March 2012

Affymetrix Inc. (2008b), *Quality control assessment in genotyping console*™ (revision 1), Affymetrix Inc., 3420 Central Expressway, Santa Clara, CA 95051, USA, available at http://www.affymetrix.com/, last accessed 27th March 2012

Affymetrix Inc. (2009), *Genotyping Console 4.0 - User Manual*, Affymetrix Inc., 3420 Central Expressway, Santa Clara, CA 95051, USA, available at http://www.affymetrix.com/, last accessed 27th March 2012

Personalis EX2155

Aglan MS, Kamel AK, and Helmy NA (2008), *Partial trisomy of the distal part of 10q: a report of two Egyptian cases*, Genetic Counselling 19(2): 199-209

Agostini M, Pucciarelli S, Enzo MV, Del Bianco P, Briarava M, Bedin C, Maretto I, Friso ML, Lonardi S, Mescoli C, Toppan P, Urso E, and Nitti D (September, 2011), *Circulating cell-free DNA: a promising marker of pathologic tumor response in rectal cancer patients receiving preoperative chemoradiotherapy*, Annals of Surgical Oncology 18(9): 2461-2468

Albain KS, Barlow WE, Shak S, Hortobagyi GN, Livingston RB, Yeh IT, Ravdin P, Bugarini R, Baehner FL, Davidson NE, Sledge GW, Winer EP, Hudis C, Ingle JN, Perez EA, Pritchard KI, Shepherd L, Gralow JR, Yoshizawa C, Allred DC, Osborne CK, and Hayes DF (January, 2010), *Prognostic and predictive value of the 21-gene recurrence score assay in postmenopausal women with node-positive, oestrogen-receptor-positive breast cancer on chemotherapy: a retrospective analysis of a randomised trial*, Lancet Oncology 11(1): 55-65

Alix-Panabières C, Schwarzenbach H, and Pantel K (February, 2012), *Circulating Tumor Cells and Circulating Tumor DNA*, Annual Review of Medicine 63: 199-215

Altschul SF, Gish W, Miller W, Myers EW, and Lipman DJ (October, 1990), *Basic local alignment search tool*, Journal of Molecular Biology 215(3): 403-410

American Cancer Society Inc. (ACS) (May, 2010), *Benign breast condition: Not all lumps are cancer*, American Cancer Society Inc., 250 Williams Street NW, Atlanta, Georgia 3030, USA, available at http://www.cancer.org/, last accessed 27th March 2012

American Cancer Society Inc. (ACS) (March, 2012), *What are the risk factors for breast cancer?*, American Cancer Society Inc., 250 Williams Street NW, Atlanta, Georgia 3030, USA, available at http://www.cancer.org/, last accessed 27th March 2012

Personalis EX2155

Andersen CL, Wiuf C, Kruhøffer M, Korsgaard M, Laurberg S, and Ørntoft TF (January, 2007), *Frequent occurrence of uniparental disomy in colorectal cancer*, Carcinogenesis 28(1): 38-48

Andrews J, Kennette W, Pilon J, Hodgson A, Tuck AB, Chambers AF, and Rodenhiser DI (January, 2010), *Multi-platform whole-genome microarray analyses refine the epigenetic signature of breast cancer metastasis with gene expression and copy number*, Public Library of Sciences One 5(1): e8665

Anker P and Stroun M (2000), *Circulating DNA in plasma or serum*, Medicina (Buenos Aires) 60(5 Part 2): 699-702

Anker P, Lyautey J, Lederrey C, and Stroun M (November, 2001), *Circulating nucleic acids in plasma or serum*, Clinica Chimica Acta 313(1-2): 143-146

Antoniou A, Pharoah PD, Narod S, Risch HA, Eyfjord JE, Hopper JL, Loman N, Olsson H, Johannsson O, Borg A, Pasini B, Radice P, Manoukian S, Eccles DM, Tang N, Olah E, Anton-Culver H, Warner E, Lubinski J, Gronwald J, Gorski B, Tulinius H, Thorlacius S, Eerola H, Nevanlinna H, Syrjäkoski K, Kallioniemi OP, Thompson D, Evans C, Peto J, Lalloo F, Evans DG, and Easton DF (May, 2003), *Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case Series unselected for family history: a combined analysis of 22 studies*, American Journal of Human Genetics 72(5): 1117-1130

Aqeilan RI, Donati V, Gaudio E, Nicoloso MS, Sundvall M, Korhonen A, Lundin J, Isola J, Sudol M, Joensuu H, Croce CM, and Elenius K (October, 2007), *Association of Wwox with ErbB4 in breast cancer*, Cancer Research 67(19): 9330-9336

Arakawa H (December, 2004), *Netrin-1 and its receptors in tumorigenesis*, Nature Reviews, Cancer 4(12): 978-987

Personalis EX2155

Armstrong K, Moye E, Williams S, Berlin JA, and Reynolds EE (April, 2007), *Screening mammography in women 40 to 49 years of age: a systematic review for the American College of Physicians*, Annals of Internal Medicine 146(7): 516-526

Arora S, Ranade AR, Tran NL, Nasser S, Sridhar S, Korn RL, Ross JT, Dhruv H, Foss KM, Sibenaller Z, Ryken T, Gotway MB, Kim S, and Weiss GJ (December, 2011), *MicroRNA-328 is associated with (non-small) cell lung cancer (NSCLC) brain metastasis and mediates NSCLC migration*, International Journal of Cancer 129(11): 2621-2631

Arteaga CL, Sliwkowski MX, Osborne CK, Perez EA, Puglisi F, and Gianni L (November, 2011), *Treatment of HER2-positive breast cancer: current status and future perspectives*, Nature Reviews Clinical Oncology 9(1): 16-32

Awada A, Bozovic-Spasojevic I, and Chow L (in press), *New therapies in HER2-positive breast cancer: A major step towards a cure of the disease?*, Cancer Treatment Reviews

Bae JS, Choi JS, Baik SH, Park WC, Song BJ, Kim JS, Lim Y, and Jung SS (April, 2010), *Genomic alterations of primary tumor and blood in invasive ductal carcinoma of breast*, World Journal of Surgical Oncology 8(32)

Balch C, Huang TH, Brown R, and Nephew KP (November, 2004), *The epigenetics of ovarian cancer drug resistance and resensitization*, American Journal of Obstetrics and Gynecology 191(5): 1552-1572

Ban H, Andoh A, Imaeda H, Kobori A, Bamba S, Tsujikawa T, Sasaki M, Saito Y, and Fujiyama Y (October, 2010), *The multidrug-resistance protein 4 polymorphism is a new factor accounting for thiopurine sensitivity in Japanese patients with inflammatory bowel disease*, Journal of Gastroenterology 45(10): 1014-1021

Personalis EX2155

Bankovic J, Stojsic J, Jovanovic D, Andjelkovic T, Milinkovic V, Ruzdijic S, and Tanic N (February, 2010), *Identification of genes associated with non-small-cell lung cancer promotion and progression*, Lung Cancer 10(2): 151-159

Barber JC, Bunyan D, Curtis M, Robinson D, Morlot S, Dermitzel A, Liehr T, Alves C, Trindade J, Paramos AI, Cooper C, Ocraft K, Taylor EJ, and Maloney VK (February, 2010), *8p23.1 duplication syndrome differentiated from copy number variation of the defensin cluster at prenatal diagnosis in four new families*, Molecular Cytogenetics 3(3)

Bärlund M, Monni O, Weaver JD, Kauraniemi P, Sauter G, Heiskanen M, Kallioniemi OP, and Kallioniemi A (December, 2002), *Cloning of BCAS3 (17q23) and BCAS4 (20q13) genes that undergo amplification, overexpression, and fusion in breast cancer*, Genes, Chromosomes, and Cancer 35(4): 311-317

Bartlett JMS, Going JJ, Mallon EA, Watters AD, Reeves JR, Stanton P, Richmond J, Donald B, Ferrier R, and Cooke TG (November, 2001), *Evaluating HER2 amplification and overexpression in breast cancer*, The Journal of Pathology 195(4): 422-428

Baum M, Buzdar A, Cuzick J, Forbes J, Houghton J, Howell A, and Sahmoud T (November, 2003), *Anastrozole alone or in combination with tamoxifen versus tamoxifen alone for adjuvant treatment of postmenopausal women with early-stage breast cancer: results of the ATAC (Arimidex, Tamoxifen Alone, or in Combination) trial efficacy and safety update analyses*, Cancer 98(9): 1779-1781

Bayoudh L, Afrit M, Daldoul O, Zarrad M, and Boussen H (January, 2012), *Trastuzumab (herceptine) in medical therapy in breast cancer*, La Tunisie Médicale 90(1): 6-12

Personalis EX2155

Beral V, Banks E, Reeves G, and Wallis M (April, 1997), *Hormone replacement therapy and high incidence of breast cancer between mammographic screens*, The Lancet 349(9058): 1103-1104

Beral V (August, 2003), *Breast cancer and hormone-replacement therapy in the Million Women Study*, The Lancet 362(9382): 419-427

Berns A (April, 2008), *A tRNA with oncogenic capacity*, Cell 133(1): 29-30

Bewley S (October, 2011), *The NHS breast screening programme needs independent review*, British Medical Journal 343: d6894

Bicciato S, Luchini A, Di Bello C (March, 2003), *PCA disjoint models for multiclass cancer analysis using gene expression data*, Bioinformatics 19(5): 571-578

Bigenwald RZ, Warner E, Gunasekara A, Hill KA, Causer PA, Messner SJ, Eisen A, Plewes DB, Narod SA, Zhang L, and Yaffe MJ (March, 2008), *Is mammography adequate for screening women with inherited BRCA mutations and low breast density?*, Cancer Epidemiology, Biomarkers, and Prevention 17(3): 706-711

Blömeke B, Bennett WP, Harris CC, and Shields PG (June, 1997), *Serum, plasma and paraffin-embedded tissues as sources of DNA for studying cancer susceptibility genes*, Carcinogenesis 18(6): 1271-1275

Blow N (August, 2007), *Tissues issues*, Nature 448: 959-962

Board RE, Williams VS, Knight L, Shaw J, Greystoke A, Ranson M, Dive C, Blackhall FH, and Hughes A (August, 2008), *Isolation and extraction of circulating tumor DNA from patients with small cell lung cancer*, Annals of the New York Academy of Sciences 1137: 98-107

Personalis EX2155

Boccardo F (February, 2008), *Switching to anastrozole after tamoxifen improves survival in postmenopausal women with breast cancer*, Nature Clinical Practice Oncology 5(2): 76-77

Bodmer D, Schepens M, Eleveld MJ, Schoenmakers EF, and Geurts van Kessel A (October, 2003), *Disruption of a novel gene, DIRC3, and expression of DIRC3-HSPBAP1 fusion transcripts in a case of familial renal cell cancer and t(2;3)(q35;q21)*, Genes, Chromosomes, and Cancer 38(2): 107-116

Boér K (February, 2010), Adjuvant chemotherapy of early stage breast cancer, Orvosi Hetilap 151(9): 344-353

Bonati MT, Finelli P, Giardino D, Gottardi G, Roberts W, and Larizza L (March, 2005), *Trisomy 15q25.2-qter in an autistic child: genotype-phenotype correlations*, American Journal of Medical Genetics Part A 133A(2): 184-188

Bowman RV, Yang IA, Semmler AB, and Fong KM (July, 2006), *Epigenetics of lung cancer*, Respirology 11(4): 355-365

Breast Cancer Surveillance Consortium (BCSC) (November, 2011), *Performance Measures for 1,960,150 Screening Mammography Examinations from 2002 to 2006 by Time (Months) Since Previous Mammography - based on BCSC data as of 2009*, BCSC, Applied Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, Executive Plaza North, Room 4111, 6130 Executive Blvd., MSC 7344, Bethesda, MD 20892-7344, USA, available at http://breastscreening.cancer.gov/, last accessed 28th March 2012

Bronte G, Terrasi M, Rizzo S, Sivestris N, Ficorella C, Cajozzo M, Di Gaudio F, Gulotta G, Siragusa S, Gebbia N, and Russo A (June, 2011), *EGFR genomic alterations in cancer: prognostic and predictive values*, Frontiers in Bioscience 3: 879-887

Personalis EX2155

Brough R, Frankum JR, Sims D, Mackay A, Mendes-Pereira AM, Bajrami I, Costa-Cabral S, Rafiq R, Ahmad AS, Cerone MA, Natrajan R, Sharpe R, Shiu KK, Wetterskog D, Dedes KJ, Lambros MB, Rawjee T, Linardopoulos S, Reis-Filho JS, Turner NC, Lord CJ, and Ashworth A (August, 2011), *Functional viability profiles of breast cancer*, Cancer Discovery 1(3): 260-273

Cailleau R, Young R, Olivé M, and Reeves WJ Jr. (September, 1974), *Breast tumor cell lines from pleural effusions*, Journal of the National Cancer Institute 53(3): 661-674

Cailleau R, Olivé M, and Cruciger QV (November, 1978), *Long-term human breast carcinoma cell lines of metastatic origin: preliminary characterization*, In Vitro 14(11): 911-915

Cancer Research UK (CRUK) (February, 2011a), *About Breast Cancer - A Quick Guide*, Cancer Research UK, Angel Building, 407 St. John Street, London EC1V 4AD, United Kingdom, available at http://info.cancerresearchuk.org/, last accessed 27th March 2012

Cancer Research UK (CRUK) (August, 2011b), *Breast cancer - UK incidence statistics*, Cancer Research UK, Angel Building, 407 St. John Street, London EC1V 4AD, United Kingdom, available at http://info.cancerresearchuk.org/, last accessed 27th March 2012

Cancer Research UK (CRUK) (February, 2011c), *Hormone therapy for secondary breast cancer*, Cancer Research UK, Angel Building, 407 St. John Street, London EC1V 4AD, United Kingdom, available at http://cancerhelp.cancerresearchuk.org/, last accessed 27th March 2012

Cancer Research UK (CRUK) (January, 2012), *Breast cancer - risk factors*, Cancer Research UK, Angel Building, 407 St. John Street, London EC1V 4AD, United Kingdom, available at http://info.cancerresearchuk.org/, last accessed 27th March 2012

Personalis EX2155

Carriaga MT and Henson DE (January, 1995), *The histologic grading of cancer*, Cancer 75(Supplement 1): 406-421

Castle JC, Biery M, Bouzek H, Xie T, Chen R, Misura K, Jackson S, Armour CD, Johnson JM, Rohl CA, and Raymond CK (April, 2010), *DNA copy number, including telomeres and mitochondria, assayed using next-generation sequencing*, BMC Genomics 11: 244

Cattell RB (1966), *The scree test for the number of factors*, Multivariate Behavioral Research 1: 629-637

Chakravarty MM, Aleong R, Leonard G, Perron M, Pike GB, Richer L, Veillette S, Pausova Z, and Paus T (May, 2011), *Automated analysis of craniofacial morphology using magnetic resonance images*, PLoS One 6(5): e20241v

Chambers AF and Goss PE (2008), *Putative growth characteristics of micrometastatic breast cancer*, Breast Cancer Research 10(6): 114

Chatfield C and Collins AJ (1980), Introduction to Multivariate Analysis, Chapman and Hall, 733 Third Avenue, NY 10017, USA, ISBN 0412160307

Checka CM, Chun JE, Schnabel FR, Lee J, and Toth H (March, 2012), *The relationship of mammographic density and age: implications for breast cancer screening*, American Journal of Roentgenology 198(3): W292-295

Chen XQ, Stroun M, Magnenat JL, Nicod LP, Kurt AM, Lyautey J, Lederrey C, and Anker P (September, 1996), *Microsatellite alterations in plasma DNA of small cell lung cancer patients*, Nature Medicine 2(9): 972-974

Chen XQ, Bonnefoi H, Pelte MF, Lyautey J, Lederrey C, Movarekhi S, Schaeffer P, Mulcahy HE, Meyer P, Stroun M, and Anker P (October, 2000), *Telomerase*

