Imperial College London

Doctoral Thesis

Statistical Approaches for Copy Number Variation Detection and Association with Complex Human Phenotypes

Author: Tisham De

A thesis submitted in fulfilment of the requirements

for the degree of Doctor of Philosophy

in the

School of Public Health

29.06.2014

Declaration of Authorship

I, Tisham De, declare that this thesis titled, 'Statistical Approaches for Copy Number Variation Detection and Association with Complex Human Phenotypes' and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given.
- I have acknowledged all main sources of help.

Signed: Tisham De

Date: 29.06.2014

"Facts are stubborn, but statistics are more pliable"

Mark Twain

Abstract

Copy number variants (CNVs) play an important role in the disease pathogenesis, including epilepsy, diabetes and many others. CNVs, are also known to affect cellular phenotypes through several phenomenon such as gene dosage.

Next generation technologies for sequencing (DNA and RNA) and metabolite profiling (metabolomics) has led to the systematic discovery and evaluation of various genomic variants and their relationship to multiple phenotypes. Such approaches often involve application of several statistical and machine learning methods for unravelling new relationships between genomic variants and phenotypes i.e. disease outcomes or quantitative traits characterized at the molecular level.

This thesis explores and develops several statistical methods for CNV detection and association with complex human phenotypes, in particular for epilepsy drug-response, epilepsy susceptibility, metabolomics and gene expression.

In more detail, chapter 3, describes a genome wide CNV association analysis for two phenotypes including epilepsy susceptibility and epilepsy drug response. I have identified several important candidate genes for these two phenotypes, including the top most associated genes, *SLC9A1* (p-value=6.69E-15) for epilepsy susceptibility and *WWOX* (p-value=1.93E-3) for epilepsy drug response. These associations were replicated in a separate Australian cohort and were further validated in lab and in-silico, leading to some positive and negative confirmation.

In chapter 4, I present CNV association with metabolomic data in the exonic regions of the *TSPAN8* gene. A strong association signal was detected in the 6th exon and 7th exon of the *TSPAN8* gene, where a large proportion of metabonomic lipid phenotypes were found to be associated with univariate (P-value=7.64E-4) and multivariate (P-value=1.33E-6) approaches. These CNVs were also found to be nominally associated with type 2 diabetes (P-value=3.32e-7). In addition, I also carried out advanced multivariate based association analysis to corroborate these results and further reported sequencing based validation results for *TSPAN8* exonic CNVs in different human populations from the 1000 genomes project.

In chapter 5, I report a genome wide CNV association analysis with gene expression in ten different regions of the human brain. I identified a novel CNV near the *DRD5* gene which was found to be strongly associated with gene expression. Further, I have reported on-going efforts to replicate and validate this finding.

Each of these different phenotype categories analysed posed its own unique challenges and required specific approaches for analysis and interpretation.

Acknowledgements

First and foremost I would like to acknowledge my supervisors Dr Lachlan JM Coin and Dr Inga Prokopenko for their immense support, guidance and help over the last three years. I could not have completed this journey without them.

A special thank you to Professor Marjo-Riitta Jarvelin who has always been most supportive for my research and PhD. I would like to thank Dr Michael R. Johnson and Dr Enrico Petretto who nurtured my interest in genetics and Bioinformatics. It was a great experience working with them. I am thankful to Matthias Heger, Tim Ovenden, James McGovern, Patricia Murphy, the School of Public Health and Imperial College London for supporting my work.

I am also thankful to Professor Phillipe Froguel and Professor Alexandra Blakemore for their kind advice that PhD is a test of survival and endurance.

Most importantly a warm thank you to my family, friends and colleagues.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

List of Figures x

List of Tables xii

Abbreviations xiv

1 Introduction 1
- 1.1 Overview of the thesis . . . 1
- 1.2 Omics background . . . 4
  - 1.2.1 Genomics . . . 4
  - 1.2.2 Metabolomics . . . 6
  - 1.2.3 Transcriptomics . . . 7
- 1.3 Genome Wide Association Studies (GWAS) . . . 9
  - 1.3.1 Single Nucleotide Polymorphism (SNP) . . . 9
  - 1.3.2 Copy Number Variation (CNV) . . . 9
  - 1.3.3 Mechanism of genomic rearrangements in humans . . . 11
    - 1.3.3.1 Non allelic homologous recombination (NAHR) . . . 11
    - 1.3.3.2 Non homologous end-joining recombination (NHEJ) . . . 14
    - 1.3.3.3 Fork stalling template switching mechanism (FoSTeS) . . . 16
  - 1.3.4 Insertion deletions (INDELS) . . . 18
  - 1.3.5 Hardy-Weinberg Equilibrium (HWE) . . . 18
  - 1.3.6 GWAS background . . . 19
  - 1.3.7 Genetic basis of GWAS . . . 20
  - 1.3.8 Significance of LD in GWAS . . . 22
  - 1.3.9 Haplotypes . . . 23
  - 1.3.10 CNV and GWAS . . . 23
  - 1.3.11 CNVs and drug response . . . 24
  - 1.3.12 Finding missing heritability in GWAS . . . 25

1.3.12.1 Missing heritability and the role of common CNVs in diseases . . . 28
1.4 Review of current approaches for CNV discovery and validation . . . 28
1.4.1 Normalization of array intensity data . . . 29
1.4.2 Genomic waves . . . 30
1.4.3 Methods for CNV genotyping for discovery . . . 30
1.4.4 Array Comparative genomic hybridisation . . . 31
1.4.5 High density SNP arrays . . . 32
1.4.6 Next generation sequencing and genome assembly comparison . . . 33
1.4.7 Lab based methods for CNV validation . . . 34
1.4.7.1 MLPA . . . 34
1.4.7.2 qPCR . . . 35
1.4.7.3 Paralogue ratio test . . . 35

2 Background Methods 36
2.1 CNV calling from genotyping arrays . . . 36
2.1.1 Detecting CNVs from genotype calls . . . 36
2.1.2 Detecting CNVs from one dimensional segmentation . . . 37
2.1.2.1 Non parametric methods for CNV detection . . . 37
2.1.2.2 Model based approaches using HMMs . . . 38
2.1.3 The hidden Markov model (HMM) . . . 39
2.1.3.1 The HMM notation . . . 39
2.1.3.2 The three problems and solutions in HMM . . . 40
2.1.3.3 Solution to the three HMM problems . . . 41
2.1.3.4 Solution to problem 1 . . . 41
2.1.3.5 Solution of problem 2 . . . 43
2.1.3.6 Solution to problem 3 . . . 44
2.1.3.7 Dynamic programming and scaling issues . . . 45
2.1.3.8 The Viterbi algorithm for state path determination . . . 46
2.1.3.9 The EM algorithm . . . 46
2.1.4 Generalised genotyping methods . . . 47
2.1.5 HMM based CNV detection methods . . . 48
2.1.6 Methods for detecting de novo CNVs . . . 50
2.1.7 Other model based approaches (GADA) . . . 50
2.2 Association approaches in genetic epidemiology . . . 51
2.2.1 Tests for multiple SNPs . . . 52
2.3 Models for association . . . 53
2.3.1 General linear regression . . . 53
2.3.1.1 Logistic regression . . . 55
2.3.1.2 Proportional odds regression . . . 56
2.3.2 Odds ratio . . . 56
2.3.2.1 The Wald statistics . . . 57
2.3.2.2 Likelihood ratio test . . . 57
2.3.2.3 Score test . . . 57
2.4 Meta analysis in GWAS . . . 58
2.5 Population stratification and cryptic relatedness in GWAS . . . 60
2.5.1 Cryptic relatedness . . . 62
2.5.2 Correcting for stratification . . . 62

2.6 Approaches for CNV genotype phenotype association . . . 66
2.7 Variable selection approaches . . . 67
2.8 Methods for clustering . . . 67
2.9 Multiple testing . . . 68
2.10 Prediction and classification . . . 69
2.11 Principal component analysis (PCA) . . . 70
2.12 Models for linkage disequilibrium (LD) . . . 70
2.13 Quality control and filtering in CNV association analysis . . . 72
2.13.1 Genotyping details and quality control for chapter 3 . . . 72
2.13.2 Genotyping platform details for discovery and control cohorts for chapter 4 . . . 73
2.13.2.1 Normalisation and quality control . . . 73
2.13.3 Analysis methods for chapter5 . . . 74

3 CNV Analysis in Epilepsy 75
3.1 Introduction . . . 75
3.1.1 Overview . . . 75
3.1.2 Background . . . 75
3.2 Patients . . . 76
3.3 Study design . . . 77
3.4 Methods . . . 77
3.4.1 Calculating LRR PCs . . . 78
3.4.2 Multiplatform integration using cnvHap . . . 81
3.4.3 Association approaches . . . 84
3.5 Results . . . 87
3.5.1 Analysis and comparison of large CNVs . . . 88
3.5.2 Epilepsy susceptibility analysis . . . 89
3.5.2.1 *NLGN1*: Missing link between autism and epilepsy? . . . 92
3.5.3 Epilepsy drug response analysis . . . 94
3.5.3.1 The *WWOX* gene . . . 96
3.5.3.2 *WWOX* and epilepsy . . . 98
3.5.4 Chromosome X results for drug remission . . . 98
3.5.5 Results from DECIPHER database . . . 99
3.5.6 Lab based qPCR validation . . . 103
3.5.7 NGS sequence validation . . . 105
3.6 Clustering based CNV detection . . . 107
3.6.1 Introduction . . . 107
3.6.2 Density based clustering approaches . . . 109
3.6.3 Results . . . 109

4 CNV Association in Metabonomics 115
4.1 Introduction . . . 115
4.1.1 Difference between metabolomics and metabonomics . . . 119
4.2 *TSPAN8* CNV and T2D . . . 120
4.3 Materials and methods . . . 122
4.3.1 Study cohorts and genotyping platforms . . . 122
4.3.2 Metabonomic measurements in NFBC 1966 and NFBC 1986 . . . 124
4.3.3 CNV analysis methods . . . 125
4.3.4 NGS based CNV identification . . . 125

4.3.5 CNV predictions using cnvHap . . . 125
4.3.6 CNV segmentation . . . 126
4.3.7 Expected CNV genotypes . . . 126
4.3.8 Association analysis . . . 126
4.3.9 Intensity based validation of CNV association . . . 127
4.3.10 Multiple testing . . . 127
4.3.11 Linkage disequilibrium . . . 128
4.4 Results . . . 128
4.4.1 CNVs in the *TSPAN8* coding regions based on sequence data . . . 128
4.4.2 CNV segmentation . . . 133
4.4.3 CNV replication and allele frequency . . . 138
4.4.4 Univariate analysis . . . 145
4.4.4.1 CNV association with metabonomic and lipid phenotypes in NFBC 1966 and NFBC 1986 . . . 145
4.4.5 Multivariate analysis . . . 159
4.4.6 Validation using raw intensity . . . 164
4.4.7 LD results for *TSPAN8* . . . 171
4.4.8 Association with glucose stratification . . . 175
4.4.9 Association with PC projection . . . 178
4.4.10 Association with lipids at chr12: 69812860 . . . 180
4.4.11 Association with lipids at CNVR5583.1 . . . 183
4.4.12 Multivariate analysis joint association and joint association with variable selection . . . 186
4.4.13 NFBC1966 PennCNV results . . . 189
4.4.14 LD results for NFBC 1966 . . . 190
4.5 Missing data . . . 193
4.6 CNV analysis in *LCE3C-LCE3B* . . . 196

5 CNV Association with Gene Expression 203
5.1 Introduction . . . 203
5.2 Section 1: CNV association with gene expression . . . 204
5.2.1 Gene expression and phenotypes . . . 204
5.2.2 eQTL studies . . . 205
5.2.3 CNV eQTL studies . . . 206
5.2.4 Population specific gene expression . . . 209
5.2.5 RNA-Seq . . . 209
5.2.6 Methods for eQTL analysis . . . 211
5.3 UKBEC Analysis . . . 212
5.3.1 Materials and Methods . . . 212
5.3.2 Results . . . 214
5.4 Section 2: CNV-eQTL methodology development . . . 221
5.5 Graphics processing unit (GPU) . . . 221
5.6 Linear systems and matrix algebra . . . 222
5.6.1 Methods for matrix decomposition . . . 223
5.6.2 Least squares solution . . . 225
5.6.3 Linear regression in matrix form . . . 225
5.7 Results . . . 228
5.8 Future work . . . 233

6 Conclusions and future work **234**
6.1 Discussion for CNV analysis in Epilepsy 234
6.1.1 Future directions 236
6.2 Discussion for CNV analysis in metabolomics 238
6.3 Discussion for CNV association with gene expression 240
6.4 Conclusions 242
6.5 Original contributions 245

A Lipid Names **278**

B Linear systems **280**

C Publications **285**

List of Figures

1.1 An example of a simple duplication . . . 10
1.2 Non allelic homologous recombination (NAHR) . . . 13
1.3 Non allelic homologous recombination (NAHR) . . . 15
1.4 Figure showing FoSTeS mechanism . . . 17
1.5 GWAS approach for detecting causal variants . . . 22

2.1 Effect of CNV on SNP genotyping data . . . 51
2.2 Population stratification in GWAS . . . 60
2.3 Population stratification in GWAS . . . 65
2.4 Genotyping QC in Australian Cohort . . . 72

3.1 Plot for LRR PCs . . . 80
3.2 Multi-platform integration through cnvHap . . . 83
3.3 Marginal scores for association . . . 85
3.4 Score for samples driving CNVs . . . 86
3.5 Example of SNPs driving association . . . 87
3.6 Quantile-quantile (QQ) plot for epilepsy susceptibility analysis . . . 90
3.7 Cluster plot for deletion in *NLGN1* gene . . . 93
3.8 Cluster plot for deletion in the *WWOX* gene . . . 96
3.9 Functional aspects of *WWOX* proteins . . . 97
3.10 CopyCaller results for the *EPHA7* gene . . . 104
3.11 NGS sequencing results for *WWOX* . . . 106
3.12 3D plot showing clustering using the DBSCSN algorithm . . . 111
3.13 3D cluster plot showing the *WWOX* CNV . . . 112
3.14 PAM clustering of *WWOX* gene . . . 114

4.1 Diversity in various biological markers . . . 116
4.2 Deconvolution of NMR spectral data . . . 117
4.3 Cluster plot of the *TSPAN8* in different cohorts . . . 121
4.4 Analysis pipeline for CNV-metabonomic association analysis reported in this chapter . . . 123
4.5 UCSC genome browser plot . . . 134
4.6 UCSC genome browser plot for control population . . . 135
4.7 Cumulative plot for M_VLDL_P in NFBC 1966 . . . 138
4.8 Allele frequency determined through genotyping array and NGS based platforms . . . 140
4.9 QQ plot for the top associated lipids . . . 153
4.10 Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1986 . 156
4.11 Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1966 . 157

4.12 Box plots for normal and deletion copy . . . 158
4.13 Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1986 using raw intensity . . . 165
4.14 Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1966 using raw intensity . . . 168
4.15 *TSPAN8* LD plot in NFBC 1986 . . . 172
4.16 Forest plot of top meta p-values . . . 179
4.17 Heatmap results for association using PennCNV . . . 189
4.18 Manhattan plot of SNPs within the *TSPAN8* gene . . . 191
4.19 Manhattan plot of SNPs within the *TSPAN8* gene in NFBC 1966 . . . 192
4.20 QQ plot for urea in NFBC 1986 . . . 197
4.21 Manhattan plot for urea in NFBC 1986 . . . 198
4.22 Heatmap for association with lipids near the *LCE3C* gene in NFBC 1986 199
4.23 Heatmap for association with lipids near the *LCE3C* gene in NFBC 1966 200
4.24 Heatmap for association with lipids near the *LCE3C* gene in joint analysis of NFBC 1966 and NFBC 1986 . . . 201
4.25 Correlation matrix plot showing the LD between the lipid associated region chr12:69809401-69812861 and CNVR5583.1 in NFBC 1986 . . . 202

5.1 Previous reports of SNP eQTLs on chromosome 4 . . . 208
5.2 RNA-Seq results for expression around *DRD5* in 3 different populations . 210
5.3 Overview of UKBEC association analysis . . . 213
5.4 Chromosome 4 duplication region in UKBEC and NABEC . . . 217
5.5 Chromosome 4 duplication in *Thousand Genomes* CEU samples . . . 218
5.6 LD plot in UKBEC around *DRD5* . . . 218
5.7 RNA-Seq results from *The 1000 Genomes CEU* population for samples having duplication at *DRD5* . . . 219
5.8 Ensembl plot of the region around chromosome 4 duplication . . . 220
5.9 Overview of GPU acceleration framework . . . 222
5.10 GPU configuration . . . 229
5.11 CPU vs GPU matrix multiplication time . . . 230
5.12 CPU vs GPU matrix inversion time . . . 230

6.1 RNA expression values for the *WWOX* gene . . . 236
6.2 Merging clustering algorithms . . . 237

List of Tables

2.1 A contingency table for calculating OR . . . 56
2.2 Heterozygosity rates in UK and the Australian cohort . . . 73

3.1 Association results for the epilepsy susceptibility analysis . . . 91
3.2 Association results for epilepsy AED remission analysis . . . 95
3.3 Patients from the decipher database showing rare chromosomal aberrations in the *WWOX* gene . . . 101
3.4 Patients from the decipher database showing rare chromosomal aberrations in *NLGN1* . . . 102
3.5 CNV breakpoints of 3 CEU samples carrying the *WWOX* deletions . . . 105
3.6 Comparison between classification and clustering . . . 107

4.1 CNV deletion frequency at *CNVR5583.1* from 1KG . . . 131
4.2 Deletion frequency at chr12 69809401 in 1KG . . . 137
4.3 Allele frequency of deletion at chr12 69809401 . . . 142
4.4 CNV frequency from genotyping arrays at *CNVR5583.1* . . . 144
4.5 Meta-analysis results at chr12 69809401 . . . 147
4.6 Univariate meta-analysis results at chr12:69809401 using LRR variance as covariate . . . 149
4.7 Univariate meta-analysis results for all metabolomic lipids at chr12:69809401 151
4.8 Inflation factor for top four lipids . . . 155
4.9 Joint association P-values from multiphen . . . 161
4.10 Association of CNV at chr12:69812681 in NFBC 1966 . . . 163
4.11 Meta-analysis results using LRR intensity values at chr12:69809401 . . . 167
4.12 Association with T2D outcome using LRR intensity values . . . 170
4.13 LD results in NFBC 1986 at CNVR5583.1 . . . 174
4.14 Univariate meta-analysis results at chr12:69809401 based on sample stratification based on blood glucose levels. LRR variance is used as a covariate 177
4.15 Association of CNV at 6th exon of *TSPAN8* with metabonomic lipids in NFBC 1966 and NFBC 1986 . . . 182
4.16 Association of deletion CNV at or near CNVR5583.1 . . . 185
4.17 Multiphen based association results using joint model in *TSPAN8* coding regions in NFBC 1986 and NFBC 1966 . . . 188
4.18 Association after imputation of missing values . . . 195

5.1 Association results from UKBEC CNV-eQTL analysis . . . 216
5.2 Comparison of linear solvers . . . 231
5.3 Comparison of LU decomposition . . . 232
5.4 Comparison of QR decomposition . . . 232

5.5 Comparison of time for matrix inversion 232

A.1 Abbreviations of metabonomic lipids 1 278
A.2 Abbreviations of metabonomic lipids 2 279

Abbreviations

aCGH	**A**rray **C**omparative **G**enomic **H**ybridisation
AED	**A**nti **E**pileptic **D**rugs
AHR	**A**llelic **H**omologous **R**ecombination
ASD	**A**utism **S**pectrum **D**isorder
BAF	**B** **A**llele **F**requency
BAM	**B**inary **A**lignment **M**ap
BLAS	**B**asic **L**inear**A**lgebra **S**ubprograms
BMI	**B**ody **M**ass **I**ndex
CBS	**C**ircular **B**inary **S**egmentation
CBZ	**C**arbamazepine
cDNA	**C**omplementary **D**eoxyribonucleic **A**cid
CNV	**C**opy **N**umber **V**ariation
CNVR	**C**opy **N**umber **V**ariation **R**egion
CUDA	**C**ompute **U**nified **D**evice **A**rchitecture
DABG	**D**ensity **A**bove **B**ackground
DBSCAN	**D**ensity -**B**ased **S**patial **C**lustering **A**pplications with **N**oise
DNA	**D**eoxyribonucleic **A**cid
DOF	**D**egrees **o**f **F**reedom
DP	**D**ynamic **P**rogramming
DSB	**D**ouble-**S**trand **B**reaks
EM	**E**xpectation **M**aximization
eQTL	**e**xpression **Q**uantitative **T**rait **L**oci
FDR	**F**alse **D**iscovery **R**ate
FoSTeS	**F**ork **S**talling and **T**emplate **S**witching
FWER	**F**amily **W**ise **E**rror **R**ate

GADA	**G**enome **A**lteration **D**etection **A**lgorithm
GLM	**G**eneralized **L**inear **M**odel
GPU	**G**raphics **P**rocessing **U**nits
GSV	**G**enomic **S**tructural **V**ariation
GWAS	**G**enome **W**ide **A**ssociation **S**tudies
HDL	**H**igh **D**ensity **L**ipoprotein
HMM	**H**idden **M**arkov **M**odel
HPLC	**H**igh **P**erformance **L**iquid **C**hromatography
HWE	**H**ardy–**W**einberg **E**quilibrium
IDL	**I**ntermediate **D**ensity **L**ipoprotein
ILAE	**I**nternational **L**eague **A**gainst **E**pilepsy
LASER	**L**ight **A**mplification by **S**timulated **E**mission of **R**adiation
LCR	**L**ow-**C**opy **R**epeats
LD	**L**inkage **D**isequilibrium
LDL	**L**ow **D**ensity **L**ipoprotein
LOH	**L**oss **o**f **H**eterozygosity
LRR	**L**og **R** **R**atio
MAF	**M**inor **A**llele **Frequency**
MEPS	**M**inimal **E**fficient **P**rocessing **S**egments
MHC	**M**ajor **H**istocompatibility **C**omplex
MLPA	**M**ultiplex **L**igation-dependent **P**robe **A**mplification
MRC	**M**edical **R**esearch **C**ouncil
MRI	**M**agnetic **R**esonance **I**maging
mRNA	**m**essenger **R**ibonucleic **A**cid
NABEC	**N**orth **A**merican **B**rain **E**xpression **C**onsortium
NAHR	**N**on-**A**llelic **H**omologous **R**ecombination
NFBC	**N**orthern **F**innish **B**irth **C**ohort
NGS	**N**ext **G**eneration **S**equencing
NHEJ	**N**onhomologous **E**nd **J**oining
NMR	**N**uclear **M**agnetic **R**esonance
OLS	**O**rdinary **L**east **S**quares **E**stimate
OR	**O**dds **R**atio
PAM	**P**artitioning **A**round **M**edoids

PCA	**P**rincipal **C**omponent **A**nalysis
PCR	**P**olymerase **C**hain **R**eaction
PEM	**P**aired **E**nd **M**apping
QC	**Q**uality **C**ontrol
qPCR	**q**uantative **P**olymerase **C**hain **R**eaction
QQ Plots	**Q**uantile –**Qunatile P**lots
RMA	**R**obust **M**ulti -**A**rray **Average**
RNA	**R**ibonucleic **A**cid
ROH	**R**uns **o**f **H**omozygosity
SNP	**S**ingle **N**ucleotide **P**olymorphism
SRO	**S**mallest **R**egion of **O**verlap
SVD	**S**ingular **V**alue **D**ecomposition
T2D	**T**ype **2 D**iabetes
UKBEC	**UK B**rain **E**xpression **C**onsortium
VLDL	**V**ery **L**arge **D**ensity **L**ipoprotein
WTCCC	**W**ellcome **T**rust **C**ase **C**ontrol **C**onsortium

Copyright Declaration

The copyright of this thesis rests with the author and is made available under a Creative Commons Attribution Non-Commercial No Derivatives licence. Researchers are free to copy, distribute or transmit the thesis on the condition that they attribute it, that they do not use it for commercial purposes and that they do not alter, transform or build upon it. For any reuse or redistribution, researchers must make clear to others the licence terms of this work.

Dedicated to my parents

Chapter 1

Introduction

1.1 Overview of the thesis

This thesis has explored various statistical approaches for detecting and associating Copy Number Variation (CNV) with complex human phenotypes. Such phenotypes analysed in this thesis can be broadly categorized into various concepts related to omics field including genomics, metabolomics and transcriptomics. These phenotypes analysed were novel drug response outcome for epilepsy, susceptibility to epilepsy in the UK population -*genomics*, well characterized metabolite signatures acting as proxy for disease outcomes like type 2 diabetes (T2D) -*metabolomics* or *metabonomics* and other molecular phenotype affecting gene expression through the RNA molecules -*transcriptomics*.

In chapter 1, I have reviewed the background and some of the concepts related to omics. Next, I have described the genetics background and some concepts related to CNVs, particularly in light of some of the recent large-scale genome-wide association studies (GWAS). In this section, I have shown how GWAS has endeavoured to find the putative role of CNVs and other genomic variations for several common and rare human diseases. I have further mentioned some of the successful discoveries from the application of such approaches. I have next described how some of these high-throughput technologies has enabled detection and validation of CNVs in large number of individuals, with high fidelity. Finally, I have discussed the merits and short-comings of such methods, and further tried to analyse the settings where they are most appropriate.

In chapter 2, I have discussed some of the current statistical methods and algorithms which are commonly used to detect and characterize CNVs from high-throughput platforms. In addition, I have further discussed the various association models used for finding relationships between the detected CNVs and various phenotypes which can be binary or quantitative. Some of the statistical frameworks used for CNV discovery and association include several parametric and non-parametric methods, Bayesian methods like hidden Markov models (HMM) and several applications of the generalized linear model (GLM). I have also covered some of the basics for important concepts related to clustering, variable selection, multiple testing and linkage disequilibrium (LD). Some of these concepts have direct application and relevance in the analysis work I have presented in the later chapters.

In chapter 3, I have reported the results from a genome-wide CNV analysis in epilepsy. This analysis was done with two study designs, one where I looked for CNVs conferring susceptibility to epilepsy in the general population and a second study design where I looked for CNVs associated with response to anti-epileptic drug (AED) medication stratified within a single epilepsy cohort. I have reported some of the novel findings from these two analyses. Next, I have discussed the validation efforts and clinical implications for these findings and lastly, I conclude with the future directions in the field of drug-response related study designs in genomics. In the end, I have also briefly discussed and reported results from a novel CNV-detection method based on the statistical concept of clustering.

In chapter 4, I have reported results from a targeted analysis of metabonomic signatures of a common CNV in the *TSPAN8* gene in two Finnish cohorts. I have further compared results for this association using univariate and multivariate association methods. Some of the novel statistical methods I have applied and reported in this section include an approach for variable selection and a method for regressing out batch effects from the association framework. I have also characterized *TSPAN8* CNVs using the 1000 genomes (1KG) next generation sequencing (NGS) results for different populations and have further reported their allele frequencies. I have also mentioned some of the other statistical issues I have addressed in this analysis such as multiple testing and imputation of missing data.

In chapter 5, I have carried out a genome-wide CNV detection and association analysis with gene expression for the UK Brain Expression Consortium (UKBEC) study. I have reported a novel CNV found to be strongly associated with gene expression. I have further validated this CNV and analysed the expression-profile around this CNV through NGS results from the 1KG data (both DNA and RNA). I have also reported ongoing work for replicating this association in a separate cohort in North American Brain Expression Consortium (NABEC). In addition, as a foundation for future development for a novel framework for analysing large scale omics dataset (such as the CNV-eQTL), I have reviewed the general theory behind matrix algebra for the commonly used linear model for association. Using this theory, I have applied and reported some of the preliminary results for compute intensive matrix operations on the parallel computing architecture of Graphics Processing Units (GPU).

In chapter 6, I have discussed some of the ongoing work and the future plans for work related to CNVs, genomics and personalised medicine. Lastly, I have summarized the important methods developed and the results obtained in this thesis.

1.2 Omics background

The suffix omics originated from the Greek language and refers to the idea of *totality* of some sort. In the field of biology this idea has translated into various specific domains related to biological molecules which can confer structural and functional characteristics of an organism. Some of the current omics field which has come into increased attention in recent times are genomics, proteomics, transcriptomics and metabolomics. These specific domains of omics have become increasingly crucial for understanding the molecular basis for several human diseases, which are believed to be manifested through unknown complex structural and molecular phenotypes. The field of *functional genomics* aims at finding genes in a given organism and its interplay with other omics such as transcriptomics, proteomics and metabolomics.

1.2.1 Genomics

The genome refers to the totality of DNA in an organism. In humans the genome is comprised of 23 chromosome pairs, where each strand of DNA comes from one parent. Though the genome of each human individual is more than 99.9% similar, the 0.1% difference accounts for much of difference that we see amongst ourselves. We now know that some of these differences can also be a causal factor for some of the heritable diseases like diabetes, obesity and epilepsy. Current genomic studies with the aid of recent high-throughput technological advances aim to find these minute differences in the human genome which can affect susceptibility to various diseases.

One of the most common forms of genetic variation often used in current large scale genomic studies is Single Nucleotide Polymorphism (SNP) [1]. SNPs are a single base pair change in the DNA at a particular location of the genome. In disease association studies SNPs are often associated with disease phenotypes or assumed to be lying in the vicinity of the true causal variant for the disease. This phenomenon has been exploited through high-throughput technologies like bead-array and NGS [2] based platforms for inferring disease causal genomic variants at the population level. These methods include gene mapping studies, fine mapping of genomic regions and the most popular approach which has been recently applied is genome-wide association studies (GWAS). GWAS aims at finding associations between the genomic variations and disease phenotypes of

interest. These phenotypes when related to diseases are often binary in form (*affected* and *unaffected*) or can be quantitative measurements like body mass index (BMI) [3] or blood serum levels [4]. All these phenotypes have been found to be heritable in humans, often through twin studies. Heritability of the phenotypes is an important aspect of GWAS studies and in this scenario refers to the phenotypic variance explained by the genotype, relative to the overall phenotypic variance [5].

There are several risk models used in the GWAS terminology at present, namely recessive, dominant, additive and multiplicative models. All these models have different encoding for SNPs [6]. In order to determine the association statistics for these models, construction of contingency tables for all the SNPs genome-wide is prohibitive (because of the billions of comparisons to be done, for e.g. 1 million SNPs for thousand people will require more than a billion comparisons) and hence in this scenario regression based models are often used.

Computational time for GWAS analysis is usually long. For example for a continuous trait analysed for 6000 individuals for 2.5 million SNPs can take up to 6 hours on a regular computer for analysis [7]. This kind of computation time is likely to increase in future with larger number of SNPs genotyped and for large sample sizes from big meta-analysis consortium. Further, the logistic regression based GWAS analysis will require more time since logistic regression involves coordinate descent methods for model convergence which are computationally intensive. However, some easy solution to this problem exists. GWAS analysis can be made parallelised easily because of the analysis of each SNP independent of each other. In addition to this, further speed up in computation time can be achieved by using optimised matrix operations and leveraging on the model sparsity [8].

One popular method in the regression setting is the univariate logistic regression model. Though this model has been widely used and has shown to work well, there are several drawbacks for this approach. This includes the problem of non-independence of the univariate tests. All SNPs used in the model, which are considered as predictors in the univariate model, are actually not independent of each other due to the existence of LD and haplotype structure. This can lead to bias in the results and the genome

wide p-value may need to be corrected and adjusted. In this scenario several alternative multivariate methods which can consider all SNPs simultaneously have shown much promise and are perhaps more appropriate in the real GWAS setting [9]. These multivariate methods usually assume an additive model of inheritance. However these methods are computationally intensive and harder to interpret.

The above mentioned analysis framework, which is usually more common for SNPs can also be extrapolated and applied for other genomic structural variants like CNVs. Unlike SNPs the number of genotypes can be more than two and can be further assigned probabilistic measures of copies in CNV confidence based on the certainty of their detection [10].