Personalis EX2155

RNA as a detection marker in the serum of breast cancer patients, Clinical Cancer Research 6(10): 3823-3826

Chen ZJ, Zhao H, He L, Shi Y, Qin Y, Shi Y, Li Z, You L, Zhao J, Liu J, Liang X, Zhao X, Zhao J, Sun Y, Zhang B, Jiang H, Zhao D, Bian Y, Gao X, Geng L, Li Y, Zhu D, Sun X, Xu JE, Hao C, Ren CE, Zhang Y, Chen S, Zhang W, Yang A, Yan J, Li Y, Ma J, and Zhao Y (January, 2011), *Genome-wide association study identifies susceptibility loci for polycystic ovary syndrome on chromosome 2p16.3, 2p21 and 9q33.3*, Nature Genetics 43(1): 55-59

Chiang PW, Beer DG, Wei WL, Orringer MB, and Kurnit DM (June, 1999), *Detection of erbB-2 amplifications in tumors and sera from esophageal carcinoma patients*, Clinical Cancer Research 5(6): 1381-1386

Chin K, DeVries S, Fridlyand J, Spellman PT, Roydasgupta R, Kuo WL, Lapuk A, Neve RM, Qian Z, Ryder T, Chen F, Feiler H, Tokuyasu T, Kingsley C, Dairkee S, Meng Z, Chew K, Pinkel D, Jain A, Ljung BM, Esserman L, Albertson DG, Waldman FM, and Gray JW (December, 2006), *Genomic and transcriptional aberrations linked to breast cancer pathophysiologies*, Cancer Cell 10(6): 529-541

Chiu HH, Yong TM, Wang J, Wang Y, Vessella RL, Ueda T, Wang YZ, and Sadar MD (June, 2010), *Induction of neuronal apoptosis inhibitory protein expression in response to androgen deprivation in prostate cancer*, Cancer Letters 292(2): 176-185

Chlebowski RT, Kuller LH, Prentice RL, Stefanick ML, Manson JE, Gass M, Aragaki AK, Ockene JK, Lane DS, Sarto GE, Rajkovic A, Schenken R, Hendrix SL, Ravdin PM, Rohan TE, Yasmeen S, and Anderson G (February, 2009), *Breast cancer after use of estrogen plus progestin in postmenopausal women*, New England Journal of Medicine 360(6): 573-587

Personalis EX2155

Choi SY, Jang JH, and Kim KR (December, 2011), *Analysis of differentially expressed genes in human rectal carcinoma using suppression subtractive hybridization*, Clinical and Experimental Medicine 11(4): 219-226

Cleator S, Heller W, and Coombes RC (March, 2007), *Triple-negative breast cancer: therapeutic options*, Lancet Oncology 8(3): 235-244

Collaborative Group on Hormonal Factors in Breast Cancer (June, 1996), *Breast cancer and hormonal contraceptives: collaborative reanalysis of individual data on 53 297 women with breast cancer and 100 239 women without breast cancer from 54 epidemiological studies*, The Lancet 347(9017): 1713-1727

Collaborative Group on Hormonal Factors in Breast Cancer (October, 2001), *Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease*, The Lancet 358(9291): 1389-1399

Conrad DF, Pinto D, Redon R, Feuk L, Gokcumen O, Zhang Y, Aerts J, Andrews TD, Barnes C, Campbell P, Fitzgerald T, Hu M, Ihm CH, Kristiansson K, Macarthur DG, Macdonald JR, Onyiah I, Pang AW, Robson S, Stirrups K, Valsesia A, Walter K, Wei J; Wellcome Trust Case Control Consortium, Tyler-Smith C, Carter NP, Lee C, Scherer SW, and Hurles ME (October, 2009), *Origins and functional impact of copy number variation in the human genome*, Nature 464(7289): 704-712

Contreras A (October, 2009), *Oncotype DX and MammaPrint Assays in Breast Cancer*, BCM Pathletters 1(2): 1-2

Coombes RC, Hall E, Gibson LJ, Paridaens R, Jassem J, Delozier T, Jones SE, Alvarez I, Bertelli G, Ortmann O, Coates AS, Bajetta E, Dodwell D, Coleman RE, Fallowfield LJ, Mickiewicz E, Andersen J, Lonning PE, Cocconi G, Stewart A, Stuart N, Snowdon CF, Carpentieri M, Massimini G, and Bliss JM (March, 2004), *A randomized trial of exemestane after two to three years of tamoxifen*

Personalis EX2155

therapy in postmenopausal women with primary breast cancer, New England Journal of Medicine 350(11): 1081-1092

Coronado GD, Beasley J, and Livaudais J (October, 2011), *Alcohol consumption and the risk of breast cancer*, Salud Pública de México 53(5): 440-447

Cristofanilli M, Budd GT, Ellis MJ, Stopeck A, Matera J, Miller MC, Reuben JM, Doyle GV, Allard WJ, Terstappen LW, and Hayes DF (August, 2004), *Circulating tumor cells, disease progression, and survival in metastatic breast cancer*, New England Journal of Medicine 351(8): 781-791

Cruz Jurado J, Richart Aznar P, García Mata J, Fernández Martínez R, Peláez Fernández I, Sampedro Gimeno T, Galve Calvo E, Murillo Jaso L, Polo Marqués E, and García Palomo A (September, 2011), *Management of patients with metastatic breast cancer*, Advances in Therapy 28(Supplement 6): 50-65

Curado MP, Edwards B, Shin HR, Ferlay J, Heanue M, Boyle P, and Storm H (2007), Cancer incidence in five continents - volume IX, International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France, ISBN 9283221605

Curtis C, Shah SP, Chin SF, Turashvili G, Rueda OM, Dunning MJ, Speed D, Lynch AG, Samarajiwa S, Yuan Y, Gräf S, Ha G, Haffari G, Bashashati A, Russell R, McKinney S; METABRIC Group; Co-chairs, Caldas C, Aparicio S; Writing committee, Curtis C, Shah SP, Caldas C, Aparicio S; Steering committee, Brenton JD, Ellis I, Huntsman D, Pinder S, Purushotham A, Murphy L, Caldas C, Aparicio S; Tissue and clinical data source sites:; University of Cambridge/Cancer Research UK Cambridge Research Institute, Caldas C, Bardwell H, Chin SF, Curtis C, Ding Z, Gräf S, Jones L, Liu B, Lynch AG, Papatheodorou I, Sammut SJ, Wishart G; British Columbia Cancer Agency, Aparicio S, Chia S, Gelmon K, Huntsman D, McKinney S, Speers C, Turashvili G, Watson P; University of Nottingham, Ellis I, Blamey R, Green A, Macmillan D, Rakha E; King's College London, Purushotham A, Gillett C, Grigoriadis A,

Personalis EX2155

Pinder S, di Rinaldis E, Tutt A; Manitoba Institute of Cell Biology, Murphy L, Parisien M, Troup S; Cancer genome/transcriptome characterization centres:; University of Cambridge/Cancer Research UK Cambridge Research Institute, Caldas C, Chin SF, Chan D, Fielding C, Maia AT, McGuire S, Osborne M, Sayalero SM, Spiteri I, Hadfield J; British Columbia Cancer Agency, Aparicio S, Turashvili G, Bell L, Chow K, Gale N, Huntsman D, Kovalik M, Ng Y, Prentice L; Data analysis subgroup:; University of Cambridge/Cancer Research UK Cambridge Research Institute, Caldas C, Tavaré S, Curtis C, Dunning MJ, Gräf S, Lynch AG, Rueda OM, Russell R, Samarajiwa S, Speed D, Markowetz F, Yuan Y, Brenton JD; British Columbia Cancer Agency, Aparicio S, Shah SP, Bashashati A, Ha G, Haffari G, McKinney S, Langerød A, Green A, Provenzano E, Wishart G, Pinder S, Watson P, Markowetz F, Murphy L, Ellis I, Purushotham A, Børresen-Dale AL, Brenton JD, Tavaré S, Caldas C, and Aparicio S (in press), *The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups*, Nature

Daphna-Iken D, Shankar DB, Lawshé A, Ornitz DM, Shackleford GM, and MacArthur CA (November, 1998), *MMTV-Fgf8 transgenic mice develop mammary and salivary gland neoplasia and ovarian stromal hyperplasia*, Oncogene 17(21): 2711-2717

Dastani Z, Pajukanta P, Marcil M, Rudzicz N, Ruel I, Bailey SD, Lee JC, Lemire M, Faith J, Platko J, Rioux J, Hudson TJ, Gaudet D, Engert JC, and Genest J (March, 2010), *Fine mapping and association studies of a high-density lipoprotein cholesterol linkage region on chromosome 16 in French-Canadian subjects*, European Journal of Human Genetics 18(3): 342-347

Davies H, Bignell GR, Cox C, Stephens P, Edkins S, Clegg S, Teague J, Woffendin H, Garnett MJ, Bottomley W, Davis N, Dicks E, Ewing R, Floyd Y, Gray K, Hall S, Hawes R, Hughes J, Kosmidou V, Menzies A, Mould C, Parker A, Stevens C, Watt S, Hooper S, Wilson R, Jayatilake H, Gusterson BA, Cooper C, Shipley J, Hargrave D, Pritchard-Jones K, Maitland N, Chenevix-Trench G, Riggins GJ, Bigner DD, Palmieri G, Cossu A, Flanagan A, Nicholson A, Ho JW, Leung SY,

Personalis EX2155

Yuen ST, Weber BL, Seigler HF, Darrow TL, Paterson H, Marais R, Marshall CJ, Wooster R, Stratton MR, and Futreal PA (June, 2002), *Mutations of the BRAF gene in human cancer*, Nature 417(6892): 949-954

Dawood S, Hu R, Homes MD, Collins LC, Schnitt SJ, Connolly J, Colditz GA, and Tamimi RM (February, 2011), *Defining breast cancer prognosis based on molecular phenotypes: results from a large cohort study*, Breast Cancer Research and Treatment 126(1): 185-192

De Carvalho DD, Binato R, Pereira WO, Leroy JM, Colassanti MD, Proto-Siqueira R, Bueno-Da-Silva AE, Zago MA, Zanichelli MA, Abdelhay E, Castro FA, Jacysyn JF, and Amarante-Mendes GP (January, 2011), *BCR-ABL-mediated upregulation of PRAME is responsible for knocking down TRAIL in CML patients*, Oncogene 30(2): 223-233

Deapen D, Liu L, Perkins C, Bernstein L, and Ross RK (June, 2002), *Rapidly rising breast cancer incidence rates among Asian-American women*, International Journal of Cancer 99(5): 747-750

Demicheli R, Terenziani M, and Bonadonna G (September, 1998), *Estimate of tumor growth time for breast cancer local recurrences: rapid growth after wake-up?*, Breast Cancer Research and Treatment 51(2): 133-137

Dempster AP, Laird NM, and Rubin DB (1977), *Maximum likelihood from incomplete data via the EM algorithm*, Journal of the royal statistical society 39(1): 1-38

Dexter DL and Leith JT (February, 1986), *Tumor heterogeneity and drug resistance*, Journal of Clinical Oncology 4(2): 244-257

Dey JH, Bianchi F, Voshol J, Bonenfant D, Oakeley EJ, and Hynes NE (May, 2010), *Targeting fibroblast growth factor receptors blocks PI3K/AKT signalling, induces apoptosis, and impairs mammary tumor outgrowth and metastasis*, Cancer Research 70(10): 4151-4162

Personalis EX2155

Donati G, Brighenti E, Vici M, Mazzini G, Treré D, Montanaro L, and Derenzini M (September, 2011), *Selective inhibition of rRNA transcription downregulates E2F-1: a new p53-independent mechanism linking cell growth to cell proliferation*, Journal of Cell Science 124(Part 17): 3017-2028

Dowsett M, Cuzick J, Wale C, Forbes J, Mallon EA, Salter J, Quinn E, Dunbier A, Baum M, Buzdar A, Howell A, Bugarini R, Baehner FL, and Shak S (April, 2010), *Prediction of risk of distant recurrence using the 21-gene recurrence score in node-negative and node-positive postmenopausal patients with breast cancer treated with anastrozole or tamoxifen: a TransATAC study*, Journal of Clinical Oncology 28(11): 1829-1834

Duclos A, Charbonnier F, Chambon P, Latouche JB, Blavier A, Redon R, Frébourg T, and Flaman JM (October, 2011), *Pitfalls in the use of DGV for CNV interpretation*, American Journal of Medical Genetics 155(10): 2593-2596

Dumur CI, Dechsukhum C, Ware JL, Cofield SS, Best AM, Wilkinson DS, Garrett CT, and Ferreira-Gonzalez A (March, 2003), *Genome-wide detection of LOH in prostate cancer using human SNP microarray technology*, Genomics 81(3): 260-269

Early Breast Cancer Trialists' Collaborative Group (EBCTCG) (May, 2005), *Effects of chemotherapy and hormonal therapy for early breast cancer on recurrence and 15-year survival: an overview of the randomised trials*, The Lancet 365(9472): 1687-717

Easton DF (August, 1999), *How many more breast cancer predisposition genes are there?*, Breast Cancer Research 1(1): 14-17

Elmore JG, Barton MB, Moceri VM, Polk S, Arena PJ, and Fletcher SW (April, 1998), *Ten-year risk of false positive screening mammograms and clinical breast examinations*, New England Journal of Medicine 338(16): 1089-1096

Personalis EX2155

Elston CW and Ellis IO (September, 2002), *Pathological prognostic factors in breast cancer. I. The value of histological grade in breast cancer: experience from a large study with long-term follow-up*, Histopathology 41(3A): 154-61.