1.2.2 Metabolomics

Metabolomics refers to all low molecular components in a biological sample which are end products of a chemical process in a cell. A related field is *metabonomics* [11] which is defined as the quantification of the metabolic response of an individual to genetic or external influence. It can also be understood as the application of systems biology approach to metabolomics. The metabolome plays an important role in the functional physiology of an individual. There are three levels of influence for the metabolome which is dynamic and varies with time. The first two levels of influence are primary level of control through the host genome and the third level of control is through the co-metabolome i.e. the overall influence of all micro-organisms like bacteria, protozoa etc. present in the body. Recently, there has been increased interest in the gut-microbiome which has shown to play an important role in shaping the physiology of an individual and also influencing the disease-susceptibility. A complete sequencing of the micro biota of obese and lean twins [12] has shown that there are substantial differences in their composition and characteristics thus indicating their influence to disease susceptibility. A deep characterization of the influence of gut-microbiome on metabolomics can reveal important facts regarding disease susceptibility for an individual and more work remains to be done in this area. The last level of control on the metabolome is through the external environment of an individual including diet and other lifestyle factors. The influence of external environment is more dynamic and more difficult to characterize, yet its influence remains important.

Several high-throughput methods have been developed in the recent years which can enable deep characterization of the metabolome. These technologies include the sophisticated nuclear magnetic resonance imaging (nMRI), Proton NMR (H NMR) and widely used mass spectrometry. These developments have been successful in better understanding the role of metabolomics in several common diseases like T2D [13], cancer [14] and cardiovascular diseases [15]. These findings have also seen the development of a large number of online resources such as databases and software. One such widely used database is the human metabolome database [16]. This database was established to catalogue all existing knowledge about metabolites and their functions.

1.2.3 Transcriptomics

Transcriptomics refers to study and characterization of all RNA molecules in the cell, especially the translated mRNA from DNA. The main purpose behind characterizing mRNA levels is to determine which genes are expressed and how they vary with time or with disease states. In cell DNA quantity remains constant but levels of RNA changes with time and with other factors, like cellular environment and tissue type. This dynamic nature of RNA is attributed to both heritability and its environment. These factors make it difficult to characterize the influence of RNA in shaping an organism's phenotype.

Several high-throughput technologies have been developed in order to quantify mRNA levels. These include microarray based methods, qPCR and NGS platforms for RNA like RNA-Seq [17]. These approaches are based on different technologies and have different merits and purposes based on the desired objective. The microarray based hybridisation approach uses gene-specific probes and is designed from the cDNA. This is used in order to hybridize the RNA molecules first and then measure the intensity.

This intensity is next processed and characterized in order to ascertain the level of gene expression. Broadly, there are two techniques for hybridization of RNA molecules. In the first technique, synthetic probes are designed with perfect match and mismatch probes for RNA [18]. There are about 16-20 base pairs which are perfect match and about 25 which are mismatch. RNA from cell is next processed and allowed to hybridize with the corresponding probes. The level of intensity is then measured by a LASER and analysed. The second technique is based on competitive hybridization consisting of two

channels. Two samples are prepared, one with red fluorescent dye and the other with green. These two samples are then brought together on a microarray plate and allowed to hybridize. The resulting relative intensities are scanned and characterized based on their corresponding wavelengths. Once these steps are done, the gene expression values can then be taken further for downstream analysis. An alternate method to hybridisation for characterizing gene expression is based on the qPCR method [19]. This method can quantify the abundance of mRNA in a sample. Though for this method, the cost of single experiment is low compared to other hybridization methods, the cost per single gene still turns out to be higher. The next-generation and most advanced method for expression profiling is RNA-Seq. The RNA-Seq method can identify pre-specified RNA sequences as well as novel sequences and requires no threshold for the quantity of RNA that it can measure and characterize. This factor is a major limitation for microarray technologies at present. The flexible dynamic range for RNA-Seq based method is another distinct advantage over the other technologies.

Though qPCR and RNA-Seq based methods are more accurate, the microarray based approach can characterize up to 30,000 genes in a single experiment. For this reason the microarray technology still remains a popular choice for some of the current large-scale transcriptomic analysis. Though sequencing based platforms are more accurate they suffer from high costs and require high computational resources for extensive bioinformatics analysis. These factors make RNA-Seq still impractical for very large cohorts. In the near future with falling cost and improved bioinformatics analysis tools, it is expected that sequencing platforms will eventually supersede microarray technology. However, it should be kept in mind that each of the above discussed methods have its own drawbacks. Each of these approaches has its own unique systematic error due to various factors like pre-processing and batch effects.

Once the gene expression data has been processed and quantified the next step involves several downstream analysis steps. A commonly observed feature of the characterized expression data (through various technologies) is that many of the gene expression measurements are highly correlated [20]. This correlation can either indicate that similarly expressed genes are present in the same biological pathway or that there is a common mechanism for their regulation. These features can be further analysed using various statistical and machine learning approaches. Some of these methods include singular

value decomposition (SVD), principal component analysis (PCA), gene networks [21] and clustering [20]. Some studies aim to find which of these genes are differentially expressed. This can shed more light in the differences between disease states or tissues [22].

1.3 Genome Wide Association Studies (GWAS)

1.3.1 Single Nucleotide Polymorphism (SNP)

SNPs are the most commonly found variation in the human genome. It is defined as a single base pair difference at a particular genomic location with allele frequency greater than 1 % in the population [23]. SNPs have also been the cornerstone of GWAS where they have been used as markers to tag the disease causing variant. This approach has been successful for several GWAS analysis and has subsequently led to the next phase of GWAS based on genetic resequencing and fine mapping. Resequencing and fine mapping helps in pinpointing the causal gene and elucidating their functional characteristics in order to understand their biological significance [24]. It is estimated that there are about eleven million SNPs present in the human genome [25], many of which are still unknown. NGS based approaches are helping in the systematic discovery of these SNPs and some of these SNPs can also be population specific. Apart from NGS, high-throughput genotyping of CNVs for large sample sizes is now possible through SNP arrays. This I have described in the following sections. The current SNP arrays have both SNP probes and CNV-specific monomorphic probes. However, improved haplotype-based methods which can model both SNP and CNV together in order to detect better association still remains limited [26]. A reanalysis of the earlier GWAS datasets with improved SNP-CNV haplotype methods remains a promising avenue to explore in the future [27].

1.3.2 Copy Number Variation (CNV)

CNVs are a form of genomic structural variation (GSV) where segments of DNA are present in abnormal copy numbers. CNV can vary from one kilo base pairs (kbp) to several mega base pairs (mb) in size. It is estimated that CNVs cover about 12% of the human genome and overlaps with about 90% of the total number of genes [28]. CNVs in the genome can arise from either inheritance from one of the parents or can also arise

through de-novo mutations. A major study of the nature of origins of CNVs was done by [29]. Another study by [28] published the first comprehensive review of CNVs in the human genome. In light of human diseases most CNVs in the human genome were found to be benign and their role in common diseases still remain uncertain. However over the past few years the number of reports of CNV-disease associations is increasing. This observation highlights the need for further investigation and further development of improved methods for CNV detection and association [30, 31].

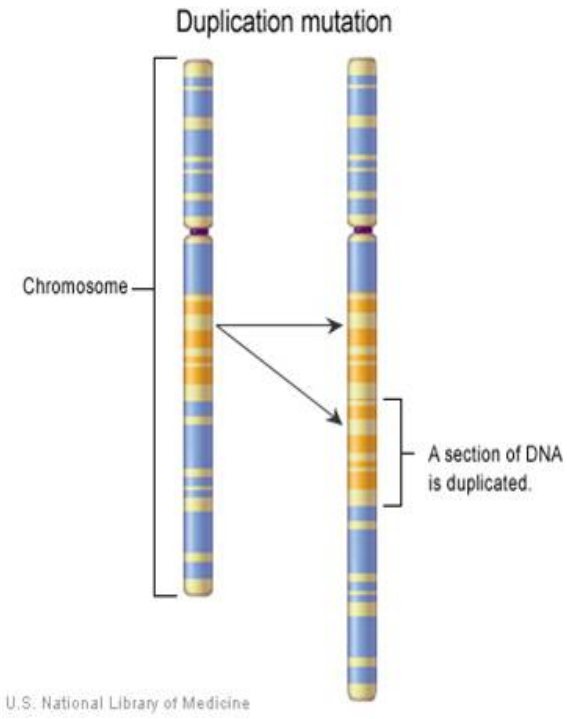


FIGURE 1.1: An example of a simple duplication

Source: National Human Genome Research Institute

1.3.3 Mechanism of genomic rearrangements in humans

Genomic re-arrangements are mutational changes in the genome such as duplication, deletion, insertion, inversion and translocation. Genomic rearrangements can range from hundreds of base pairs to few mega bases. However, these rearrangements are distinctive from other small-scale gene mutations like indels and point mutations. Genomic rearrangements have distinctive mechanism of formation and the way they effect upstream or downstream of a gene, and this can manifest as different disease phenotypes [32, 33]. Usually, disease causing variants can be recurrent with fixed breakpoints and length or be non-recurrent with no fixed size, but a smallest region of overlap (SRO). SRO in this case encompasses the disease causing variant.

There are three major mechanism for genomic rearrangements Non allelic homologous recombination (NAHR), non-homologous end-joining recombination (NHEJ) and Fork Stalling and Template Switching models (FoSTeS). These re arrangements when occur in germ lines, it leads to genomic disorders and when they occur in somatic cells, it can lead to cancer. CNVs can also be neutral in function. Some examples for these include changes in CNVs of dosage sensitive genes or CNVs disrupting genes leading to the creation of fusion genes. Rearrangements which are in the size range of 3-5 Kb can be visually seen, where as for smaller variants, other laboratory based techniques are needed.

1.3.3.1 Non allelic homologous recombination (NAHR)

NAHR have been found to occur preferentially within certain hotspot regions of the genome, usually within low-copy repeats (LCR) [34]. Many such DNA rearrangements have been observed in different individuals and the most common one observed are between two low-copy repeats (low copy repeats or segmental duplications). LCR are defined as region specific DNA blocks of 10-300 kb in length where they have 95-97 % similarity to each other. One study by Bailey and Eichler [35] have recently reported the evolution and distribution of LCRs in the human genome.

Due to the high sequence similarity non-allelic copies of LCR can sometimes get misaligned during mitosis and meiosis. The subsequent crossover in the misaligned state can result in observed genomic rearrangements in progeny cells. The non-allelic copies

thus act as mediators of homologous recombination and are responsible for the observed breakpoint clustering. When the LCR are located on the same chromosome and are in direct orientation they result in deletion or duplication and when in opposite orientation then they result in inversion [36]. When NAHR occurs on different chromosomes then they results in translocation.

Some other studies have shown that strand exchanges during NAHR cluster amongst certain genomic hotspots and also lie in the same genomic regions as allelic homologous recombination (AHR) [37, 38]. In addition, DNA double strand breaks have been found near NAHR indicating a functional link between the two. NAHR can take place when there is high similarity of identity of certain minimum length. This minimum length is known as minimal efficient processing segments (MEPs) [39, 40]. Several studies have shown the importance of MEPS for intra and inter chromosomal mitotic recombination. NAHR can also occur in both meiotic and mitotic cells. When they occur in germ line cells it leads to constitutional genomic rearrangements which leads to genomic disorders which can transmitted (germ line) or sporadic (de novo). An example of transmitted NAHR is *CMT1A* and *HNPP* caused by recurrent deletion/duplication at 17p12 locus [41] and an example of sporadic genomic disorder is Potocki-Lupski syndrome caused by reciprocal deletion/duplication at 17p11.2 [42].

NAHR, when it occurs in mitosis it results in mosaic population of somatic cells. Such somatic rearrangements have been related to many cancers. In addition, genomic rearrangements in somatic cells leading to the formation of CNVs have been shown in monozygotic twins and also in blood of healthy people. Such mosaic manifestations can result in many genomic diseases. One such example is *NF1* deletions causing neurofibromatosis [43]. Some studies have shown that mitotic and meiotic NAHR do not share the same spot. At present more technical analysis are required to characterize the similarities and differences between these two types. MEPs can be 300-500 bps in humans and it can differ from mitosis and meosis. It has been found that most of genomic re arrangements causing genomic disorders take place between 10-400 Kb in length and requires $> 96\%$ sequence similarity.

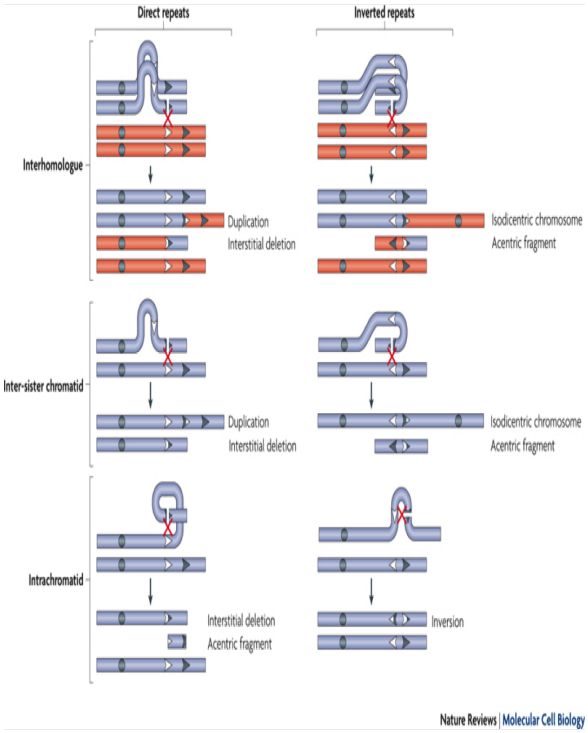


FIGURE 1.2: Non allelic homologous recombination (NAHR)

This figure shows six chromosomal outcomes of NAHR between repeats located on the same chromosome, which can be in direct or inverted orientation to each other. Such crossover recombination between repeated DNA sequences at non-allelic locations can generate duplication, deletion or inversions. Red and blue colour refer to homologous chromosomes and sister chromatids are sister chromatids are shown in the same colour. LCRs are shown by white and black arrows.
source: doi:10.1038/nrm2849 [44]

1.3.3.2 Non homologous end-joining recombination (NHEJ)

The other mechanism of non-recurrent rearrangements can take place by NHEJ [45]. There are four main steps in NHEJ [46, 47]: 1. Detection of DSB 2. Molecular bridging of the two broken DNA 3. Modification of the ends of DNA to make them compatible 4. Final ligation step. Unlike NAHR, NEHJ does not require LCR and MEPS and in addition it leaves an information scar at the rejoining site as during the editing there is addition or cleavage of several nucleotides on both ends. NHEJ is mainly used by eukaryotic cells to repair DNA double strand breaks and has been deeply characterized mammals as well as bacteria. NHEJ is central in the intrinsic physiological DSBs like V(D)J recombinations and also to DSB caused by extrinsic pathological DSB like ionization radiation or reactive oxidative stress. For some inherited diseases like SCID, NHEJ can account for about 15 % of the disease incidence [48, 49].

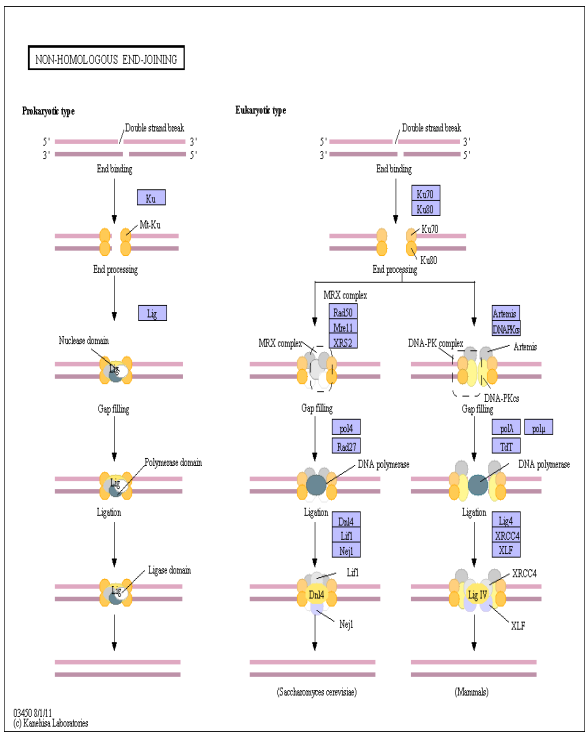


FIGURE 1.3: Non allelic homologous recombination (NAHR)

NHEJ are involved in repairing DNA double strand breaks in all stages of the cell cycle without requiring the need for sequence homology and thus is error prone. NHEJ works by a heterodimer (KU) with double-stranded DNA ends and then recruits DNA-Pkcs, processes the ends and next recruits DNA ligase.
Source: http://www.genome.jp/kegg-bin/show_pathway?ko03450

1.3.3.3 Fork stalling template switching mechanism (FoSTeS)

In past, NAHR and NHEJ were characterized through fluorescence in situ hybridization (FISH) which provided a resolution of about 150kb-200 kb. Next, with the advent of array based CGH allowed unprecedented resolution to observe the complex details of the third major mechanism of human re arrangements Fork Stalling and Template Switching (FoSTeS) [50].

Based on the findings in E.Coli Lee et al. [50] first proposed the FoSTeS DNA replication model. In this model, during DNA replication the replication fork stalls at a particular position and the lagging strand detaches from the original replication template, gets transferred and then anneals (due to the microhomology at the 3 end) to a nearby replication fork and starts DNA synthesis again. The transfer and annealing depends on the sequence similarity between the invaded and original site. When the switching and annealing takes place in another fork located downstream it results in deletion and when the switching takes places up stream it results in duplication. The fragment from the new replication fork can be in direct or inverted orientation to the original position and this depends on several factors such as whether the lagging or the leading strand was invaded and direction of fork progression [50].

This entire process of detaching, invading, annealing and extension can occur multiple times in different time points, thus showing the poor processing of the DNA polymerase and occurrence of complex genomic re arrangements. FoSTeS model is the only explanation for many complex re arrangements like at DMD locus [51], it also explains some other features like complex chromosome rearrangements (CCR, unveiled by cytogentic data), tandem duplications in *PLP1* and *LMNB* region, which were previously thought to be caused by combining HR and NHEJ, but now can be explained better through FoSTeS [52]. Results from several studies have further indicated that like NAHR and NHEJ, FoSTeS is also probably influenced by local genomic architectures.

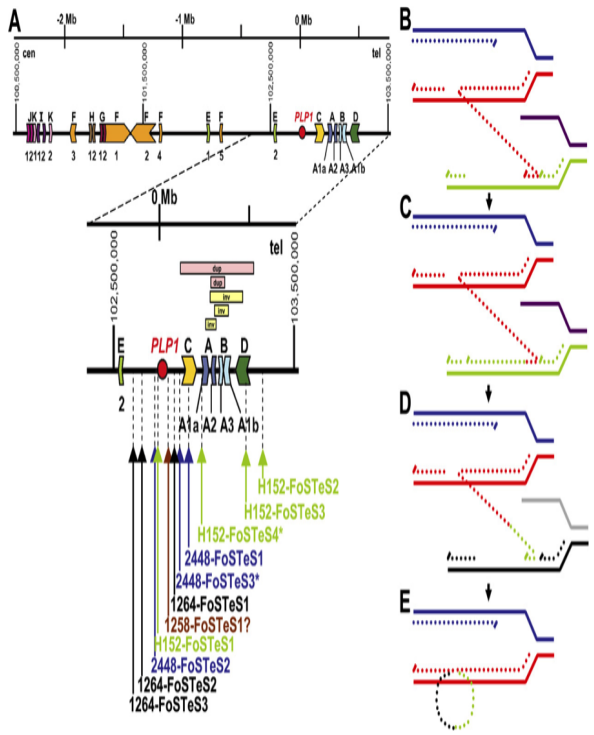


FIGURE 1.4: Figure showing FoSTeS mechanism

A detailed overview of the FoSTeS. A) shows location of FoSTeS in reference to LCRs. Out of the seven FoSTeS events, two (2448-FoSTeS1 and H152-FoSTeS3) occur within an LCR; and in additional one event (H152-FoSTeS4) occurs in the centromeric region within LCR-PMDC. In B), one replication fork (dark blue and red, solid lines) with a lagging strand (red, dotted line) invades a second fork (purple and green, solid lines), and next followed by (C) DNA synthesis (green, dotted line). After the disintegration of the fork (D), the original fork (dark blue and red, solid lines) with its lagging strand (red and green, dotted lines) invades a third fork (gray and black, solid lines). Newly synthesized DNA is represented by dotted lines. Serial replication fork disengaging and lagging strand invasion can occur several times before (E) resumes the the replication on the original template. [50]

1.3.4 Insertion deletions (INDELS)

Indels are an important class of genetic variation caused by mutations leading to net gain or loss of up to 50 nucleotides. It is the second most prevalent variation in the human genome and is believed to be as polymorphic as SNPs [53]. Analysis of the 1000 genomes (1KG) pilot-project data lead to the detection of about 1.3 million INDELS (less than 50 bp) out of which about 890 were found to be functionally disruptive [25]. Previously, detection of indels for large sample size was difficult due to technical reasons. At present, with the development of NGS platforms and other bioinformatics tools there is now a new opportunity to detect indels at a higher resolution. This can now help in characterizing their functional role to a greater extent.

1.3.5 Hardy-Weinberg Equilibrium (HWE)

The concept of evolution was first proposed by English mathematician Godfrey Hardy and a German physician Wilhelm Weinberg in the 20th century. The emergence of new species from existing species can occur through small changes within a species from one generation to another, occurring over a long period of time. Evolution can also be generally understood as the sum total of the genetically inherited changes in members of a population of a particular species. The effects of evolution affects all members of the population but it is the population as a whole that evolves. Evolution results in the change in allele frequency in the gene pool of a population. Hardy, Weinberg and the other population geneticists came to postulate that evolution will not occur in a particular species until certain conditions are met. These conditions include:

1. mutation is not occurring
2. natural selection is not occurring
3. population is infinitely large
4. all members of the population breed
5. matings between individuals are totally random
6. there is no migration in or out of the population

But these conditions are rarely applicable to any species. Hence, it can be safely assumed that evolution is taking place in all populations all the time. Hardy and Weinberg next came up with a simple equation which can be used to find the probable genotype frequencies in a given population and to track their frequencies over one generation to another. This idea came to be known as the Hardy-Weinberg principle and was formulated as the Hardy-Weinberg equation:

$$p^2 + 2pq + q^2 = 1 \tag{1.1}$$

Here, p and q refer to the frequency of dominant and recessive alleles (for e.g. A and a) in a given population. Then the expected frequencies for homozygotes are given by: $f(A) = p$, therefore $f(AA) = p^2$ $f(a) = q$, then $f(aa) = q^2$ and for heterozygote $f(Aa) = 2pq$

p^2, q^2 and $2pq$ are known as Hardy-Weinberg proportions and by applying binomial expansion and putting the constraint that the sum of all genotype possibilities must be equal to 1, equation 1.1 can be proven. Further, if the union of gametes from one generation to the other is random it can be shown that allele frequencies from one generation to another is constant [54].

1.3.6 GWAS background

The human genome project [55] started in the year 2000. The main aim of this project was to characterize the entire DNA sequence of a human genome and further understand it from a physical and functional point of view. This project subsequently led to the discovery of many genetic variants associated with different disease outcomes and further helped in conceiving the possibility of personalised medicine. The initial success of the human genome project and the new aim of personalised medicine next paved way for the International Hapmap study [23]. The Hapmap study aimed at uncovering the haplotype map of the human genome and elucidated the common patterns of genetic variation found across different populations in the world [56]. The results obtained from the Hapmap study helped the scientific community next to propose the *common disease common variant hypothesis* [57, 58]. This hypothesis asks whether the true disease causal variants can be captured through the phenomenon of LD with common genetic variation like SNPs [59]. Several initial success stories confirming the feasibility of such kind of

hypothesis driven genomic studies helped in heralding the era of genome-wide association studies (GWAS). This mainly started around the year 2005 [59, 60]. The GWAS success proved to be a catalyst for several commercial companies like Illumina and Affymetrix to come up with high-density genotyping microarray chips and also led to the development of advanced sequencing (NGS) based platforms [61, 62]. Some of these genotyping platforms can interrogate more than two million genomic loci simultaneously for large sample sizes, while the NGS technology has enabled sequencing of entire genomes in a matter of days and at a fraction of cost compared to the previous sequencing-based methods which were considerably more expensive and took months or even years to complete.

Most importantly from personalised medicine point of view many of these findings from the GWAS era were successfully translated into gene-therapies and promising drug-targets for several diseases [63, 64]. However, in spite of these successes it was observed that for many diseases the common SNPs accounted for only a partial fraction of the total observed heritability. This unaccounted heritability problem was referred to as the *missing heritability* problem [5]. Many geneticists and scientists now believe that some fraction of this missing heritability can be accounted for by CNVs, epigenetic factors and to some extent by gene expression [5].

1.3.7 Genetic basis of GWAS

Most human diseases cluster in families and are known to be affected by several genetic and environmental factors. Since Mendels first experiments elucidating the inheritance pattern in plants, and later the work of Galton and Fisher, family based studies have dominated genetic studies. In family based study designs, Mendels laws are used to compute expectations for disease risk in the offspring based on the parental genotypes. Some popular methods for family based study rely on trios (two parents and an affected offspring) and this approach were first introduced by Rubenstein et al. (1981) and Falk and Rubinstein (1987) [65]. Later, a detailed analysis of such designs was reported by Spielman et al. (1993) [66], Ott (1989) [67], and Terwilliger and Ott (1992) [68].

Next, with the progress of the Human Genome Project, elucidation of genomic markers spanning the entire genome have enabled discovery of many Mendelian disease traits

for many genes through Genome Wide Association Analysis (GWAS). This association analysis is fundamentally different from the family based linkage analysis since this approach does not require family based design, and inferences on genetic association are made from unrelated individuals [69].

GWAS involves systematic scanning of biomarkers (SNPs) across the genomes of people in order to identify, treat and prevent common complex diseases such as asthma, cancer, diabetes, cardiovascular and neuropsychiatric diseases. The genetic basis of this approach relies on the fact that causal genomic polymorphism are more likely to be found in people with disease, compared to those without. The main strategy in this approach involves two groups of participants: one with the disease (cases) and similar group of people without the disease (controls). DNA is obtained from each of these group, usually though blood or saliva sample, from which, next, cells are harvested. The DNA from these cells are then purified and scanned through high throughput genotyping platforms, which surveys each individuals genome for strategically selected markers (SNPs). This usually proceeds by analysing the DNA from cases and controls on a SNP genotyping chip, which can have up-to 300,000 to 2,500,000 probes interrogating different loci in the genome. If certain variants are found to be significantly more frequent in cases compared to the controls, then that particular variant is said to be associated with the disease outcome.

GWAS approach is an important improvement over the earlier family based candidate gene studies, which were often limited by low sample size and power to detect relevant makers at the desired significance threshold [70]. In addition, for complex multi factorial diseases like diabetes and epilepsy, the true disease-causal-variant may have modest odds ratio and effect size. In this scenario, family based linkage studies have low power and resolution. GWAS, in this respect proves to be a powerful alternative approach, where thousands of individuals can be analysed for millions of markers simultaneously. The success of GWAS has led to the identification of hundreds of new susceptibility loci for several complex multi-factorial diseases.

GWAS, in addition allows pathway and haplotype based analysis, which can allow better understanding of the biological context and history of such disease causing variants [71, 72]. However, it should be bared in mind that in reality the detected variant itself

may not be the true causal variant for the disease, but often serves as a pointer to the actual variant lying in linkage disequilibrium (LD) with the genotyped marker. Due to this, results from the GWAS often require additional follow up sequencing experiments in order to pin point the exact genetic variation involved in the pathogenesis of the disease.

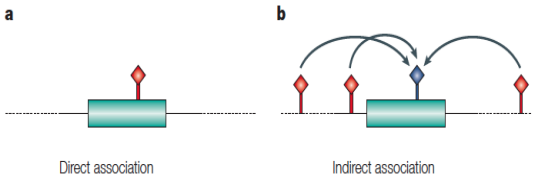


FIGURE 1.5: GWAS approach for detecting causal variants

There are two approaches for detecting candidate SNPs in GWAS analysis. In a) the candidate causal SNP is tested for association directly and in b) the candidate SNP tested is in LD with other SNPs around it, one of which is the causal one. [73]

1.3.8 Significance of LD in GWAS

LD refers to the correlation between two genomic loci (SNP or CNV) observed more often in a population than expected from random haplotype formation derived from population allele frequency. This concept is the backbone for modern GWAS approaches which is based on the assumption that common SNPs in the human genome can tag the disease causal variants through LD. Studies of LD have further helped in understanding the human haplotype diversity between and within populations. The effect of positive selection on LD has been reviewed by [74, 75].

The relationship between SNPs with common CNVs has come to much attention recently. This relationship is often termed as *CNVs tagged by SNPs*. In a recent landmark study of CNVs [76], the authors found that most common CNVs are tightly correlated with nearby SNPs through LD. An interesting feature about this finding is that this gives rise to the possibility of using SNPs as a proxy for discovering CNVs. If the feasibility of using such an approach is successfully validated then in the future we will have the possibility to develop better models of CNV-SNP haplotype and use them for large-scale

association studies. This might also help in reducing the cost of CNV genotyping for large studies.

Nonetheless, more work needs to be done in this area by improved modelling of the SNPs and CNVS. Models for LD for SNPs outside and within the CNVS also need to be differentiated, since correlation structure varies between the two. Several methods have been proposed to deal with this issue [77]. Some models for duplications have also been proposed and are being developed at present. The landmark CNV study by [76] had concluded that since most common CNVs were well tagged by common SNPs their association to diseases have indirectly been explored through earlier GWAS studies. However, new metrics of LD for CNVs and new algorithms may lead to the discovery of new CNV regions. These have the potential to reveal novel associations with diseases. An effort to re-analyse previous GWAS data with new methods and in light of NGS can achieve this goal.

1.3.9 Haplotypes

The human genome is diploid which means it has two chromosomes, each derived from a parent. The ordered list of alleles on each of the chromosome is referred to as the haplotype. Due to linkage, alleles within genomic regions tend to be transmitted together from one generation to the next. This gives rise to a block-like structure in the genome. Many haplotype-based approaches for GWAS use this information for phasing [78].

Phasing refers to the differentiation of the maternal and paternal allele information from diploid genetic data. This is particularly useful for studying phenomenon like allele-specific gene expression and improved association test by inferring ancestral haplotype [79]. This method has shown to give greater power to detect disease causing mutations. Importance of haplotype information in next-generation GWAS has been recently reviewed in an important study [80].

1.3.10 CNV and GWAS

CNVs are prevalent in the human genome. CNVs are also found to be associated with several human phenotypes including disease outcomes [31, 81]. In the recent years many

methods have been developed for detecting and associating CNVs with various phenotypes. These approaches are similar to the methods used for SNPs [10], however there are some important differences between the two. I have discussed in details several of these differences in the later chapters. One of the commonly used methods for CNV detection prior to NGS was the aCGH platform, which was also used by the landmark Wellcome Trust Case Control Consortium (WTCCC) CNV analysis [76]. Recently, the SNP genotyping arrays used to detect SNPs were further extended by the inclusion of non-polymorphic CNV detection probes. This allowed simultaneous profiling of SNPs and CNVs through a common genotyping chip. This development enabled CNV detection for large sample sizes easy to do and more cost-effective. Since then several sophisticated CNV detection methods (often HMM based) have also been developed to call CNVs from such SNP/CNV genotyping platforms. These include the widely used CNV-calling algorithms like PennCNV [82] and cnvHap [10]. These methods have further been able to borrow many of the standard concepts in GWAS such as imputation, phasing and inflation and successfully applied them for CNV based analysis [10]. In spite of these developments, compared to the success of GWAS, there has been paucity in the number of CNV-disease associations. Some of the reasons for this include the fact that there are several artefacts prevalent in CNV-association analysis. These include batch effects, DNA source and also high false positive rates for the current CNV detection methods [27, 76]. Another important factor is that most SNPs discarded in QC process are the ones which have allelic imbalance due to the presence of CNVs. This fact is often overlooked and needs to be taken into account.