Enmark E and Gustafsson JA (August, 1999), *Oestrogen receptors - an overview, Journal of International Medicine* 246(2): 133-138

Espinosa E, Gámez-Pozo A, Sánchez-Navarro I, Pinto A, Castañeda CA, Ciruelos E, Feliu J, and Vara JA (in press), *The present and future of gene profiling in breast cancer*, Cancer Metastasis Reviews

Eswarakumar VP, Lax I, Schlessinger J (April, 2005), *Cellular signaling by fibroblast growth factor receptors*, Cytokine & Growth Factor Reviews, 16(2): 139-149

Evron E, Dooley WC, Umbricht CB, Rosenthal D, Sacchi N, Gabrielson E, Soito AB, Hung DT, Ljung B, Davidson NE, and Sukumar S (April, 2001), *Detection of breast cancer cells in ductal lavage fluid by methylation-specific PCR*, Lancet 357(9265): 1335-1336

Fang M, Toher J, Morgan M, Davison J, Tannenbaum S, and Claffey K (May, 2011), *Genomic differences between estrogen receptor (ER)-positive and ER-negative human breast carcinoma identified by single nucleotide polymorphism array comparative genome hybridization analysis*, Cancer 117(10): 2024-2034

Farace F, Massard C, Vimond N, Drusch F, Jacques N, Billiot F, Laplanche A, Chauchereau A, Lacroix L, Planchard D, Le Moulec S, André F, Fizazi K, Soria JC, and Vielh P (August, 2011), *A direct comparison of CellSearch and ISET for circulating tumour-cell detection in patients with metastatic carcinomas*, British Journal of Cancer 105(6): 847-853

Feng Y, Zhou HX, Li JH, Li ZY, Cheng WG, Jin ML, and He L (August, 2011), *Expression of deleted in liver cancer 1 and phosphorelated focal adhesion*

Personalis EX2155

kinase in breast cancer, Journal of Southern Medical University 31(8): 1448-1451

Ferlay J, Shin HR, Bray F, Forman D, Mathers C, and Parkin DM (2010), GLOBOCAN 2008: Cancer Incidence and Mortality Worldwide - IARC CancerBase Number 10 (version 1.2), International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France, ISBN 9789283224471, http://globocan.iarc.fr

Field SF, Howson JM, Maier LM, Walker S, Walker NM, Smyth DJ, Armour JA, Clayton DG, and Todd JA (October, 2009), *Experimental aspects of copy number variant assays at CCL3L1*, Nature Medicine 15(10): 1115-1117

Fischer U, Jänicke RU, and Schulze-Osthoff K (January, 2003), *Many cuts to ruin: a comprehensive update of caspase substrates*, Cell Death and Differentiation 10(1): 76-100

Fitzgibbon J, Iqbal S, Davies A, O'Shea D, Carlotti E, Chaplin T, Matthews J, Raghavan M, Norton A, Lister TA, and Young BD (July, 2007), *Genome-wide detection of recurring sites of uniparental disomy in follicular and transformed follicular lymphoma*, Leukemia 21(7): 1514-1520

Fogh J and Trempe G (1975), Human tumor cells in vitro, Plenum Press, New York/London, ISBN 0306308109

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, and Stratton MR (April, 2008), *The Catalogue of Somatic Mutations in Cancer (COSMIC)*, Current Protocols in Human Genetics, Chapter 10, Unit 10.11

Forcet C, Stein E, Pays L, Corset V, Llambi F, Tessier-Lavigne M, and Mehlen P (May, 2002), *Netrin-1-mediated axon outgrowth requires deleted in colorectal cancer-dependent MAPK activation*, Nature 417(6887): 443-447

Personalis EX2155

Fujita PA, Rhead B, Zweig AS, Hinrichs AS, Karolchik D, Cline MS, Goldman M, Barber GP, Clawson H, Coelho A, Diekhans M, Dreszer TR, Giardine BM, Harte RA, Hillman-Jackson J, Hsu F, Kirkup V, Kuhn RM, Learned K, Li CH, Meyer LR, Pohl A, Raney BJ, Rosenbloom KR, Smith KE, Haussler D, and Kent WJ (January, 2011), *The UCSC Genome Browser database: update 2011*, Database issue 39(D876-882)

Gahan PB (April, 2007), *Circulating nucleic acids in plasma and serum: roles in diagnosis and prognosis in diabetes and cancer*, Infectious Disorders - Drug Targets 8(2): 100-108

Gajdos C, Tartter PI, and Bleiweiss IJ (November, 1999), *Lymphatic invasion, tumor size, and age are independent predictors of axillary lymph node metastases in women with T1 breast cancers*, Annals of Surgery, 230(5): 692-696

Gang F, Guorong L, An Z, Anne GP, Christian G, and Jacques T (February, 2010), *Prediction of clear cell renal cell carcinoma by integrity of cell-free DNA in serum*, Urology 75(2): 262-265

Gangnus R, Langer S, Breit E, Pantel K, and Speicher MR (May, 2004), *Genomic profiling of viable and proliferative micrometastatic cells from early-stage breast cancer patients*, Clinical Cancer Research 10(10): 3457-3464

García-Closas M, Brinton LA, Lissowska J, Chatterjee N, Peplonska B, Anderson WF, Szeszenia-Dabrowska N, Bardin-Mikolajczak A, Zatonski W, Blair A, Kalaylioglu Z, Rymkiewicz G, Mazepa-Sikora D, Kordek R, Lukaszek S, and Sherman ME (July, 2006), *Established breast cancer risk factors by clinically important tumour characteristics*, British Journal of Cancer 95(1): 123-129

García-Olmo D, García-Olmo DC, Ontañón J, Martinez E, and Vallejo M (October, 1999), *Tumor DNA circulating in the plasma might play a role in metastasis.*

Personalis EX2155

The hypothesis of the genometastasis, Histology and Histopathology 14(4): 1159-1164

García-Olmo D, García-Olmo DC, Ontañón J, and Martinez E (January, 2000), *Horizontal transfer of DNA and the "genometastasis hypothesis"*, Blood 95(2): 724-725

García-Olmo D and García-Olmo DC (September, 2001), *Functionality of circulating DNA: the hypothesis of genometastasis*, Annals of the New York Academy of Sciences, 945: 265-275

García-Olmo DC, Samos J, Picazo MG, Asensio AI, Toboso I, and García-Olmo D (December, 2008), *Release of cell-free DNA into the bloodstream leads to high levels of non-tumor plasma DNA during tumor progression in rats*, Cancer Letters 272(1): 133-140

Garnett MJ and Marais R (October, 2004), *Guilty as charged: B-RAF is a human oncogene*, Cancer Cell 6(4): 313-319

Garshasbi M, Hadavi V, Habibi H, Kahrizi K, Kariminejad R, Behjati F, Tzschach A, Najmabadi H, Ropers HH, and Kuss AW (May, 2008), *A defect in the TUSC3 gene is associated with autosomal recessive mental retardation*, American Journal of Human Genetics 82(5): 1158-1164

Gaudet MM, Press MF, Haile RW, Lynch CF, Glaser SL, Schildkraut J, Gammon MD, Douglas Thompson W, and Bernstein JL (November, 2011), *Risk factors by molecular subtypes of breast cancer across a population-based study of women 56 years or younger*, Breast Cancer Research and Treatment 130(2): 587-597

Gautier L, Cope L, Bolstad BM, and Irizarry RA (February, 2004), *affy - analysis of Affymetrix GeneChip data at the probe level*, Bioinformatics 20(3): 307-315

Personalis EX2155

Gee JM, Robertson JF, Ellis IO, and Nicholson RI (July, 2001), *Phosphorylation of ERK1/2 mitogen-activated protein kinase is associated with poor response to anti-hormonal therapy and decreased patient survival in clinical breast cancer*, International Journal of Cancer 95(4): 247-275

Gee JM, Robertson JF, Gutteridge E, Ellis IO, Pinder SE, Rubini M, and Nicholson RI (July, 2005), *Epidermal growth factor receptor/HER2/insulin-like growth factor receptor signalling and oestrogen receptor activity in clinical breast cancer*, Endocrine-Related Cancer 12(Supplement 1): 99-111

Gene Ontology Consortium (May, 2005), *Gene ontology: tool for the unification of biology*, Nature Genetics 25(1): 25-29

George RE, Attiyeh EF, Li S, Moreau LA, Neuberg D, Li C, Fox EA, Meyerson M, Diller L, Fortina P, Look AT, and Maris JM (February, 2007), *Genome-wide analysis of neuroblastomas using high-density single nucleotide polymorphism arrays*, PLoS One 2(2): e255

Geyer FC, Rodrigues DN, Weigelt B, and Reis-Filho JS (January, 2011), *Molecular classification of estrogen receptor-positive/luminal breast cancers*, Advances in Anatomical Pathology 19(1): 39-53

Giamas G, Filipović A, Jacob J, Messier W, Zhang H, Yang D, Zhang W, Shifa BA, Photiou A, Tralau-Stewart C, Castellano L, Green AR, Coombes RC, Ellis IO, Ali S, Lenz HJ, and Stebbing J (June, 2011), *Kinome screening for regulators of the estrogen receptor identifies LMTK3 as a new therapeutic target in breast cancer*, Nature Medicine 17(6): 715-719

Gillan TL, Hughes R, Godbout R, and Grundy PE (September, 2003), *The Simpson-Golabi-Behmel gene, GPC3, is not involved in sporadic Wilms tumorigenesis*, American Journal of Medical Genetics Part A 122A(1): 30-36

Personalis EX2155

Gilley D, Tanaka H, Herbert BS (May, 2005), *Telomere dysfunction in aging and cancer*, The International Journal of Biochemistry and Cell Biology 37(5): 1000-1013

Glass AG, Lacey JV Jr, Carreon JD, and Hoover RN (August, 2007), *Breast cancer incidence, 1980-2006: combined roles of menopausal hormone therapy, screening mammography, and estrogen receptor status*, Journal of the National Cancer Institute 99(15): 1152-1161

Glenn TC (September, 2011), *Field guide to next-generation DNA sequencers*, Molecular Ecology Resources 11(5): 759-769

Goebel G, Zitt M, Zitt M, and Müller HM (2005), *Circulating nucleic acids in plasma or serum (CNAPS) as prognostic and predictive markers in patients with solid neoplasias*, Disease Markers 21(3): 105-120

Goodison S, Kawai K, Hihara J, Jiang P, Yang M, Urquidi V, Hoffman RM, and Tarin D (September, 2003), *Prolonged dormancy and site-specific growth potential of cancer cells spontaneously disseminated from nonmetastatic breast tumors as revealed by labeling with green fluorescent protein*, Clinical Cancer Research 9(10 Part 1): 3808-3814

Gormally E, Caboux E, Vineis P, and Hainaut P (May-June, 2007), *Circulating free DNA in plasma or serum as biomarker of carcinogenesis: practical aspects and biological significance*, Mutation Research 635(2-3): 105-117

Goss PE, Ingle JN, Martino S, Robert NJ, Muss HB, Piccart MJ, Castiglione M, Tu D, Shepherd LE, Pritchard KI, Livingston RB, Davidson NE, Norton L, Perez EA, Abrams JS, Therasse P, Palmer MJ, and Pater JL (November, 2003), *A randomized trial of letrozole in postmenopausal women after five years of tamoxifen therapy for early-stage breast cancer*, New England Journal of Medicine 349(19): 1793-1802

Personalis EX2155

Goss PE, Ingle JN, Martino S, Robert NJ, Muss HB, Piccart MJ, Castiglione M, Tu D, Shepherd LE, Pritchard KI, Livingston RB, Davidson NE, Norton L, Perez EA, Abrams JS, Cameron DA, Palmer MJ, and Pater JL (September, 2005), *Randomized trial of letrozole following tamoxifen as extended adjuvant therapy in receptor-positive breast cancer: updated findings from NCIC CTG MA.17*, Journal of the National Cancer Institute, 97(17): 1262-1271

Goss PE and Chambers AF (December, 2010), *Does tumour dormancy offer a therapeutic target?*, Nature Reviews, Cancer 10(12): 871-877

Gøtzsche PC and Nielsen M (January, 2011), *Screening for breast cancer with mammography*, Cochrane Database of System Reviews 1: CD001877

Gøtzsche PC and Jørgensen KJ, (September, 2011), *The breast screening programme and misinforming the public*, Journal of the Royal Society of Medicine 104(9): 361-369

Gourley C, Paige AJ, Taylor KJ, Scott D, Francis NJ, Rush R, Aldaz CM, Smyth JF, and Gabra H (June, 2005), *WWOX mRNA expression profile in epithelial ovarian cancer supports the role of WWOX variant 1 as a tumour suppressor, although the role of variant 4 remains unclear*, International Journal of Oncology 26(6): 1681-1689

Graubert TA, Cahan P, Edwin D, Selzer RR, Richmond TA, Eis PS, Shannon WD, Li X, McLeod HL, Cheverud JM, and Ley TJ (January, 2007), *A high-resolution map of segmental DNA copy number variation in the mouse genome*, PLoS Genetics 3(1)

Grayson BL and Aune TM (April, 2011), *A comparison of genomic copy number calls by Partek Genomics Suite, Genotyping Console and Birdsuite algorithms to quantitative PCR*, BioData Mining 4(8)

Personalis EX2155

Graziano V and De Laurenzi V (August, 2011), *Role of p63 in cancer development*, Biochimica et Biophysica Acta 1816(1): 57-66

Greenman CD, Bignell G, Butler A, Edkins S, Hinton J, Beare D, Swamy S, Santarius T, Chen L, Widaa S, Futreal PA, and Stratton MR (January, 2010), *PICNIC: an algorithm to predict absolute allelic copy number variation with microarray cancer data*, Biostatistics 11(1): 164-175

Grünblatt E, Zehetmayer S, Bartl J, Löffler C, Wichart I, Rainer MK, Jungwirth S, Bauer P, Danielczyk W, Tragl KH, Riederer P, and Fischer P (January, 2009), *Genetic risk factors and markers for Alzheimer's disease and/or depression in the VITA study*, Journal of Psychiatric Research 43(3): 298-308

Gruss HJ and Dower SK (June, 1995), *Tumor necrosis factor ligand superfamily: involvement in the pathology of malignant lymphomas*, Blood 85(12): 3378-3404

Guervós MA, Marcos CA, Hermsen M, Nuño AS, Suárez C, and Llorente JL (2007), *Deletions of N33, STK11 and TP53 are involved in the development of lymph node metastasis in larynx and pharynx carcinomas*, Cellular Oncology 29(4): 327-334

Gururaj AE, Singh RR, Rayala SK, Holm C, den Hollander P, Zhang H, Balasenthil S, Talukder AH, Landberg G, and Kumar R (April, 2006), *MTA1, a transcriptional activator of breast cancer amplified sequence 3*, Proceedings of the National Academy of Sciences of the USA 103(17): 6670-6675

Gururaj AE, Peng S, Vadlamudi RK, and Kumar R (August, 2007), *Estrogen induces expression of BCAS3, a novel estrogen receptor-alpha coactivator, through proline-, glutamic acid-, and leucine-rich protein-1 (PELP1)*, Molecular Endocrinology 21(8): 1847-1860

Personalis EX2155

Hackshaw A (November, 2009), *Luteinizing hormone-releasing hormone (LHRH) agonists in the treatment of breast cancer*, Expert Opinion on Pharmacotherapy 10(16): 2633-2639

Haimovich AD (December, 2011), *Methods, challenges, and promise of next-generation sequencing in cancer biology*, The Yale Journal of Biology and Medicine 84(4): 439-446

Hamady M, Walker JJ, Harris JK, Gold NJ, and Knight R (March, 2008), *Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex*, Nature Methods 5(3): 235-237

Hannemann J, Kristel P, van Tinteren H, Bontenbal M, van Hoesel QG, Smit WM, Nooij MA, Voest EE, van der Wall E, Hupperets P, de Vries EG, Rodenhuis S, and van de Vijver MJ (November, 2006), *Molecular subtypes of breast cancer and amplification of topoisomerase II alpha: predictive role in dose intensive adjuvant chemotherapy*, British Journal of Cancer 95(10): 1334-1341

Hardiman G (July, 2004), *Microarray platforms - comparisons and contrasts*, Pharmacogenomics 5(5): 487-502

Hartmann LC, Sellers TA, Frost MH, Lingle WL, Degnim AC, Ghosh K, Vierkant RA, Maloney SD, Pankratz VS, Hillman DW, Suman VJ, Johnson J, Blake C, Tlsty T, Vachon CM, Melton LJ 3rd, and Visscher DW (July, 2005), *Benign breast disease and the risk of breast cancer*, New England Journal of Medicine 353(3): 229-237

Hautbergue GM, Hung ML, Walsh MJ, Snijders AP, Chang CT, Jones R, Ponting CP, Dickman MJ, and Wilson SA (December, 2009), *UIF, a New mRNA export adaptor that works together with REF/ALY, requires FACT for recruitment to mRNA*, Current Biology 19(22): 1918-1924

Personalis EX2155

Haybittle JL, Blamey RW, Elston CW, Johnson J, Doyle PJ, Campbell FC, Nicholson RI, and Griffiths K (March, 1982), *A prognostic index in primary breast cancer*, British Journal of Cancer 45(3): 361-366

Heering J, Erlmann P, and Olayioye MA (September, 2009), *Simultaneous loss of the DLC1 and PTEN tumor suppressors enhances breast cancer cell migration*, Experimental Cell Research 315(15): 2505-2514

Heidinger BJ, Blount JD, Boner W, Griffiths K, Metcalfe NB, and Monaghan P (2012), *Telomere length in early life predicts lifespan*, Proceedings of the National Academy of Sciences of the USA

Héry C, Ferlay J, Boniol M, and Autier P (February, 2008), *Changes in breast cancer incidence and mortality in middle-aged and elderly women in 28 countries with Caucasian majority populations*, Annals of Oncology 19(5): 1009-1018

Hesketh T, Lu L, and Xing ZW (September, 2005), *The effect of China's one-child family policy after 25 years*, New England Journal of Cancer, 353(11): 1171-1176

Hezova R, Ehrmann J, and Kolar Z (June, 2007), *WWOX, a new potential tumor suppressor gene*, Biomedical Papers of the Medical Faculty of the University of Palacký, Olomouc, Czech Republic 151(1): 11-15

Hicks C, Kumar R, Pannuti A, and Miele L (February, 2012), *Integrative analysis of response to tamoxifen Treatment in ER-positive breast cancer using GWAS information and transcription profiling*, Breast Cancer: Basic and Clinical Research 6: 47-66

Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS, and Manolio TA (June, 2009), *Potential etiologic and functional implications of genome-wide association loci for human diseases and traits*, Proceedings of the National Academy of Sciences of the USA 106(23): 9362-9367

Personalis EX2155

Hofvind S, Thoresen S, and Tretli S (October, 2004), *The cumulative risk of a false-positive recall in the Norwegian Breast Cancer Screening Program*, Cancer 101(7): 1501-1507

Holdenrieder S, Stieber P, Chan LY, Geiger S, Kremer A, Nagel D, and Lo YM (August, 2005), *Cell-free DNA in serum and plasma: comparison of ELISA and quantitative PCR*, Clinical Chemistry 51(8): 1544-1546

Holland DG, Burleigh A, Git A, Goldgraben MA, Perez-Mancera PA, Chin SF, Hurtado A, Bruna A, Ali HR, Greenwood W, Dunning MJ, Samarajiwa S, Menon S, Rueda OM, Lynch AG, McKinney S, Ellis IO, Eaves CJ, Carroll JS, Curtis C, Aparicio S, and Caldas C (March, 2011), *ZNF703 is a common Luminal B breast cancer oncogene that differentially regulates luminal and basal progenitors in human mammary epithelium*, EMBO Molecular Medicine 3(3): 167-180

Holm K, Hegardt C, Staaf J, Vallon-Christersson J, Jönsson G, Olsson H, Borg A, and Ringnér M (June, 2010), *Molecular subtypes of breast cancer are associated with characteristic DNA methylation patterns*, Breast Cancer Research 12(3): R36

Hoque MO, Lee CC, Cairns P, Schoenberg M, and Sidransky D (May, 2003), *Genome-wide genetic characterization of bladder cancer: a comparison of high density single-nucleotide polymorphism arrays and PCR-based microsatellite analysis*, Cancer Research 63(9): 2216-2222

Hoque MO, Feng Q, Toure P, Dem A, Critchlow CW, Hawes SE, Wood T, Jeronimo C, Rosenbaum E, Stern J, Yu M, Trink B, Kiviat NB, and Sidransky D (September, 2006), *Detection of aberrant methylation of four genes in plasma DNA for the detection of breast cancer*, Journal of Clinical Oncology 24(26): 4262-4269

Personalis EX2155

Horner MJ, Ries LAG, Krapcho M, Neyman N, Aminou R, Howlader N, Altekruse SF, Feuer EJ, Huang L, Mariotto A, Miller BA, Lewis DR, Eisner MP, Stinchcomb DG, and Edwards BK (2009), *SEER Cancer Statistics Review 1975-2006*, http://seer.cancer.gov/csr/1975_2006/

Hosey AM, Gorski JJ, Murray MM, Quinn JE, Chung WY, Stewart GE, James CR, Farragher SM, Mulligan JM, Scott AN, Dervan PA, Johnston PG, Couch FJ, Daly PA, Kay E, McCann A, Mullan PB, and Harkin DP (November, 2007), *Molecular basis for estrogen receptor alpha deficiency in BRCA1-linked breast cancer*, Journal of the National Cancer Institute 99(22): 1683-1694

Hostetter G, Kim SY, Savage S, Gooden GC, Barrett M, Zhang J, Alla L, Watanabe A, Einspahr J, Prasad A, Nickoloff BJ, Carpten J, Trent J, Alberts D, and Bittner M (January, 1010), *Random DNA fragmentation allows detection of single-copy, single-exon alterations of copy number by oligonucleotide array CGH in clinical FFPE samples*, Nucleic Acids Research 38(2): e9

Howe JR, Klimstra DS, and Cordon-Cardo C (July, 1997), *DNA extraction from paraffin-embedded tissues using a salting-out procedure: a reliable method for PCR amplification of archival material*, Histology and Histopathology 12(3): 595-601

Hu DG and Mackenzie PI (August, 2009), *Estrogen receptor alpha, fos-related antigen-2, and c-Jun coordinately regulate human UDP glucuronosyltransferase 2B15 and 2B17 expression in response to 17beta-estradiol in MCF-7 cells*, Molecular Pharmacology 76(2): 425-439

Huang ZH, Li LH, and Hua D (November, 2006), *Quantitative analysis of plasma circulating DNA at diagnosis and during follow-up of breast cancer patients*, Cancer Letters 243(1): 64-70

Personalis EX2155

Huang DW, Sherman BT, and Lempicki RA (January, 2009a), *Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists*, Nucleic Acids Research 37(1): 1-13

Huang DW, Sherman BT, and Lempicki RA (2009b), *Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources*, Nature Protocols 4(1): 44-57

Huang XS, Xiao L, Li X, Xie Y, Jiang HO, Tan C, Wang L, and Zhang JX (May-June, 2010), *Two neighboring microdeletions of 5q13.2 in a child with oculo-auriculo-vertebral spectrum*, European Journal of Medical Genetics 53(3): 153-158

Hughes S, Yoshimoto M, Beheshti B, Houlston RS, Squire JA, and Evans A (March, 2006), *The use of whole genome amplification to study chromosomal changes in prostate cancer: insights into genome-wide signature of preneoplasia associated with cancer progression*, BioMed Central Genomics 7(65)

Huggins R, Li LH, Lin YC, Yu AL, and Yang HC (2008), *Nonparametric estimation of LOH using Affymetrix SNP genotyping arrays for unpaired samples*, Journal of Human Genetics 53(11-12): 983-990

Huss S, Wardelmann E, Goltz D, Binot E, Hartmann W, Merkelbach-Bruse S, Büttner R, and Schildhaus HU (in press), *Activating PDGFRA mutations in inflammatory fibroid polyps occur in exons 12, 14 and 18 and are associated with tumour localization*, Histopathology

Hutchinson L (November, 2004), *Can exemestane improve adjuvant treatment for postmenopausal women with primary breast cancer?*, Nature Clinical Practice Oncology 1(1): 24-25

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, and Lee C (September, 2004), *Detection of large-scale variation in the human genome*, Nature Genetics 36(9): 949-951

Personalis EX2155

Ingermann AR, Yang YF, Han J, Mikami A, Garza AE, Mohanraj L, Fan L, Idowu M, Ware JL, Kim HS, Lee DY, and Oh Y (September, 2010), *Identification of a novel cell death receptor mediating IGFBP-3-induced anti-tumor effects in breast and prostate cancer*, The Journal of Biological Chemistry 285(39): 30233-30246

International Consortium for Systemic Lupus Erythematosus Genetics, Harley JB, Alarcón-Riquelme ME, Criswell LA, Jacob CO, Kimberly RP, Moser KL, Tsao BP, Vyse TJ, Langefeld CD, Nath SK, Guthridge JM, Cobb BL, Mirel DB, Marion MC, Williams AH, Divers J, Wang W, Frank SG, Namjou B, Gabriel SB, Lee AT, Gregersen PK, Behrens TW, Taylor KE, Fernando M, Zidovetzki R, Gaffney PM, Edberg JC, Rioux JD, Ojwang JO, James JA, Merrill JT, Gilkeson GS, Seldin MF, Yin H, Baechler EC, Li QZ, Wakeland EK, Bruner GR, Kaufman KM, and Kelly JA (February, 2008), *Genome-wide association scan in women with systemic lupus erythematosus identifies susceptibility variants in ITGAM, PXK, KIAA1542 and other loci*, Nature Genetics 40(2): 204-210

International Union Against Cancer (1960), *Clinical Stage Classification*, British Journal of Surgery 47: 330

Irvin MR, Wineinger NE, Rice TK, Pajewski NM, Kabagambe EK, Gu CC, Pankow J, North KE, Wilk JB, Freedman BI, Franceschini N, Broeckel U, Tiwari HK, and Arnett DK (2011), *Genome-wide detection of allele specific copy number variation associated with insulin resistance in African Americans from the HyperGEN study*, PLoS One 6(8): e24052

Jarick I, Vogel CI, Scherag S, Schäfer H, Hebebrand J, Hinney A, and Scherag A (February, 2011), *Novel common copy number variation for early onset extreme obesity on chromosome 11q11 identified by a genome-wide analysis*, Human Molecular Genetics 20(4): 840-852

Personalis EX2155

Jasmine F, Ahsan H, Andrulis IL, John EM, Chang-Claude J, and Kibriya MG (December, 2008), *Whole-genome amplification enables accurate genotyping for microarray-based high-density single nucleotide polymorphism array*, Cancer Epidemiology, Biomarkers, and Prevention 17(12): 3499-3508

Jedrzejowska M, Milewski M, Zimowski J, Borkowska J, Kostera-Pruszczyk A, Sielska D, Jurek M, and Hausmanowa-Petrusewicz I (2009), *Phenotype modifiers of spinal muscular atrophy: the number of SMN2 gene copies*, deletion in the NAIP gene and probably gender influence the course of the disease, Acta Biochimica Polonica 56(1): 103-108

Jiang WW, Zahurak M, Goldenberg D, Milman Y, Park HL, Westra WH, Koch W, Sidransky D, and Califano J (December, 2006), *Increased plasma DNA integrity index in head and neck cancer patients*, International Journal of Cancer 119(11): 2673-2676

Jin G, Sun J, Liu W, Zhang Z, Chu LW, Kim ST, Sun J, Feng J, Duggan D, Carpten JD, Wiklund F, Grönberg H, Isaacs WB, Zheng SL, Xu J (May, 2011), *Genome-wide copy-number variation analysis identifies common genetic variants at 20p13 associated with aggressiveness of prostate cancer*, Carcinogenesis 32(7): 1057-1062

Johannsdottir HK, Jonsson G, Johannesdottir G, Agnarsson BA, Eerola H, Arason A, Heikkila P, Egilsson V, Olsson H, Johannsson OT, Nevanlinna H, Borg A, and Barkardottir RB (September, 2006), *Chromosome 5 imbalance mapping in breast tumors from BRCA1 and BRCA2 mutation carriers and sporadic breast tumors*, International Journal of Cancer 119(5): 1052-1060

Joliffe IT (2002), Principal Components Analysis (second edition), Springer, 233 Spring Street, NY 10013, USA, ISBN 0387954422

Personalis EX2155

Jung K, Fleischhacker M, and Rabien A (November, 2010), *Cell-free DNA in the blood as a solid tumor biomarker--a critical appraisal of the literature*, Clinica Chimica Acta 411(21-22): 1611-1624

Kai L, Wang J, Ivanovic M, Chung YT, Laskin WB, Schulze-Hoepfner F, Mirochnik Y, Satcher RL Jr, and Levenson AS (February, 2011), *Targeting prostate cancer angiogenesis through metastasis-associated protein 1 (MTA1)*, The Prostate 71(3): 268-280

Kaiser HF (1960), *The application of electronic computers to factor analysis*, Educational and Psychological Measurement 20: 141-151

Kanehisa M and Goto S (January, 2000), *KEGG: Kyoto Encyclopedia of Genes and Genomes*, Nucleic Acids Research 28(1): 27-30

Karolchik D, Hinrichs AS, Furey TS, Roskin KM, Sugnet CW, Haussler D, and Kent WJ (January, 2004), *The UCSC table browser data retrieval tool*, Nucleic Acids Research 32

Karpuzcu M and Şeneş S (1987), *Design of monitoring systems for water quality by principal component analysis and a case study*, Proceedings of the International Symposium on Environmental Management 1: 673-690

Karrison TG, Ferguson DJ, and Meier P (January, 1999), *Dormancy of mammary carcinoma after mastectomy*, Journal of the National Cancer Institute 91(1): 80-85

Katoh Y and Katoh M (May, 2006), *Comparative integromics on BMP/GDF family*, International Journal of Molecular Medicine 17(5): 951-955

Key TJ, Verkasalo PK, and Banks E (March, 2001), *Epidemiology of breast cancer*, Lancet Oncology 2(3): 133-140

Personalis EX2155

Keydar I, Chen L, Karby S, Weiss FR, Delarea J, Radu M, Chaitcik S, and Brenner HJ (May, 1979), *Establishment and characterization of a cell line of human breast carcinoma origin*, European Journal of Cancer 15(5): 659-670

Kibriya MG, Jasmine F, Argos M, Andrulis IL, John EM, Chang-Claude J, and Ahsan H (April, 2009), *A pilot genome-wide association study of early-onset breast cancer*, Breast Cancer Research and Treatment 114(3): 463-477

Kim ST, Cheng Y, Hsu FC, Jin T, Kader AK, Zheng SL, Isaacs WB, Xu J, and Sun J (December, 2010), *Prostate cancer risk-associated variants reported from genome-wide association studies: meta-analysis and their contribution to genetic variation*, The Prostate 70(16): 1729-1738

Kim JH, Kang HS, Kim TW, and Kim SJ (October, 2011), *Differential methylation hybridization profiling identifies involvement of STAT1-mediated pathways in breast cancer*, International Journal of Oncology 39(4): 955-963

Kirov G, Grozeva D, Norton N, Ivanov D, Mantripragada KK, Holmans P, International Schizophrenia Consortium, Wellcome Trust Case Control Consortium, Craddock N, Owen MJ, and O'Donovan MC (January, 2009), *Support for the involvement of large copy number variants in the pathogenesis of schizophrenia*, Human Molecular Genetics 18(8): 1497-1503

Köchert K, Ullrich K, Kreher S, Aster JC, Kitagawa M, Jöhrens K, Anagnostopoulos I, Jundt F, Lamprecht B, Zimber-Strobl U, Stein H, Janz M, Dörken B, and Mathas S. (November, 2010), *High-level expression of Mastermind-like 2 contributes to aberrant activation of the NOTCH signaling pathway in human lymphomas*, Oncogene 30(15): 1831-1840

Koivisto M, Herva R, and Linna SL (1981), *Serial duplication of 10 (q21 to q22) in a mentally retarded boy with congenital malformations*, Human Genetics 57(2): 224-225

Personalis EX2155

Korabecna M, Opatrna S, Wirth J, Rulcova K, Eiselt J, Sefrna F, and Horinek A (August, 2008), *Cell-free plasma DNA during peritoneal dialysis and hemodialysis and in patients with chronic kidney disease*, Annals of the New York Academy of Sciences 1137: 296-301

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, Lee C, Nizzari MM, Gabriel SB, Purcell S, Daly MJ, and Altshuler D (October, 2008), *Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs*, Nature Genetics 40(10): 1253-1260

Kösters JP and Gøtzsche PC (2003), *Regular self-examination or clinical examination for early detection of breast cancer*, Cochrane Database of Systematic Reviews 2003(2): CD003373

Koyama S, Wada-Hiraike O, Nakagawa S, Tanikawa M, Hiraike H, Miyamoto Y, Sone K, Oda K, Fukuhara H, Nakagawa K, Kato S, Yano T, and Taketani Y (February, 2010), *Repression of estrogen receptor beta function by putative tumor suppressor DBC1*, Biochemical and Biophysical Research Communications 392(3): 357-362

Kruglyak L and Nickerson DA (March, 2001), *Variation is the spice of life*, Nature Genetics 27(3): 234-236

Kupfer SS, Anderson JR, Hooker S, Skol A, Kittles RA, Keku TO, Sandler RS, and Ellis NA (November, 2010), *Genetic heterogeneity in colorectal cancer associations between African and European Americans*, Gastroenterology 139(5): 1677-1685

Kunimoto Y, Nakano S, Kataoka H, Shimada Y, Oshimura M, and Kitano H (October, 2011), *Deleted in Esophageal Cancer 1(DEC1) is down-regulated and contributes to migration in head and neck squamous cell carcinoma cell lines*, Journal for Otorhinolaryngology and its Related Specialities 73(1): 17-23