1.3.11 CNVs and drug response

An initial investigation has indicated that CNVs look promising for explaining inter-individual drug-response. On doing an analysis of the overlap between common CNVs in the genome and the known common drug-target from the DrugBank database [83], it was found that the current standard genotyping platform with a million probe covers about 91% of the known drug-targets regions defined in the DrugBank database. The high incidence of overlap observed is promising. This can perhaps in the future help in developing personalized medicine based methods for treatment and diagnosis for a person based on his or her CNV profile. This can also prove particularly interesting for cancer, where CNVs are a commonly found genetic aberration. More work in this

direction, especially in the light of NGS data can reveal interesting results and remains to be explored.

1.3.12 Finding missing heritability in GWAS

GWAS is primarily based on the common disease common variant hypothesis. The hypothesis postulates that the common complex multifactorial diseases like type 2 diabetes and obesity can be explained by common allelic variants ($> 1-5\%$) present in the human population [84, 85]. However, the results from numerous studies based on this approach have failed to reveal the true genetic causes for common complex diseases. So far, except for age related macular degeneration, where several variants of large effect size have been found to increase disease risk fold by 2 or more, other variants found through GWAS alone or in combination, confers only small increment of disease risk, usually about 1.1-1.5 in most cases [86]. This lack of explanation through common genetic variation for complex common diseases is often referred to as the missing heritability problem in GWAS. In the GWAS terminology, heritability is the proportion of phenotypic variance in a population explained by additive genetic factors. One classic example is heritability of human height, which through family based studies, has been estimated to be 80% heritable. In spite of the high heritability, all the GWAS loci found to date to be associated with human height, explain only about 5% of the phenotypic variance [87].

The main question is why so much heritability is unexplained by GWAS findings. There are many plausible explanations to the missing heritability in GWAS. One possible explanation is that there are large number of variants with small effect size, many of which are yet to be found [88]. Another important possibility is the existence of rarer variants, with possible large effect sizes. At present, the capability to detect rare variants through current genotyping platforms remain poor since the available arrays are more suited to variants present greater than 5% of the population. Further, current platforms are also inadequate for detecting structural variants and the association models are often inadequate for accounting for gene-gene interactions and the shared environmental factors amongst subjects in the study. Another related observation from GWAS studies is that most of disease associated variants occur in the protein coding regions of the genome, but 80% of the variants which have been found to be associated with common diseases, fall outside the coding regions [89].

This makes it important to include both coding and non-coding regions in the analysis framework for GWAS, but such coverage is lacking from the current commonly used genotyping chips in GWAS.

Potential sources of missing heritability

The allelic architecture for different traits including the number, type, effect, size and frequency of various susceptibility variants can vary . In this scenario, the missing heritability can take different forms for different diseases [57]. Age related macular degeneration is a good example where a handful of common variants show a large effect, explaining much of the heritability. For other diseases like Crohns disease, the proportion of heritability explained by GWAS candidates is far less, in spite of a larger number of variants found to be associated with the disease [90]. There is no ready explanation for such differences in the allelic architecture except for some possible reasons which can be attributed to the stage of investigation or the sample size (It has been shown that increasing sample size increases number of variants detected).

There have been several reasons put forward by population genetics to explain this missing heritability. Some theories suggest that in variants which have high predisposition to the disease outcome, would normally have low reproductive fitness and hence selected out. This might explain the relative lack of variants found to be associated for several neuropsychiatric diseases like autism [91]. Selection might also be responsible keeping genetic effect size low since individuals with extreme traits would have been removed by long term stabilization selection [92]. Selection has also been found to be affect the ability of the GWAS to detect different loci for different complex diseases, particularly in case the some diseases are affected more by selection than other diseases. One such example is that immune and infectious agents which have been recognized as the strongest selection pressure in human evolution, and immune related genes have been found to be the most associated variants for diseases like Crohns disease and other immune genic diseases [86].

In addition another important of current GWAS has been that the study design have incompletely surveyed the potential causal variants in each gene. The marker SNPs used in the study may underestimate the actual risk associated with the actual causal

variant. One such example in this case is for lipid associated genes where 11 out of 30 genes carrying common variants associated with lipid levels also carried known rare alleles of large effect size identified in mendelian dyslipidemias including *PCSK9*, *ABCA1* and *LDLR* [93, 94].

Another important point to keep in mind is that GWAS studies have focused on European populations where the genetic diversity is low compared to the populations from African ancestry [95] [96]. Thus extending coverage to non-African populations can reveal new variants for common diseases. In addition, studying isolated populations enriched for rare disease causing variants can further help in understanding the missing heritability problem [97]. Most importantly one should keep in mind that the current heritability estimates determined from family based studies (narrow sense heritability) can be inflated due to several factors like dominance and epistasis, gene-gene interactions and gene-environmental interactions.

Rare variants and unexplained heritability Due to the modest explanation of missing heritability through the initial wave of GWAS studies, there is now more emphasis on rare variants. Rare variants which vary from 0.5 % - 5% can not be captured through genotyping arrays nor do they have large enough effect size to be captured through family based linkage analysis [98]. It is expected that low frequency variants can have substantial effect sizes, increasing disease risk two-to-three folds without showing any Mendelian segregation and thus contributing substantially to missing heritability [98]. For example for type 2 diabetes, it is estimated that twenty variants with allele frequency of 1% and odds ratio of 3 can explain most of the familial aggregation.

Detection of low frequency and rare variants is likely to be captured efficiently through next generation sequencing. Sequencing a larger fraction of people with extreme phenotypes for quantitative traits like lipid levels or age of onset of disease, can help in further identifying common and rare variants. Such extreme traits are more likely to carry LOF alleles. Projects like the 1000 genomes [25], which aims at capturing all variation at 1% allele frequency in different population from around the world, will also hasten the discovery of rare variants and their role in explaining heritability.

In addition to this, structural variation and CNVs like indels, inversions and translocations can account for some of the missing heritability further. CNVs (1Kb or longer)

are present in variable numbers in patients but most of the differences arise from a limited number of CNVs with allele frequency $> 5\%$. Also, large and rare CNVs and affecting many genes have been found to be highly penetrant [99], where as common disease associated CNVs are more likely to be smaller 20-40 kbs. Except for cnvHap, there are limited methods available to detect and genotype small and common CNVs.

1.3.12.1 Missing heritability and the role of common CNVs in diseases

Though GWAS has been widely successful it explains modest fraction of the trait heritability, often referred to as the *missing heritability*. This gap in heritability perhaps can be explained partly by CNVs and by epigenetic factors [98] and gene expression. The role of CNVs has been increasingly highlighted by various recent studies. There is also a rapid increase on reports of the role of common CNVs in common diseases in the last five years [100]. In addition to this, studying the inter-population differences in CNVs can perhaps help in explaining adaptation, differences in susceptibility to diseases, gene expression and also selection. A previous study specifically examined and showed that CNVs can explain up to 20% of the variation in gene expression [101]. This was attributed mainly due to the variable copy number at the regulatory regions of the genome [101]. The role of CNVs in relation to transcriptomics, heritability and pathway needs further method development. This might perhaps better explain the missing heritability.

1.4 Review of current approaches for CNV discovery and validation

Whole genome CNV studies have become more common with the possibility of genome-wide CNV profiling using high-density SNP arrays for large sample sizes. Though NGS may promise to replace array-based approaches, sequencing large number of samples for various diseases is not common at present (though we can expect it will happen in the near future). In this report I have mainly focused on the issues related to array-based techniques such as CNV genotyping. However, I will also briefly review some aspects of NGS and CNV detection from it.

At present the popular methods for CNV genotyping for large sample are commonly using high-density SNP arrays. Some of the widely known problems related to high-density SNP platforms are the high false-positive rates and the high signal to noise ratio. The WTCCC CNV study ascertained that some of the reasons for such problems are due to several factors like homology, source of DNA (in particular difference between cell lines and whole blood) and batch effects [27]. Thus it is important to keep these aspects in consideration while conducting any genome-wide CNV study and further to take measures to correct for them. In my case study while performing the genome-wide scan for CNVs in epilepsy I took several measures to correct for these factors. I have described them in details in chapter 3.

1.4.1 Normalization of array intensity data

Intensity data from genotyping arrays (used for clustering both SNP and CNVs) are quite noisy and subject to the quality of the DNA and its hybridization with the test probes. Different technology platforms use different methods for summarizing the intensity values while taking into consideration various factors. I will not delve into the details of the low level data-processing used by different technologies but refer to the commonly summarized and reported intensity values. These values are usually denoted by X and Y intensity values corresponding to the two alleles A and B. These values are then further processed to derive two values namely Log R Ratio (LRR) and the B allele frequency (BAF) [102]. The LRR value directly correlates with the total intensity measurement and thus can reflect presence of CNV, whereas the BAF relates to the frequency of the B allele in the population. The generalized formula for LRR and BAF are as follows for X and Y intensity measurements

$$LRR = log_2(X + Y)$$
$$BAF = Y/(X + Y)$$

However in reality the derivation of the actual LRR and BAF is much more complicated than the above mentioned formulae, since many other factors are taken into account and included.

The technology used for detecting CNVs may be different for various platforms. We expect different kinds of nuisance factors from the measured intensity. Some of these

nuisance signals can arise due to several factors like sequence content of the probes in the genomic region, biases due to allele specificity or DNA amplification, errors due to scanning and human errors in handling of the samples. Though some of these factors cannot be accounted for by any statistical model, there are some common aspects that can be dealt with for all platforms. A common feature for many normalisation pipelines is standardisation of signal measurements of probes within a particular sample and also within all samples. For CNVs, often the population distribution of the intensity measurement is examined for each probe based on the assumption that copy number alterations are relatively sparse. The assumption of CNV-sparsity has helped in removing some of the biases due to experimental variability. This is mostly acceptable for most cases except for tumour samples, where the lack of normal copy number regions for the reference sample is not present [103].

1.4.2 Genomic waves

In the results reported by the WTCCC CNV study [28] one of the major technical artefacts found was spatial auto-correlation or wave effect. The study had reported that the hybridisation intensity of the DNA was related to the clone used at that position. They further added that these wave effects correlated with the GC content of the clones. It was observed that regions of low GC content were correlated with peaks of the wave and the high GC content regions correlated with the troughs. Several follow-up studies have reported that it is important to remove such wave and GC effects and proposed several methods to correct for them. These studies also confirmed that correcting for this bias leads to much better CNV clustering quality [104]. Usually, some methods to correct wave effects involve fitting a *loess* curve to the total intensity values while making sure that the real CNV do not get affected by the fit. In addition, the GC content is also regressing out the model for improved CNV detection. These quality control (QC) steps are now a common practice in most CNV-calling pipelines.

1.4.3 Methods for CNV genotyping for discovery

There are several methods available for the discovery and validation of CNVs in a population-based study. CNV discovery is usually carried out by CNV genotyping using high-throughput genotyping arrays or more recently trough NGS. These methods

can be applied in a locus specific manner or genome-wide, depending on the context of the study and the planned coverage. For whole genome coverage most commonly used methods are array comparative genomic hybridisation (aCGH) and high-density SNP arrays. Recently, NGS can now be used for whole genome analysis for thousands of individuals [105].

1.4.4 Array Comparative genomic hybridisation

Amongst the bead-array based platforms, the array comparative genomic hybridisation (aCGH) platform has proven to be most sensitive for detecting CNVs greater than 100 kbps. This has led to the creation of several gold-standard datasets. This includes the first comprehensive genome-wide characterization of CNVs in the human genome [28]. In general, the aCGH platforms utilize two different coloured dyes or fluorescent markers. The test and reference DNA is labelled with these fluorescent markers and then simultaneously allowed to hybridize on solid glass. The difference between the two intensity between these dyes is then measured. Next, from these values the copy number gain or loss information can be determined.

At present there are two types of aCGH platforms. One of these is based on bacterial artificial clone (BAC) probes and the other is based on oligo-probes [106]. The main difference between the two is that BAC probes are more appropriate for large structural rearrangements while the second type based on oligo-probes is more appropriate for smaller deletions and duplications. Both these methods have advantages and drawbacks.

The BAC probes provide better coverage of the genome with high signal to noise ratio but it fails to capture smaller variants less than 50 kbps [107]. It should be kept in mind that the smaller variants are also known to be more prevalent, hence they are most likely to be missed by BAC. In addition, another drawback of BAC is that it can lead to overestimation of CNV size due to the presence of large insert sizes. Due to these reasons this method is more appropriate when we are looking for large CNVs, but ineffective for smaller ones. The second type of aCGH is oligo-nucleotide arrays. The oligo-aCGH has good genomic coverage but they suffer from poor signal to noise ratio which leads to higher variability between studies. One of the main advantages of the oligo-method because is its smaller probes size, hence it can capture smaller variants better. The

oligo-aCGH platform was used by the landmark Wellcome trust CNV study for eight common diseases [76] and lead to the discovery of several common CNVs associated with diseases like T2D and psoriasis. A common drawback for both BAC and oligo-aCGH is that in general they turn out to be expensive for large sample sizes. Further, the aCGH platforms do not measure the allelic information like BAF which can prove to be useful for indicating the presence of CNVs in many regions of the genome.

1.4.5 High density SNP arrays

The most popular method for whole genome coverage for CNVs in large cohorts is by using high-density SNP arrays. This has been made possible in some of the current platforms by incorporating CNV specific monomorphic probes in addition to the normal SNP probes. Some examples of these kind of chips are the recent Illumina Omni 2.5M chip, Affymetrix 660K platform and the older Illumina 660K bead-array chip. These chips cover significant portion of the genome and have shown to have high signal to noise ratio. Some of the popular CNV detection methods for these kind of platforms is carried out using model based approaches like HMM [10, 82]. However other non-parametric methods have also been proposed and applied. In the HMM models the intensity and the allelic content of a SNP is jointly modelled to give detailed CNV-genotype output. These genotypes are usually used further for association-analysis using methods like regression.

Like some of the previous studies, in my analysis I found that monomorphic CNV probes from the high-density chips were noisy (while using cnvHap), and led to a high false positive rate. Due to this problem, I relied mainly on the intensity probes and genotype data from normal probes. Some reviews have suggested that each of the monomorphic probes in a CNV region should be given different weights (for e.g. using principal components based prior data) to reduce over calling. The rationale behind this is that each of the CNV probe in a particular genomic region is essentially measuring the same CNV and hence should not be considered independent from each other.

1.4.6 Next generation sequencing and genome assembly comparison

With the recent development of NGS based approaches, several new methods for capturing GSV at a single base pair level has emerged. Thus sequencing is now helping in discovering new variants like CNVs, indels and inversions at unprecedented high resolution. Many of these variants have been found to have new biological relevance for diseases and other phenotypes [108]. Though sequencing for thousands of samples is still quite expensive, the price for sequencing is getting cheaper every day. It is expected that in the near future sequencing for large cohorts will be possible, some of which have already started through projects like the UK10K (`http://www.uk10k.org/`).

There are two broad categories for inferring CNVs from NGS methods. One is based on genome alignment and the other is based on genome assembly. The genome assembly based approach genomes for a population is assembled de-novo and then aligned to reference genome and then compared. Deletions are identified as gaps in the sample sequence, where as duplications are identified by gaps in the reference sequence [109]. One advantage of this method is that it assumes reference sample as haploid due to which homozygous events are easier to detect. However, on the other hand one of the main drawbacks for this method is that DNA used for assembly is usually drawn from different sources and thus not ideal for sequence comparison.

Alignment based approaches are based on short-reads produced by NGS platforms like Roche 454 and the Illumina Genome Analyzer [110, 111]. These methods avoid bacterial cloning and create millions of short-reads which are then mapped to the reference genome using a commonly used technique called paired end mapping (PEM). In PEM library, fragments of 300-500 base pairs are created and then the difference between these fragments and the reference genome can indicate the kind of CNV present. One advantage of this method is that it allows the detection of inversions and CNV breakpoints at a single base pair level. But on the downside, it cannot identify variants in complex genomic regions like the major histocompatibility complex (MHC) [112]. One alternate strategy to PEM is using read-depth approach where overlapping intervals of sequences are used as a quantitative measure of CNV. In this method change of read-depth across multiple genomic windows indicates deletion or duplication. However one

of the limitations of using the read-depth approach is that it is not able to ascertain rearrangements or structural variations which are highly repetitive. This becomes problematic for detecting insertions deletions and their precise locations. One useful strategy for overcoming this problem could be by using combined depth and PEM for enhanced CNV detection and has been applied through the cnvHitseq algorithm [113, 114].

1.4.7 Lab based methods for CNV validation

Prior to NGS, the most common approaches for CNV discovery in large population cohort were based on high-throughput methods like SNP chips and aCGH platforms. However, it is well known that CNVs discovered through such platforms can suffer from many confounding factors leading to many false positives. Thus, it is a common practice to validate the CNV findings in wet-laboratory before publishing. Based on the context and the application of the study design there are various methods for validating CNVs. When the target loci to be validated are small, quantitative polymerase chain reaction (qPCR) and multiplex ligation-dependent probe amplification (MLPA) methods can be appropriate. However, if the number of loci to be tested are large then custom made aCGH arrays or exome-capture based sequencing can be a more suitable approach for validation. Another advantage of doing a lab-based method for validation is that it allows comparison of different CNV-calling algorithms. However, I do not cover the details for this possibility here. In my analysis, one of the main aims was to discover and validate small, common CNVs associated with several diseases like epilepsy. For this purpose, I initially considered a custom aCGH or exome-capture method for about 50 loci or alternately use a qPCR for a handful of test regions.

1.4.7.1 MLPA

MLPA is a kind of PCR which can identify multiple targets using only one primer. It can interrogate about 40 genomic regions simultaneously for CNVs. It relies on the oligonucleotide probes to generate locus specific amplicons [115]. These amplicons are then resolved using capillary electrophoresis where deletions are detected as reduced peaks and duplications are detected as enhanced peaks. MLPA was been used successfully in determining some of the large rare deletions in obesity like in the study by [99].

1.4.7.2 qPCR

qPCR is a kind of PCR where DNA is amplified and quantified at the same time. It is one of most commonly used methods for detecting CNVs at targeted regions. This method is relatively high-throughput and technically easy and efficient to carry out. However, this method suffers from a drawback in that like MLPA, it cannot be used simultaneously for several targets in a single reaction. Further, precise measurement of CNV break-points is not possible as it provides CNV measurement as a continuous measurement [116].

1.4.7.3 Paralogue ratio test

Paralogue ratio test is used for accurate and high-throughput detection of CNVs at targeted loci. This method involves simultaneous amplification of a target and a reference DNA which is simultaneously amplified using the same primers present in the same reaction medium. The difference between the target and reference DNA is then quantified to detect the presence of a CNV. This is mainly done by comparing the internal differences in measurements for the target and reference sequences. [116, 117].

Chapter 2

Background Methods

2.1 CNV calling from genotyping arrays

Previously in Chapter 1, I have described how genotyping arrays have enabled high-throughput analysis of SNPs and CNVs in GWAS. These include genotyping of millions of markers in thousands of individuals simultaneously and subsequent laboratory-based validation methods to confirm the CNVs. In this chapter I have reviewed the various analytical frameworks and methods after genotyping, mainly involving CNV-calling and association. I have reviewed several models and methods for calling CNVs from genotyping platforms. I have avoided describing methods that relate to calling CNVs from NGS platforms, since it is beyond the work done in this thesis. I have further described several commonly used association models for both CNVs and SNPs. Lastly, I have touched up on several common topics related to genomic studies like methods for clustering, variable selection, prediction and classification

The various methods used to call CNVs from the bead-array technology can be broadly classified into the following categories.

2.1.1 Detecting CNVs from genotype calls

The expected genotype calls from SNP arrays are usually AA, AB and BB. In the presence of a deletion and duplication the intended calls falls outside these three classes

and shows up as improper genotype call or deviation from Hardy-Weinberg equilibrium (HWE). This deviation from expected SNP genotypes has been used successfully to detect CNVs [118, 119]. Here the authors have reported the detection of CNVs based on no calls or as nulls. However, one known problem with such kind of approaches is that stretches of homozygous diploid genotypes can sometimes be incorrectly detected as deletions. In addition, the sensitivity to detect duplication through such methods remains low.

2.1.2 Detecting CNVs from one dimensional segmentation

This approach was originally developed for detecting CNVs from the aCGH platform [120]. In this approach, the total intensity value LRR is modelled in one dimension through various models including both non-parametric and model based approaches. Model based approach includes methods like the HMM. One common aspect in these methods is that deletion and duplications are usually seen as downward or upward shifts in the intensity values and thus makes it easy to interpret results or visually inspect them. Further details of non-parametric approaches and model based methods are explained in the next sections.

2.1.2.1 Non parametric methods for CNV detection

Non-parametric CNV calling methods are more generic and assume minimal structure in the intensity data and the underlying noise distributions. The two commonly used methods which use non-parametric approaches are Smith-Waterman Array [121] (SW-Array) and Circular Binary Segmentation (CBS) [120]. SW-Array uses a modified version of Smith-Waterman dynamic programming approach to identify regions of CNVs spanning several probes. However, dependence amongst the probes is not modelled through this method thus leading to high false positive rates. Prior to CNV calling such approaches were used for identifying genomic regions with unusual properties. CBS does not make any assumption in the noise distribution in the intensity data and splits the data into several child segments. A similarity statistic is computed for these child segments and then compared to the reference distributions. The reference distributions are calculated through a permutations-based method. Based on the likelihood of the similarity statistic a threshold is decided for accepting or rejecting the child segments. This process

continues till the sub-segments reach some stopping criteria and results in regions of homogeneous intensity data.

However, the above methods still perform poorly when considering sensitivity of CNV detection and further these models do not incorporate the population level allele frequency information (BAF) into the model. Further, some of these non-parametric approaches have also been extended to CNV discovery using SNP arrays [122, 123]. However a systematic comparison of such methods with the current approaches still remains to be done.

2.1.2.2 Model based approaches using HMMs

Model-based approaches assume an underlying distribution in the intensity data. One such application of model-based approach is dChip [124]. This method uses a HMM-based approach with the assumption that the observed intensity follows a t-distribution. Usually HMM-based models are based on the fact that a CNV spans several contiguous probes, thus creating a linear spatial structure. The genotype at each of these probes is the observed sequence and the hidden state represents the copy number at a particular probe. The most probable path for the hidden state path is next inferred through a dynamic programming (DP) approach like the Viterbi algorithm.

One commonly used approach by which the HMM model is improved is by modifying the transition rate matrix. One such scenario is when the switching probability is controlled by the distance between probes i.e. a large distance between probes makes transition from one CNV state to another easier compared to closely spaced probes. Improving modelling accuracy through the transition matrix was first implemented in the dCHip and CNAG software [125] and subsequently in several other recent software.

In next section, I have described in details the HMM model, some important issues related to it, the dynamic programming method and the EM algorithm for estimating the model parameters. An understanding of these concepts proved important for me in order to modify and apply the cnvHap algorithm for my analyses in the later chapters.

2.1.3 The hidden Markov model (HMM)

HMM has been used widely in machine learning applications including speech recognition [126] and recently for biological sequence analysis [127]. For CNV detection applications the HMM model has been applied in several matured and widely used CNV detection methods like PennCNV [82] and cnvHap [10]. In several analyses presented in the later chapters, I used a flexible haplotype HMM framework implemented through the cnvHap algorithm. I incorporated several modifications and made important observations in the HMM framework which I think has led to an improvement in the CNV analysis. Through the understanding of this model, I have also explored the possibility of imputing CNVs from one dataset to another and discussed the possibility of combining several CNV detection platforms together. This can also include the combination of genotyping arrays with NGS data in the future [113]. Further to these thoughts, in the next section I discuss in detail the general framework behind the HMM [128–130].

2.1.3.1 The HMM notation

One of the most important aspects of HMMs is its notation. The following is a general notation for a general HMM.

- T= length of the observation sequence
- N= number of states in the model
- M= number of observation symbols
- Q= $\left\{q_0, q_1, ..., q_{N-1}\right\}$ = distinct states of the Markov process
- V= $\left\{0, 1, ...M-1\right\}$ = set of possible observations
- A=state transition probabilities
- B= observation probability matrix

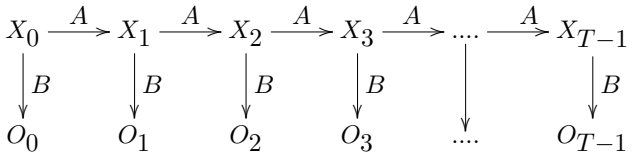
- π= initial state distribution

- O=$(O_0, O_1, ...O_{T-1})$=observation sequence

Further, here I have discussed only a first-order HMM applied to a sequence of stochastic processes with the following assumptions.

1. Random values hold a finite set
2. The state at any point is only dependent on the previous state (since we assume first order Markov process)
3. At each point only the output sequence is observable and it is dependent on the hidden state at the same point.

One of the common ways to conceptualise a HMM is through a trellis diagram which shows the relationship between the observation and the hidden variables.



Here, in the above diagram the observations are denoted by $O = (O_0, O_1, ..., O_{T-1})$ and X_i represents the hidden states corresponding to the observed states. Only the observed states are seen and they are related to the hidden states through the B matrix. Matrix A refers to the transition matrix for the hidden matrix and is usually given or assumed before solving the HMM problem.

2.1.3.2 The three problems and solutions in HMM

In general, for any given HMM problem there are three different problems which we need to solve. For all HMM applications, depending on the specific objective at hand, one must try to solve one of the three problems described below:

- Given the model $\lambda = (A, B, \pi)$ and a sequence of observations O, find $P(O \mid \lambda)$
- Given model $\lambda = (A, B, \pi)$ and an observation sequence O, find the underlying Markov process.
- Given an observation sequence O and dimensions N and M find the model $\lambda = (A, B, \pi)$ that maximizes the observation sequence O

For the application of CNV detection using HMMs, we usually try to answer the second question i.e given the observed intensity measurements we try and infer the underlying copy number classes.

2.1.3.3 Solution to the three HMM problems

Here, I have described the various solutions related to the HMM model. The various derivations described below can help in determining some of the modifications I have tried out in cnvHap later. I have also discussed how I tackled some of the problems that I encountered during my analysis. Though for the CNV analysis we were mainly interested in the solution to the second problem, the derivation of the first to the third problem follows a sequential order. Hence, I have reviewed the general derivation for problem 1 and 2 and then briefly mentioned the important steps for problem 3.

2.1.3.4 Solution to problem 1

The initial model of any HMM is denoted by $\lambda = (A, B, \pi)$ and $O = (O_0, O_1, ..., O_{T-1})$ is a series of observations then the main objective is to find $P(O \mid \lambda)$.

Let $X = (x_0, x_1, ..., x_{T-1})$ be a state sequence, then by definition of B we have

$$P(O \mid X, \lambda) = b_{x0}(O_0) b_{x1}(O_1) ... b_{xT-1}(O_{T-1}) \tag{2.1}$$

Then by the definition of initial distribution π and state transition probabilities A, it follows that

$$P(X \mid \lambda) = \pi_{x0} a_{x0,x1}, a_{x1,x2} ... a_{xT-2,XT-1} \tag{2.2}$$

Since it is known that

$$P(O, X \mid \lambda) = \frac{P(O \cap X \cap \lambda)}{P(\lambda)} \tag{2.3}$$

and

$$P(O \mid X, \lambda)P(X \mid \lambda) = \frac{P(O \cap X \cap \lambda)}{P(X \cap \lambda)} \cdot \frac{P(X \cap \lambda)}{P(\lambda)} = \frac{P(O \cap X \cap \lambda)}{P(\lambda)} \tag{2.4}$$

we have

$$P(O, X \mid \lambda) = P(O \mid X, \lambda)P(X \mid \lambda) \tag{2.5}$$

Then summing over all possible sequences we obtain

$$\begin{aligned} P(O \mid \lambda) &= \sum_X P(O, X \mid \lambda) \\ &= \sum_X P(O \mid X, \lambda)P(X \mid \lambda) \\ &= \sum_X \pi_{x0} b_{xo}(O_0) a_{xo,x1} b_{x1}(O_1) ... a_{xT-2,xT-1} b_{xT-1}(O_{T-1}) \end{aligned} \tag{2.6}$$

However, the computation of equation 2.6 is infeasible because it requires about $2TN^T$ multiplications, which usually is computationally intense. The main strength of HMMs comes from the result that the above formulae can be calculated by an efficient formula as below i.e.

To find $P(O \mid \lambda)$ we use the forward algorithm or $\alpha - pass$. Given $t = 0, 1, ...T-1$ and $i = 0, 1, ...N-1$, we define

$$\alpha_t(i) = P(O_0, O_1, ...O_t, x_t = q_i \mid \lambda) \tag{2.7}$$

Then $\alpha_t(i)$ is the probability of the partial observation sequence up to time t, where the underlying Markov process is in state q_i at time t. The crucial insight is that $\alpha_t(i)$ can be computed recursively as follows:

1. let $\alpha_0(i) = \pi_i b_i(O_0)$, for $i = 0, 1, ...N-1$

2. for $t = 1, 2, ...T-1$ and $i = 0, 1, ...N-1$ we compute

$$\alpha_t(i) = \left[\sum_{j=0}^{N-1} \alpha_{t-1}(j)a_{ji}\right] b_i(O_t) \text{ then it is clear from (2.7) that}$$
$$P(O \mid \lambda) = \sum_{i=0}^{N-1} \alpha_{T-1}(i)$$

The forward algorithm now requires about N^2T multiplications as opposed to more than $2TN^T$ using the naive approach.

2.1.3.5 Solution of problem 2

Here the goal is to find the *most likely* sequence of observations O. The alternate option of finding the most likely states can result in a different interpretation. For HMMs, we want to maximize expected number of correct states which is in contrast to DP and these two need not always concur. The difference is that DP aims to find the highest scoring overall path. First we derive the backward algorithm or $beta - pass$ which is analogous to α-pass but now it starts from the back and moves towards the beginning

For $t = 0, 1, ...T-1$ and $i = 0, 1, ...N-1$

$$\beta_t(i) = P(O_{t+1}, O_{t+2}, ...O_{T-1} \mid x_t = q_i \mid \lambda) \tag{2.8}$$

Then $\beta_t(i)$ can be computed recursively and efficiently as follows:

1. $\beta_t(i) = 1$, for $i = 0, 1, ...N - 1$

2. For t=T-2, T-3, ..., 0 and i=0,1,...N-1 compute

$$\beta_t(i) = \sum_{j=0}^{N-1} a_{ij} b_j(O_{t+1}) \beta_{t+1}(j) \tag{2.9}$$

For $t = 0, 1, ...T - 2$ and $i = 0, 1, ...N - 1$, define

$$\gamma_t(i) = P(x_t = q_i \mid O, \gamma) \tag{2.10}$$

Since $\alpha_t(i)$ measures the relevant probability up to time t and $\beta_t(i)$ measures the relevant probability after time t we get

$$\gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{P(O \mid \lambda)} \tag{2.11}$$

It was already seen that the denominator $P(O \mid \lambda)$ can be obtained by summing $\alpha_{T-1}(i)$ over i. From the definition of $\gamma_t(i)$ it follows that most likely state at time t is the state q_i, for which $\gamma_t(i)$ is maximum where the maximum is taken over the index i.

2.1.3.6 Solution to problem 3

This requires us to adjust the model parameters to best fit the observations, the sizes of matrices (N and M) are fixed but the elements of A, B and π are to be determined. Without going into much detail in this problem, the overall algorithm for this step is as follows

1. Initialize $\lambda = (A, B, \pi)$

2. compute $\alpha_t(i), \beta_t(i), \gamma_t(i, j)$ and $\gamma_t(i)$

3. re estimate the model $\gamma = (A, B, \pi)$

4. if the $P(O \mid \lambda)$ increases, go to step 2

If $P(O \mid \lambda)$ does not increase by at least some threshold or set up to maximum number of iterations then the iteration is stopped.

2.1.3.7 Dynamic programming and scaling issues

In this section, I have reviewed the relationship between DP and HMMs. In short the DP is $\alpha - pass$ where "sum" is replaced by "max". For the given π, A and B, the DP can be estimated as follows.

1. let $\delta_0(i) = \pi_i b_i(O_0)$, for $i = 0, 1, ...N-1$
2. For $t = 1, 2, ...T-1$ and $i = 0, 1....N-1$ we need to compute

$$\delta_t(i) = max_{j\epsilon 0,1,..N-1} \left[\delta_{t-1}(j) a_{ij} b_j(O_t)\right]$$

At each successive t, the DP determines the probability of the best path ending at each of the states at i =0,1,...,N-1. Consequently the best overall path is given by

$$max_{j\epsilon 0,1,...N-1} \left[\delta_{t-1}(j)\right] \tag{2.12}$$

It is to be noted that equation 2.10 only gives the optimal probability and not the state path itself. The DP procedure can be augmented to recover the optimal path by tracking back from the highest scoring final state.

Scaling The three problems in HMMs require computations of products of probabilities and some of these often tend to zero as the probabilities increase exponentially as T increases [129]. In the field of statistical computing this issue is often referred to as the problem of *underflow*. To solve this issue we often scale the numbers and simultaneously take care that the results and formulae remain valid. One of the ways to prevent the *underflow* problem is to compute the following formula.

$$log \left[P(O \mid \lambda) = -\sum j = 0^{T-1} log c_j\right] \tag{2.13}$$

where c_j is the scaling factor which is used for $\beta_t(i)$ as well as $\alpha_t(i)$ and initial value is defined as $c_0 = 1/\sum j = 0^{N-1} \alpha_0(j)$ Using c_t at time t such that $\hat{\beta}_t(i) = c_t \beta_t(i)$ and

$\gamma_t(i,j)$ and $\gamma_t(i)$ are computed using $\hat{\beta}_t(i)$ in place of $\beta_t(i)$. The resulting gammas and di-gammas are then used to re-estimate π, A and B. From this it can be easily shown that the re estimated values are the same as the original estimates.

2.1.3.8 The Viterbi algorithm for state path determination

The main motivation behind the Viterbi algorithm is that there are several paths through the hidden states for a given observation sequence. Howeve, all these paths do not have the same probability values. Hence, the Viterbi algorithm is used to compute the most probable path. The principle used for this algorithm is similar to DP for calculating the probability [131]. It requires the knowledge of parameters and in particular, the output sequences. It then finds the state sequence that is most likely to have generated the output sequence. It can be summarized as

$$p_l(i,x) = e_l(i) max_k(p_k(j,x-1).p_{kl}) \tag{2.14}$$

where $e_l(i)$ = probability to observe element i in state l and $p_k(j,x-1)$ the probability of the most probable path ending at position $x-1$ in state k with element j. p_{kl} = refers to the probability of the transition from state l to state k.

2.1.3.9 The EM algorithm

The idea behind the application of the EM algorithm is that, while finding the maximum-likelihood is straightforward in the presence of latent variables or missing variables, the sum for the likelihood becomes intractable [132]. The general likelihood function is denoted by

$$L(\theta,x) = log p_\theta[x] = log \sum_z p_\theta(x,z) = log \sum_z p_\theta[z \mid x] p_\theta[x] \tag{2.15}$$

The EM approach for this formula can be interpreted as, first use the current parameters to re-estimate the latent variables and then use those re-estimated parameter values for the calculating parameters for the entire model. The formulae for these steps are as follows.

First, the expectation step for the EM theorem for any given function x and parameter values θ θ' is given by

$$L(\theta', x) - L(\theta, x) \geq \sum_z p_\theta[Z \mid x] log p_{\theta'}[x, z] - \sum_z p_{\theta'}[Z \mid x] log p_\theta[x, z] \tag{2.16}$$

Next, we can use this to maximize expectation

$$\sum_z p_\theta[z \mid x] log p_{\theta'}[x, z] = E_{z\ p_\theta[\cdot|x]} \left[log p'_\theta[x, z] \right] \tag{2.17}$$

The main iterations steps are as follows.

- compute $p_\theta[z \mid x]$ for current value of θ
- compute expectation and find maximizing θ'

The meta-algorithm which can also summarize the above steps is:

1. E step : $q^{t+1} \leftarrow p^t_\theta[z \mid x]$

   Interpretation: posterior probability of latent variables given observed variables and current parameters

2. M step : $\theta^{t+1} \leftarrow\mid argmax_\theta \sum_z q^{t+1}(z \mid x) log p_\theta[x, z]$

   Interpretation: maximum likelihood parameter estimation given all data

The EM algorithm applied to HMMs is often known as Baum-Welch algorithm. Several variants and extensions for this method exist. One important point to be noted for this algorithm is that it is sensitive to initial values and hence their choice plays an important role in the final results.

2.1.4 Generalised genotyping methods

The segmentation methods discussed above do not take the allele frequency information i.e. BAF, into consideration while calling CNVs. BAF undergoes changes in the presence

of CNV and can be incorporated to get extended CNV-genotype information. In these approaches, CNV genotypes are clustered in the two dimensional feature space of total intensity LRR and BAF in order to derive the allele-specific CNV-genotypes [133]. In this feature-space additional genotype-classes are possible for different copy number classes. In normal genomic regions with no copy number variation the total intensity LRR is normalised to have zero mean and the distribution of the BAF is trimodal where each cluster corresponds to AA, BB or AB. In a deletion region the LRR decreases as it directly corresponds to the DNA copy number and the BAF is bimodal with only two genotypes allowed (A, B). In regions of copy number gains there is increased value of total intensity and the BAF has more than two modes (AAA, AAB, ABB, BBB etc.). I describe in details some of the implementations of such generalized model through software like cnvHap.

2.1.5 HMM based CNV detection methods

In the following section, I have described some of the applications of the HMM model implemented in the commonly used software for detecting CNVs. I have discussed the general principle behind these methods, discussed some of their merits and demerits and also briefly touched on the recommended parameters for running them.

CNV detection methods from high density genotyping platforms are currently classified into two categories: *wide* or *deep*. The *deep* category of algorithm considers all samples simultaneously for modelling while keeping the number of probes restricted while the *wide* category of algorithm processes each sample independently within fixed intensity clusters. The commonly used *wide* algorithm usually either uses a HMM or a segmentation algorithm.

PennCNV and QuantiSNP Some of the widely used HMM-based CNV detection methods are PennCNV [82] and QuantiSNP [133]. Both these methods are based on first-order Markov models. This means that the copy number call at a particular probe is dependent on the copy number at the preceding probe. The total number of copy number calls possible for both these algorithms is restricted to six and have similar approaches for detection. One of the main differences between the two methods is based on the probability transition matrix for the HMM. In QuantiSNP the state transition

matrix is the same for all copy number classes while in PennCNV the transition rate varies for different classes. For example, in PennCNV, the transition rate from a normal copy to a single copy deletion is more likely than transition from a single copy to copy number three.

Some of the differences between these two are as follows. The emission probabilities for PennCNV are related to uniform and Gaussian distributions and is considered to be known. In QuantiSNP, a Bayesian framework where the unknown hyper-parameters are estimated through a maximum likelihood method is used. The means for the Gaussian distributions for PennCNV is set to $-3.53, -0.66, 0, 0.40, 0.68$ (polymorphic probes) and for QuantiSNP it is set to $-6, -0.55, 0, 0.36, 0.63$. PennCNV further has a separate *deep* module where the CNV breakpoints are specified by the useR with prior probabilities for deletion and duplication. Next, for each copy number class a likelihood is calculated based on the posterior probabilities. This module is more suited for detecting common CNVs.

cnvHap The cnvHap algorithm is an advanced haplotype-based CNV detection and genotyping algorithm [10]. The cnvHap algorithm unlike the others uses a haploid HMM and has integrated *wide* and *deep* modules. The emission distributions are Gaussian and depends on the hidden state g. This is used except for copy number class of zero, when a uniform distribution is assumed.

$$LRR \sim N(\mu_{r,g}, \sigma^2_{r,g})$$

$$\begin{cases} I_{0 \leq b \leq 1} \;\; N(\mu_{b,g}, \sigma^2_{b,g}) & if \;\; CN(g) > 0 \\ U(0,1) & if \;\; CN(g) = 0 \end{cases}$$

One of the recommended initial cluster mean positions $\mu_{r,g}$ is $-4.5, -0.55, 0, .31, 0.53$. These values can be adjusted based on the given platform. Some of the other cnvHap features include a Baum-Welch algorithm for determining the haplotype HMM parameters and the Viterbi algorithm for finding the most likely sequence of allele-specific copy number states.

cnvPartition The cnvPartition algorithm was developed by Illumina for its Beadstudio software [134]. It is a *wide* based CNV detection method and models LRR and BAF

as a Gaussian distribution model. It models up to four copies for each allele and in total considers fifteen different allele-specific copy number states. An estimate for each probe to belong to each of the copy number class is determined using a maximum likelihood based method. The breakpoints are segmented based on the change in copy number class along the probes.

2.1.6 Methods for detecting de novo CNVs

De novo CNVs have been found to be quite important from disease and non-disease point of view. For example, rare de novo CNVs have shown to play an important role in Autism [135], while identical twins were found to have different genomes based on differences in the presence of de novo CNVs [136].

Given the importance of de novo CNVs, there has been development of several software and methods for detecting and characterizing de novo CNVs. Some of the array based methods for detecting CNVs include DNA-Chip (dchip) [124], Partek (http://www.partek.com), Quanti-SNP [133] and PennCNV [82]. Some of the earlier methods for CNV segmentation such as the circular binary segmentation algorithm has been adapted to detect de novo CNVs [137]. Recently there has been development of sequencing based methods for detecting rare CNVs such as the commonly used plinkseq (https://atgu.mgh.harvard.edu/plinkseq/). Some advanced stochastic based inference of de novo CNVs have also been developed and recently published [138].

2.1.7 Other model based approaches (GADA)

Some other parametric models which are not based on HMMs have also been proposed. One such example is genome alteration detection analysis (GADA) [139]. GADA uses a linear combination of step functions for representing the copy number classes and then uses a sparse Bayesian learning to infer the basis function with presumed prior distribution. CNV breakpoints are then determined using a backward elimination process. Some other methods similar to GADA are ITALICS [140] and GLAD [141].

The various commonly used statistical frameworks and methods used to discover and associate CNVs with complex phenotypes is reviewed in the following sections.

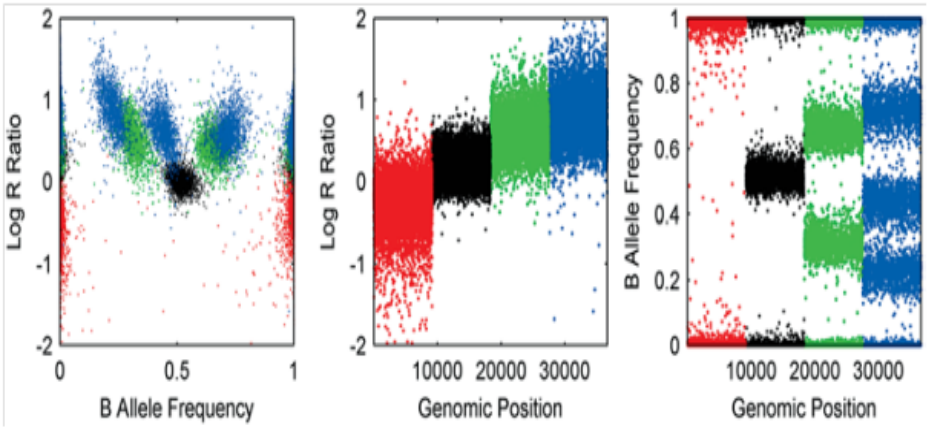


FIGURE 2.1: Effect of CNV on SNP genotyping data

In presence of CNV there is change in the LRR values and also the allele frequencies in the intensity data [103]. Above the colour labels are Red: copy number 1, black: copy number 2, green: copy number 3, blue: copy number 4

2.2 Association approaches in genetic epidemiology

Association models in genetic epidemiology aims at finding genetic polymorphisms, which are systematically different between cases and controls. This can lead to the determination of risk enhancing as well as protective alleles. Some of the main statistical challenges in such association approaches are to correct for the suggestive causal variants, which can arise randomly through chance. This would mainly involve separating spurious signals from the true causal variants [142].

Association approaches in GWAS mainly follow three approaches including case-control, quantitative (continuous) and categorical disease outcomes. In case control approach, there are three genotypes for the two groups of cases and controls. The commonly used tests are pearson test (2 d.o.f.) and Fishers exact test. In most cases latter is preferred. In additive models, heterozygous is considered intermediate of the two homozygous states and some of the methods used are Cochran-armitage test. The Cochran armitage test does not assume HWE and tests the hypothesis of zero slope for the line that fits the three genotype risk estimates. Compared to these, the Bayesian approach assumes a priori on the nature of disease risk not used much because of intensive computations resources required. For continuous outcomes, ANOVA based methods are used. For regression analysis, normality of outcome is assumed for the trait and a transformation is used to achieve this if it is not. Due to this for case control analysis a logistic regression is used

since nothing lies outside the range of 0-1. Logistic regression is also more flexible to help in incorporating multiple testing and also including multiple SNPs. It can also include complex epistasis and environmental interactions or also incorporating covariates like age of onset and sex and is a general procedure for generating tests that are asymptotically equivalent to likelihood based tests and it is based on the derivative of likelihood with respect to parameter of interest where the unknown parameter is set to zero. This test also has the possibility of adding covariates into the model. Ordered and categorical outcomes includes several categories of outcomes analysed using multinomial regression. This can be in two forms ordered and unordered [142].

2.2.1 Tests for multiple SNPs

SNPs perhaps lying candidate genes have little recombination. Most association based on single SNPs avoid the nature of correction existing in the SNPs due to LD and avoid leveraging the joint nature of their distribution. One method to correct the correlation nature of SNPs is use step wise selection or Bayesian shrinkage methods. Such multi-SNP analysis with highly correlated SNPs can also use haplotype based methods. For this computation methods exit to infer haplotype structure but there is some uncertainty associated with it. But this is simpler in regions of high LD which means less complexity. Some other haplotype based methods try and ascertain the proportion of haplotype amongst cases and controls with assumptions of HWE and near additive measures of disease risk. In some cases there are excess of haplotypes in cases compared to controls.

Some of the other fundamental issues involved in genetic association analysis are

1. Setting tight regulation of statistical significance.
2. Finding patterns of polymorphisms which can only generated by causal genetic mechanisms, like from evolutionary and recombination processes.
3. Dealing with systematic errors dealing with missing values.

In addition to this one also needs to correct for hardy Weinberg equilibrium (HWE). HWE can arise through several factors like inbreeding, population stratification, selection and through disease association. All these factors need to be accounted for in association analysis. Deviation from HWE can also occurs in the presence of deletion

or tandem duplication polymorphism. The usual practice of setting HWE in GWAS is $\alpha = 10^{-3}/10^{-4}$. Some other aspects of GWAS include correcting for study center related batch effects or unusual patterns of missing data, deciding how many individuals and genetic markers to include in the study. Patterns in missing data can be estimated through several methods. One approach is the maximum likelihood (ML) approach, which can find the best predictor of missing genotype, given the data. Another approach is to randomly select from a probability distribution through multiple imputation. Advisable to remove those SNPs which have more missing data. LD is crucial for designing association studies. There are about 11 million single nucleotide variants, but only a handful is actually associated with the phenotype of interest. LD is usually calculated through several measure and some of the common ones are D and r^2. Both these methods have their own merits and demerits. Fine scaling estimates can help in summarizing LD in a region, as recombination is the most important phenomenon behind LD. The human genome is block like with little or no recombination within blocks but several block boundaries having hotspots of recombination [143]. This also leads to the important concept of SNP tagging which means to retain as much genetic variation as possible from all the SNPs. The SNPs selected from the Hapmap study will help in increasing selecting population specific SNPs and reducing SNPs can help in increasing statistical power.

2.3 Models for association

2.3.1 General linear regression

The general linear regression model is used to characterize the relationship between two variables of interest. Intuitively, the main purpose for this model can be compared to correlation, but the regression model has several advantages [144]. The regression model, in addition to the purpose of determining the strength of association, also allows finding the direction of effect. This information can be useful for performing meta-analysis and for comparing results between different genetic studies. In the case of quantitative outcomes the weights calculated through the regression model can also be used to predict unobserved values. This can be a useful feature in predicting missing phenotypes in case there is a large number of correlated phenotypes available.

The general representation of the regression model seeks to capture the relationship between a dependent variable Y_i and one or more explanatory variables denoted by $X_{i_1,...i_n}$. The model also accounts for an error term e_i which is unobserved noise. The general form of the model is as follows:

$$Y_i = \beta_1 x_{i1} + ... + \beta_p x_{ip} + \varepsilon_i = X_i^T \beta + \varepsilon_i, \quad i = 1, ...n$$

or it is also sometimes represented as

$$Y = X\beta + \varepsilon$$

There are several assumptions related to the predictor and response variables in the regression model. Some of these assumptions are enumerated below.

- linearity of the relationship between dependent and independent variables
- errors are independently distributed
- errors have constant variance
- errors have normal distribution

Though there are several modified versions of the regression model where these assumptions have been relaxed or modified, but in general violations of these assumptions lead to biased estimates of parameters. Once the weights of the regression model have been estimated then a hypothesis test is carried out to check the significance of association. This usually relates to how a unit change in predictor variable leads to the corresponding change in the outcome variable.

Current regression models for genomic studies often include two kinds of multivariate models, one where all the columns of response variable (Y_i) are tested together, and one other where in which the columns of Y_i are tested independently. Further, there are two other generalizations of the general regression model which is commonly used i.e. multiple regression model and the generalized regression model (GLM). The multiple

regression model restricts the number of dependent variable to one and is denoted by

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + ... + \beta_p X_{ip} + \varepsilon_i$$

In GLMs, the response variable can have arbitrary distributions rather than normal distribution [145]. In addition, an arbitrary function (link function) makes the response variable linearly vary with predicted values. It also generalizes various statistical models like Poisson regression, logistic regression and the linear model. Several methods have been developed for parameter estimation for GLM, including iterative re-weighted least squares, maximum likelihood, Bayesian approaches and least squares estimates for variance stabilized responses.

2.3.1.1 Logistic regression

The logistic regression model is used in statistical modelling when the outcome variable is binary or categorical. This model is useful in case-control genetic studies and has been used widely in GWAS analysis [146]. It can also be considered as a special case of GLM and the link function for this case is a logit function. The model is represented as below

$$logit(p) = ln(\frac{p}{1-p}) = \beta_0 + \beta_1 X + e_i)$$

This allows the input values to vary from $-\infty$ to $+\infty$ but the logit function always returns values between 0 and 1. Ordinary regression parameters are estimated by maximum likelihood whereas for logistic regression, closed form expression of these parameters does not exist and instead an iterative process like the Newton's method is used [147]. In iterative process, the solution begins with a value and is revised until the improvement in the numerical solution is minimal. At this point the process stops and is said to have converged.

2.3.1.2 Proportional odds regression

The proportional-odds model, also known as the ordered logit model, is used for ordinal dependent variables. Compared to logistic regression where the response variable is dichotomous, in ordered logit model the dependent variable allows more than two ordered response categories [148]. The model cannot be estimated through ordinary least squares (OLS) and is estimated through maximum likelihood.

2.3.2 Odds ratio

The odds ratio (OR) refers to the factor by which risk of a disease is increased in exposed individuals compared to unexposed individuals. For example for the table below,

TABLE 2.1: A contingency table for calculating OR

	Exposed	Non-exposed
Cases	A	B
Controls	C	D

OR is defined as the following.

$$OR = \frac{AD}{BC}$$

OR is non-Gaussian but log(OR) is approximately normally distributed. Variance for the $log(OR)$ is defined as

$$var(log(OR)) = \frac{1}{A} + \frac{1}{B} + \frac{1}{C} + \frac{1}{D}$$

The above formula can then be used to calculate the confidence intervals of OR.

$$CI(log(OR))_{1-\alpha} = log(OR) + Z_{\alpha/2}\sqrt{var(log(OR))}$$

where the $Z_{\alpha/2}$ is Gaussian distribution quantile of $1-\alpha/2$ order. In genetic epidemiology a particular exposed group is defined by the genetic marker with risk related to categories of the genotypes AA, BB and AB.

2.3.2.1 The Wald statistics

The Wald statistics considers the difference between estimated $\hat{\beta}$ (the regression coefficient) and zero and it follows a chi-square distribution with one degree of freedom.

$$W = \frac{\hat{\beta^2}}{\hat{var}(\hat{\beta})} \sim \chi_1^2$$

2.3.2.2 Likelihood ratio test

The likelihood ratio test ascertains the likelihood of two models (L_1 and L_2) which differ by at least one variable. The second model should contain all the variables in the first model and the difference in the number of variables between the two defines the degrees of freedom of the chi-square test. For example, suppose we are interested in a variable X_{k+1}, model 1 does not contain X_{k+1} and model 2 does in a logistic regression model.

$$Model\ 1\text{: } logit(P) = \alpha + \sum_{i=1}^{k} \beta_i X_i$$
$$Model\ 2\text{: } logit(P) = \alpha + \sum_{i=1}^{k+1} \beta_i X_i$$

In the above case since there is a difference of one variable between the two models, the likelihood ration test follows a chi-square distribution with 1 d.o.f shown below

$$LRT = 2ln\frac{L_2}{L_1} \sim \chi_1^2$$

2.3.2.3 Score test

This test is based on the tangent of the likelihood function at zero. If $L(\beta)$ is the likelihood function of β (regression coefficient), G is the partial derivative of $log(L(\beta))$ and H is the second partial derivative of the function $log(L(\beta))$ then

$$G(\beta) = \frac{\partial logL(\beta)}{\partial \beta}$$
$$H(\beta) = \frac{\partial^2 logL(\beta)}{\partial^2 \beta}$$

The statistics then follows a chi-square distribution with 1 d.o.f

$$S(\beta = 0) = \frac{G(\beta = 0)^2}{H(\beta = 0)} \sim \chi_1^2$$

2.4 Meta analysis in GWAS

Although GWAS has identified thousands of variants associated with several complex diseases, these variants explain little of the risk variability. The effect size of common genetic alleles are often small and the detection signals require large sample size to reach the appropriate significance level [149]. In this scenario a single GWAS are often underpowered and necessities statistical approaches to combine several studies in order to increase power and reduce false positives. One such popular approach is meta analysis in GWAS, where summary results from different association analysis studies are combined together. Over the last few years through meta analysis several hundred results have been published and has led to the identification several new validated new genetic risk loci [149]. In the following sections I have discussed some of the widely used meta-analysis approaches including some recent and alternative methods that have been proposed.

P-value and Z-scores

Meta-analysis using P-values have been used in the past but then later discontinued. This approach has several limitations including dealing with heterogeneity. Some popular examples for this approach include Fishers approach for combining pvalues and Z scores. Results from both these approach are highly correlated but the main difference is that Fisherss method used the average of log(Pi) (where Pi is the P value from the ith study) across all studies where as the Z score in based on the average of Z values [150]. The advantage of using Z scores is that it can take direction of effect into account and has the possibility of introducing weights. Both these methods are well suited for

low frequency and rare variants, where the association tests used provides the P values and Z scores rather than the effect sizes [151].

Fixed effects model

Fixed effect based meta-analysis has been the most popular approach for GWAS data so far for discovering SNP-phenotype associations. Fixed effects meta-analysis is based on the assumption that the true effect of each risk allele is same in every data set analysed. Though this assumption is not valid in every case, this approach has the maximum power to discover novel risk loci compared to other method like the random-effects model. There are different approaches in the fixed-effect meta-analysis. One the most commonly used method is where each study is weighted according to the inverse of its squared error. Most of the bioinformatics software carry out inverse variance meta-analysis. An alternate fixed-effect model giving similar results to the fixed-effect model is Cochran-Mantel-Haenszel approach, but is used less often [150].

Random effects model

In random effects model, the effect size is allowed to vary from one study to another. The differences in the effect sizes can vary based on different factors inherent in the study like age and environmental factors. If somehow it was possible to conduct infinite number of studies then the true effect sizes for these studies will be distributed about some mean. The effect size in random effects model is thus assumed to represent a random sample from a particular distribution and thus referred to as random effects model. Though random effects are more appropriate than fixed effect model it is usually not used in discovery projects due to its limited power compared to fixed effect model. However random effects model becomes more appropriate when there is a need to consider the generalizability of the observed association and its uncertainty across different populations [152]. Some recent random effects model include improvement of discovery power in the presence of between-study heterogeneity in effect sizes. In contrast to traditional random effects model Han and Eskin assume that there is no heterogeneity in the effect sizes [153]. Their approach can essentially divided into two parts, one that is similar to fixed effect statistic and the other is a test statistic for heterogeneity.

2.5 Population stratification and cryptic relatedness in GWAS

The idea behind genetic association studies is identify allelic state which is correlated with disease phenotypes of interest, but this is often accompanied with lot of false positives. This was mainly because of insufficient recombination between these studies to refine causal locations. The other approach is based on linkage studies. Find loci where there is correlation between phenotype of interest and the pattern of transmission of DNA sequence over generations in known pedigree. While relatedness is crucial for linkage studies it is usually a problem for association studies. Since previous linkage studies were affected by insufficient power, GWAS proved to be a powerful alternate method for finding genomic loci associated with complex diseases [154]. GWAS aims to find loci with weaker effect size where as family based studies are better for detecting rare variants with large effect size. Also, association studies tries to find genotype distribution, which are systematically different between two study populations, pedigree can also generate differences between cases and controls at all loci where there is no strong selection [155].

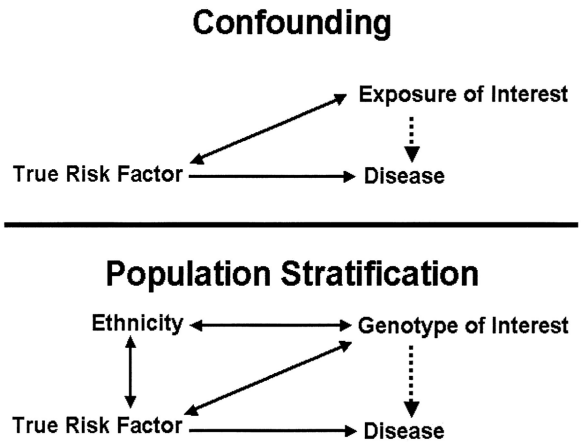


FIGURE 2.2: Population stratification in GWAS

Diagram showing confounding and stratification in GWAS setting. Solid arrows show direction of causal relationship, broken lines with bidirectional arrows shows correlation which is not causal. Dotted unidirectional arrow shows confounded association. In a classic situation of confounding, the exposure phenotype correlates with the true risk factor but is not true causal for the disease, leading to the problem of stratification [156].

There are three possible reasons why in an unmatched study phenotypes of study subjects might vary systematically and lead to the problem of population stratification.

1. Disease prevalence varies across sub populations in accordance with frequencies of causal alleles.
2. sub-population prevalence may vary because of different environmental risks.
3. ascertainment bias- differences in sampling strategies between cases and controls correlated with ancestry.

Stratification occurs when there are large-scale systematic differences in ancestry, which can be caused by immigration or within individuals with more recent ancestors than one would expect from random mating. One related and popular way of understanding population structure is based on the island model comprising of sub populations. In the island model mating between individuals occurs preferentially between individuals of the same island, but this concept again is not completely accurate to the true kinship patterns in a given population. Shared ancestry refers to the patterns of kinship amongst groups of individuals [157]. This can broadly related to shared ancestry based on geographic location, but this is not always accurate because there can be underlying patterns of kinship based on social and religious groups [158].

The main cause of population structure is LD between unlinked loci in the genome. This can generate spurious marker-phenotype associations. This happens when proportion of cases varies across sub populations and consequently alleles which also happen to vary across these sub populations will show spurious associations with the phenotype. The simplest model of population structure assumes an island model comprising of subpopulations. Mating occurs preferentially between pairs of individuals of the same island but the island model is not always accurate.

One of causes of population structure can be LD between unlinked markers/loci which can consequently generate spurious marker-phenotype associations. Therefore, if the proportion of cases vary across subpopulations. Alleles which vary in frequency between the sub populations will show associations with the phenotype. Standard GWAS approaches do not distinguish such alleles which vary across subpopulations due to genetic drift or natural selection with the true causal alleles involved in determining the phenotype [158].

One common sense method to deal with this problem would be to chose controls for the association analysis from the sub population itself, however this captures certain fractions of population structure and may miss the smaller effects due to other complex underlying structure. Such small effects can lead to substantial increase of false positives when the sample size is quite large, which is often the case for current GWAS studies. A simulation study by Marchini et al. (2004a) [159] concluded that even in a population of modest level of structure the most significantly associated SNP can have their pvalue reduced by a factor of 3, thus overestimating the significance of association.

2.5.1 Cryptic relatedness

Another closely related confounding factor to population structure is cryptic relatedness. Cryptic relatedness happens when there are unknown relatives even in a small group of unrelated individuals. While population structure refers to remote common ancestry among large groups of people, cryptic relatedness refers to common ancestry amongst smaller groups of people, sometimes even just a pair. Some studies have shown that cryptic relatedness amongst outbred populations can be almost minimal but for isolated populations it can be significant [160]. Using pedigree and empirical genotype data it has been shown that cryptic relatedness can reduce association p value by a factor of 4.

2.5.2 Correcting for stratification

Various methods for correction association analysis for confounding factors

1. Family Based Tests of Linkage and Association (FBTLA): FBTLA tests for differences in genotypes of affected children and those expected under Mendelian randomization of alleles of the unaffected parent. If an allele is risk-enhancing or in LD with the true causal variant it will be over transmitted to the children. This approach is thus a test for both linkage and also association and is robust to the confounder of population stratification. However, the main drawback of family based studies is to obtain enough sample size for a well powered association analysis. In addition in family based analysis, for every case-control pair, three individuals must be genotyped, leading to higher cost. Using family based analysis thus becomes a more difficult choice for association analysis, given that methods which can correct for stratification in GWAS. However, family based analysis can

help in finding parent of origin effect which is currently not possible in GWAS settings.

2. Genomic control: Genomic control is an easy and computationally efficient way for correcting for population structure. This correction was developed for the Armitage test statistic which is asymptotically equivalent to score test in logistic regression [161]. In the absence of any population structure it has an asymptotic chi square null distribution. In GWAS analysis confounding due to population structure or ascertainment bias is reflected as deviation from the x=y line a QQ plot, often referred to as inflation factor lambda. lambda can also be understood as the correction factor for type 1 error for the Armitage test, which corrects for the error but leaves the ranking of genomic markers unchanged. If the effect of stratification can be assumed to be similar for different loci, lambda can be calculated using the formula $lambda = median(Y_1, Y_2, ..Y_L)$ where L is number of unlinked markers and the denominator is derived from a gamma distribution as a robust estimator of lambda. Some other estimators of lambda have also been suggested [162].

3. Structured association methods island model:

   Structured association is based on the island model of population stratification. It assumes that the ancestor of each individual is drawn from one or more of such islands. These methods use a Markov process to model variation along a chromosome and then carry out a stratified test for association. Some example of the association test include Mantel-Haenzel or logistic regression with admixture proportion of each subpopulation included in the model as covariates. Some of the software packages implementing this method include ADMIXMAP [163] and STRUCTURE [164].

4. Regression control: Some studies have shown that it is possible to correct for population structure in the case-control logistic regression framework used in GWAS. The idea is to include genotype information informative of the ancestry as a covariate in the association model. Balding et al have proposed using a set of 10^2 widely spaced SNPs which are non causal or random set of SNPs [165]. These null SNPs are informative of the underlying pedigree, which is the confounding factor leading to inflation in association analysis. Several studies have shown that these null SNPs can remove any effect of population structure and further suggested

methods like backward elimination and shrinkage to avoid any over fitting in the model.

5. PC adjustment: Some related methods to regression control have been proposed which can control for population structure in quantitative trait association by using principal components of genome-wide SNPs as covariates in the association model [165]. Instead of using a large number of SNPs and risking over fitting the PC regression based approach uses linear combination of SNPs obtained through a PC based approach. However, may more SNPs (10^4) are needed to obtain these PCs. Price et al 2006 have presented a similar method focusing on its application to case-control GWAS [166].

6. Linear mixed models: In addition to the above methods several model based on linear mixed models have also been proposed. Yu et al have extended linear mixed model used to map quantitative trait loci in outbred population using kinship matrix to also correct for population structure. Several software are available which can implement this approach such as EMMA [167] and GRAMMAR [168].