Personalis EX2155

Kunz G (March, 2011), *Use of a genomic test (MammaPrint™) in daily clinical practice to assist in risk stratification of young breast cancer patients*, Archives in Gynecology and Obstetrics 283(3): 597-602

Kwei KA, Kung Y, Salari K, Holcomb IN, and Pollack JR (June, 2010), *Genomic instability in breast cancer: pathogenesis and clinical implications*, Molecular Oncology 4(3): 255-266

LaFramboise T (July, 2009), *Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances*, Nucleic Acids Research 37(13): 4181-4193

Langdon JA, Lamont JM, Scott DK, Dyer S, Prebble E, Bown N, Grundy RG, Ellison DW, and Clifford SC (January, 2006), *Combined genome-wide allelotyping and copy number analysis identify frequent genetic losses without copy number reduction in medulloblastoma*, Genes, Chromosomes, and Cancer 45(1): 47-60

Lebeau A, Deimling D, Kaltz C, Sendelhofert A, Iff A, Luthardt B, Untch M, and Löhrs U (January, 2001), *Her-2/neu analysis in archival tissue samples of human breast cancer: comparison of immunohistochemistry and fluorescence in situ hybridization*, Journal of Clinical Oncology 19(2): 354-363

Ledergerber C and Dessimoz C (September, 2011), *Base-calling for next-generation sequencing platforms*, Briefings in Bioinformatics 12(5): 489-497

Le Guellec S, Perallon R, Alunni JP, Charitansky H, Leaha C, Gonzalez AM, Chateau MC, Simony-Lafontaine J, Jacot W, Gutowski M, Penault-Llorca F, Dalenc F, and Lacroix-Triki M (December, 2011), *Neoadjuvant treatment of breast cancer: implications for the pathologist*, Annales de Pathologie 31(6): 442-454

Personalis EX2155

Lee TH, Montalvo L, Chrebtow V, and Busch MP (February, 2001), *Quantitation of genomic DNA in plasma and serum samples: higher concentrations of genomic DNA found in serum than in plasma*, Transfusion 41(2): 276-282

Leon SA, Shapiro B, Sklaroff DM, and Yaros MJ (March, 1977), *Free DNA in the serum of cancer patients and the effect of therapy*, Cancer Research 37(3): 646-650

Leong AS and Zhuang Z (June, 2011), *The changing role of pathology in breast cancer diagnosis and treatment*, Pathobiology 78(2): 99-114

Levenson AS and Jordan VC (August, 1997), *MCF-7: the first hormone-responsive breast cancer cell line*, Cancer Research 57(15): 3071-3078

Levenson VV (June, 2007), *Biomarkers for early detection of breast cancer: what, when, and where?*, Biochimica et Biophysica Acta 1770(6): 847-856

Li CN, Hsu HL, Wu TL, Tsao KC, Sun CF, and Wu JT (2003), *Cell-free DNA is released from tumor cells upon cell death: a study of tissue cultures of tumor cell lines*, Journal of Clinical Laboratory Analysis 17(4): 103-107

Li L, Fu YC, Xu JJ, Chen XC, Lin XH, and Luo LL (November, 2011), *Caloric restriction promotes the reproductive capacity of female rats via modulating the level of insulin-like growth factor-1 (IGF-1)*, General and Comparative Endocrinology 174(2): 232-237

Li T, Yang Y, Li X, Xu C, and Meng L (February, 2012), *EGFR- and AKT-mediated reduction in PTEN expression contributes to tyrphostin resistance and is reversed by mTOR inhibition in endometrial cancer cells*, Molecular and Cellular Biochemistry 361(1-2): 19-29

Personalis EX2155

Lianidou ES and Markou A (July, 2011), *Circulating tumor cells in breast cancer: detection systems, molecular characterization, and future challenges*, Clinical Chemistry 57(9): 1242-1255

Lieberfarb ME, Lin M, Lechpammer M, Li C, Tanenbaum DM, Febbo PG, Wright RL, Shim J, Kantoff PW, Loda M, Meyerson M, and Sellers WR (August, 2003), *Genome-wide loss of heterozygosity analysis from laser capture microdissected prostate cancer using single nucleotide polymorphic allele (SNP) arrays and a novel bioinformatics platform dChipSNP*, Cancer Research 63(16): 4781-4785

Lin Z and Altman RB (November, 2004), *Finding haplotype tagging SNPs by use of principal components analysis*, American Journal of Human Genetics 75(5): 850-861

Lindström S, Vachon CM, Li J, Varghese J, Thompson D, Warren R, Brown J, Leyland J, Audley T, Wareham NJ, Loos RJ, Paterson AD, Rommens J, Waggott D, Martin LJ, Scott CG, Pankratz VS, Hankinson SE, Hazra A, Hunter DJ, Hopper JL, Southey MC, Chanock SJ, Silva Idos S, Liu J, Eriksson L, Couch FJ, Stone J, Apicella C, Czene K, Kraft P, Hall P, Easton DF, Boyd NF, and Tamimi RM (March, 2011), *Common variants in ZNF365 are associated with both mammographic density and breast cancer risk*, Nature Genetics 43(3): 185-187

Linos E, Spanos D, Rosner BA, Linos K, Hesketh T, Qu JD, Gao YT, Zheng W, and Colditz GA (October, 2008), *Effects of reproductive and demographic changes on breast cancer incidence in China: a modelling analysis*, Journal of the National Cancer Institute 100(19): 1352-1360

Liu J, Hoppman N, O'Connell JR, Wang H, Streeten EA, McLenithan JC, Mitchell BD, and Shuldiner AR (in press), *A functional haplotype in EIF2AK3, an ER stress sensor, is associated with lower bone mineral density*, Journal of Bone and Mineral Research

Personalis EX2155

Lorain S, Quivy JP, Monier-Gavelle F, Scamps C, Lécluse Y, Almouzni G, and Lipinski M (September, 1998), *Core histones and HIRIP3, a novel histone-binding protein, directly interact with WD repeat protein HIRA*, Molecular and Cellular Biology 18(9): 5546-5556

Lu T, Hano H, Meng C, Nagatsuma K, Chiba S, and Ikegami M (February, 2007), *Frequent loss of heterozygosity in two distinct regions, 8p23.1 and 8p22, in hepatocellular carcinoma*, World Journal of Gastroenterology 13(7): 1090-1097

Maathuis M (October, 2008), *Principal component Analysis (PCA)*, http://stat.ethz.ch/~maathuis/teaching/fall08/Notes3.pdf

Madsen T (2007), *Multivariate data analysis with PCA, CA and MS*, http://www.archaeoinfo.dk/PDF%20files/Multivariate%20data%20analysis.pdf

Mahloch JL (1974), *Multivariate Techniques for Water Quality Analysis*, Journal of the Environmental Engineering Division 100(5): 1119-1132

Maller J, George S, Purcell S, Fagerness J, Altshuler D, Daly MJ, and Seddon JM (September, 2006), *Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration*, Nature Genetics 38(9): 1055-1059

Mandel P and Metais P (1948), *Les acides nucléiques du plasma sanguin chez l'homme*, Comptes Rendus de l'Académie des Sciences de Paris 142: 241–243

Mandelblatt JS, Cronin KA, Bailey S, Berry DA, de Koning HJ, Draisma G, Huang H, Lee SJ, Munsell M, Plevritis SK, Ravdin P, Schechter CB, Sigal B, Stoto MA, Stout NK, van Ravesteyn NT, Venier J, Zelen M, and Feuer EJ (November, 2009), *Effects of mammography screening under different screening schedules: model estimates of potential benefits and harms*, Annals of Internal Medicine 151(10): 738747

Personalis EX2155

Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ, McCarthy MI, Ramos EM, Cardon LR, Chakravarti A, Cho JH, Guttmacher AE, Kong A, Kruglyak L, Mardis E, Rotimi CN, Slatkin M, Valle D, Whittemore AS, Boehnke M, Clark AG, Eichler EE, Gibson G, Haines JL, Mackay TF, McCarroll SA, and Visscher PM (October, 2009), *Finding the missing heritability of complex diseases*, Nature 461(7265): 747-753

Manuvakhova M, Thottassery JV, Hays S, Qu Z, Rentz SS, Westbrook L, and Kern FG (September, 2007), *Expression of the SNT-1/FRS2 phosphotyrosine binding domain inhibits activation of MAP kinase and PI3-kinase pathways and antiestrogen resistant growth induced by FGF-1 in human breast carcinoma cells*, Oncogene 25(44): 6003-6014

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, and Rothberg JM (September, 2005), *Genome sequencing in microfabricated high-density picolitre reactors*, Nature 437(7057): 376-380

Mazlum N, Özer A, and Mazlum S (1999), *Interpretation of Water Quality Data by Principal Components*, Turkish Journal of Engineering and Environmental Sciences 23: 19-26

Mazumdar A, Wang RA, Mishra SK, Adam L, Bagheri-Yarmand R, Mandal M, Vadlamudi RK, and Kumar R (January, 2001), *Transcriptional repression of oestrogen receptor by metastasis-associated protein 1 corepressor*, Nature Cell Biology 3(1): 30-37

Personalis EX2155

McCormack VA and dos Santos Silva I (June, 2006), *Breast density and parenchymal patterns as markers of breast cancer risk: a meta-analysis*, Cancer Epidemiology, Biomarkers, and Prevention 15(6): 1159-1169

McKian KP, Reynolds CA, Visscher DW, Nassar A, Radisky DC, Vierkant RA, Degnim AC, Boughey JC, Ghosh K, Anderson SS, Minot D, Caudill JL, Vachon CM, Frost MH, Pankratz VS, and Hartmann LC (December, 2009), *Novel breast tissue feature strongly associated with risk of breast cancer*, Journal of Clinical Oncology 27(35): 5893-5898

Meldrum C, Doyle MA, and Tothill RW (November, 2011), *Next-generation sequencing for cancer diagnostics: a practical perspective*, The Clinical Biochemist Reviews 32(4): 177-195

Meng Z, Zaykin DV, Xu CF, Wagner M, and Ehm MG (July, 2003), *Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes*, American Journal of Human Genetics 73(1): 115-130

Meng S, Tripathy D, Frenkel EP, Shete S, Naftalis EZ, Huth JF, Beitsch PD, Leitch M, Hoover S, Euhus D, Haley B, Morrison L, Fleming TP, Herlyn D, Terstappen LW, Fehm T, Tucker TF, Lane N, Wang J, and Uhr JW (December, 2004), *Circulating tumor cells in patients with breast cancer dormancy*, Clinical Cancer Research 10(24): 8152-8162

Merla G, Ucla C, Guipponi M, and Reymond A (May, 2002), *Identification of additional transcripts in the Williams-Beuren syndrome critical region*, Human Genetics 110(5): 429-438

Mervis CB (2008), *7q11.23 Duplication Syndrome*, available at http://louisville.edu/psychology/mervis/research/dup.html, last accessed 28th March 2012

Personalis EX2155

Minchin RF, Carpenter D, and Orr RJ (March, 2001), *Polyinosinic acid and polycationic liposomes attenuate the hepatic clearance of circulating plasmid DNA*, The Journal of Pharmacology and Experimental Therapeutics 296(3): 1006-1012

Mistry M, Parkin DM, Ahmad AS, and Sasieni P (October, 2011), *Cancer incidence in the United Kingdom: projections to the year 2030*, British Journal of Cancer 105(11): 1795-803

Moch H, Blank PR, Dietel M, Elmberger G, Kerr KM, Palacios J, Penault-Llorca F, Rossi G, and Szucs TD (January, 2012), *Personalized cancer medicine and the future of pathology*, Virchows Archiv 460(1): 3-8

Modi WS (April, 2004), *CCL3L1 and CCL4L1 chemokine genes are located in a segmental duplication at chromosome 17q12*, Genomics 83(4): 735-738

Møller H, Fairley L, Coupland V, Okello C, Green M, Forman D, Møller B, and Bray F (May, 2007), *The future burden of cancer in England: incidence and numbers of new patients in 2020*, British Journal of Cancer 96(9): 1484-1488

Moore TM, Garg R, Johnson C, Coptcoat MJ, Ridley AJ, and Morris JD (February, 2000), *PSK, a novel STE20-like kinase derived from prostatic carcinoma that activates the c-Jun N-terminal kinase mitogen-activated protein kinase pathway and regulates actin cytoskeletal organization*, The Journal of Biological Chemistry 275(6): 4311-4322

Mortusewicz O, Fouquerel E, Amé JC, Leonhardt H, and Schreiber V (March, 2011), *PARG is recruited to DNA damage sites through poly (ADP-ribose)- and PCNA-dependent mechanisms*, Nucleic Acids Research 39(12): 5045-5056

Mount DW (2004), Bioinformatics: Sequence and Genome Analysis (second edition), Cold Spring Harbor Laboratory Press, 1 Bungtown Road, Cold Spring Harbor, NY 11724, USA, ISBN 0879697121

Personalis EX2155

Mühlethaler-Mottet A, Flahaut M, Bourloud KB, Nardou K, Coulon A, Liberman J, Thome M, and Gross N (March, 2011), *Individual caspase-10 isoforms play distinct and opposing roles in the initiation of death receptor-mediated tumour cell apoptosis*, Cell Death and Disease 2: e125

Mukohara T (in press), *Role of HER2-Targeted Agents in Adjuvant Treatment for Breast Cancer*, Chemotherapy Research and Practice

Murray C (February, 1995), *Tumour dormancy: not so sleepy after all*, Nature Medicine 1(2): 117-118

Murtagh MJ, Demir I, Harris JR, and Burton PR (June, 2011), *Realizing the promise of population biobanks: a new model for translation*, Human Genetics 130(3): 333-145

Musolino A, Naldi N, Bortesi B, Pezzuolo D, Capelletti M, Missale G, Laccabue D, Zerbini A, Camisa R, Bisagni G, Neri TM, and Ardizzoni A (April, 2008), *Immunoglobulin G fragment C receptor polymorphisms and clinical efficacy of trastuzumab-based therapy in patients with HER-2/neu-positive metastatic breast cancer*, Journal of Clinical Oncology 26(11): 1789-1796

Närvä E, Autio R, Rahkonen N, Kong L, Harrison N, Kitsberg D, Borghese L, Itskovitz-Eldor J, Rasool O, Dvorak P, Hovatta O, Otonkoski T, Tuuri T, Cui W, Brüstle O, Baker D, Maltby E, Moore HD, Benvenisty N, Andrews PW, Yli-Harja O, and Lahesmaa R (April, 2010), *High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity*, Nature Biotechnology 28(4): 371-377

Nawroz H, Koch W, Anker P, Stroun M, Sidransky D (September, 1996), *Microsatellite alterations in serum DNA of head and neck cancer patients*, Nature Medicine 2(9): 1035-1037

Personalis EX2155

Naylor TL, Greshock J, Wang Y, Colligon T, Yu QC, Clemmer V, Zaks TZ, and Weber BL (November, 2005), *High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization*, Breast Cancer Research 7(6): R1186-1198

Nilsson M, Meza-Zepeda LA, Mertens F, Forus A, Myklebost O, and Mandahl N (April, 2004), *Amplification of chromosome 1 sequences in lipomatous tumors and other sarcomas*, International Journal of Cancer 109(3): 363-369

Nishiwaki T, Daigo Y, Kawasoe T, and Nakamura Y (February, 2000), *Isolation and mutational analysis of a novel human cDNA, DEC1 (deleted in esophageal cancer 1), derived from the tumor suppressor locus in 9q32*, Genes, Chromosomes, and Cancer 27(2): 169-176

Nishiyama H, Takahashi T, Kakehi Y, Habuchi T, and Knowles MA (October, 1999), *Homozygous deletion at the 9q32-33 candidate tumor suppressor locus in primary human bladder cancer*, Genes, Chromosomes, and Cancer 26(2): 171-175