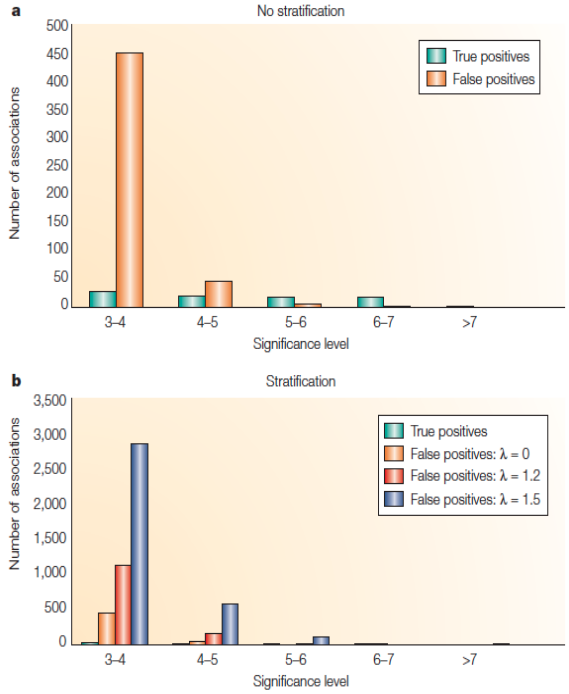


FIGURE 2.3: Population stratification in GWAS

Figure showing the effect of population stratification in whole-genome association studies. a) shows the expected number of false-positives in GWAS determined through simulations. Simulations involved 1,000 cases and 1,000 controls, for which 500,000 independent tests were carried out. Out of this only, 100 represent true associations, ranging from 15% and odds ratio ranging from 1.2-1.5. In b) the same scenario as a) is simulated with $\lambda = 1.2$ and $\lambda = 1.5$. The effect of stratification is modelled on the false-positives by modelling the effect in the association testing. [73]

2.6 Approaches for CNV genotype phenotype association

Association statistics for CNVs remain difficult at present, with often high inflation factors reported for many of the currently used methods. Some approaches usually test for the presence of a CNV in a particular region, and then association is performed with a simple chi-square statistic or Fishers test between cases and controls. However, these methods fail to accommodate other covariates like sex, batch effects and population structure. Also what is missing at present is association of quantitative measurement of CNV (like expected genotypes) and phenotype with allele frequency information further incorporated as a covariate in the model. A penalized regression method for clustering has been shown to remove batch-effects in the paper by [10]. Such direct association methods can prove useful in removing batch effects and other con founders further development of methods in this direction needs further research.

In the landmark WTCCC study of CNVs in eight common diseases [76] several different methods were explored for case-control association statistics. Some of these included location-shift based methods for intensity using Mann-Whitney test, Gaussian mixture models (where individuals were assigned to different CNV classes on the basis of posterior probability) and likelihood ratio testing based approaches. However, for most of these association statistics a very high inflation was observed, and it was further concluded that there is a further need to create a better statistical model for CNV-association at the population level [27].

To address the above problem, this thesis explored a novel method for associating CNV with binary outcome in a case-control setting. I have used a logistic regression based framework for regressing expected CNV genotype (which is a probabilistic measure of integer copy number) as the predictor variable and the case-control phenotype as the outcome.

2.7 Variable selection approaches

Variable selection, which is also known as *feature selection* in machine learning, is the process of selecting a subset of predictors which best describes the data. In many cases the two central objectives consist of predicting and ranking variables. In such cases, the idea of *minimal optimal* and *all relevant* are important aspects to consider. Highly correlated phenotypes often occur in micro-array and metabonomic data [169]. In such cases the meaning of *minimal optimal* and *all relevant* will mean different things and will include different variables. Hence, care should be taken in deciding which method to use for variable selection for a particular application.

2.8 Methods for clustering

Clustering refers to the method of classifying data into groups based on similarity. It includes several steps such as data-transformation, pre-processing and application of distance based methods [170]. One of the commonly used distance-based approach is Euclidean distance, but one drawback for this method is that it is not scale-invariant. Some other distance-based methods are max-Manhattan, Canberra and correlation based distance (Pearson correlation based distance). However, the disadvantage with Pearson correlation is that it is sensitive to outliers.

Clustering algorithms are usually classified based on applied algorithm and there is no mathematically correct way of clustering. Also, there are drawbacks for each of the clustering methods and one model cannot fit all. Some of the major classification of clustering methods are connectivity-based clustering (Hierarchical clustering), centroid-based clustering (k-means), recent models based on distribution (Gaussian mixture models and EM-based clustering), and density based clustering. One common problem in the field of clustering is the *curse of dimensionality* [171] which basically refers to the fact that as the data becomes more and more high-dimensional the sparsity among the data points increases, leading to the problem that more and more evidence is needed to make statistically sound inferences from such models.

2.9 Multiple testing

In the frequentist model based hypothesis testing, the decision to reject the null hypothesis is taken with a significance level of α, which is also known as the type 1 error. If the observed statistics falls into the rejection region corresponding to α, the hypothesis is rejected. However, if the same experiment is repeated m number of times, then the overall significance level for all the experiments will exceed the α level since no control is applied over the global error-rate of the experiment. One approach to correct for this is to apply a Family Wise Error Rate (FWER) control. The Bonferroni-Correction is the most basic and commonly used method for applying FWER to calculate a proper nominal α'

$$\alpha' = \frac{\alpha}{m}$$

In GWAS, another commonly used correction is the Sidak method:

$$\alpha' = 1 - (1 - \alpha)^{\frac{1}{m}}$$

Here m is the number of tests performed. However, there are several experiments like gene expression studies with numerous probes not reaching the FWER level. In this case the false discovery rate (FDR) is applied and is more appropriate [172]. FDR is essentially the average proportion of falsely rejected tests. This method allows errors but also controls for them. One commonly used method for FDR is Benjamini-Hochberg sequential method for ranking corrected p-values [172]. In this method for higher p-value will mean it has lower rank and the correction will be small. The p-values in this method are ordered from lower to greater such that

$$p(k) \leq K\frac{\alpha}{m}$$

In the perspective of GWAS, FWER is still preferred to FDR approaches to avoid any false positives. Based on earlier studies by [173] the GWAS significance was estimated to be $5E-8$ [174], which is based on the average number of SNPs tested. Some of the other issues which arise when a region is known to be associated with the outcome. In this case the question is whether they should be included in the association model for finding

new loci associated with the disease. An alternate approach is to consider only the top N SNPs which can be $10, 20$ or 100. Those top SNPs will be tested for replication in second or more cohorts, but such approaches usually require a large consortium [175]. One problem with such studies is that many of the populations are genotyped on different platforms which can then potentially lead to a loss of power.

2.10 Prediction and classification

Statistical prediction is closely related to the concept of variable selection which looks into *all relevant* or *minimal optimal* [176] variables for selecting the best possible subset of predictors. Once we are satisfied with a model either through a variable selection or through prior choice, the next issue is of determining the prediction rule. Prediction rule is derived from the training dataset of known relationship between X and Y. Prediction rule in a simple regression model is the linear combination of β and the explanatory variable X_i.
Let

$$f(x) = \hat{\beta}(x)$$

Then the predicted function is defined as

$$\hat{y_l} = \hat{f}(x_l)$$

Then we need to define a loss function which quantifies the divergence between true value of Y and $\hat{Y} = \hat{f}(x)$. The squared error loss function is defined as

$$L(Y, \hat{f}(x)) = (Y - \hat{f}(x))^2$$

It is important that the prediction rule does not fit the training data too closely since it will then affect its generalizatoin to other datasets and may not perform well in real world applications. The power to generalize is an important characteristic for a good classifier. The general steps in prediction are as follows:

- Train data to fit model

- Validate data to estimate extra parameters of prediction
- Test data to generalize properties

The training error is the average loss over specific training sample

$$err = \frac{1}{n} \sum_{lxtr} \Big[L(y_l, \hat{f}(x_l)) \Big]$$

However, training error is not always adequate for assessing the performance of prediction function since it is overly optimistic about how it can generalize to new data. Prediction rules based on low training error on the other hand tends to over fit. Another problem with prediction error is the problem of *bias variance trade-off*. Some of the successful applications of prediction in genomic studies have been in detecting various cancer subtypes and now this approach is being increasingly used for various other purposes [177].

2.11 Principal component analysis (PCA)

PCA is a dimension reduction technique which enables high-dimensional data to be projected to lower-dimensional space in order to detect patterns easily, predict an outcome to classify samples points or to visualize the data more efficiently. The projection into a p-dimensional space produces p principal-components (PCs) but the main idea is to find the optimal m PCs which can explain maximum variation in the data $m \leq\leq p$.

The main steps for this method are to calculate *eigen* values and their corresponding *eigenvectors* of the variance-covariance matrix of the original data matrix. Once the *eigenvectors* are calculated, it is used to transform the original data matrix to obtain the PCs for each observation. An alternate method to calculate PCs by singular value decomposition (SVD) also exits [178].

2.12 Models for linkage disequilibrium (LD)

LD has an important relation to CNVs and has previously helped in several applications including elucidating haplotype diversity [179] and determining regions of positive

selection in the genome [75]. At present there are contradictory reports over the extent of LD between CNVs and SNPs. While the Wellcome Trust CNV study [76] reported that most CNVs characterized in their study were in high LD with surrounding SNPs, some earlier studies like [119, 180] reported contradictory results. At present with the advent of high-density genotype data from the 1KG project will require better methods to assess LD and will also lead to new high-resolution map of LD in different populations. New methods for calculating LD which can accommodate GSVs with SNPs will be necessary in the near future.

There are three methods for calculating LD denoted by D, D' and r^2. D is calculated based on the assumption of independent segregation between any two loci given by

$$D = P_{AB} - P_A \cdot P_B$$

However, there are still limitations of using D in genome-wide applications due its sensitivity to allele frequencies. Alternately, other LD measures such as D' and r^2 have been proposed and are used more often.
D' is defined as

$$D' = \begin{cases} \dfrac{D}{min(P_A \cdot (1 - P_B), (1 - P_A)P_B)} & D > 0 \\ \dfrac{D}{min(P_A \cdot P_B, (1 - P_A) \cdot (1 - P_B))} & D < 0 \end{cases}$$

The most popular LD measure is the r^2, also known as the correlation coefficient. The r^2 measure is determined as follows:

$$r^2 = \frac{D^2}{P_A(1 - P_A)P_B(1 - P_B)} \tag{2.18}$$

Here, $r^2 = 1$ is interpreted as perfect LD, which means that one marker gives complete information for the other. One advantage of r^2 is that it is less influenced by sample size and is not affected by recombination. One important application of r^2 is to select tag -SNPs in GWAS.

2.13 Quality control and filtering in CNV association analysis

2.13.1 Genotyping details and quality control for chapter 3

The BC1958 controls were genotyped on the Illumina 1M chip. Further details have been reported earlier [181]. Genotyping of the UK and Australian cohorts was done at the Wellcome Trust Sanger Institute, Cambridge. All samples from both the cohorts were genotyped on the Illumina 660K quad chip. Samples were also simultaneously genotyped on a Sequenom genotyping platform for 26 SNPs and then compared with the Illumina chip results. Any sample which had less then 90% concordance between the Sequenom platform and the genotyping chip were discarded.

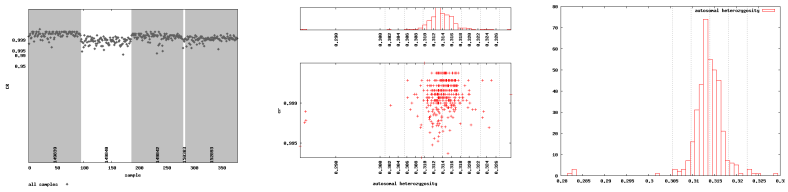


FIGURE 2.4: Genotyping QC in Australian Cohort

Left plot shows the call rate for each sample. Centre figure plots call rate versus autosomal heterozygosity. Rightmost figure shows the histogram of the heterozygosity for all samples.

Further, filtering criteria were applied based on the SNP call-rate > 95 %, $MAF \geq 0.05$ and a Hardy-Weinberg cut-off p-value of $1xE-4$. The cut-off for heterozygosity was different for the UK and Australian cohorts and is shown in Table 3.1. These steps

	Min	Max
UK	0.3057	0.3227
Australia	0.3054	0.3054

TABLE 2.2: Heterozygosity rates in UK and the Australian cohort

provided us with a high quality list of samples as well as SNPs which were used for further downstream analysis including our CNV work.

2.13.2 Genotyping platform details for discovery and control cohorts for chapter 4

The genotyping details for the various cohorts are as follows. NFBC 1966 was genotyped on Illumina Infinium 370cnvDuo array (N=5402), NFBC 1986 was genotyped on custom Illumina Cardiometabochip (N=6143), BC1958 was genotyped on Illumina 1M-Duo chip (N=2867) and NBS genotyped on the Illumina 1M-Duo chip (N=2737).

2.13.2.1 Normalisation and quality control

The normalization process I applied for the genotyping cohorts is as follows. Cohorts genotyped on the Illumina platforms were processed through the Illumina Beadstudio or Genomestudio 2.0 software. LRR, BAF and sample genotypes were exported from the Beadstudio/Genomestudio softwares for the CNV analysis. Prior to CNV-calling data normalisation was done in a genotyping plate specific manner in order to adjust for batch effects. For every genotyping plate, samples were corrected for LRR median correction and LRR variance. Wave effects were accounted for by fitting a localised loess function

with a 500 kb window. Processed LRR and BAF were next used for CNV-calling by cnvHap.

2.13.3 Analysis methods for chapter5

LRR and BAF intensity measurements were exported from the genome studio software for all the UKBEC samples for CNV analysis. Prior to CNV calling, specific pre-processing steps based on GC and loess was performed in a similar manner as described in Chapter 3. CNV-calling was done using cnvHap with recommended parameter settings for Illumina genotyping platforms. For this analysis, I applied an approach for defining copy number variation regions (CNVR) as defined and implemented in the cnvPipe software. CNVR was based on the idea of *minimal region approach* which is defined as the largest region where samples having a CNV do not change in CNV-state in one or more cohorts. This approach is relevant for regions of the genome where there are multiple overlapping CNV segments. In this case, the cnvPipe method can simplify this complex overlapping CNV segments into a simple minimal region. CNV genotypes are then calculated for these minimal regions, and have the possibility to be taken further for meta-analysis. After defining these minimal regions, an eQTL association analysis was done for them using the matrix eQTL package in R. A linear model for association between the marker genotypes and normalised gene expression values was done for each brain region plus the average across all these regions (referred to as average-all). The nominal p-values calculated were next converted into FDR based on the number of test performed and were calculated using the Benjamini-Hochberg procedure. Associations were defined as significant whenever FDR $< 1\%$

Chapter 3

CNV Analysis in Epilepsy

3.1 Introduction

3.1.1 Overview

In this chapter I have explored the role of GSVs like CNVs in conferring disease phenotypes, in particular for traits related to epilepsy. These phenotypes include general susceptibility to epilepsy in a case-control setting and also a complex drug-response phenotype defined as remission to anti-epileptic drug (AED) therapy. I have described the results from these two analyses, the validation approaches for the top results, have discussed some of the biological and clinical implications and lastly touched upon the future directions for drug-response analysis.

3.1.2 Background

Epilepsy is one of the most common and heritable neurological diseases often characterised by recurrent unprovoked epileptic seizures [182]. In clinical practice, prognosis is referred to as the future outcome and progress of any disease. Prognosis for epilepsy is highly variable, including for patients with similar types of epilepsy and epilepsy syndromes. The variable prognosis may have some genomic causal factors but its true causal reasons remain unknown. However, overall about 60 % of all epilepsy patients

achieve long term remission of seizures on AED therapy, but there is still a large proportion of patients (about 20-30 %) who do not respond to medication and never achieve remission of epilepsy seizures [183]. There are about 20 anti-epileptic drugs in the market but it has been observed that resistance to any of them usually means resistance to all. This phenomenon has been termed as multi-drug resistance. The mechanisms of multi-drug resistance are poorly understood and may include pharmacokinetic and pharmacodynamic aspects of drug-response, as well as variable inherent disease severity.

Achieving control of seizures for epilepsy patients is a crucial aspect for decreasing their risk of death and for improving their quality of life [184]. Although in recent years there has been progress in elucidating the molecular and genetic basis of epilepsy susceptibility, there has been limited progress evaluating the role of genetic factors in prognosis of epilepsy, with to date only a single GWAS of drug-response in epilepsy thus mandating further investigation [184].

3.2 Patients

In my analysis I have used data from two cohorts of newly treated epilepsy patients. This consisted of a UK based discovery cohort coordinated by Dr Michael Johnson at Imperial College London, and an Australian replication cohort coordinated by Professor Terry OBrien at the University of Melbourne [184]. Details of the cohorts are contained within [184]. The UK cohort consisted of 916 patients who had earlier participated in the Standard And New Anti-epileptic Drug (SANAD) randomized controlled trial which consisted of two treatment arms. Arm A consisted primarily of patients with focal epilepsy treated with carbamazepine (CBZ) while Arm B consisted of patients with generalized epilepsy [185].

The Australian replication cohort consisted of about 380 patients recruited from two epilepsy clinics from hospitals in Australia, including the Royal Melbourne Hospital and the Austin Hospital in Victoria. The selection criteria for these cohorts were identical to the UK cohort except that here all cases were aged 10 or over (5 or over in the UK study) [184]. All patients gave written informed consent. Some of the baseline covariate information collected and considered was gender, age, cranial computed tomography

(delta CT) or magnetic resonance imaging (MRI) and several other features described here.

3.3 Study design

In this case study for CNV analysis for epilepsy, I have carried out two approaches. First, I have carried out an analysis of susceptibility to epilepsy where I looked for CNVs at the genome-wide level in a case-control study design. This approach aimed at finding common CNVs which conferred risk for epilepsy. To this end I used the cnvHap algorithm. One of the capabilities of the cnvHap algorithm is that it can integrate several different datasets genotyped on different platforms into a single probabilistic model for CNV detection. This approach as described in the original study is more sensitive to detecting common CNVs by using population level information for model training. In the second analysis, I have carried out a drug-response association analysis within the epilepsy cohort, where all patients were stratified based on whether they responded to AED therapy, defined as achieving 12 consecutive months of freedom from seizures, in accordance with International League against Epilepsy (ILAE) [186]. Definition of treatment response was: patients who achieved twelve month remission were categorized as responders while patients who failed to remit were categorized as non-responders. In our association analysis framework we considered the UK cohort as the discovery cohort and the Australian cohort as the replication cohort.

3.4 Methods

In this section, I have described the overall analysis pipeline for the CNV-analysis for epilepsy. These steps were applied for both UK and Australian case cohorts, and the 1958 British Birth Cohort (BC1958) and the Helsinki Birth Cohort Study (HBCS) which were the control cohorts for the susceptibility analysis. In the first section, I have described the quality control (QC) steps applied to the discovery (UK) and replication cohorts (Australian). This included the genotyping details and the various QC measures applied to the Australian cohort. These QC steps were important to get a list of high-quality SNPs for further downstream analysis. I have also applied some custom normalization steps, which are necessary prior to CNV-calling. This includes pre-calculating scores for

regressing out GC effect. A loess function was further fit to the intensity data to cancel out wave effects commonly encountered in genotyping datasets and corresponding offsets were calculated. These offsets and GC scores were regressed out while CNV-calling was performed through a linear model in the cnvHap framework.

Next, I have addressed an important problem related to batch effects commonly encountered in CNV studies. Batch effects are caused due to the biases in the various genotyping plates and are known to cause spurious associations. To address this problem, I have described a novel way of calculating genome-wide LRR PCs implemented through a fast iterative orthogonalization process. Related processes have shown earlier to capture platform-specific biases in an exome-sequencing study for psoriasis, and here we applied it for SNP genotyping platforms. This method provided us with LRR PCs which can capture the essential bias in genotyping platforms due to differential hybridization on different plates. The LRR PCs were used as covariates in our association regression model. A population stratification based QC based on kinship matrix was also undertaken in a prior analysis in order to remove individuals who were closely related. Further, genotyping PCs were also calculated and adjusted for in the regression model this approach is furhter explained in detail in a later section. To assess the CNV-calling quality comparison, I have tested the allele frequency of the CNVs of interest in several other control cohorts including National blood survey (NBS), BC1958 genotyped on the aCGH platform and the HBCS. A principal component analysis of SNP genotypes had shown in an earlier analysis that the UK and Helsinki cohorts were genetically quite distant and were consequently not appropriate for association analysis. I therefore excluded the HBCS cohort as our controls, and used the BC1958 for this purpose, although the HBCS cohort was used to compare some of the control population allele frequencies. The aCGH WTCCC datasets including BC1958 were also useful for comparing allele frequencies and provided a preliminary in-silico validation for some of our CNVs.

3.4.1 Calculating LRR PCs

I have implemented a method for calculating global PCs for SNP genotyping platforms similar to the approach described earlier for exome-sequencing data (Figure 3.1) [187]. Briefly, I have first calculated local PCs using a fast iterative scaling algorithm and then derived global PCs from this. The iterative method works by splitting all SNPs into

regions i. Then, for each individual $k = 1...N_i$ and LRR values were calculated every jth probe and the first PC for each matrix given below is calculated

$$R_i = \{r_{ijk}\}$$

This method is based on the algorithm described in [188]. It first centralizes the matrix R_i and then derives the first principal component FPC_i by iteratively updating the following equations until the algorithm converges.

$$t = \sum_{j=1 \cdot M} < r_{ij}, FPC_i > r_{ij}$$
$$FPC_i = \frac{t}{< t, t >}$$

The convergence threshold I applied for my analysis was $1E-3$. This process has low memory usage and avoids the computationally expensive calculation of the covariance matrix RR^T. Next, using the SVD algorithm global PCs were calculated comprising of all the local PCs for H global principal components, which I refer to as $GPC_1, ...GPC_H$. In my analysis I considered the first ten PCs to be used as co-variates in the association model. These global PCs help in capturing the batch or plate effects in the genotyping platforms and I used them as covariates in our subsequent association analysis. One approach which is similar to the earlier work, will be to project out the global PCs from the local PCs and use these modified local PCs for association. This possibility of trying to associate local PCs for analysis still remains to be explored.

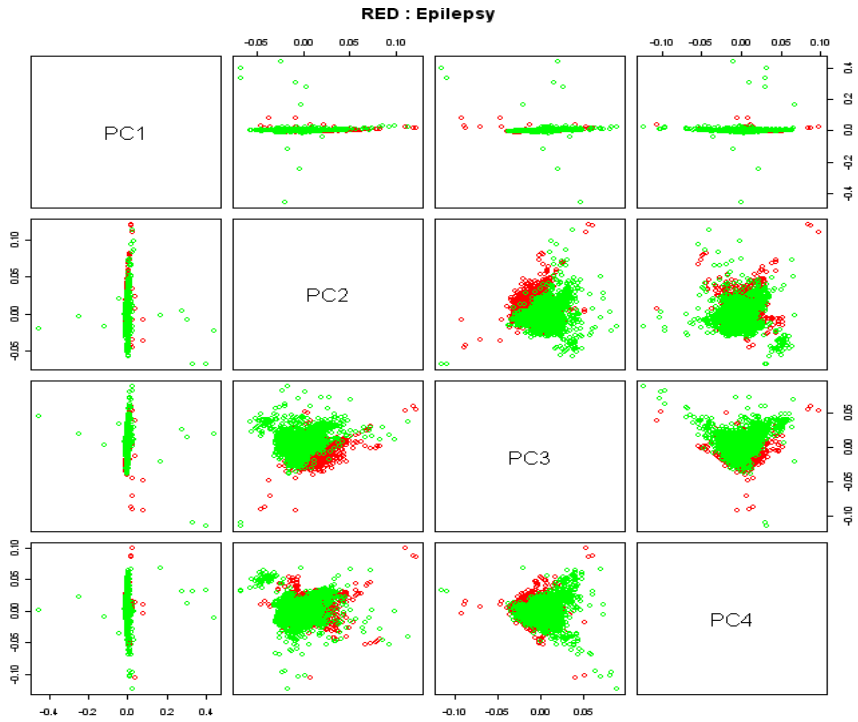


FIGURE 3.1: Plot for LRR PCs

LRR PCs were calculated by the iterative method described in 3.4.2 for epilepsy and BC58 in a joint manner. Samples in red denote epilepsy as cases and BC58 as controls and is denoted by green.

3.4.2 Multiplatform integration using cnvHap

cnvHap has shown to outperform other algorithms for calling CNVs on both aCGH and Illumina platforms and has increased sensitivity for CNV-calling and genotyping when different platforms are combined into a single probabilistic model [10]. The basic idea of how cnvHap achieves this multi-platform integration is described briefly in the following paragraph.

First considering all the probes in all the platforms, a meta set of probes is created with their genomic locations as the spatial coordinates for the HMM. Next, based on the emission probability of the corresponding genotyping platform, the likelihood of each sample for different copy number classes is calculated. The emission probabilities for different platforms have been defined using different basis functions. The following basis function works well for the corresponding platforms but in principle different functions can be used. Further exploration in this area remains to be done.
Equation 3.1 shows the emission probability for Illumina arrays. Here I_m refers to two orthogonal measurements of intensity LRR and BAF and is denoted as I_m=(BAF_m,LRR_m).

$$P(I_m \mid \mu(g), \sum(g)) = P(LRR_m \mid N(\mu_{rm}(g), \sigma_{rm}(g)))$$

$$\begin{cases} P(BAF_m \mid U(BAF_{min}, BAFmax)) & if CN(g) = 0 \\ \\ P(BAF_m \mid TN(\mu_{bm}(g), \sigma_{bm}(g)) & otherwise \end{cases} \tag{3.1}$$

N refers to the normal distribution and TN is truncated normal to lie between $0 \leq \text{BAF}_m \leq 1$.

The parameters from equations 3.1, namely $\mu_{rm}(g)$, $\mu_{bm}(g)$, $\sigma_{rm}(g)$ and $\sigma_{bm}(g)$ are functions of the hidden genotype g and are modelled through a linear model given below

$$\begin{pmatrix} \mu_{rm} \\ \mu_{rm}^2 \\ \mu_{bm} \\ \mu_{bm}^2 \end{pmatrix} = \beta * \begin{pmatrix} f_0(g) = 1 \\ f_1(g) = log(CN(g)/2) \\ f_2(g) = log(CN(g)/2)^2 \\ f_3(g) = bfrac(g) \\ f_4(g) = bfrac(g) * (1 - bfrac(g)) \\ f_5(g) = bfrac(g) * (bfrac(g) - 0.5) * ()bfrac(g) - 1) \end{pmatrix} \tag{3.2}$$

The emission probability for Agilent arrays is denoted by array of total intensity values (LRR_m) except that there is no term for BAF. BAF in this case is randomly assigned between 0 and 1. The emission probability for Agilent arrays is defined by equation 3.3.

$$P(I_m \mid \mu_m(g), \sum_m(g)) = P(LRR_m \mid N(\mu_{rm}(g), \sigma_{rm}(g))) \tag{3.3}$$

If an individual did not have measurements at a particular probe, then a uniform fuzzy distribution is used for determining the genotype distribution which includes the possibility of any CNV genotype at that location. For probes which were measured by both platforms being integrated, the combined likelihood is calculated as the product of the two platform specific likelihoods. In our case, there were only two platforms, but in principle using this framework more than two platforms can be combined (Figure 3.2). This approach further allows mapping CNVs from one chip to another and then later can be compared and verified. The way to do this is to include all the probes from one chip and then mask the actual values of the other and use uniform fuzzy values. I have tried this approach in chapter 4 where I used PennCNV CNV calls in the cnvHap framework, for visualization and comparison of results for the *TSPAN8* CNV.

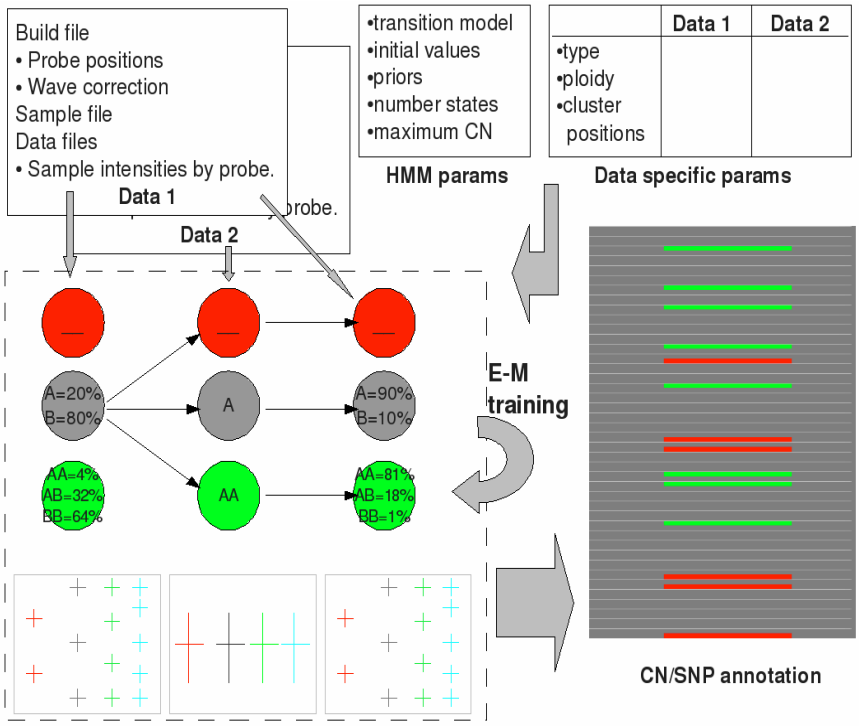


FIGURE 3.2: Multi-platform integration through cnvHap

Here, data 1 and data 2 are two different platforms which are integrated through probe locations. Data 1 does not have probe from data 2. In the final results the CNV call for data 1 takes into account information through data 2, thus creating integrated calls (Figure provided by Dr Lachlan Coin).

In the first analysis the objective was to find common CNVs conferring susceptibility to epilepsy. However, the cases and the BC1958 controls were genotyped separately. In this scenario cnvHap proved to be an ideal framework to use since it can accommodate situations where cases and controls have been genotyped on different genotyping arrays. As a test of combining different platforms for CNV association analysis effectively, I successfully replicated the CNV association for *TSPAN8* gene in the WTCCC T2D cases and BC1958 controls as a part of the WTCCC CNV study [76]. These cohorts were genotyped on the aCGH platform and showed positive association with T2D in the region $CNVR5583.1$, which was shown to be the top CNV associated with T2D earlier.

In addition, I also found that, by including high quality CNV dataset through the multi-platform framework described above, this improved CNV-clustering quality and also indicated that this might lead to improved association statistics. This indicates that inclusion of high quality datasets might have the possibility of improving CNV calling and genotyping, but requires further analysis. One possibility to achieve this could be to replicate the WTCCC CNV associations. I have the datasets to do this and I aim to explore this in the future. I further postulate that inclusion of high-quality sequencing data with genotyping data could show improved results. I echo this point again in chapter 4 in the light of CNV breakpoints. Methodology development and framework to do this remains to be done.

3.4.3 Association approaches

The two phenotypes used in our study design were binary (a) susceptibility to epilepsy and (b) remission of epileptic seizures to AED therapy (a binary outcome related to drug-response). In the susceptibility analysis I split the WTCCC controls randomly between the UK and the Australian cohorts with two thirds of control samples used for the discovery association and one third for replication. For the drug-response analysis cases and controls were stratified within the UK and Australian epilepsy cohorts based on response to AED (i.e. drug response cases and controls were selected from the same epilepsy cohort). Prior to CNV association analysis, for each sample I tested the total number of CNV-calls made by cnvHap. The main motivation to do this was to remove spurious samples which had excessive number of CNV calls which can results in wrong p-values.

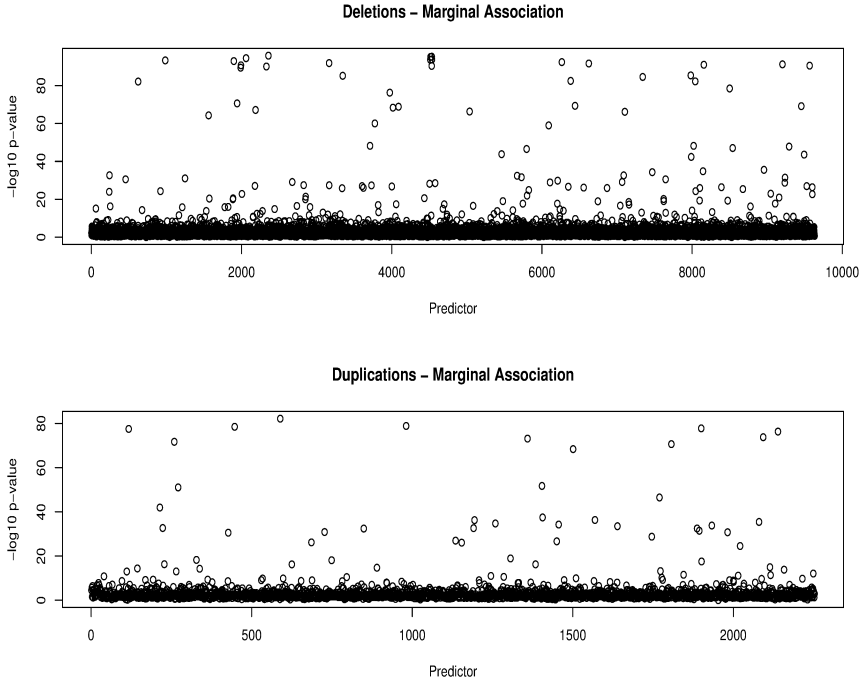


FIGURE 3.3: Marginal scores for association

Marginal scores shows the marginal test of association for deletions and duplications (frequentist) which are similar to the p-values calculated through association.

I first derived probabilistic measures of CNV-genotypes for each sample for all probes having CNV calls. Next, I created an association statistic based on a likelihood function. This gave me a score for each sample, which reflected the total number of CNVs. Samples with high score were subsequently removed. The above work was done in collaboration with Doug Speed from UCL (Figure 3.3, 3.4, 3.5).

Next, association of the epilepsy related phenotypes and the cnvHap derived expected genotypes was performed using a simple logistic regression model. The covariates used in the association model for the two studies were age, sex, first 5 population PCs and first 10 LRR PCs. Based on the top hits from the discovery cohort, I further short-listed CNVs which replicated in the replication cohort with a p-value threshold of 0.05. Next using the METAL software [189], I performed a meta-analysis of the top CNVs using inverse-variance fixed effects model with genomic control. I further applied a filter to this list based on the genomic location of the CNVs, focusing only on genic and exonic regions. This was done to prioritize and focus on the functional aspects of the CNVs in

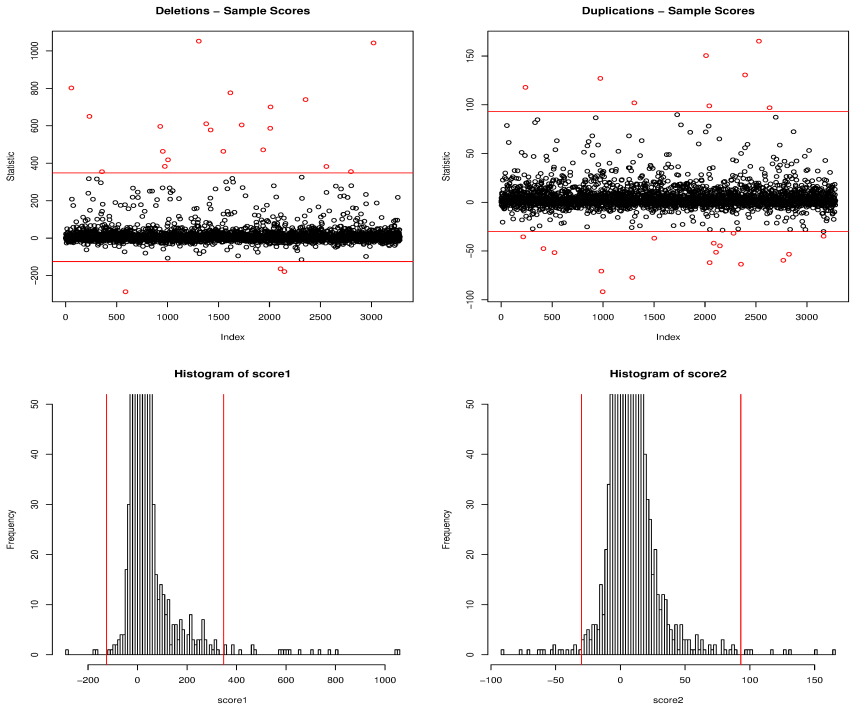


FIGURE 3.4: Score for samples driving CNVs

These scores are the scores per SNP for deletions and duplications in every sample. The idea is to look for samples which stick out, as this way indicates that a group of samples drives a lot of strong associations

the first instance. cnvHap produces a CNV annotation on a probe-by-probe basis, hence theoretically has the capability to detect single-probe level. I confirmed such a possibility where cnvHap was able to detect a CNV at single-probe level on chromosome 12 in an adult-obesity cohort at a genomic location where through sequencing results (1KG CEU population) it was subsequently shown to contain a CNV. But in general, single probe CNVs are harder to call through HMMs (due to their linear spatial structure) and more likely to have higher false positive rates. Thus it will be unwise to rely on HMM models for single probes CNVs, unless a proper parameter optimization and genome-wide analysis is done to assess the efficacy of such an approach. Results from sequencing can help in creating such a test-harness and remains to be done in future. Subsequently, I have discarded all single-probe CNV calls and considered CNVs spanning a minimum two consecutive probes. I have further assessed the clustering quality of these CNVs through visual inspection and further compared the clustering quality in the replication and other control cohorts. I further applied a filter to consider probes where the joint

allele frequency between the cases and controls combined was greater than 0.01. These steps gave me a final list of top CNVs associated with epilepsy susceptibility and epilepsy drug-response. These were next taken further for downstream analysis.

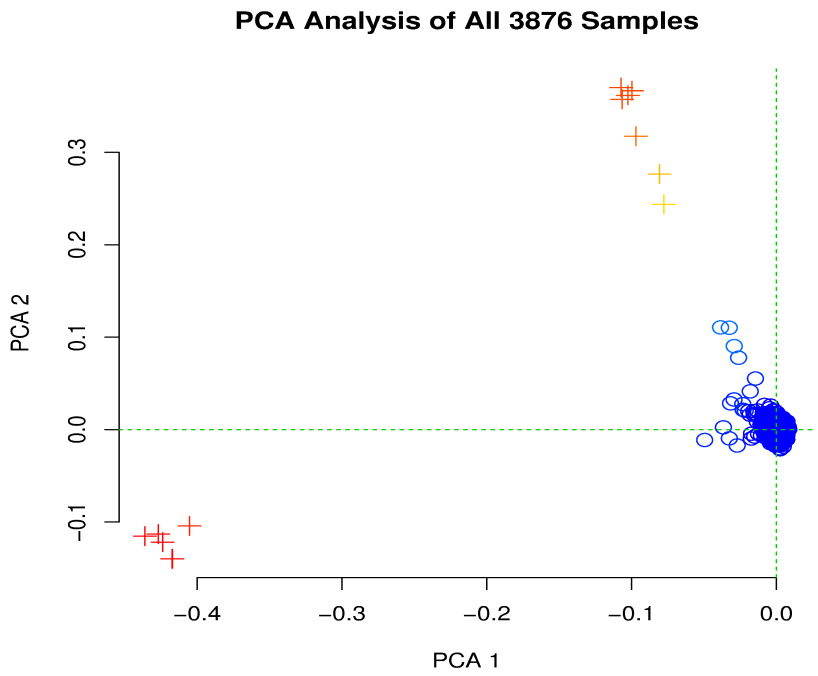


FIGURE 3.5: Example of SNPs driving association

The plot shows the first two PCs, coloured according to the extremity of the individual scores, so it can be seen that some extreme population samples separate out, denoted as red coloured crosses. Similar approach was applied to our epilepsy analysis.

3.5 Results

In this section I have described the results for susceptibility and drug-response analysis. First, I have reported the results for detection of large CNVs for epilepsy susceptibility. I have then described the results for whole-genome analysis of susceptibility and drug-response. I have discussed the biological functions for some of the top genes and the validation steps carried out for them. In the last section I have also explored a novel CNV detection and genotyping method based on the concept of clustering. I have next re-analysed some of the CNVs discovered through cnvHap for the susceptibility and

drug-response analysis and then discussed some of the future directions and possibilities of analysis.

3.5.1 Analysis and comparison of large CNVs

In the first instance I aimed at discovering large CNVs contributing to risk for epilepsy. For defining large CNVs, I set the criteria for the CNVs to be spanning at least 10 consecutive probes and then compared our results with another CNV study for epilepsy, performed using on the aCGH platform [190]. This approach undertaken in collaboration with Dr Mefford, would permit validation of any positive results, and perhaps save further lab work. With this objective in mind I started looking into large CNVs spanning more than > 900 Kbps in our discovery cohorts and compared those overlapping with Dr. Meffords results. For association, I performed a Fisher test for calculating p-values. Comparison of CNV genotypes was not possible since we did not have the raw data of Dr Meffords dataset and hence ab-initio CNV calling for them was not possible. This comparison produced a list of four CNVs out of which only one, a chromosome 10 duplication had consistent p-value for association with epilepsy and was taken further for analysis. The results for this CNV gave us reasonable confidence but there was one potential drawback. Our platforms were Illumina where as Dr. Meffrod's dataset was aCGH. Though it is possible to combine aCGH and Illumina through cnvHap, it is advisable to do so only with proper genome wide calibration. Exploring this possibility was not feasible in our case. Therefore as an extra validation step for the chromosome 10 duplication, I downloaded and analysed the aCGH data for BC1958. I aimed to compare the CNV calls from this dataset with Dr Meffords calls. I argue that this comparison might be a slightly better approach since both datasets were genotyped on aCGH platforms and hence the lack of parameter optimization will not make much difference.

However, results from this association did not give a significant p-value (< 0.05) for the chromosome 10 duplication or for any of the other 3 CNVs I initially discovered. Therefore, I concluded that I had found a common CNV and the association I saw earlier in our discovery cohort might have been driven through slightly different allele frequencies between the two different datasets. One possible reason for this could be under calling in the control datasets. I further re-analysed these results again by setting

more stringent parameters for the mean clustering position in cnvHap. Though I did rediscover the chromosome 10 duplication, there was no association with epilepsy susceptibility for either the UK or the Australian cohort. Further, none of the previously reported CNV associations with epilepsy was identified in either the UK or Australian epilepsy cohorts.

However, it should be noted that testing these parameters for some other known CNVs leads to a restricted allele frequency. This indicated that the parameters applied may not be completely accurate. I concluded that these results remain inconclusive unless further lab based validation or replication is done. I have highlighted that reproducing associations through different genotyping platforms is feasible, but remains challenging. At present there is no exact method for adjusting parameters for our approach since these values vary on a case-to-case basis. Hence, I conclude that the chromosome 10 duplication is probably real but its true nature of association with epilepsy remains unknown for now.

3.5.2 Epilepsy susceptibility analysis

In the earlier GWAS for our UK epilepsy population the samples were found to be predominantly Caucasian and very few samples were discarded due to relatedness. However, the Australian cohort was found to be more heterogeneous and large number of samples had to be discarded due to population heterogeneity. This made the replication cohort size decrease from 400 to about 280 and further posed power issues for replicating some of the CNV associations. Because of this I short-listed the top CNVs from the UK cohort, and then looked for replication in the Australian cohort. An alternative approach for discovering CNV associations with epilepsy susceptibility is to create a single list of top CNVs associated with epilepsy using a meta-analysis approach genome-wide. Table 3.2 shows the top CNVs associated with susceptibility after meta-analysing the top 40 CNVs with the UK as discovery cohort and the Australian dataset as the replication cohort and further restricting the CNVs to genic regions. The top gene associated with susceptibility was found to be *SLC31A1* (Table 3.1). This gene has a known role in copper transport and is known to play a role in Wilsons disease [191]. On doing a literature review I found an interesting relationship between this gene and epilepsy, suggesting further experimental validation of this association. Some other top genes found to be

associated with epilepsy were *EPHA7*, *ABCB4*, *ADAM22* and *NLGN1*. Brief literature review for these genes show that several of these genes have earlier been reported to be involved in neuronal activity and seizures across different species including humans and rats. Some interesting examples include are described below.

A patient carrying a micro-deletion in 6q16.1 encompassing *EPHA7* and SNP *rs12661215* (also present in our results) was found to have mild neurological abnormalities [192]. SNP variation within the *ABCB4* gene was shown to be associated with epilepsy drug-response [193, 194], while mutations in the *ADAM22* gene has been shown to be involved in causing seizures by interacting with the *LG1* gene in rats [195]. The *NLGN1* gene which is involved in important function of signal transmission in neurons has also been linked to one of the possible causes for epilepsy in autism patients [196, 197]. Amongst these top hits, the *NLGN1* gene proved to have strong clustering quality in all the datasets analysed and further proved to be the biologically most relevant candidate. A detailed analysis of the *NLGN1* gene function and its plausible shared role in epilepsy and autism spectrum disorder (ASD) is discussed in details in the following section. The presence of several biologically relevant candidate genes in our top results shows the inherent complexity of epilepsy and further suggests that it has a possible shared aetiology with several other diseases.

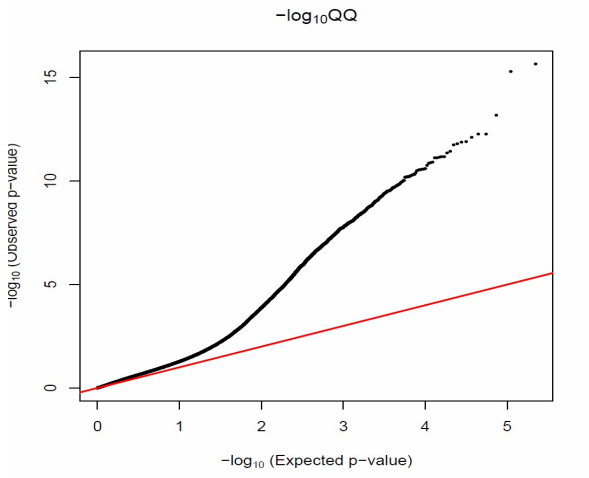


FIGURE 3.6: Quantile-quantile (QQ) plot for epilepsy susceptibility analysis

The QQ plot was found to be inflated for the susceptibility analysis. We considered only the top hits in this case. Inflation factor λ equals to 1.4 in this analysis.

TABLE 3.1: Association results for the epilepsy susceptibility analysis

Probe Name	Chromosome	Position	Control Freq	UK Freq	Z Score	Meta p-val	Direction	Gene
rs2233916	9	115060957	0.0089, 0.052	NA, 0.087	-7.79	6.69E-15	–	SLC31A2
rs344257	8	63124625	0.0084,0.06	0.004,0.048	-7.357	1.88E-13	–	Intergenic
rs12661215	6	94155840	0.0051,0.025	0.002,0.464	-7.243	4.38E-13	–	EPHA7
rs17149647	7	86924867	0.0009,0.041	0.0021,0.386	-7.042	1.90E-12	–	ABCB4
rs1688888	7	87443554	0.0032,0.037	0.0021,0.083	-6.454	1.09E-10	–	ADAM22
rs9878356	3	175379796	0.0094,0.06	0.013,0.03	-6.311	2.78E-10	–	NLGN1
rs13294756	9	2724038	0.0037,0.037	NA,0.151	-6.294	3.08E-10	–	KCNV2

3.5.2.1 *NLGN1*: Missing link between autism and epilepsy?

Approximately 70% of ASD patients show mental retardation and about 30% of ASD cases have epilepsy as a co-morbidity. There is emerging evidence to suggest overlapping aetiology between epilepsy and ASD and further elucidation of their molecular basis is crucial. The neuroligin membrane protein has known implication in causing ASD.

The *NLGN1* gene codes for cell adhesion molecules, located in the post-synaptic part of synapse. They interact with beta-neurexins, which lead to the formation of the synapses, and are crucial for synapse functions in neuron cells. The neuroligin gene family consists of *NLGN1* at 3q26, *NLGN2* at 17p13, *NLGN3* at Xq13, *NLGN4* at Xp22 and *NLGN4Y* at Yq11. While mutations in the two X-linked neuroglins genes *NLGN3* and *NLGN4* have already been implicated in causing autism, some of the recent studies have further shown that *NLGN1* and *NLGN3* are the current best candidates for autism.

Two important studies [198] and [199] have recently further confirmed this hypothesis. Some of the results from these studies have shown the link between activity-induced regulation of *NLGL1* by its proteolytic cleavage and seizures. This effect was shown by injecting pilocarpine in the brain of the mice (forebrain [199] and hippocampus [198]) leading to seizures and increased levels of NTF of *NLGN1* suggesting a link between the two. The link between neurexin and neurogilin has shown to be associated with autism but one review highlights that the regulation of *NLGN1* by proteolytic cleavage can uncover the missing genetic link between epilepsy and autism [200]. The functional role of *NLGN1* CNVs in light of the above cellular mechanisms remains to be further characterized.

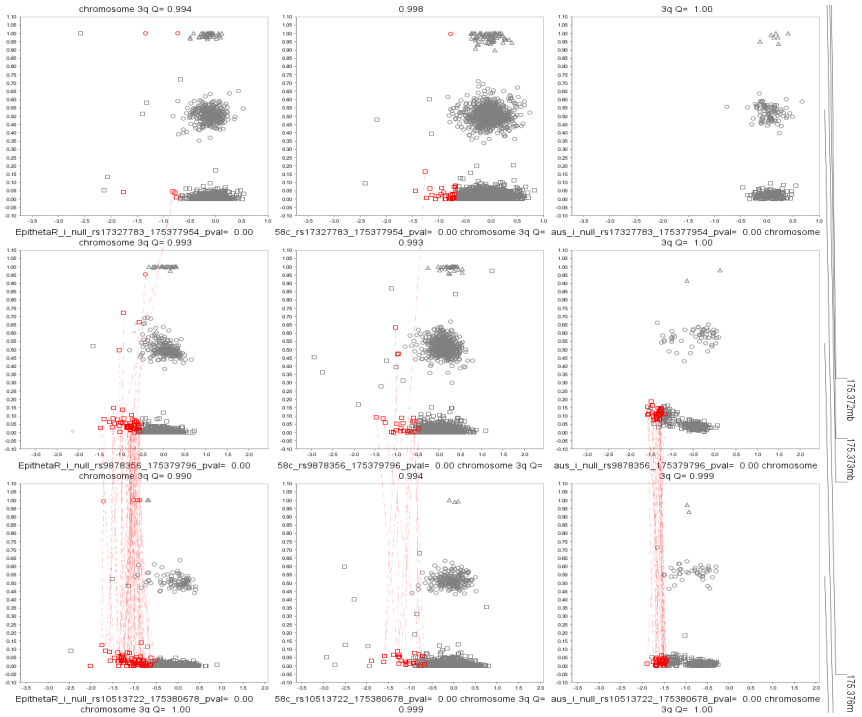


FIGURE 3.7: Cluster plot for deletion in *NLGN1* gene

Two dimensional cluster plot of LRR (X axis) and BAF (Y axis) showing deletion in the *NLGN1* gene. Each point is a sample. Homozygous, heterozygous and normal copy numbers are denoted by pink, red and grey respectively. Samples having CNV spanning multiple probes are connected by lines.

3.5.3 Epilepsy drug response analysis

In this analysis the phenotype of interest was 12-month remission of epilepsy (yes/no) in accordance with the ILAE definition of treatment response. Based on this criteria and earlier described quality control measures, I had a samples size of 647 for the UK discovery cohort and 254 for the Australian replication cohort. The low sample size in the Australian cohort was due to high population heterogeneity (as described earlier), decreasing our power to replicate hits from the UK discovery cohort. The association approach for the remission phenotype was similar to the susceptibility analysis described earlier, except that the cases and controls were drawn from the same epilepsy cohort, and thus genotyped on the same platform. I applied a filtering criteria of combined allele frequency between cases and control of 0.01 and further considered CNVs spanning at least 3 SNPs. This gave me a final list of 6 CNVs, out of which 4 were intergenic, 1 pseudo-gene and 1 genic-CNV. The genic-CNV was present in the *WWOX* gene (Table 3.2). I only considered the *WWOX* CNV for functional follow up. However, I failed to replicate this association in the Australian cohort (p-value > 0.05). Further analysis revealed that I had only 40% power to replicate the finding from the primary discovery cohort. To achieve a power of 80% I would have required a replication cohort size more than 600. This power calculation was performed using *pbsize2* function in the R package gap [201]. In this function the population prevalence of epilepsy was put as 0.005 with an additive disease risk model.

My current analysis and results remain promising but inconclusive. Current efforts are ongoing in recruiting more patients for the Australian cohort and in the future I plan to rerun this analysis. In the following section I have discussed the functional aspects of *WWOX* gene in greater details.

TABLE 3.2: Association results for epilepsy AED remission analysis

CNV region	P val Min	Beta	SNP	Remission Frequency	Not remissed Freq	Gene
chr3:163618455-163625724	8.52E-04	-0.932	rs10936279	0.067	0.166	Intergenic
chr4:161276893-161289109	8.01E-04	-1.46	rs10857326	0.022	0.084	Intergenic
chr10:114103579-114106565	7.24E-03	-0.294	rs12245271	0.147	0.269	GUCY2G
chr14:69090072-69091816	7.03E-03	-0.611	rs11851273	0.0198	0.069	Intergenic
chr16:76931145-76941622	1.93E-03	0.331	rs9940910	0.359	0.184	WWOX
chr22:16153400-16153450	4.06E-03	0.629	rs9606674	0.117	0.074	Intergenic

3.5.3.1 The *WWOX* gene

WWOX is a cytoplasmic protein which is involved in several cellular processes like cell differentiation, growth and tumor suppression [202, 203]. The WWOX gene lies within one of the most fragile sites in the human genome where several translocations and deletions have found to be associated with different types of cancer [203]. The *WWOX* gene encodes a 414 amino-acid protein consisting of two WW domains and a short-chain of hydrogenase-reductase domain. *WWOX* is also known to be highly expressed in several tissues, including cerebral-cortex and the cerebellum [204]. The WW domain of the *WWOX* gene is specialized in protein-protein interactions especially for proline-rich amino-acid sequences. Some of the interacting partners of *WWOX* identified are *p73*, *SIMPLE/LITAF*, *RUNX2* and *ERBB4*. The possible role of *WWOX* mutations in epileptic seizures is reported in the next section.

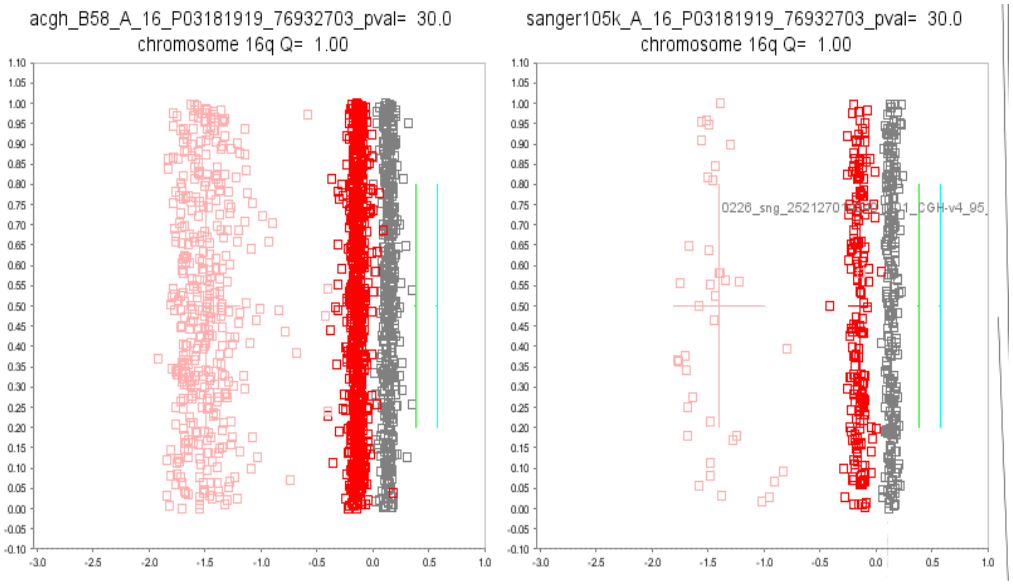


FIGURE 3.8: Cluster plot for deletion in the *WWOX* gene

A very good separation is observed in the plots above, indicating that this CNV is likely to be real. Samples here show homozygous deletions (pink) and heterozygous deletions in red. The left plot corresponds to WTCCC CNV study of BC1958 cohort and right plot corresponds to WTCCC CNV discovery study [76].

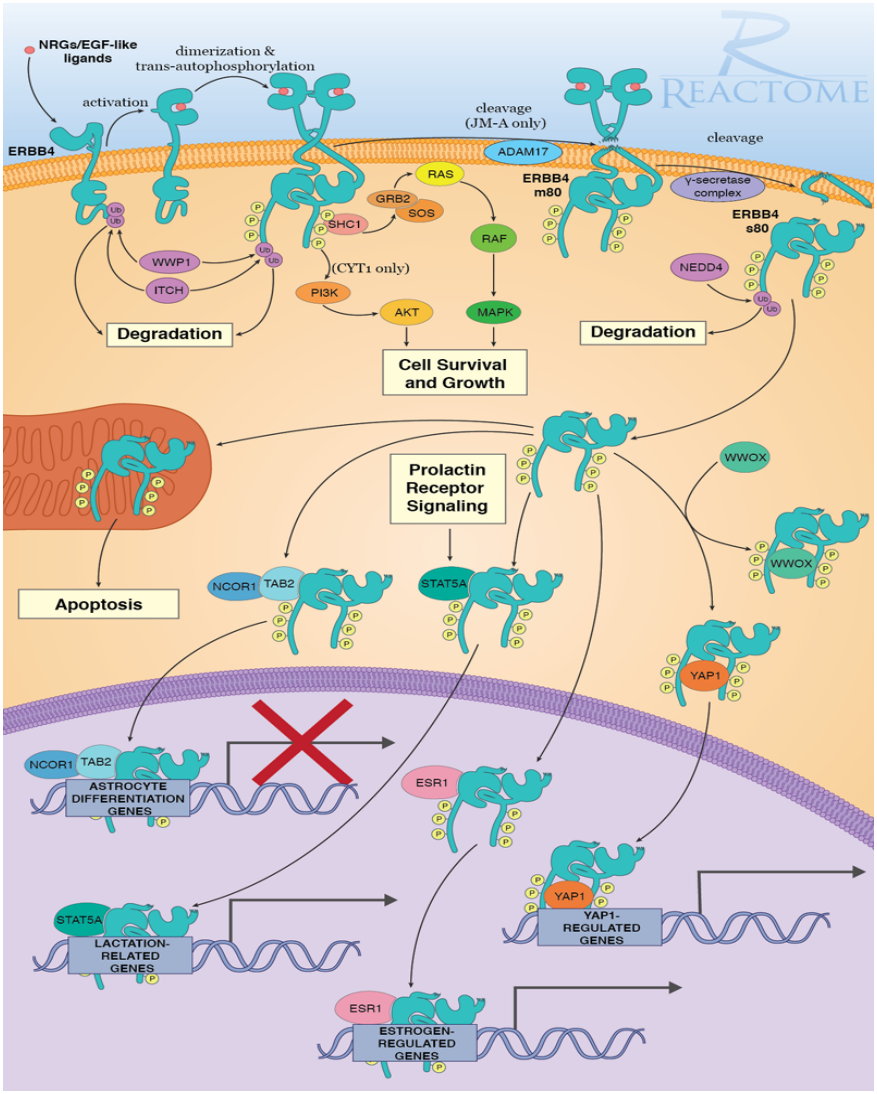


FIGURE 3.9: Functional aspects of *WWOX* proteins

WWOX is a cytosolic protein which binds to ERBB4s80 using the WW-domain binding motifs in the C-tail of ERBB4. This leads to the formation of ERBB4s80:*WWOX* complex and this competes with the formation of ERBB4:YAP1 complex which prevents translocation of ERBB4s80 complex to the nucleus. Feng et al. has shown that *WWOX* binds with the same affinity to s80CYT1 and s80CYT2 complex, and identified PY3 as the most important *WWOX* binding WW-domain (Source: http://www.reactome.org/).

3.5.3.2 *WWOX* and epilepsy

A recent study created of a new strain of rats with a new phenotype caused by a 13 bp deletion in the exon 9 of the *WWOX* gene, which they report as lethal dwarfism with epilepsy (LDE) [205]. This phenotype is characterized by high lethality, male hypogonadism and significant occurrences of epileptic seizures. They also report that viability and epileptic seizures are probably specific for LDE rats. Another study by [206] report that germ-line mutations in the *WWOX* gene led to several neuropsychiatric diseases like epilepsy. This study highlighted that variants in the *WWOX* gene should be investigated further in patients with epilepsy, mental retardation, and complicated hereditary spastic paraplegia (cHSP). They indicated that investigation of pediatric cohorts with early onset ataxia, generalized epilepsy and mental retardation will help in elucidating the functional role of *WWOX* mutations in greater details. In light of the above findings, the *WWOX* deletion is probably a good functional candidate for our epilepsy drug-response analysis.

3.5.4 Chromosome X results for drug remission

In this section, I have discussed the results for chromosome X, which required a different approach to autosomes. CNV-calling in this case was done on a gender-specific manner where the clustering for males and females performed separately. I also carried out an analysis to detect gender from intensity data and found that 6 samples had wrong assignment and consequently were excluded from the analysis. First, I carried out a chromosome X analysis only for the drug-response phenotype only. For the susceptibility phenotype, there was an added complexity of adjusting for mean-clustering positions for different platforms in addition to gender and therefore susceptibility as a phenotype was not considered for the time being. I have highlighted that several other shortcomings for chromosome X can affect CNV detection, arising mainly from the special characteristics of the intensity data for chromosome X, and thus require an additional calibration of the HMM parameters. One way to correct for this will be to create a test-harness for calling CNVs in a gender-specific manner for chromosome-X under different HMM parameters and then comparing calls to already known CNV regions. We will then need to choose appropriate parameters with highest sensitivity and specificity.

This gender-specific analysis for chromosome-X remains to be done in the near future. In the drug-response analysis the samples were split into two groups; treatment responders and non-responders. Analysis was done separately for males and females. At a p-value threshold of 0.05, the association results indicated that few CNVs were associated with drug-response outcome for males whereas no significant association was found for females. The top hit association for males in the *all patients* category was found to be the *ZDHHC15* gene. An earlier study has reported aberrations in this gene associated with seizures [207]. However, I indicate that these results remain preliminary at present unless they are replicated. Replication for the chromosome X results in the Australian cohort still remains to be done.

3.5.5 Results from DECIPHER database

The DECIPHER database (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resources) [208] (`http://decipher.sanger.ac.uk/`) was set up to catalogue rare deletions and duplications which alter gene dosage. The Sanger Institute initiated this project in order to make the analysis and comparison of various neuro-developmental disorders for several patients easier across the world. The database helps in giving more certainty to the physician about the pathogenicity of such kind of genomic alterations and further sheds light on their functionality with respect to the disease phenotypes. This knowledge further has the potential to define new syndromes. Some of the useful features of the DECIPHER database is a catalogue of all common human CNVs. Rare and pathogenic genomic alterations having special clinical significance have been well documented with rich annotations of associated genes and other aberrations. Thus the DECIPHER database helps bridge an important link between CNVs and their functional effects on disease phenotypes and is a useful resource for the genomic community.

I have analysed some of the top hits from my results in the DECIPHER database to get some further insights regarding the phenotypic characteristics of the CNVs I found associated with epilepsy susceptibility and drug-response outcomes. These results indicated that there have been several earlier neuro-degenerative symptoms in patients who had aberrations in the genes of our interest. Some of these findings are reported below. For the *NLGN1* gene, 6/8 patients reported in the DECIPHER database had

some kind of intellectual disability and interestingly one patient (DECIPHER id 257342) also manifested seizures. Most of the patients who carried CNVs in *NLGN1* gene had intellectual disability (Table 3.4). For the CNVs in the *WWOX* gene region 6/15 patients had been reported to have intellectual disability and further 2 manifested some kind of seizures (Table 3.3). It is to be noted that intellectual disability sometimes is closely related to autism and thus indirectly links to our premise that these all those outcomes might have some underlying functional relationship with epilepsy.