Nonaka M, Ma BY, Imaeda H, Kawabe K, Kawasaki N, Hodohara K, Kawasaki N, Andoh A, Fujiyama Y, and Kawasaki T (June, 2011), *Dendritic cell-specific intercellular adhesion molecule 3-grabbing non-integrin (DC-SIGN) recognizes a novel ligand, Mac-2-binding protein, characteristically expressed on human colorectal carcinomas*, The Journal of Biological Chemistry 286(25): 22403-22413

O'Neill JS, Elton RA, and Miller WR (March, 1988), *Aromatase activity in adipose tissue from breast quadrants: a link with tumour site*, British Medical Journal 296(6624): 741-743

Oikawa M, Yoshiura K, Kondo H, Miura S, Nagayasu T, and Nakashima M (December, 2011), *Significance of genomic instability in breast cancer in atomic bomb*

Personalis EX2155

survivors: analysis of microarray-comparative genomic hybridization, Radiation Oncology 6(168)

Ozeri-Galai E, Lebofsky R, Rahat A, Bester AC, Bensimon A, and Kerem B (July, 2011), *Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites*, Molecular Cell 43(1): 122-131

Paige AJ, Taylor KJ, Taylor C, Hillier SG, Farrington S, Scott D, Porteous DJ, Smyth JF, Gabra H, and Watson JE (September, 2001), *WWOX: a candidate tumor suppressor gene involved in multiple tumor types*, Proceedings of the National Academy of Sciences of the USA 98(20): 11417-11422

Page K, Powles T, Slade MJ, De Bella MT, Walker RA, Coombes RC, and Shaw JA (September, 2006), *The importance of careful blood processing in isolation of cell-free DNA*, Annals of the New York Academy of Sciences 1075: 313-317

Page K, Hava N, Ward B, Brown J, Guttery DS, Ruangpratheep C, Blighe K, Sharma A, Walker RA, Coombes RC, and Shaw JA (April, 2011), *Detection of HER2 amplification in circulating free DNA in patients with breast cancer*, British Journal of Cancer 104(8): 1342-1348

Palmieri C, Coombes RC, Kim S-B, and Cleator S (July, 2008), *Ethnicity and breast cancer research*, The Lancet 372(9634): 188-189

Pangon L, Sigglekow ND, Larance M, Al-Sohaily S, Mladenova DN, Selinger CI, Musgrove EA, and Kohonen-Corish MR (September, 2010), *The "Mutated in Colorectal Cancer" Protein Is a Novel Target of the UV-Induced DNA Damage Checkpoint*, Genes and Cancer 1(9): 917-926

Papadopoulou E, Davilas E, Sotiriou V, Georgakopoulos E, and Georgakopoulou S (September, 2006), *Cell-free DNA and RNA in plasma as a new molecular marker for prostate and breast cancer*, Annals of the New York Academy of Sciences 1075: 235-243

Personalis EX2155

Park I, Son HK, Che ZM, and Kim J (February, 2012), *A novel gain-of-function mutation of TGF-β receptor II promotes cancer progression via delayed receptor internalization in oral squamous cell carcinoma*, Cancer Letters 315(2): 161-169

Parker JS, Mullins M, Cheang MC, Leung S, Voduc D, Vickery T, Davies S, Fauron C, He X, Hu Z, Quackenbush JF, Stijleman IJ, Palazzo J, Marron JS, Nobel AB, Mardis E, Nielsen TO, Ellis MJ, Perou CM, and Bernard PS (March, 2009, *Supervised risk predictor of breast cancer based on intrinsic subtypes*, Journal of Clinical Oncology 27(8): 1160-1167

Parkin DM, Boyd L, and Walker LC (December, 2011), *The fraction of cancer attributable to lifestyle and environmental factors in the UK in 2010*, British Journal of Cancer 105(Supplementary 2): 77-81

Partek Inc. (2009), *Partek® Documentation - turning data into discovery®*, Partek Inc., 12747 Olive Boulevard, Suite 205, St. Louis, MO 63141, USA, available at http://www.partek.com/, last accessed 27th March 2012

Parsons KJ, Cooper WJ, and Albertson RC (November, 2009), *Limits of principal components analysis for producing a common trait space: implications for inferring selection, contingency, and chance in evolution*, PLoS One 4(11): 7957

Paschou P, Ziv E, Burchard EG, Choudhry S, Rodriguez-Cintron W, Mahoney MW, and Drineas P (September, 2007), *PCA-correlated SNPs for structure identification in worldwide human populations*, PLoS Genetics 3(9):1672-86

Paschou P, Lewis J, Javed A, and Drineas P (December, 2010), *Ancestry informative markers for fine-scale individual assignment to worldwide populations*, Journal of Medical Genetics 47(12): 835-847

Personalis EX2155

Paterlini-Brechot P and Benali NL (August, 2007), *Circulating tumor cells (CTC) detection: clinical impact and future directions*, Cancer Letters, 253(2): 180-204

Patil CK, Mian IS, and Campisi J (October, 2005), *The thorny path linking cellular senescence to organismal aging*, Mechanisms of Ageing and Development 126(10): 1040-1045

Pauletti G, Dandekar S, Rong H, Ramos L, Peng H, Seshadri R, and Slamon DJ (November, 2000), *Assessment of methods for tissue-based detection of the HER-2/neu alteration in human breast cancer: a direct comparison of fluorescence in situ hybridization and immunohistochemistry*, Journal of Clinical Oncology 18(21): 3651-3664

Payne R, Hava N, Page K, Blighe K, Ward B, Slade M, Brown J, Guttery D, Zaidi SAA, Stebbing J, Jacob J, Tat T, Yagüe E, Shaw JA, and Coombes RC (January, 2012), The presence of disseminated tumour cells in the bone marrow is inversely related to circulating-free DNA in plasma in breast cancer dormancy, British Journal of Cancer 106(2): 375-382

Pereg Y, Liu BY, O'Rourke KM, Sagolla M, Dey A, Komuves L, French DM, and Dixit VM (April, 2010), *Ubiquitin hydrolase Dub3 promotes oncogenic transformation by stabilizing Cdc25A*, Nature Cell Biology 12(4): 400-406

Perotti D, De Vecchi G, Testi MA, Lualdi E, Modena P, Mondini P, Ravagnani F, Collini P, Di Renzo F, Spreafico F, Terenziani M, Sozzi G, Fossati-Bellani F, and Radice P (November, 2004), *Germline mutations of the POU6F2 gene in Wilms tumors with loss of heterozygosity on chromosome 7p14*, Human Mutations 24(5): 400-407

Perou CM, Sørlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, Fluge O, Pergamenschikov A, Williams C, Zhu SX, Lønning PE, Børresen-Dale AL, Brown PO, Botstein D (August, 2000), Molecular portraits of human breast tumours, Nature 406(6797): 747-752

Personalis EX2155

Pestrin M, Bessi S, Galardi F, Truglia M, Biggeri A, Biagioni C, Cappadona S, Biganzoli L, Giannini A, and di Leo A (2009), *Correlation of HER2 status between primary tumors and corresponding circulating tumor cells in advanced breast cancer patients*, Breast Cancer Research and Treatment 118: 523-530

Pharoah PD, Day NE, Duffy S, Easton DF, and Ponder BA (May, 1997), *Family history and the risk of breast cancer: a systematic review and meta-analysis*, International Journal of Cancer 71(5): 800-809

Pietras A, von Stedingk K, Lindgren D, Påhlman S, and Axelson H (May, 2011), *JAG2 induction in hypoxic tumor cells alters Notch signaling and enhances endothelial cell tube formation*, Molecular Cancer Research 9(5): 626-636

Pinzani P, Salvianti F, Zaccara S, Massi D, De Giorgi V, Pazzagli M, and Orlando C (November, 2011), *Circulating cell-free DNA in plasma of melanoma patients: qualitative and quantitative considerations*, Clinica Chimica Acta 412(23-24): 2141-2145

Piotrowski A, Bruder CE, Andersson R, Diaz de Ståhl T, Menzel U, Sandgren J, Poplawski A, von Tell D, Crasto C, Bogdan A, Bartoszewski R, Bebok Z, Krzyzanowski M, Jankowski Z, Partridge EC, Komorowski J, and Dumanski JP (September, 2008), *Somatic mosaicism for copy number variation in differentiated human tissues*, Human Mutation 29(9): 1118-1124

Porreca GJ, Zhang K, Li JB, Xie B, Austin D, Vassallo SL, LeProust EM, Peck BJ, Emig CJ, Dahl F, Gao Y, Church GM, and Shendure J (November, 2007), *Multiplex amplification of large sets of human exons*, Nature Methods 4(11): 931-936

Pruitt KD, Tatusova T, and Maglott DR (January, 2005), *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins*, Nucleic Acids Research 1(33) (Database Issue; D501-4)

Personalis EX2155

Qian Y, Jung YS, and Chen X (April, 2011), *ΔNp63, a target of DEC1 and histone deacetylase 2, modulates the efficacy of histone deacetylase inhibitors in growth suppression and keratinocyte differentiation*, The Journal of Biological Chemistry 286(14): 12033-12041

Quackenbush J (December, 2002), *Microarray data normalization and transformation*, Nature Genetics 32: 496-501

Quintarelli C, Dotti G, Hasan ST, De Angelis B, Hoyos V, Errichiello S, Mims M, Luciano L, Shafer J, Leen AM, Heslop HE, Rooney CM, Pane F, Brenner MK, and Savoldo B (March, 2011), *High-avidity cytotoxic T lymphocytes specific for a new PRAME-derived peptide can target leukemic and leukemic-precursor cells*, Blood 117(12): 3353-3362

R Development Core Team (2008), *R: a language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria, ISBN 3900051070, http://www.R-project.org/

Radhakrishnan A, Badhrinarayanan N, Biswas J, and Krishnakumar S (October, 2009), *Analysis of chromosomal aberration (1, 3, and 8) and association of microRNAs in uveal melanoma*, Molecular Vision 15: 2146-2154

Rakha EA and Ellis IO (July, 2011), *Modern classification of breast cancer: should we stick with morphology or convert to molecular profile characteristics*, Advances in Anatomical Pathology, 18)4): 255-267

Ray R, Mitra N, Baldwin D, Guo M, Patterson F, Heitjan DF, Jepson C, Wileyto EP, Wei J, Payne T, Ma JZ, Li MD, and Lerman C (May, 2010), *Convergent evidence that choline acetyltransferase gene variation is associated with prospective smoking cessation and nicotine dependence*, Neuropsychopharmacology 35(6): 1374-1382

Personalis EX2155

Redon R, Ishikawa S, Fitch KR, Feuk L, Perry GH, Andrews TD, Fiegler H, Shapero MH, Carson AR, Chen W, Cho EU, Dallaire S, Freeman JL, González JR, Gratacós M, Huang J, Kalaitzopoulos D, Komura D, MacDonald JR, Marshall CR, Mei R, Montgomery, Nishimura K, Okamura K, Shen F, Somerville MJ, Tchinda J, Valsesia A, Woodwark C, Yang F, Zhang J, Zerjal T, Zhang J, Armengols L, Conrad DF, Estivill X, Tyler-Smith C, Carter NP, Aburatani H, Lee C, Jones KW, Scherer SW, and Hurles ME (November, 2006), *Global variation in copy number in the human genome*, Nature 444(7118): 444-454

Reich DE, Cargill M, Bolk S, Ireland J, Sabeti PC, Richter DJ, Lavery T, Kouyoumjian R, Farhadian SF, Ward R, and Lander ES (May, 2001), *Linkage disequilibrium in the human genome*, Nature 411(6834): 199-204

Rème T, Hose D, De Vos J, Vassal A, Poulain PO, Pantesco V, Goldschmidt H, and Klein B (January, 2008), *A new method for class prediction based on signed-rank algorithms applied to Affymetrix microarray experiments*, BMC Bioinformatics 9 (16)

Rennstam K, Ringberg A, Cunliffe HE, Olsson H, Landberg G, and Hedenfalk I (January, 2010), *Genomic alterations in histopathologically normal breast tissue from BRCA1 mutation carriers may be caused by BRCA1 haploinsufficiency*, Genes, Chromosomes, and Cancer 49(1): 78-90

Richards M (October, 2011), *An independent review is under way*, British Medical Journal 343: d6843

Robinson D, Holmberg L, and Møller H (August, 2008), *The occurrence of invasive cancers following a diagnosis of breast carcinoma in situ*, British Journal of Cancer 99(4): 611-615

Rook MS, Delach SM, Deyneko G, Worlock A, and Wolfe JL (January, 2004), *Whole genome amplification of DNA from laser capture-microdissected tissue for high-*

Personalis EX2155

throughput single nucleotide polymorphism and short tandem repeat genotyping, The American Journal of Pathology 164(1): 23-33

Rosen PP and Groshen S (August, 1990), *Factors influencing survival and prognosis in early breast carcinoma (T1N0M0-T1N1M0). Assessment of 644 patients with median follow-up of 18 years*, The Surgical Clinics of North America 70(4): 937-962

Ross CW, Ouillette PD, Saddler CM, Shedden KA, and Malek SN (August, 2007), *Comprehensive analysis of copy number and allele status identifies multiple chromosome defects underlying follicular lymphoma pathogenesis*, Clinical Cancer Research 13(16): 4777-4785

Ross JS and Fletcher JA (1998), *The HER-2/neu oncogene in breast cancer: prognostic factor, predictive factor, and target for therapy*, Stem Cells 16(6): 413-428

Ross JS (July, 2009), *Multigene classifiers, prognostic factors, and predictors of breast cancer clinical outcome*, Advances in Anatomic Pathology 16(4): 204-215

Rozen S and Skaletsky HJ, *Primer3 on the WWW for general users and for biologist programmers*, In: Krawetz S and Misener S (2000), Bioinformatics: Methods and Protocols (Methods in Molecular Biology), Humana Press, Totowa, NJ USA, ISBN 0896037320

Rummukainen J, Kytölä S, Karhu R, Farnebo F, Larsson C, and Isola JJ (April, 2001), *Aberrations of chromosome 8 in 16 breast cancer cell lines by comparative genomic hybridization, fluorescence in situ hybridization, and spectral karyotyping*, Cancer Genetics and Cytogenetics 126(1): 1-7

Rumsby M, Schmitt J, Sharrard M, Rodrigues G, Stower M, and Maitland N (February, 2011), *Human prostate cell lines from normal and tumourigenic epithelia differ in the pattern and control of choline lipid headgroups released into the medium on stimulation of protein kinase C*, British Journal of Cancer 104(4): 673-684

Personalis EX2155

Sabatti C and Risch N (April, 2002), *Homozygosity and linkage disequilibrium*, Genetics 160(4): 1707-1719

Sachelarie I, Grossbard ML, Chadha M, Feldman S, Ghesani M, and Blum RH (June, 2006), *Primary systemic therapy of breast cancer*, Oncologist 11(6): 574-589

Saha Roy S and Vadlamudi RK (in press), *Role of estrogen receptor signaling in breast cancer metastasis*, International Journal of Breast Cancer

Sahin AA (May, 2000), *Biologic and clinical significance of HER-2/neu (cerbB-2) in breast cancer*, Advances in Atomic Pathology 7(3): 158-166

Sakazume S, Okamoto N, Yamamoto T, Kurosawa K, Numabe H, Ohashi Y, Kako Y, Nagai T, and Ohashi H (August, 2007), *GPC3 mutations in seven patients with Simpson-Golabi-Behmel syndrome*, American Journal of Medical Genetics Part A 143A(15): 1703-1707

Sanpaolo P, Barbieri V, and Genovesi D (October, 2011), *Prognostic value of breast cancer subtypes on breast cancer specific survival, distant metastases and local relapse rates in conservatively managed early stage breast cancer: a retrospective clinical study*, European Journal of Surgical Oncology 37(10): 876-882