TABLE 3.3: Patients from the decipher database showing rare chromosomal aberrations in the *WWOX* gene

Patient ID	CNV description	Length	Phenotype Description
248495	loss 16:77843209-78961183	1.12	Non neurological disabilities
249906	gain 16:60340063-90294753	29.95	
250849	loss 16:78424175-82686053	4.26	Non neurological disabilities
253240	gain 16:72643701-90124279	17.48	
253977	loss 16:78064369-80913149	2.85	Non neurological disabilities
256542	gain 16:78186861-90267387	12.08	Non neurological disabilities
257432	gain 16:78021803-78472169	0.45	Intellectual disability
260535	loss 16:78361637-78394618	0.03	Intellectual disability, Short stature
263035	loss 16:78186845-78244864	0.06	Seizures, Intellectual disability and other impairments
265334	loss 16:78814326-79018563	0.2	Hyperactivity, Intellectual disability
270840	gain 16:77966858-78658425	0.69	Intellectual disability, Scoliosis
272477	loss 16:76180903-82314260	6.13	Non neurological disabilities
277950	loss 16:78403602-78455851	0.05	Seizures, Global developmental delay and other impairments
280976	gain 16:77647054-78283673	0.64	Global developmental delay, Abnormal facial shape
284999	gain 16:77850811-79451199	1.6	Autism spectrum disorder

TABLE 3.4: Patients from the decipher database showing rare chromosomal aberrations in *NLGN1*

Patient ID	CNV description	Length	Phenotype Description
251819	gain 3:163643791-180703745	17.06	Intellectual disability
282426	gain 3:168719148-197837020	29.12	NA
249927	loss 3:170492528-179161826	8.67	NA
280844	loss 3:172281099-178521936	6.24	Intellectual disability, moderate
257342	loss 3:172538343-173614890	1.08	Intellectual disability, Febrile seizures
270579	loss 3:172849041-173539490	0.69	Delayed speech and language development, Intellectual disability etc
283752	loss 3:173064931-173645205	0.58	Global developmental delay
255971	loss 3:173732084-174013501	0.28	Cognitive impairment, Attention deficit hyperactivity disorder

3.5.6 Lab based qPCR validation

We aimed at lab validation of some of our top CNV results to confirm our epilepsy findings. We had several options for validation available to us, including custom aCGH genotyping platform, sequencing and also the $TaqMan$ qPCR protocol for CNVs. After a detailed cost-benefit analysis we decided to go ahead with the $TaqMan$ qPCR protocol. As an initial check we selected three genes from the susceptibility analysis results including *ADAM22*, *NLGN1* and *EPHA7*. We made the selection based on the clustering quality, raw data and the certainty of CNV calls in both discovery and the replication cohorts. In the following section I have briefly reviewed the methodology behind the qPCR protocol. I further indicate that at present lab-based methods for validating smaller classes of CNVs may be prone to errors and therefore should be carefully considered.

The $TaqMan$ copy number assay protocol detects CNVs by the relative quantification of genomic DNA and target DNA using a real-time PCR with custom pre-designed assays. Each of the copy number assays is designed to detect specific genomic locations or sequences that consist of a reference and a target sequence. The reference sequence is usually a known genomic location with two normal copies of a sequence in the diploid genome, e.g. the *RNase PH1 RNA* gene. The next step involves the reaction mixture consisting of the test and reference DNA to be run on a real-time PCR using absolute quantification settings. After the PCR reaction, results containing the ΔCT values are exported from the PCR machine and imported into the CopyCaller software for post PCR analysis. The CopyCaller software first determines number of copies of the target sequence in each DNA sample and then the ΔCT between the cycles of the target and reference assays. Then a custom inbuilt method compares the ΔCT values of a test sample to the reference sample with known number of copies to determine the absolute copy number calls.

qPCR analyses for these three CNVs failed to detect any CNV association for epilepsy for our genomic regions of interest (Figure 3.10). There are several potential reasons why these CNVs did not validate. One important drawback was that the DNA concentrations for some UK and Australian samples received from the Sanger Institute were not present in the desired uniform concentration levels. This could have affected the PCR cycles and also the results to some extent.

To determine whether the possibility of DNA concentration or any other experimental factors had affected my results, I tried to carry out an in-silico analysis of the laboratory results. Some studies, such as by Fernandez et al. 2011 [209] have reported that input DNA amount and template integrity can affect the stability of qPCR based approach for validating CNVs. They reported that in a test set of three genes, qPCR results were prone to false copy number assignment when the DNA concentration varied. They reported that even a 4% change in amplification efficiency of the PCR cycles can lead up to 400 % error in the calculation of delta CT (also denoted as ΔCt) [210]. ΔCt is the difference in total DNA concentration from qPCR cycles for the test and reference DNA, and is used to calculate copy number state for each sample.

Since ΔCt and LRR both correlate to the underlying copy number, they should be correlated.To determine this and to indirectly confirm that data from the PCR and array data concur, I correlated ΔCt from the qPCR output with raw intensity LRR values from the genotyping arrays. However, on applying this approach, I found a very low correlation ($r^2 < 0.25$) for all three genes analysed in the experiments. These factors make the current experimental results inconclusive and will require further validation in the future before any valid conclusions can be made.

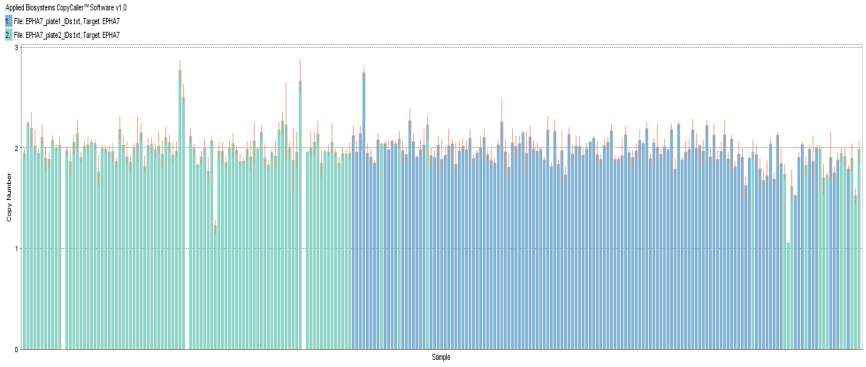


FIGURE 3.10: CopyCaller results for the *EPHA7* gene

Here samples are represented in the X axis and copy number in the Y axis. Samples were plated in two plates denoted by the green and blue. Except one sample, all samples had a copy number call close to two.

3.5.7 NGS sequence validation

In addition to experimental validation, I have also explored the possibility of indirect validation of the *WWOX* CNV association with epilepsy outcome using NGS data from the 1KG data. Amongst most of our hits, only the *WWOX* CNV had a high enough allele frequency to make us confident that it can be replicated through this approach. I downloaded the sequence alignment files (bam format) for 42 CEU individuals and used the cnvHitSeq algorithm for calling CNVs in this dataset. The *WWOX* CNV was successfully replicated in these samples. *WWOX* was found to be a heterozygous-deletion spanning the region $chr16 : 76929145 - 76942386$ (hg18) (Table 3.5, Figure 3.11). The detected allele frequency from sequencing results was 0.42 compared to an allele frequency of 0.1926 from the array-data for the UK cohort. The allele frequency in the Australian replication cohort was found to be 0.1375. All the CNV breakpoints had average certainty greater than 0.9 with greater than 50% reciprocal overlap with the cnvHap derived breakpoints. These results conclusively proved the existence of the *WWOX* CNV in our epilepsy datasets.

Table 3.5: CNV breakpoints of 3 CEU samples carrying the *WWOX* deletions

Sample Id	Chromosome	Start	End	Length	Type	Certainty
NA06984	16	78371644	78384885	13241	1	0.999
NA11843	16	78371644	78384885	13241	1	0.998
NA12005	16	78371644	78384885	13241	1	0.993

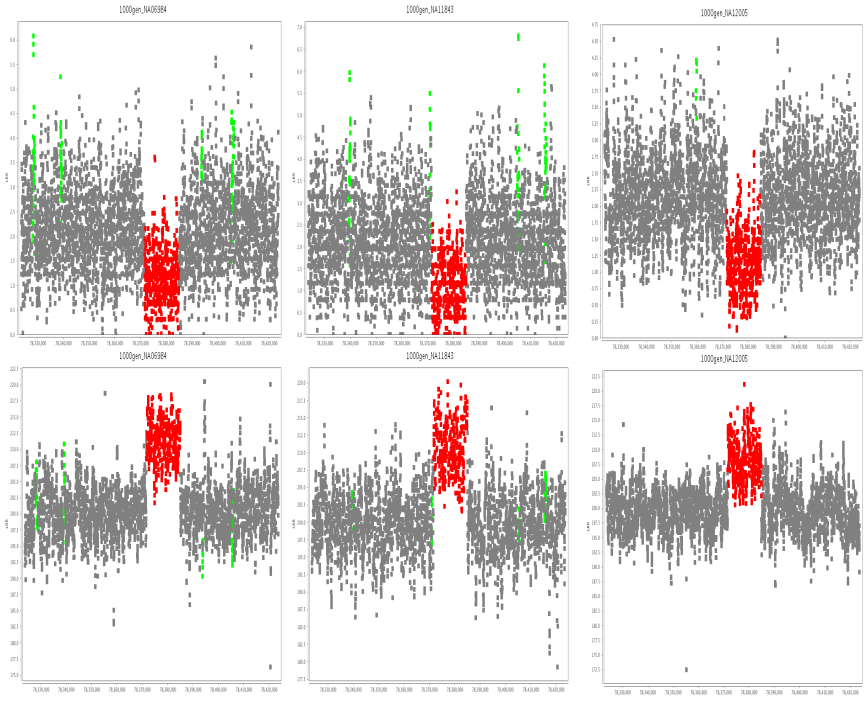


FIGURE 3.11: NGS sequencing results for *WWOX*

The plot shows a deletion in three samples (*NA06984, NA11843, NA12005*) from the *1000 genomes* CEU samples. Here, the top row corresponds to read-depth information and the bottom row corresponds to the read-pair information. It can seen that in presence of deletion the mean read-depth goes below the mean for normal copy 2 and read pair is above the average read-pair of 200.

3.6 Clustering based CNV detection

In the following section I have presented and developed some novel ideas related to CNV detection using a clustering approach. I have successfully applied these approaches for re-discovering some of the epilepsy CNVs. In the last section, I have briefly described how such clustering approaches can be developed further in the future.

3.6.1 Introduction

Classification and clustering can intuitively have similar meanings but there are important differences between the two (Table 3.6). Classification is a supervised learning method where a model is created from a prior labelled (classified) dataset and then used to predict unknown classes. Clustering on the other hand is an unsupervised learning method that aims to find natural groupings in the data with no prior information. Table 3.7 shows important differences between supervised classification and clustering.

TABLE 3.6: Comparison between classification and clustering

	classification	**clustering**
Model	$Y = F(x)$	generate model
Labelled training set	available	unavailable
Variable set	X_i, Y_i	X_i
Learning objective	$Y = G(x)$	unavailable
Goal	$E < F(x) - G(x) > 0$	unavailable
accracy	RMSE	unavailable

In spite of the problem highlighted above, several subjective criteria might be considered for determining high-quality clusters. Some of the criteria for determining good quality clusters can be for example based on the following points.

- High similarity within the same class
- Low similarity between classes
- Identify hidden patterns which otherwise can not be detected

At present, clustering methods have become a popular method for data analysis in particularly in the omics field. In general the concept of clustering can be categorized into the following four classes:

1. Partition algorithms, e.g k-means
2. Hierarchical algorithms, such as single-link, complete-link, group-average, ward.
3. Model-based algorithms e.g the EM algorithm
4. Density-based algorithms, e.g DBSCSN

Two of the most commonly used clustering algorithms are k-means [211] and density based spatial clustering of applications with noise (DBSCSN) [212]. While k-means is widely used because of its ease of implementation, it suffers from several limitations. These include the need for prior specification of the total number of clusters and its inability to reach an optimal global minimum (sum of squares in this case). A more robust version of the k-means algorithm known as partition around medoids (PAM) has also been proposed (Figure 3.15). PAM uses medoids in the algorithm instead of centroids and is considered to be more robust to outliers. PAM also has some limitations. These include the need of specifying the number of clusters and no specific criteria for determining the initial configuration of means of the medoids. Density based clustering approaches like DBSCSN [213] on the other hand has several advantages. DBSCSN is discussed in details in the next section

3.6.2 Density based clustering approaches

The DBSCSN algorithm tries to find clusters in unclassified data using density of data points in a region. The main idea behind DBSCSN is that for each cluster, the neighbourhood of a given radius (often denoted as Eps) has to have certain number of minimum points (denoted as MinPts). The main DBSCSN algorithm separates the data points into three classes:

1. Core points: These are points which lie within a cluster. A point lies in an interior region if there are enough points in the neighbourhood.
2. Border points: A border point is not a core point but it lies but falls within the neighbourhood of a core point
3. Noise points: A noise point is any point which is not a core point or a border point.

In the main algorithm of DBSCSN, it first starts with an arbitrary point (p) in a given data (D) and then retrieves all instances of D with respect to Eps and MinPts. This approach uses the RTree method [214] to locate points with Eps distance from the centre of core points of the clusters. DBSCSN can determine the total number of clusters on its own and further can fit arbitrary shapes inherent in the data. Here, the density around an object is measured by the number of objects within a hyperspace. The algorithm requires two parameters Eps and minPts. As a common practice the minimum points specified is 4 for most applications, including in my analysis presented here.

Some improved versions of DBSCSN have also been developed such as incremental DBSCSN, generalized DBSCSN (GDBSCSN) [215] and parallel version of DBSCSN (PDBSCSN) [216]. Other sophisticated methods which can automatically determine the optimal number of minimum points also exist. DBSCSN can identify outliers and has shown promise to perform better than many other clustering algorithms.

3.6.3 Results

In this section, I have briefly reviewed the application of DBSCSN and k-means clustering algorithms for detecting CNVs from Illumina and aCGH platforms. Using DBSCSN,

I have applied two approaches for detecting CNVs and also tested their capability for genotyping. In the first approach, I have applied the DBSCSN clustering algorithm on a probe by probe basis while in the second approach I have applied it on a moving window of 3 consecutive probes for all samples together (Figure 3.12, 3.13). In the latter approach, as the window is moved along the chromosome, whenever the first probe having a CNV is encountered by the model, separation of intensity data into clusters is seen. As the window moves along, separation of LRR clusters continues as long as the CNV segment exists. In this way every probe gets clustered two times, once taking in consideration the probe preceding it and once without. Preliminary results show that this method can be used to detect change in CNV breakpoints and has the possibility to incorporate CNV genotyping of the samples as well. The main motivation of applying the latter approach i.e. clustering over three consecutive SNPs is to capture both the *deep* and *wide* elements for CNV detection. While detecting CNV regions and genotyping samples simultaneously will be to difficult to apply for this kind of clustering approach, I propose a two-staged analysis in this scenario. In the first stage it will involve the application of the DBSCSN algorithm using the three window SNPs for a chromosome-wide run in order to identify CNV containing regions. Once these regions have been defined, further refinement of breakpoints, probabilistic assignment of genotypes can be done by re-clustering the data on a probe-by-probe basis and further comparing results using one or more different clustering algorithms. Further in my approach, I propose to use a meta-algorithm to determine the CNV-breakpoints after the initial clustering has been performed. This can be based on number of factors such as the genotyping consistency between consecutive probes, distance between the probes, previously known CNV at that particular probe, etc. Some advantages of this method is that it will not get affected by the problem of distance between the probes and possibly can be benefit from being able to detect single-probe CNVs. Model-based algorithms for detecting CNV break-points often perform inadequately when the probe density is low for a particular genotyping platform or when we apply a criteria of n SNPs for filtering high quality CNVs.

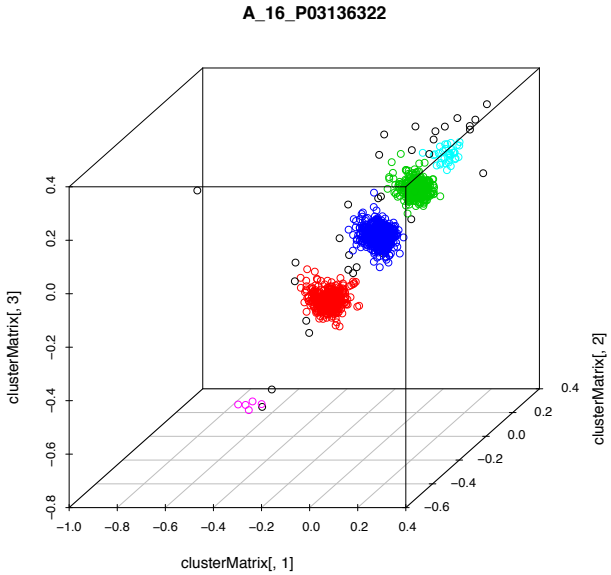


FIGURE 3.12: 3D plot showing clustering using the DBSCSN algorithm

DBSCSN is run over 3 consecutive probes in the WTCCC aCGH dataset. The algorithm detects 5 different clusters by itself and further also delineates outliers (shown as black circles).

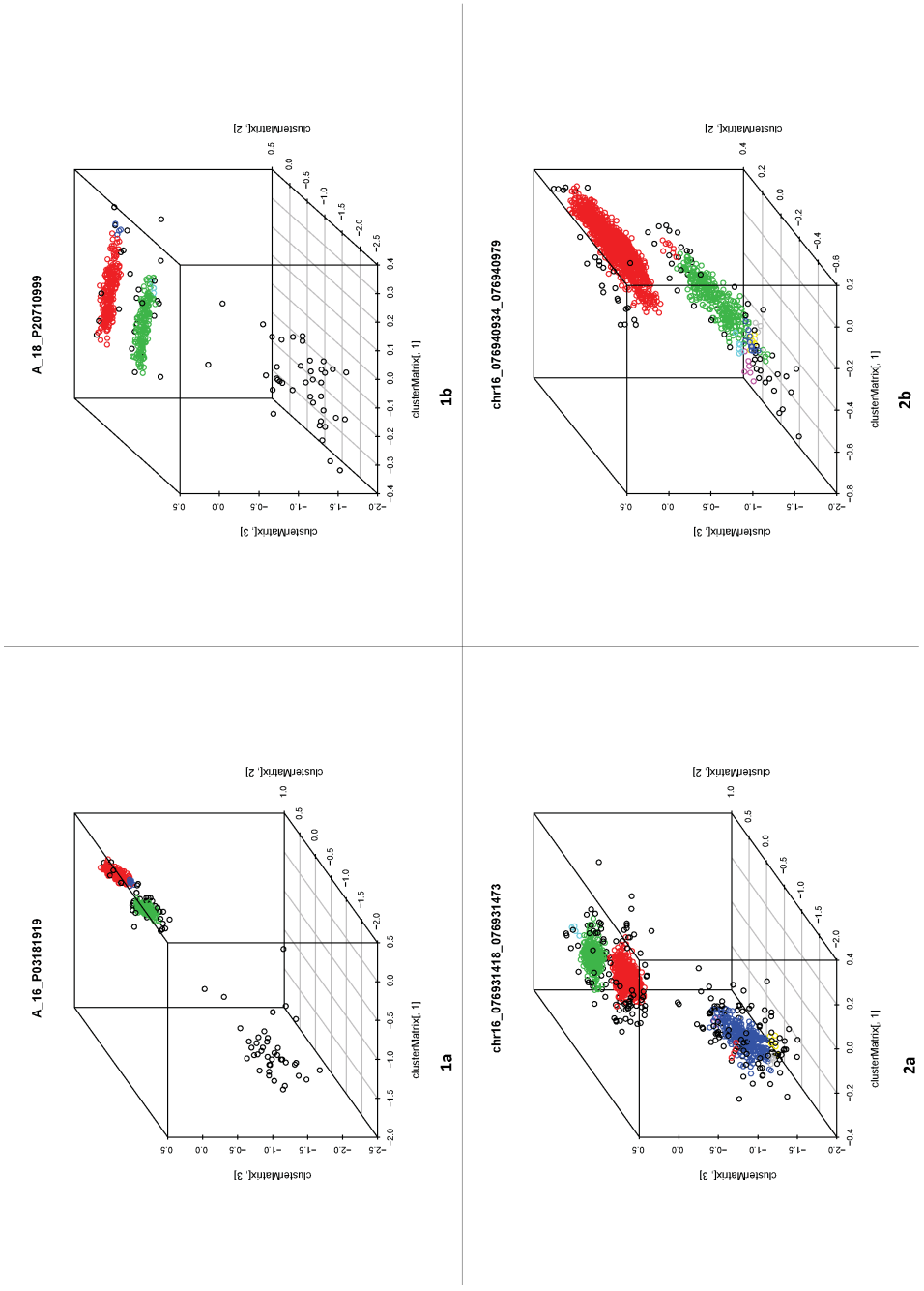


FIGURE 3.13: 3D cluster plot showing the *WWOX* CNV

The datasets used are aCGH BC58 (1a, 1b) and aCGH Sanger105k chips (2a, 2b). Sanger105k chip was the initial high quality CNV discovery platform used by the WTCCC CNV study containing known CNV regions.

The possible advantages of clustering-based CNV genotyping will be that the selection criteria will be based on the clustering quality which might be more accurate. As I previously mentioned single-probe CNV detection is possible in some cases through genotyping platforms, but is more likely to contain more false positives. Some preliminary analysis of genotyping array and then comparison to sequencing results have also shown that CNVs can be detected on a single-probe level to some extent, but more work needs to be done in this area. Further, DBSCSN can also detect outliers from the clusters and thus can provide a list of high quality samples with CNV calls. These samples can then be taken further for validation and association analysis with more confidence. Though current methods are not tailored to detect CNVs at single-probe level, this possibility remains to be explored. If single-probe CNV detection can be implemented then the resolution to detect CNVs from genotyping arrays will be increased many folds and might lead to novel discoveries. DBSCSN has low computational footprint and can be scaled up easily for large datasets. Some of the preliminary results look promising for the application of clustering in CNV detection as well as for CNV genotyping. In addition, the method seems to work well for both SNP genotyping probes and aCGH probes.

At first, I have tried and tested this method for the epilepsy *WWOX* CNV. Both the approaches of DBSCSN algorithm applied to single and multiple probes performed well for detecting this particular CNV. However, the clustering quality was better when a moving window of three SNPs is applied. For *WWOX* the total number of samples with heterozygous samples detected by the clustering approach were less than those detected by cnvHap, but the genotyping concordance was 100%.

I indicate that although the number of samples detected were smaller by DBSCSN, and its capability to remove outliers from different clusters, the ones excluded must be low-quality samples for CNV detection, and thus less reliable. While the k-means algorithm also seemed to work well when applied to *WWOX*, it seemed to perform better for single-probes than multi-probes. Further k-means is not capable of removing outliers, hence we expect the results to contain more false positives.

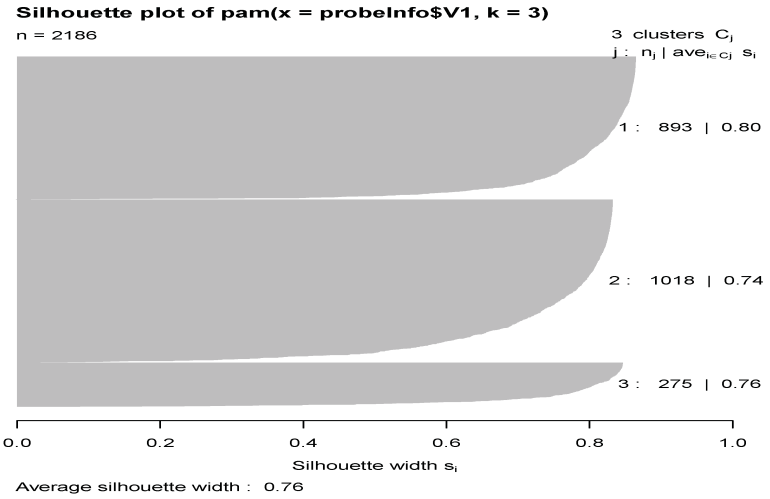


FIGURE 3.14: PAM clustering of *WWOX* gene

Silhoutte plot showing results from PAM clustering of LRR in the *WWOX* gene in BC58 aCGH platform

Chapter 4

CNV Association in Metabonomics

4.1 Introduction

The metabolome can be considered as the complete collection of small molecules in an organism. These molecules or metabolites are usually found inside cells, tissues or organs. There exists is a wide diversity of metabolites. These can be classified based on sizes such as large or small, or as endogenous or exogenous or based on the nature of their existence like transient or theoretical. However, amongst all these classifications, the category based on size is usually found to be the most important from a biological and physiological point of view [217]. In addition, from a functional point-of-view, metabolites play an important role and there are several levels of metabolite regulation. These regulation can include genomic and transcriptomic influence. Such influence can sometimes be multiplicative such that even a single base-pair change in the genome can lead to about 10,000 fold-change in the metabolite level . Also, compared to genome and transcriptome, the metabolome has the most dynamic nature and unlike the other two it has direct relation to systems biology [218]. From a clinical and diagnostic point of view the role of metabolites is also very important. At present about 95% of the diagnostic assays use some kind of small molecules, about 90 % of the diagnostic assays are small molecules themselves and further 30% of the genetic disorders have direct involvement in small molecular regulation. At present, metabolomics is catching up

with large-scale systematic studies like genomics and transcriptomics [219, 220]. This has become increasingly important from a diagnostic and therapeutic point of view. This development has been made possible with the recent advent of high-throughput metabolomic technologies like HPLC, mass spectrometry and NMR spectroscopy.

In the following section, I have described in details the technology behind the NMR spectroscopy technique. Later in this chapter, I will use lipid profiles characterized through NMR-spectroscopy in order to carry out their association with CNVs for diseases like T2D and psoriasis.

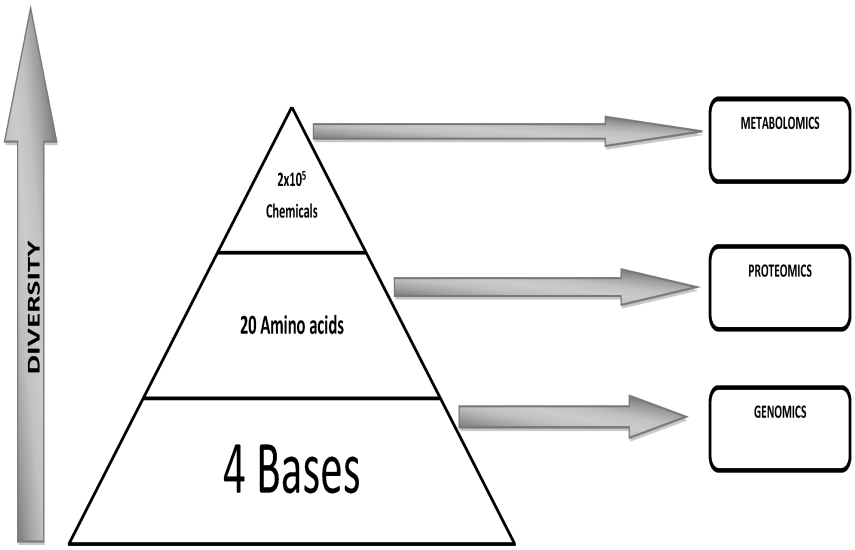


FIGURE 4.1: Diversity in various biological markers

NMR-spectroscopy NMR-spectroscopy is a spectral method to determine the unique structure of compound molecule using the carbon-hydrogen molecular framework. This

approach exploits the charged nature of the atomic nucleus which has its own magnetic field. In general these fields are random but when an external field is applied then they align with or against the applied external magnetic field. Protons which align with the external magnetic field are in lower energy state and those which align against are present in a higher energy state. When a radiation with the same energy difference between these states is applied to the given compound then a flip of spin states takes place and this leads to the emission of electromagnetic radiation. The emission makes the nuclei return to its original state and this process is also referred to as *relaxation*. Different nuclei absorb radiation at different wavelengths for resonance and this unique wavelength indicates the chemical and electronic fingerprint for the given compound.

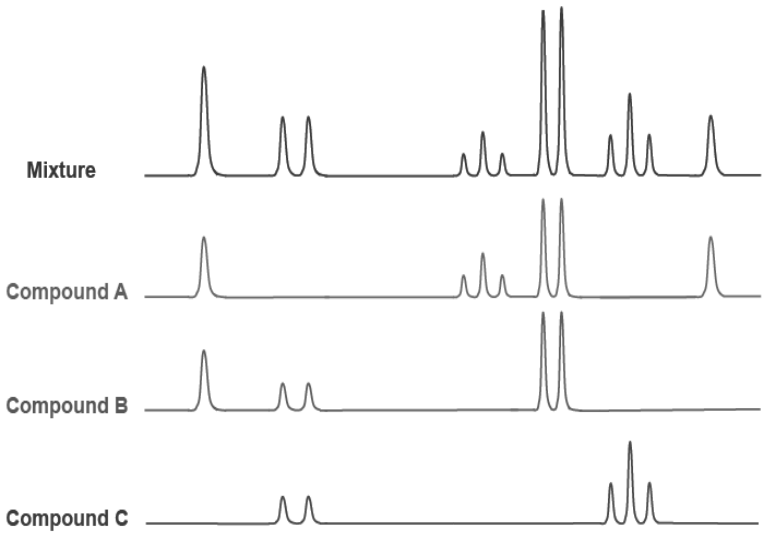


FIGURE 4.2: Deconvolution of NMR spectral data

Not all nuclei can be used for NMR. Usually hydrogen (H) and carbon (C) are the most important NMR-active nuclei. For proton NMR (H ^{1}NMR) the absorption frequency

for a given compound defines its unique chemical signature. Usually the detected signal is plotted as frequency vs intensity. In these plots the position (chemical shift delta) and the pattern (splitting) of the NMR signals give the most important information about the chemical environment of the nuclei. However, interpretation and processing of such signal-data pose its own unique challenges and problems [221]. Some of the problems include calibrating chemical shift, disentangling signal-data for different compounds, removing water signal and making baseline corrections (Figure 4.2). Some of the key informatics challenges in metabolomics are as follows:

- QC steps related to spectra including data integrity and quality
- Data alignment and normalization
- Data reduction and classification
- Determination of the significance of detection and identification of metabolites

Once a list of chemical compounds with their concentrations have been characterised, the next step is to further map these metabolic profiles to biomarkers and disease outcomes.

Metabolomics and disease The relationship between metabolomics and disease-outcome has been well characterized. The metabolome has greater influence on disease phenotype than the genome and transcriptome. Metabolomics has been related to type 1 diabetes (T1D) progression [222], pneumonia [223] HIV [224] and cancer [225].

A successful application of metabolomics has been proposed for diagnosing prostate cancer [14]. In this study a knock-down of the enzyme that generated sarcosine from glycine attenuated prostate cancer while addition of sarcosine dehydrogenase induced invasive cancer. In general, metabolomics can show up to 0.85-90 area under the curve (AUC) receiver operating characteristics (ROC) compared to 0.50-0.60 AUC-ROC for genomics. It can be used for both diagnostic as well as predictive purposes. A good example of this is for kidney rejection where metabolomics has shown up to 90% prediction and prognosis accuracy [226].

In the last decade, thanks to GWAS the genetic contribution to diseases has received increased attention and has led to several successful discoveries for common and rare diseases. In order to get the complete picture for the true mechanistic causes of disease

phenotypes, metabolomics has to be taken into consideration and related to these findings. Further, it has been observed through genomics studies that for complex diseases, there exists a large amount of unexplained heterogeneity in the observed disease phenotypes. For polygenic diseases the heritability (h^2) rarely exceeds 50% [5]. Most GWAS can explain only about 20% of the disease incidence. Thus from these observations it can be postulated that a significant fraction of this heterogeneity in disease phenotypes and missing heritability can perhaps be explained further through metabolomics.

In this chapter, I have explored the role of genomic variant like CNVs in regulating metabolomic lipids in large population cohorts. Previous association studies have mainly focussed on SNP association with metabolites and their heritability [217]. There has been limited investigation on the putative role of CNVs in lipids.

4.1.1 Difference between metabolomics and metabonomics

Metabolomics is defined as the measurement of concentrations and location of all metabolites in a given cell of an organism [227]. Such metabolites are often small and can be transformed from one state to another in the process of metabolism, for example by acting as substrates or by-products of enzymes [228].