Santos DC, Zaphiropoulos PG, Neto CF, Pimentel ER, Sanches JA Jr, and Ruiz IR (July, 2011), *PTCH1 gene mutations in exon 17 and loss of heterozygosity on D9S180 microsatellite in sporadic and inherited human basal cell carcinomas*, International Journal of Dermatology 50(7): 838-843

Scherer SW, Lee C, Birney E, Altshuler DM, Eichler EE, Carter NP, Hurles ME, and Feuk L (July, 2007), *Challenges and standards in integrating surveys of structural variation*, Nature Genetics 39(Supplement 7): S7-15

Personalis EX2155

Schirrmacher V (August, 2001), *T-cell immunity in the induction and maintenance of a tumour dormant state*, Seminars in Cancer Biology 11(4): 285-295

Schmidt-Kittler O, Ragg T, Daskalakis A, Granzow M, Ahr A, Blankenstein TJ, Kaufmann M, Diebold J, Arnholdt H, Muller P, Bischoff J, Harich D, Schlimok G, Riethmuller G, Eils R, and Klein CA (June, 2003), *From latent disseminated cells to overt metastasis: genetic analysis of systemic breast cancer progression*, Proceedings of the National Academy of Sciences 100(13): 7737-7742

Schülke JP, Wochnik GM, Lang-Rollin I, Gassen NC, Knapp RT, Berning B, Yassouridis A, and Rein T (July, 2010), *Differential impact of tetratricopeptide repeat proteins on the steroid hormone receptors*, Public Library of Sciences One 5(7): e11717

Schuster-Böckler B, Conrad D, and Bateman A (March, 2010), *Dosage sensitivity shapes the evolution of copy-number varied regions*, PLoS One 5(3): e9474

Schwarzenbach H, Hoon DS, and Pantel K (June, 2011), *Cell-free nucleic acids as biomarkers in cancer patients*, Nature Reviews, Cancer 11(6): 426-437

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Månér S, Massa H, Walker M, Chi M, Navin N, Lucito R, Healy J, Hicks J, Ye K, Reiner A, Gilliam TC, Trask B, Patterson N, Zetterberg A, and Wigler M (July, 2004), *Large-scale copy number polymorphism in the human genome*, Science 305(5683): 525-528

Sebat J, Lakshmi B, Malhotra D, Troge J, Lese-Martin C, Walsh T, Yamrom B, Yoon S, Krasnitz A, Kendall J, Leotta A, Pai D, Zhang R, Lee YH, Hicks J, Spence SJ, Lee AT, Puura K, Lehtimäki T, Ledbetter D, Gregersen PK, Bregman J, Sutcliffe JS, Jobanputra V, Chung W, Warburton D, King MC, Skuse D, Geschwind DH, Gilliam TC, Ye K, and Wigler M (April, 2007), *Strong association of de novo copy number mutations with autism*, Science 316(5823): 445-449

Personalis EX2155

Sellar GC, Watt KP, Rabiasz GJ, Stronach EA, Li L, Miller EP, Massie CE, Miller J, Contreras-Moreira B, Scott D, Brown I, Williams AR, Bates PA, Smyth JF, and Gabra H (July, 2003), *OPCML at 11q25 is epigenetically inactivated and has tumor-suppressor function in epithelial ovarian cancer*, Nature Genetics 34(3): 337-343

Sen S (January, 2000), *Aneuploidy and cancer*, Current Opinion in Oncology 12(1): 82-88

Senée V, Vattem KM, Delépine M, Rainbow LA, Haton C, Lecoq A, Shaw NJ, Robert JJ, Rooman R, Diatloff-Zito C, Michaud JL, Bin-Abbas B, Taha D, Zabel B, Franceschini P, Topaloglu AK, Lathrop GM, Barrett TG, Nicolino M, Wek RC, and Julier C (July, 2004), *Wolcott-Rallison Syndrome: clinical, genetic, and functional study of EIF2AK3 mutations and suggestion of genetic heterogeneity*, Diabetes 53(7): 1876-1883

Shao ZM and Nguyen M (September, 2006), *Tumor-specific DNA in plasma of breast cancer patients*, Anticancer Drugs 13(4): 353-357

Shapiro B, Chakrabarty M, Cohn EM, and Leon SA (June, 1983), *Determination of circulating DNA levels in patients with benign or malignant gastrointestinal disease*, Cancer 51(11): 2116-2120

Sharma G, Mirza S, Parshad R, Srivastava A, Gupta SD, Pandya P, and Ralhan R (February, 2011), *Clinical significance of maspin promoter methylation and loss of its protein expression in invasive ductal breast carcinoma: correlation with VEGF-A and MTA1 expression*, Tumour Biology 32(1): 23-32

Shaw JA, Walsh T, Chappell SA, Carey N, Johnson K, and Walker RA (June, 1996), *Microsatellite instability in early sporadic breast cancer*, British Journal of Cancer 73(11): 1393-1397

Personalis EX2155

Shaw JA, Smith BM, Walsh T, Johnson S, Primrose L, Slade MJ, Walker RA, and Coombes RC (March, 2000), *Microsatellite alterations in plasma DNA of primary breast cancer patients*, Clinical Cancer Research 6(3): 1119-1124

Shaw JA, Page K, Blighe K, Hava N, Guttery D, Ward B, Brown J, Ruangpratheep C, Stebbing J, Payne R, Palmieri C, Cleator S, Walker RA, and Coombes RC (February, 2012), *Genomic analysis of circulating cell free DNA infers breast cancer dormancy*, Genome Research 22(2): 220-231

Shawi M and Autexier C (January/February, 2008), *Telomerase, senescence and ageing*, Mechanisms of Ageing and Development 129(1-2): 3-10

Sheh A, Ge Z, Parry NM, Muthupalani S, Rager JE, Raczynski AR, Mobley MW, McCabe AF, Fry RC, Wang T, and Fox JG (June, 2011), *17β-estradiol and Tamoxifen prevent gastric cancer by modulating leukocyte recruitment and oncogenic pathways in Helicobacter pylori-infected INS-GAS male mice*, Cancer Prevention Research 4(9): 1426-1435

Shimokawa T, Masutani M, Nozaki T, Nakagama H, Araki S, Aoki Y, and Sugimura T (December, 1998), *The human poly(ADP-ribose) glycohydrolase maps to chromosome 10q11.23-21.1 by fluorescence in situ hybridization*, Human Cell 11(4): 243-246

Shuen AY and Foulkes WD (April, 2011), *Inherited mutations in breast cancer genes--risk and response*, Journal of Mammary Gland Biology and Neoplasia 16(1): 3-15

Sicher K and Waterhouse JA (December, 1973), *Evaluation of TNM classification of carcinoma of the breast*, British Journal of Cancer 28(6): 580-588

Sinha S, Singh RK, Alam N, Roy A, Roychoudhury S, and Panda CK (November, 2008), *Alterations in candidate genes PHF2, FANCC, PTCH1 and XPA at*

Personalis EX2155

chromosomal 9q22.3 region: pathological significance in early- and late-onset breast carcinoma, Molecular Cancer 7(84)

Slade MJ, Payne R, Riethdorf S, Ward B, Zaidi SA, Stebbing J, Palmieri C, Sinnett HD, Kulinskaya E, Pitfield T, McCormack RT, Pantel K, Coombes RC (January, 2009), *Comparison of bone marrow, disseminated tumour cells and blood-circulating tumour cells in breast cancer patients after primary treatment*, British Journal of Cancer 100(1): 160-166

Slamon DJ, Clark GM, Wong SG, Levin WJ, Ullrich A, and McGuire WL (January, 1987), *Human breast cancer: correlation of relapse and survival with amplification of the HER-2/neu oncogene*, Science 235(4785): 177-182

Smart CE, Wronski A, French JD, Edwards SL, Asselin-Labat ML, Waddell N, Peters K, Brewster BL, Brooks K, Simpson K, Manning N, Lakhani SR, Grimmond S, Lindeman GJ, Visvader JE, and Brown MA (March, 2011), *Analysis of Brca1-deficient mouse mammary glands reveals reciprocal regulation of Brca1 and c-kit*, Oncogene 30(13): 1597-1607

Smeets SJ, Harjes U, van Wieringen WN, Sie D, Brakenhoff RH, Meijer GA, and Ylstra B (June, 2011), *To DNA or not to DNA? That is the question, when it comes to molecular subtyping for the clinic!*, Clinical Cancer Research 17(15): 4959-4964

Smith LI (February, 2002), *A tutorial on Principal Components Analysis*, http://www.cs.otago.ac.nz/cosc453/student_tutorials/principal_components.pdf

Smith SA and Ponder BA (October, 1993), *Predisposing genes in breast and ovarian cancer: an overview*, Tumori 79(5): 291-296

Sorlie T, Tibshirani R, Parker J, Hastie T, Marron JS, Nobel A, Deng S, Johnsen H, Pesich R, Geisler S, Demeter J, Perou CM, Lønning PE, Brown PO, Børresen-Dale AL, and Botstein D (July, 2003), *Repeated observation of breast tumor*

Personalis EX2155

subtypes in independent gene expression data sets, Proceedings of the National Academy of Sciences of the USA 100(14): 8418-8423

Soule HD, Vazguez J, Long A, Albert S, and Brennan M (November, 1973), *A human cell line from a pleural effusion derived from a breast carcinoma*, Journal of the National Cancer Institute 51(5): 1409-1416

Sparano JA and Paik S (February, 2008), *Development of the 21-gene assay and its application in clinical practice and clinical trials*, Journal of Clinical Oncology 26(5): 721-728

Srivastava S, Verma M, and Henson DE (May, 2001), *Biomarkers for early detection of colon cancer*, Clinical Cancer Research 7(5): 1118-1126

Stead LA, Lash TL, Sobieraj JE, Chi DD, Westrup JL, Charlot M, Blanchard RA, Lee JC, King TC, and Rosenberg CL (2009), *Triple-negative breast cancers are increased in black women regardless of age or body mass index*, Breast Cancer Research 11(2): R18

Stickeler E (June, 2011), *Prognostic and predictive markers for treatment decisions in early breast cancer*, Breast Care (Basel) 6(3): 193-198

Stjernqvist S, Rydén T, Greenman CD (May, 2011), *Model-integrated estimation of normal tissue contamination for cancer SNP allelic copy number data*, Cancer Informatics 10: 159-173

Stroun M, Anker P, Lyautey J, Lederrey C, and Maurice PA (June, 1987), *Isolation and characterization of DNA from the plasma of cancer patients*, European Journal of Clinical Oncology 23(6): 707-712

Stroun M, Lyautey J, Lederrey C, Olson-Sand A, and Anker P (November, 2001), *About the possible origin and mechanism of circulating DNA apoptosis and active DNA release*, Clinica Chimica Acta 313(1-2): 139-142

Personalis EX2155

Sunami E, Vu AT, Nguyen SL, Giuliano AE, and Hoon DS (August, 2008), *Quantification of LINE1 in circulating DNA as a molecular biomarker of breast cancer*, Annals of the New York Academy of Sciences 1137: 171-174

Suzuki N, Kamataki A, Yamaki J, and Homma Y (January, 2008), *Characterization of circulating DNA in healthy human plasma*, Clinica Chimica Acta 387(1-2): 55-58

Szoke J and Udvarhelyi N (January, 2012), *Modern pathologic diagnostics in breast cancer*, Orvosi Hetilap 153(1): 22-30

Tagami M, Kusuhara S, Imai H, Uemura A, Honda S, Tsukahara Y, and Negi A (October, 2010), *MRP4 knockdown enhances migration, suppresses apoptosis, and produces aggregated morphology in human retinal vascular endothelial cells*, Biochemical and Biophysical Research Communications 400(4): 593-598

Tam GW, Redon R, Carter NP, and Grant SG (December, 2009), *The role of DNA copy number variation in schizophrenia*, Biological Psychiatry 66(11): 1005-1012

Tamkovich SN, Vlassov VV, and Laktionov PP (2008), *Circulating DNA in the blood and its application in medical diagnosis*, Molecular Biology 42(1): 12-23

Tategu M, Nakagawa H, Hayashi R, and Yoshida K (October, 2008), *Transcriptional co-factor CDCA4 participates in the regulation of JUN oncogene expression*, Biochimie 90(10): 1515-1522

Tay PN, Tan P, Lan Y, Leung CH, Laban M, Tan TC, Ni H, Manikandan J, Rashid SB, Yan B, Yap CT, Lim LH, Lim YC, and Hooi SC (October, 2010), *Palladin, an actin-associated protein, is required for adherens junction formation and intercellular adhesion in HCT116 colorectal cancer cells*, International Journal of Oncology 37(4): 909-926

Personalis EX2155

Taylor CR and Levenson RM (October, 2006), *Quantification of immunohistochemistry--issues concerning methods, utility and semiquantitative assessment II*, Histopathology 49(4): 411-424

Teh MT, Blaydon D, Chaplin T, Foot NJ, Skoulakis S, Raghavan M, Harwood CA, Proby CM, Philpott MP, Young BD, and Kelsell DP (October, 2005), *Genomewide single nucleotide polymorphism microarray mapping in basal cell carcinomas unveils uniparental disomy as a key somatic event*, Cancer Research 65(19): 8597-8603

Thompson PA, Brewster AM, Kim-Anh D, Baladandayuthapani V, Broom BM, Edgerton ME, Hahn KM, Murray JL, Sahin A, Tsavachidis S, Wang Y, Zhang L, Hortobagyi GN, Mills GB, and Bondy ML (August, 2011), *Selective genomic copy number imbalances and probability of recurrence in early-stage breast cancer*, Public Library of Sciences One 6(8): e23543

Tomlinson D, Knowles M, and Speirs V (in press), *Mechanisms of FGFR3 actions in endocrine resistant breast cancer*, International Journal of Cancer

Tsang JC and Lo YM (April, 2007), *Circulating nucleic acids in plasma/serum*, Pathology 39(2): 197-207

Tuefferd M, De Bondt A, Van Den Wyngaert I, Talloen W, Verbeke T, Carvalho B, Clevert DA, Alifano M, Raghavan N, Amaratunga D, Göhlmann H, Broët P, and Camilleri-Broët S (November, 2008), *Genome-wide copy number alterations detection in fresh frozen and matched FFPE samples using SNP 6.0 arrays*, Genes, Chromosomes, and Cancer 47(11): 957-964

Tung N, Wang Y, Collins LC, Kaplan J, Li H, Gelman R, Comander AH, Gallagher B, Fetten K, Krag K, Stoeckert KA, Legare RD, Sgroi D, Ryan PD, Garber JE, and Schnitt SJ (February, 2010), *Estrogen receptor positive breast cancers in BRCA1 mutation carriers: clinical risk factors and pathologic features*, Breast Cancer Research 12(1): R12

Personalis EX2155

Turner N, Pearson A, Sharpe R, Lambros M, Geyer F, Lopez-Garcia MA, Natrajan R, Marchio C, Iorns E, Mackay A, Gillett C, Grigoriadis A, Tutt A, Reis-Filho JS, and Ashworth A (March, 2010), *FGFR1 amplification drives endocrine therapy resistance and is a therapeutic target in breast cancer*, Cancer Research 70(5): 2085-2094

Tzvetkov MV, Becker C, Kulle B, Nürnberg P, Brockmöller J, and Wojnowski L (February, 2005), *Genome-wide single-nucleotide polymorphism arrays demonstrate high fidelity of multiple displacement-based whole-genome amplification*, Electrophoresis 26(3): 710-715

Uchida K (2006), *Gene Amplification and Cancer*, Encyclopedia of Life Sciences, Wiley Online Library, http://onlinelibrary.wiley.com/doi/10.1038/npg.els.0006047/full

Ullrich A and Schlessinger J (April, 1990), *Signal transduction by receptors with tyrosine kinase activity*, Cell 61(2): 203-212

Umetani N, Giuliano AE, Hiramatsu SH, Amersi F, Nakagawa T, Martino S, and Hoon DS (September, 2006), *Prediction of breast tumor progression by integrity of free circulating DNA in serum*, Journal of Clinical Oncology 24(26): 4270-4276

US Food and Drug Administration (US FDA) (June, 2007), *Test determines risk of breast cancer returning*, http://www.fda.gov/ForConsumers/ConsumerUpdates/ucm048477.htm

Vainio H, Bianchini F, and Heseltine E (October, 2002), IARC Handbooks of Cancer Prevention, Volume 7: Breast Cancer Screening, International Agency for Research on Cancer (IARC), 150 Cours Albert Thomas, 69372 Lyon CEDEX 08, France, ISBN 9283230078

Personalis EX2155

Van de Wiel MA, Picard F, van Wieringen WN, and Ylstra B (January, 2011), *Preprocessing and downstream analysis of microarray DNA copy number profiles*, Briefings in Bioinformatics 12(1): 10-21

Van der Vaart M and Pretorius PJ (September, 2008), *Characterization of circulating DNA in healthy human plasma*, Clinica Chimica Acta 395(1-2): 186

Vecchi M, Confalonieri S, Nuciforo P, Viganò MA, Capra M, Bianchi M, Nicosia D, Bianchi F, Galimberti V, Viale G, Palermo G, Riccardi A, Campanini R, Daidone MG, Pierotti MA, Pece S, Di and Fiore PP (April, 2008), *Breast cancer metastases are molecularly distinct from their primary tumors*, Oncogene 27(15): 2148-2158

Venter DJ, Ramus SJ, Hammet FM, de Silva M, Hutchins AM, Petrovic V, Price G, and Armes JE (July, 2005), *Complex CGH alterations on chromosome arm 8p at candidate tumor suppressor gene loci in breast cancer cell lines*, Cancer Genetics and Cytogenetics 160(2): 134-140

Vogel CL, Cobleigh MA, Tripathy D, Gutheil JC, Harris LN, Fehrenbacher L, Slamon DJ, Murphy M, Novotny WF, Burchmore M, Shak S, Stewart SJ, and Press M (February, 2002), *Efficacy and safety of trastuzumab as a single agent in first-line treatment of HER2-overexpressing metastatic breast cancer*, Journal of Clinical Oncology 20(3): 719-726

Von Holst S, Picelli S, Edler D, Lenander C, Dalén J, Hjern F, Lundqvist N, Lindforss U, Påhlman L, Smedh K, Törnqvist A, Holm J, Janson M, Andersson M, Ekelund S, Olsson L, Ghazi S, Papadogiannakis N, Tenesa A, Farrington SM, Campbell H, Dunlop MG, and Lindblom A (August, 2010), *Association studies on 11 published colorectal cancer risk loci*, British Journal of Cancer 103(4): 575-580

Personalis EX2155

Walker RA (June, 2003), Prognostic and Predictive Factors in Breast Cancer (second edition), Martin Dunitz, Taylor and Francis Group, 11 New Fetter Lane, London EC4P 4EE, United Kingdom, ISBN 1841843008

Wall PK, Leebens-Mack J, Chanderbali AS, Barakat A, Wolcott E, Liang H, Landherr L, Tomsho LP, Hu Y, Carlson JE, Ma H, Schuster SC, Soltis DE, Soltis PS, Altman N, and dePamphilis CW (August, 2009), *Comparison of next generation sequencing technologies for transcriptome characterization*, BMC Genomics 10: 347

Wamelink MM, Struys EA, and Jakobs C (December, 2008), *The biochemistry, metabolism and inherited defects of the pentose phosphate pathway: a review*, Journal of Inherited Metabolic Disease 31(6): 703-717

Wang BG, Huang HY, Chen YC, Bristow RE, Kassauei K, Cheng CC, Roden R, Sokoll LJ, Chan DW, and Shih IeM (July, 2003), *Increased plasma DNA integrity in cancer patients*, Cancer Research 63(14): 3966-3968

Wang J and Dreyfuss G (November, 2001), *Characterization of functional domains of the SMN protein in vivo*, The Journal of Biological Chemistry, 276(48): 45387-45393

Wang JY, Hsieh JS, Chang MY, Huang TJ, Chen FM, Cheng TL, Alexandersen K, Huang YS, Tzou WS, and Lin SR (July, 2004), *Molecular detection of APC, K-ras, and p53 mutations in the serum of colorectal cancer patients as circulating biomarkers*, World Journal of Surgery 28(7): 721-726

Wang K, Li M, Hadley D, Liu R, Glessner J, Grant SF, Hakonarson H, and Bucan M (November, 2007), *PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data*, Genome Research 17(11): 1665-1674

Personalis EX2155

Wang X, Chao L, Ma G, Chen L, Zang Y, and Sun J (February, 2011), *The prognostic significance of WWOX expression in patients with breast cancer and its association with the basal-like phenotype*, Journal of Cancer Research and Clinical Oncology 137(2): 271-278

Wat MJ, Enciso VB, Wiszniewski W, Resnick T, Bader P, Roeder ER, Freedenberg D, Brown C, Stankiewicz P, Cheung SW, and Scott DA (November, 2010), *Recurrent microdeletions of 15q25.2 are associated with increased risk of congenital diaphragmatic hernia, cognitive deficits and possibly Diamond-Blackfan anaemia*, Journal of Medical Genetics 47(11): 777-781

Weichert J, Schröer A, Amari F, Siebert R, Caliebe A, Nagel I, Gillessen-Kaesbach G, Mohrmann I, and Hellenbroich Y (May-June, 2011), *A 1 Mb-sized microdeletion Xq26.2 encompassing the GPC3 gene in a fetus with Simpson-Golabi-Behmel syndrome Report, antenatal findings and review*, European Journal of Medical Genetics 54(3): 343-347

Wells D, Alfarawati S, and Fragouli E (December, 2008), *Use of comprehensive chromosomal screening for embryo assessment: microarrays and CGH*, Molecular Human Reproduction 14(12): 703-710

Williams NM, Zaharieva I, Martin A, Langley K, Mantripragada K, Fossdal R, Stefansson H, Stefansson K, Magnusson P, Gudmundsson OO, Gustafsson O, Holmans P, Owen MJ, O'Donovan M, and Thapar A (October, 2010), *Rare chromosomal deletions and duplications in attention-deficit hyperactivity disorder: a genome-wide analysis*, The Lancet 376(9750): 1401-1408

Wülfing P, Borchard J, Buerger H, Heidl S, Zänker KS, Kiesel L, and Brandt B (March, 2006), *HER2-positive circulating tumor cells indicate poor clinical outcome in stage I to III breast cancer patients*, Clinical Cancer Research 12(6): 1715-1720

Xiong F, Wu C, Bi X, Yu D, Huang L, Xu J, Zhang T, Zhai K, Chang J, Tan W, Cai J, and Lin D (July, 2010), *Risk of genome-wide association study-identified genetic*

Personalis EX2155

variants for colorectal cancer in a Chinese population, Cancer Epidemiology, Biomarkers, and Prevention 19(7): 1855-1861

Xue X, Teare MD, Holen I, Zhu YM, and Woll PJ (June, 2009), *Optimizing the yield and utility of circulating cell-free DNA from plasma and serum*, Clinica Chimica Acta 404(2): 100-104

Yagi T, Baroja-Fernández E, Yamamoto R, Muñoz FJ, Akazawa T, Hong KS, and Pozueta-Romero J (March, 2003), *Cloning, expression and characterization of a mammalian Nudix hydrolase-like enzyme that cleaves the pyrophosphate bond of UDP-glucose*, The Biochemical Journal 370(Part 2): 409-415

Yamamoto G, Nannya Y, Kato M, Sanada M, Levine RL, Kawamata N, Hangaishi A, Kurokawa M, Chiba S, Gilliland DG, Koeffler HP, and Ogawa S (July, 2007), *Highly sensitive method for genomewide detection of allelic composition in nonpaired, primary tumor specimens by use of affymetrix single-nucleotide-polymorphism genotyping microarrays*, American Journal of Human Genetics 81(1): 114-126

Yang TL, Chen XD, Guo Y, Lei SF, Wang JT, Zhou Q, Pan F, Chen Y, Zhang ZX, Dong SS, Xu XH, Yan H, Liu X, Qiu C, Zhu XZ, Chen T, Li M, Zhang H, Zhang L, Drees BM, Hamilton JJ, Papasian CJ, Recker RR, Song XP, Cheng J, and Deng HW (December, 2008), *Genome-wide copy-number-variation study identified a susceptibility gene, UGT2B17, for osteoporosis*, American Journal of Human Genetics 83(6): 663-674

Yang XR, Chang-Claude J, Goode EL, Couch FJ, Nevanlinna H, Milne RL, Gaudet M, Schmidt MK, Broeks A, Cox A, Fasching PA, Hein R, Spurdle AB, Blows F, Driver K, Flesch-Janys D, Heinz J, Sinn P, Vrieling A, Heikkinen T, Aittomäki K, Heikkilä P, Blomqvist C, Lissowska J, Peplonska B, Chanock S, Figueroa J, Brinton L, Hall P, Czene K, Humphreys K, Darabi H, Liu J, Van 't Veer LJ, van Leeuwen FE, Andrulis IL, Glendon G, Knight JA, Mulligan AM, O'Malley FP, Weerasooriya N, John EM, Beckmann MW, Hartmann A, Weihbrecht SB,

Personalis EX2155

Wachter DL, Jud SM, Loehberg CR, Baglietto L, English DR, Giles GG, McLean CA, Severi G, Lambrechts D, Vandorpe T, Weltens C, Paridaens R, Smeets A, Neven P, Wildiers H, Wang X, Olson JE, Cafourek V, Fredericksen Z, Kosel M, Vachon C, Cramp HE, Connley D, Cross SS, Balasubramanian SP, Reed MW, Dörk T, Bremer M, Meyer A, Karstens JH, Ay A, Park-Simon TW, Hillemanns P, Arias Pérez JI, Menéndez Rodríguez P, Zamora P, Benítez J, Ko YD, Fischer HP, Hamann U, Pesch B, Brüning T, Justenhoven C, Brauch H, Eccles DM, Tapper WJ, Gerty SM, Sawyer EJ, Tomlinson IP, Jones A, Kerin M, Miller N, McInerney N, Anton-Culver H, Ziogas A, Shen CY, Hsiung CN, Wu PE, Yang SL, Yu JC, Chen ST, Hsu GC, Haiman CA, Henderson BE, Le Marchand L, Kolonel LN, Lindblom A, Margolin S, Jakubowska A, Lubiński J, Huzarski T, Byrski T, Górski B, Gronwald J, Hooning MJ, Hollestelle A, van den Ouweland AM, Jager A, Kriege M, Tilanus-Linthorst MM, Collée M, Wang-Gohrke S, Pylkäs K, Jukkola-Vuorinen A, Mononen K, Grip M, Hirvikoski P, Winqvist R, Mannermaa A, Kosma VM, Kauppinen J, Kataja V, Auvinen P, Soini Y, Sironen R, Bojesen SE, Ørsted DD, Kaur-Knudsen D, Flyger H, Nordestgaard BG, Holland H, Chenevix-Trench G, Manoukian S, Barile M, Radice P, Hankinson SE, Hunter DJ, Tamimi R, Sangrajrang S, Brennan P, McKay J, Odefrey F, Gaborieau V, Devilee P, Huijts PE, Tollenaar RA, Seynaeve C, Dite GS, Apicella C, Hopper JL, Hammet F, Tsimiklis H, Smith LD, Southey MC, Humphreys MK, Easton D, Pharoah P, Sherman ME, and Garcia-Closas M (February, 2011), *Associations of breast cancer risk factors with tumor subtypes: a pooled analysis from the Breast Cancer Association Consortium studies*, Journal of the National Cancer Institute 103(3): 250-263

Yang YM, Liu H, and Li WB (July, 2010), *Expression of Mta-1 and VEGF and their correlation in the endometrial cancer*, Chinese Journal of Cellular and Molecular Immunology 26(7): 682-684

Yoshihara K, Tajima A, Adachi S, Quan J, Sekine M, Kase H, Yahata T, Inoue I, and Tanaka K (March, 2011), *Germline copy number variations in BRCA1-*

Personalis EX2155

associated ovarian cancer patients, Genes, Chromosomes, and Cancer 50(3): 167-177

Yu LR, Lv JQ, Jin LY, Ding SD, Ma XY, Wang JJ, and Zhu XQ (2011), *Over-expression of protein kinase C isoforms (α, δ, θ and ζ) in squamous cervical cancer*, Neoplasma 58(6): 491-498

Zahl PH and Mæhlen J (February, 2011), *Overdiagnosis of breast cancer after 14 years of mammography screening*, Tidsskrift for den Norske Laegeforening 132(4): 414-417

Zanetti-Dällenbach R, Wight E, Fan AX, Lapaire O, Hahn S, Holzgreve W, and Zhong XY (March-April, 2008), *Positive correlation of cell-free DNA in plasma/serum in patients with malignant and benign breast disease*, Anticancer Research 28(2A): 921-925

Zhao H, Lo YH, Yu L, and Wang SC (May, 2011), *Overcoming resistance to fulvestrant (ICI182,780) by downregulating the c-ABL proto-oncogene in breast cancer*, Molecular Carcinogenesis 50(5): 383-389

Zhang M, Mai W, Li C, Cho S, Hao C, Moeckel G, Zhao R, Kim I, Wang J, Xiong H, Wang H, Sato Y, Wu Y, Nakanuma Y, Lilova M, Pei Y, Harris R, Li S, Coffey R, Sun L, Wu D, Chen X, Breyer M, Zhao Z, McKanna J, and Wu G (February, 2004), *PKHD1 protein encoded by the gene for autosomal recessive polycystic kidney disease associates with basal bodies and primary cilia in renal epithelial cells*, Proceedings of the National Academy of Sciences of the USA 101(8): 2311-2316

Zhang W, Gordon M, Schultheis AM, Yang DY, Nagashima F, Azuma M, Chang HM, Borucka E, Lurje G, Sherrod AE, Iqbal S, Groshen S, and Lenz HJ (August, 2007), *FCGR2A and FCGR3A polymorphisms associated with clinical outcome of epidermal growth factor receptor-expressing metastatic colorectal cancer*

Personalis EX2155

patients treated with single-agent cetuximab, Journal of Clinical Oncology 25(24): 3712-3718

Zhong XY, Ladewig A, Schmid S, Wight E, Hahn S, and Holzgreve W (October, 2007), Archives of Gynecology and Obstetrics 276(4): 327-331

Zhou Y, Feng X, and Koh DW (July, 2011), *Synergistic cytotoxicity of N-methyl-N'-nitro-N-nitrosoguanidine and absence of poly(ADP-ribose) glycohydrolase involves chromatin decondensation*, International Journal of Oncology 39(1): 121-127

Zhu X, Guo Y, Li X, Ding Y, and Chen L (August, 2010), *Metastasis-associated protein 1 nuclear expression is associated with tumor progression and clinical outcome in patients with non-small cell lung cancer*, Journal of Thoracic Oncology 5(8): 1159-1166

Ziegler RG, Hoover RN, Pike MC, Hildesheim A, Nomura AM, West DW, Wu-Williams AH, Kolonel LN, Horn-Ross PL, Rosenthal JF, and Hyer MB (November, 1993), *Migration patterns and breast cancer risk in Asian-American women*, Journal of the National Cancer Institute 85(22): 1819-1827

Zihni C, Mitsopoulos C, Tavares IA, Baum B, Ridley AJ, and Morris JD (March, 2007), *Prostate-derived sterile 20-like kinase 1-alpha induces apoptosis. JNK- and caspase-dependent nuclear localization is a requirement for membrane blebbing*, The Journal of Biological Chemistry 289(9): 6484-6493

Zippelius A and Pantel K (April, 2000), *RT-PCR-based detection of occult disseminated tumor cells in peripheral blood and bone marrow of patients with solid tumors - An overview*, Annals of the New York Academy of Sciences 906: 110-123

Personalis EX2155