Metabonomics is a subset of metabolomics, and is defined as the quantitative measurement of various metabolic responses of a living organism in response to various genetic and environmental modifications. These external influences can be due to physiology of an individual or differences in population, disease or other environmental factors. Metabonomics not only includes intracellular molecules but also includes components of extracellular biofluids [229]. Many of these molecules have been characterized earlier using older technology, but metabonomics specifically relates to the recently introduced high throughput technology like NMR spectroscopy. Such detailed characteristics of metabolite levels obtained through high throughput technologies have been found to correlate well with the pathological state of the organism, leading to improved understanding of the biology and disease prognosis [229]. Metabonomics also extends the idea of integrated analysis with genomics and proteomics, leading to a systems biology like approach and analytical methods applied in it. Some examples of such integrated analysis can include changes in the metabolic pathway induced by ingestion of a toxin,

which in turn can trigger expression of certain other genes [230, 231]. These genes can affect expression of other genes leading change in concentration of other proteins and transcription factors. Such interplay of the genome, proteome and metabolome often requires a systems biology approach which is the hallmark metabonomics [232].

4.2 *TSPAN8* CNV and T2D

Dysregulation of lipid metabolites or *dislipidemia* is a well-known symptom for T2D. In the following sections I have reported a novel analysis approach for relating CNVs, lipid profiles and T2D disease outcome.

A deletion CNV in the 3rd exon of the *TSPAN8* gene was reported to be strongly associated with T2D in the WTCCC CNV study [76]. This CNV was coded as CNVR5583.1, and I have referred to this CNV in using the same notation in the following sections. Here, I have characterized CNVs in the coding region of the *TSPAN8* gene including CNVR5583.1 through the 1000 genomes (1KG) sequencing data and further analysed these CNVs through genotyping arrays in several Finnish and British cohorts. Next, I have tested the association of the 1KG derived CNVs with high-throughput metabolomic and lipid phenotypes.

Through my results I have successfully shown the association with replication of two novel CNVs in the exonic region chr12:69809401-69812861 encompassing the 6th and 7th exon of the *TSPAN8* gene, where a large proportion of metabonomic and lipid phenotypes were found to be significantly associated using univariate (p-value=7.64E-4) and multivariate (p-value=1.33E-06) association approaches (Table 4.5). These CNVs were further were found to be nominally associated with T2D outcome (p-value=3.32e-7 at chr12: 69809401, using raw intensity data for a T2D case-control cohort) (Table 4.12) and were also found to be in LD with CNVR5583.1 with $r^2 > 0.5$ (Table 4.13). Thus, results from my analysis show an important and novel role of T2D associated CNVs in lipid regulation in humans and defines a new phenotype for T2D (Figure 4.4). These results are promising for uncovering the hidden genetic and metabolomic basis for T2D.

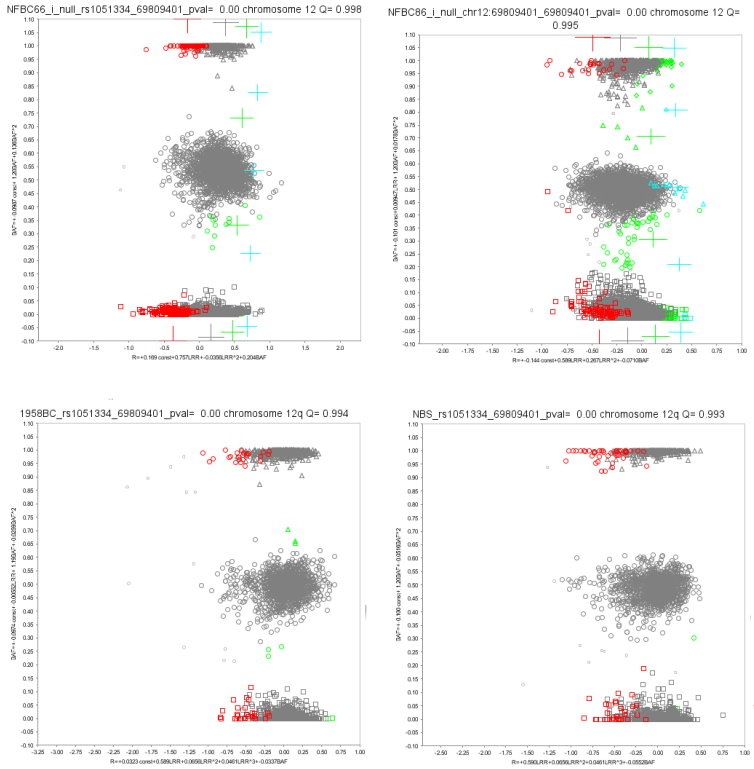


FIGURE 4.3: Cluster plot of the *TSPAN8* in different cohorts

Two-dimensional cluster plots plot for CNV at chr12 69809401 in 4 different cohorts, a) NFBC 1966 b) NFBC 1986 c) BC 1958 d) NBS. X and Y axes in the plot corresponds to LRR and BAF respectively. Each point on the plot represents a sample; shape of the point refers to the number of B-alleles in the SNP genotype; colour represents CN state ($pink = 0$, $red = 1$, $grey = 2$, $green = 3$, $lightblue = 4$).

4.3 Materials and methods

4.3.1 Study cohorts and genotyping platforms

All data for different cohorts reported in my analysis including data from the 1KG project, Northern Finnish Birth Cohort 1966 (NFBC 1966) [233], Northern Finnish Birth Cohort 1986 (NFBC 1986), 1958 British Birth Cohort (BC1958) [181], National Blood Survey (NBS) (http://www.wtccc.org.uk/) and Helsinki Birth Cohort (HBCS) [234] have received prior ethical approval and consent from all study subjects involved in the respective studies. Further details, including aims and methods for the above studies have been reported in earlier publications. In my analysis, I have referred to NFBC 1966 and NFBC 1986 as the primary discovery cohorts and BC1958, NBS and HBCS as control population cohorts. Genotyping details for the discovery and control cohorts are reported in the section 2.13.2.

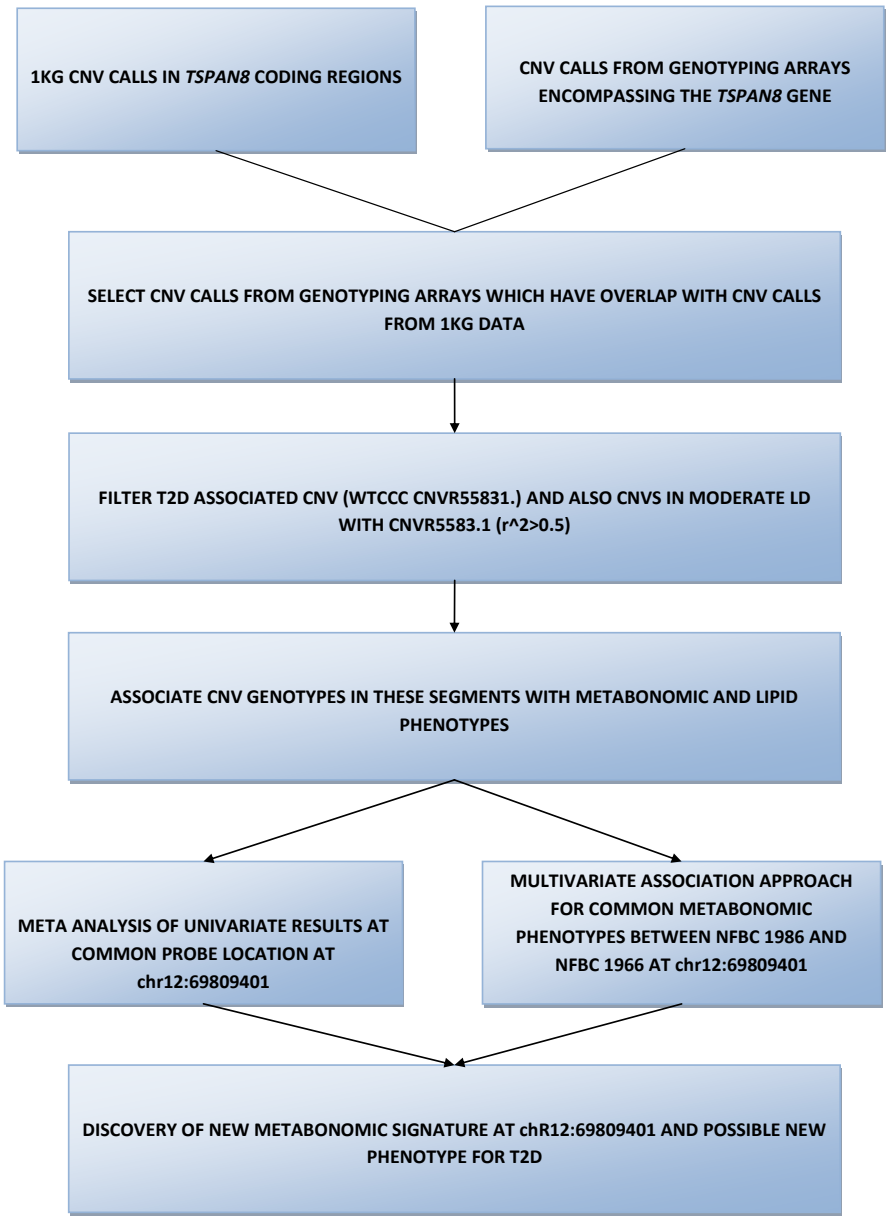


FIGURE 4.4: Analysis pipeline for CNV-metabonomic association analysis reported in this chapter

4.3.2 Metabonomic measurements in NFBC 1966 and NFBC 1986

Lipoprotein measurements and characterization for NFBC 1966 (Number of lipids, N=137) and NFBC 1986 (Number of lipids, N=93) cohorts was carried out using high-throughput H^1NMR spectroscopy. These measurements included lipoprotein subclass, particle and lipid concentrations. These data were acquired from native serum samples at 37.0 C using the NMR platform (Bruker AVANCE III spectrometer) operating at 500 MHz. In addition, a NOESY-presat pulse sequence was applied with mixing time of 10 ms and irradiation field of 25 Hz for suppressing the water peak. The spectra was recorded at 80k data points using 8 transients which was acquired automatically and then calibrated with 90μ pulse. The acquisition time was 2.7s, with relaxation delay of 3.0 s. The spectral data was processed and phase corrected in an automated manner. Then, quantification of particle concentrations and various lipid components in 14 lipoprotein subclasses was obtained using a computationally efficient framework of the regression approach as presented in [235]. The lipoprotein subclasses were next calibrated and cross-validated via HPLC with the following criterias: chylomicrons and extremely large VLDL particles (with particle diameters greater than 75 nm). Five different VLDL subclasses were characterized including: very large VLDL (average particle diameter of 64.0 nm), large VLDL (53.6 nm), medium VLDL (44.5 nm), small VLDL (36.8 nm), and very small VLDL (31.3 nm), IDL (28.6 nm), three LDL subclasses: large LDL (25.5 nm), medium LDL (23.0 nm), and small LDL (18.7), and four HDL subclasses: very large HDL (14.3 nm), large HDL (12.1 nm), medium HDL (10.9 nm), and small HDL (8.7 nm). Apolipoprotein-B and apolipoprotein A-1 were calculated using an extended Friedewald algorithm based on triglycerides, total cholesterol and HDL cholesterol measures quantified from the spectral data. Further details, aims and methods for characterization of various lipids and their sub-categories have been reported in earlier studies [217, 236, 237]. A complete list of metabonomic phenotypes and their base-line characteristics have also been reported earlier [237]. Between NFBC 1966 and NFBC 1986 there were 90 common metabonomic lipids which I subsequently used for replication and meta-analysis

4.3.3 CNV analysis methods

In the following sections, I have described the methods used for the CNV analysis in a sequential manner (Figure 4.4). The main steps involved are identification of CNVs from sequencing data, normalisation and quality control for array data, CNV predictions and segmentation using cnvHap and application of several different association approaches (univariate and multi-variate). I have also touched on a separate validation approach for the association analysis using raw intensity data, multiple testing and LD measurements.

4.3.4 NGS based CNV identification

I first generated CNV-calls using the cnvHitseq algorithm [113] in the *TSPAN8* coding regions from NGS low coverage data for 17 different populations available from the 1KG project website (http://www.1000genomes.org/). cnvHitseq is an integrative probabilistic model for genotyping and discovering CNVs from NGS platforms. This algorithm incorporates various signatures from sequencing data, like read-depth, read-pair and allele frequency information into a single probabilistic model to provide an improved sensitivity for CNV detection. This method works well for even low-coverage data. I have performed the normalisation of the sequence data prior to CNV analysis using cnvHitSeq in the following manner. First, I downloaded the sequencing files in binary alignment format (bam) for the different population from the 1KG website. For each population, I normalized the samples using a sliding window of 25 base-pairs and further corrected for wave-effects and GC content. Next, I ran the cnvHitseq algorithm with a combination of read-depth and split read information, with an initial transition probability of 0.15, 15 EM training iterations and setting the weighted insert option as false in the parameter file of cnvHitseq. Next, using cnvHap I re-characterized the sequence derived CNVs in the discovery and control cohorts using a probe-by-probe CNV-call approach.

4.3.5 CNV predictions using cnvHap

In this section, I have briefly described the CNV analysis for the discovery and control cohorts. CNVs in the *TSPAN8* gene were called in various cohorts using cnvHap. Briefly, this method uses a haplotype hidden Markov model (HMM) based method for discovering and genotyping CNVs from various bead-array based high-throughput

SNP/CNV genotyping platforms, like Illumina and Agilent aCGH arrays. The haplotype HMM of cnvHap uses combined CNV/SNP and population-level information to improve its CNV predictions and has specific emission parameters for different platforms. In my analysis, an Illumina platform specific emission probability was applied and cnvHap was used in its population mode where all samples were used to train the HMM model.

4.3.6 CNV segmentation

As I previously described, cnvHap calculates the most probable linear sequence of copy number states for each sample, which in this case is the CNV segmentation output of cnvHap and reported in the results later.

4.3.7 Expected CNV genotypes

The haplotype HMM of cnvHap calculates the probability of deletion and duplication for each sample at each probe which I refer to as the expected genotypes. For example, if a sample has CNV genotype assigned as 1 at a particular probe location, and this has a deletion probability of 0.8, then the expected CNV genotype is calculated as

$$1*0.8+2*0.2=1.2$$

The expected CNV genotypes were calculated separately for deletions and duplications for every sample at each probe location. I next used these expected genotypes for all subsequent association analyses.

4.3.8 Association analysis

Using the multiphen software [238], I carried out both univariate and multivariate approaches for association of expected CNV genotypes with metabonomic and lipid phenotypes in the discovery and control population cohorts. All phenotypes were quantile normalised for the univariate association, and no transformation was performed for the multivariate analysis (not required for the multiphen approach). In the univariate analysis, p-values for association were calculated using CNV genotypes at each probe as the predictor, and the metabonomic or the lipid phenotypes as outcome. At common

genomic probe locations, meta-analysis for the 90 common lipids in NFBC 1966 and NFBC 1986 was performed using the inverse-variance fixed effect model.

For multivariate analysis, I used the multiphen software with joint-model option. In the joint-model approach, multiphen uses an ordinal probit regression to regress CNV genotype on multiple phenotypes simultaneously, and obtains a joint p-value for all phenotypes together. In addition to this, I further implemented a backward variable selection method in the multiphen package to further determine the best subset of phenotypes associated with the CNV, giving a minimum joint p-value for multiple phenotypes. For both univariate and multivariate analysis, gender and LRR-variance were used as covariates. A new method for correcting batch effects in CNV-association analysis based on LRR principal components was successfully tested and is further described in later sections. Briefly, in this method, LRR PCs are projected out from the response variable in order to reduce inflation and correct for batch effects.

4.3.9 Intensity based validation of CNV association

Several recent studies have shown that CNV phenotypes association can be captured using raw-intensity alone, without calling CNV genotypes [239]. I used a similar LRR based linear model to further validate my CNV phenotype association results, independent of any CNV-calling. To assess such an approach, I first associated CNVR5583.1 and the exonic CNVs described here with T2D disease outcome in a separate T2D cohort (DESIR: Data from Epidemiological Study on the Insulin Resistance syndrome comprising of 4,663 individuals) [240] and further validated CNV lipid association in NFBC 1966 and NFBC 1986.

4.3.10 Multiple testing

Previous studies have reported high degree of correlation prevailing amongst metabonomic and lipid phenotypes. To address this correlation structure, several alternative methods to Bonferroni correction have been proposed like the Sidak-Nyholt correction [241, 242]. Briefly, to estimate the number of effective tests M_{eff} in the presence of

correlation structure in the phenotypes, the following formula is used

$$M_{eff} = 1 + (M - 1)(1 - \frac{var(\lambda_{obs})}{M}) \tag{4.1}$$

where $var(\lambda_{obs})$ is variance of the eigen decomposition of the correlation matrix of metabonomic or lipid phenotypes. The net effective number of test is M_{eff} is then put into the Sidak formula $1 - (1 - \alpha)1/M_{eff}$ to get the significance threshold. On applying the Sidak-Nyholt correction to NFBC 1966 and NFBC 1986 metabonomic phenotypes, the multiple testing nominal p-value threshold obtained was 7.7E-4 and 7.5E-4, respectively.

4.3.11 Linkage disequilibrium

I calculated the LD measure (correlation coefficient r^2) for the total copy number and the number of B alleles. This method is based on the correlation coefficient r^2 between the expected CNV genotypes calculated by cnvHap at a particular probe and the SNP genotype around that CNV region. This indicates how well the CNV is tagged by SNPs around it. In accordance with the conclusions from the WTCCC CNV study, CNVs are expected to be well tagged by SNPs around them.

4.4 Results

4.4.1 CNVs in the *TSPAN8* coding regions based on sequence data

As described in the methods section, I looked into the 1000 genomes NGS low-coverage data for 17 different populations to identify CNVs in the coding regions of the *TSPAN8* gene (Table 4.1, 4.2). I have identified several CNVs overlapping with exonic regions, but I have focussed on CNVR5583.1 and regions showing maximum association with metabonomic lipids. In this respect, the region chr12: 69809401-69812861 was particularly interesting since there were maximal number of lipids found to be associated at a p-value threshold of 0.05. I have shown in a later section, the meta-analysis and validation results for this CNV (Table 4.5).

On sequence validation for the CNVR5583.1 region (which was earlier reported to be a common CNV) I found that all the 17 1KG populations carry this deletion but with varying allele frequencies (Table 4.1). The WTCCC CNV study had earlier reported a MAF of 0.39 for CNVR5583.1. In the British samples of the 1KG NGS data (GBR) CNVs in the CNVR5583.1 region showed a MAF=0.33. On analysing the Finnish samples (FIN) from 1KG data I found that this population carried the deletion in CNVR5583.1 with an allele frequency of 0.38 and this was maximum amongst all other populations (Table 4.1). At the chr12:69809401 locus, I further found that 11/17 populations had a deletion there with maximum allele frequency of 0.1475 (FIN) (Table 4.2). In GBR, the CNV allele frequency was found to be 0.021 (Table 4.2).

Table 4.1 CNV deletion frequency and CNV breakpoint information for different populations from the 1000 genomes NGS which overlap with CNVR5583.1.

TABLE 4.1: CNV deletion frequency at *CNVR5583.1* from 1KG

POPULATION CODE	POPULATION NAME	SAMPLE ID	CHR	START	END	LENGTH	TYPE	AVG CERTAINTY	OVERLAP WITH CNVR5583.1	ALLELE FREQUENCY	SAMPL SIZE
		HG01880	12	69819208	69820101	901	1	0.982	73.94%		
		HG01890	12	69819441	69820101	660	1	0.993	49.60%		
		HG01914	12	69819281	69820101	820	1	0.913	65.76%		
		HG02010	12	69819421	69820101	680	1	0.994	51.62%		
ACB	African Caribbean in Barbados	HG02281	12	69819281	69820101	820	1	0.983	65.76%	0.173076923	52
		HG02339	12	69819461	69820320	859	1	0.974	47.58%		
		HG02343	12	69819481	69820101	620	1	0.997	45.56%		
		HG02420	12	69819501	69820101	600	1	0.98	43.54%		
		HG02427	12	69819501	69820101	600	1	0.985	43.54%		
		NA19700	12	69819360	69819941	581	1	0.986	57.78%		
		NA19701	12	69819281	69819960	679	1	0.983	65.76%		
		NA19704	12	69819341	69820141	800	1	0.97	59.70%		
		NA19711	12	69819061	69820161	1100	1	0.994	79.18%		
		NA19712	12	69819121	69820161	840	1	0.986	61.72%		
ASW	HapMap African ancestry individuals from SW US	NA19713	12	69819301	69820000	699	1	0.978	63.74%	0.22	50
		NA19900	12	69819381	69820161	780	1	0.971	55.66%		
		NA19914	12	69819261	69820161	900	1	0.978	67.78%		
		NA19916	12	69819221	69820161	940	1	0.989	71.82%		
		NA19917	12	69819441	69820161	720	1	0.975	49.60%		
		NA19982	12	69819361	69819861	500	1	0.966	50.51%		
		HG00864	12	69819261	69820061	800	1	0.955	67.78%		
		HG01795	12	69819341	69820220	879	1	0.976	59.70%		
		HG01796	12	69819541	69820300	759	1	0.974	39.49%		
		HG01797	12	69819221	69820301	1080	1	0.965	65.83%		
		HG01799	12	69819580	69820320	740	1	0.963	35.56%		
		HG01805	12	69819181	69820220	1039	1	0.981	72.28%		
		HG01806	12	69819281	69820221	940	1	0.944	65.76%		
		HG01809	12	69819360	69820100	740	1	0.976	57.78%		
		HG01811	12	69819381	69820261	880	1	0.972	55.66%		
		HG01816	12	69819180	69820180	1000	1	0.985	75.20%		
		HG01817	12	69819221	69820180	959	1	0.98	71.82%		
		HG02156	12	69818741	69819541	800	1	0.97	60.51%		
		HG02165	12	69819320	69820140	820	1	0.977	61.82%		
CDX	Chinese Dai in Xishuangbanna, China	HG02166	12	69818321	69819761	1440	1	0.988	56.88%	0.341463415	82
		HG02178	12	69818660	69820061	1401	1	0.975	70.66%		
		HG02185	12	69819120	69820180	1060	1	0.96	76.60%		
		HG02186	12	69819160	69820060	900	1	0.982	77.98%		
		HG02188	12	69818740	69820241	1501	1	0.998	65.96%		
		HG02351	12	69819320	69820221	901	1	0.973	61.82%		
		HG02353	12	69819581	69820221	640	1	0.957	35.45%		
		HG02356	12	69819461	69820301	540	1	0.826	47.58%		
		HG02363	12	69819181	69820240	1059	1	0.98	70.92%		
		HG02364	12	69819261	69820201	1000	1	0.992	67.10%		
		HG02382	12	69819181	69820081	900	1	0.986	75.86%		
		HG02384	12	69819181	69820101	920	1	0.989	75.86%		
		HG02385	12	69819521	69820221	700	1	0.982	41.52%		
		HG02391	12	69819141	69820340	1199	1	0.985	65.97%		
		HG02409	12	69819161	69820221	1060	1	0.986	72.74%		
		NA06989	12	69819064	69820204	1140	1	0.982	76.14%		
CEU	CEPH individuals	NA11843	12	69818943	69820144	1201	1	0.972	82.35%	0.147058824	34
		NA12003	12	69819064	69820103	1039	1	0.988	83.54%		
		NA12058	12	69819624	69820144	520	1	0.952	31.11%		
		NA12342	12	69819104	69820144	1040	1	0.982	79.62%		
		HG00403	12	69819301	69820320	1019	1	0.983	61.92%		
		HG00449	12	69819600	69820321	721	1	0.965	33.54%		
		HG00513	12	69819160	69820001	841	1	0.993	77.98%		
		HG00524	12	69819661	69820300	639	1	0.969	27.37%		
		HG00533	12	69819281	69820101	820	1	0.978	65.76%		
		HG00560	12	69819181	69820100	919	1	0.991	75.86%		
CHS	Han Chinese South	HG00566	12	69819221	69820240	1019	1	0.983	69.77%	0.223880597	67
		HG00596	12	69819440	69820001	561	1	0.959	49.70%		
		HG00610	12	69818741	69820081	1340	1	0.985	73.88%		
		HG00629	12	69819160	69819940	780	1	0.971	77.98%		
		HG00651	12	69819341	69820301	960	1	0.971	59.70%		
		HG00653	12	69819321	69820101	780	1	0.961	61.72%		
		HG00689	12	69819521	69820321	800	1	0.976	41.52%		
		HG00693	12	69819240	69820340	1100	1	0.985	62.91%		
		HG00707	12	69819581	69820120	539	1	0.945	35.45%		
		HG01125	12	69819641	69820221	580	1	0.944	29.39%		
		HG01136	12	69819521	69820220	699	1	0.991	41.52%		
CLM	Colombian in Medellin, Colombia	HG01137	12	69819521	69820181	660	1	0.982	41.52%	0.127272727	55
		HG01359	12	69819521	69820200	679	1	0.972	41.52%		
		HG01462	12	69819521	69820180	659	1	0.977	41.52%		
		HG01489	12	69819541	69820200	659	1	0.964	39.49%		
		HG01491	12	69819501	69820101	600	1	0.981	43.54%		
		HG00173	12	69819364	69820163	799	1	0.983	57.37%		
		HG00179	12	69819244	69819964	720	1	0.97	69.49%		
		HG00180	12	69819023	69819963	940	1	0.984	91.82%		
		HG00182	12	69819763	69820284	521	1	0.962	17.07%		
		HG00183	12	69818964	69820004	1040	1	0.978	91.08%		
		HG00189	12	69819404	69820124	720	1	0.971	53.33%		
		HG00266	12	69819324	69820004	680	1	0.982	61.41%		
		HG00281	12	69819304	69820304	1200	1	0.985	69.00%		
		HG00282	12	69819184	69820184	1000	1	0.993	74.80%		
		HG00284	12	69819103	69820343	1240	1	0.995	66.85%		
		HG00306	12	69819504	69820664	1160	1	0.986	36.90%		
FINS	HapMap Finnish individuals from Finland	HG00310	12	69819504	69820143	639	1	0.989	43.23%	0.37704918	61
		HG00313	12	69819224	69820144	920	1	0.984	71.52%		
		HG00315	12	69819061	69819924	861	1	0.991	86.97%		
		HG00319	12	69819144	69820324	1180	1	0.99	66.78%		
		HG00323	12	69819923	69820523	600	1	0.959	0.91%		
		HG00325	12	69819164	69820324	1160	1	0.987	66.21%		
		HG00326	12	69819224	69820104	880	1	0.991	71.52%		
		HG00327	12	69819104	69820504	1400	1	0.995	59.14%		
		HG00330	12	69819224	69820183	959	1	0.975	71.52%		
		HG00372	12	69819644	69820183	539	1	0.968	29.09%		
		HG00375	12	69819243	69820223	980	1	0.989	69.60%		
		HG00380	12	69818844	69820023	1179	1	0.996	83.97%		
		HG00108	12	69819261	69820023	760	1	0.958	67.58%		
		HG00111	12	69819104	69820184	1080	1	0.979	76.67%		
		HG00116	12	69819364	69820303	939	1	0.949	57.37%		
		HG00125	12	69819684	69820343	659	1	0.954	25.05%		
		HG00131	12	69819384	69820383	999	1	0.978	54.85%		
		HG00133	12	69819303	69820124	821	1	0.975	61.54%		
GBR	British individuals from England and Scotland (GBR)	HG00136	12	69819184	69820464	1280	1	0.983	58.44%	0.304347826	46
		HG00138	12	69819523	69820243	720	1	0.939	41.31%		
		HG00142	12	69819303	69820523	720	1	0.981	61.54%		
		HG00154	12	69819223	69820224	1001	1	0.958	70.83%		
		HG00155	12	69819364	69820343	879	1	0.96	57.37%		
		HG00160	12	69819364	69820244	880	1	0.948	57.37%		
		HG00245	12	69819284	69820044	760	1	0.987	65.45%		
		HG00262	12	69819243	69820224	981	1	0.973	69.60%		
		NA20858	12	69819284	69819924	640	1	0.972	64.65%		
		NA20867	12	69819624	69820303	679	1	0.954	31.11%		
		NA20869	12	69819424	69820124	700	1	0.984	51.33%		
		NA20885	12	69819524	69820164	640	1	0.956	41.21%		
GIH	HapMap Gujarati India individuals from Texas	NA20890	12	69819304	69820343	1039	1	0.972	60.44%	0.148648649	74
		NA20892	12	69819524	69820404	880	1	0.982	41.21%		
		NA20902	12	69819564	69820363	799	1	0.965	37.17%		
		NA20910	12	69819284	69820064	780	1	0.973	65.45%		
		NA21088	12	69819524	69820044	520	1	0.974	41.21%		
		NA21090	12	69818644	69820663	2019	1	0.994	49.03%		
		NA21123	12	69819284	69820103	819	1	0.991	65.45%		
		HG01500	12	69819161	69820001	840	1	0.984	77.88%		
		HG01501	12	69819340	69819880	540	1	0.902	54.55%		
		HG01603	12	69819440	69820000	560	1	0.978	49.70%		
		HG01607	12	69819460	69820061	601	1	0.975	47.68%		
		HG01673	12	69819081	69820261	1180	1	0.976	72.12%		
		HG01675	12	69819260	69820200	940	1	0.985	67.58%		
		HG01676	12	69819501	69820161	660	1	0.98	43.54%		
		HG01679	12	69819760	69820281	521	1	0.948	17.37%		
		HG01682	12	69819240	69819960	720	1	0.947	69.90%		
		HG01695	12	69819240	69820061	821	1	0.956	69.90%		
		HG01697	12	69819461	69820380	919	1	0.952	47.58%		
IBS	Iberian populations in Spain	HG01705	12	69819680	69820181	501	1	0.923	25.45%	0.302631579	76
		HG01707	12	69819541	69820180	639	1	0.976	39.49%		
		HG01710	12	69819681	69820240	559	1	0.953	25.15%		
		HG01746	12	69819381	69820121	740	1	0.973	55.66%		
		HG01747	12	69819481	69820200	719	1	0.987	45.56%		
		HG01770	12	69819341	69820121	780	1	0.922	59.70%		
		HG01777	12	69819401	69820080	679	1	0.965	53.64%		
		HG01781	12	69819361	69819881	520	1	0.979	52.53%		
		HG02223	12	69819341	69820400	1059	1	0.982	55.81%		
		HG02231	12	69819361	69819980	619	1	0.982	57.68%		
		HG02232	12	69819781	69820301	520	1	0.934	15.25%		
		HG02235	12	69819521	69820160	639	1	0.975	41.52%		
		NA18970	12	69819121	69820040	919	1	0.987	81.92%		
		NA18975	12	69815400	69824001	8601	3	0.956	11.51%		
		NA18990	12	69819181	69820041	860	1	0.97	75.86%		
JPT	JPT Japanese individuals	NA19054	12	69819241	69820100	859	1	0.983	69.80%	0.123076923	65
		NA19057	12	69819281	69820160	879	1	0.993	65.76%		
		NA19063	12	69819461	69820040	579	1	0.992	47.58%		
		NA19068	12	69819200	69820041	841	1	0.972	73.94%		
		NA19075	12	69819361	69820040	679	3	0.964	57.68%		
		HG01597	12	69819421	69820280	859	1	0.98	51.62%		
		HG01843	12	69819620	69820200	580	1	0.948	31.52%		
		HG01844	12	69819201	69820061	860	1	0.964	71.84%		
		HG01848	12	69819601	69820161	560	1	0.968	33.43%		
		HG01862	12	69819521	69820120	599	1	0.969	41.52%		
KHV	Kinh in Ho Chi minh City, Vietnam	HG01865	12	69819060	69820180	1120	1	0.986	77.80%	0.144736842	76
		HG01866	12	69819421	69820161	740	1	0.96	51.62%		
		HG02023	12	69819260	69819881	621	1	0.955	62.73%		
		HG02046	12	69818741	69819340	599	1	0.89	40.20%		
		HG02048	12	69819421	69820001	580	1	0.982	51.62%		
		HG02078	12	69819161	69820063	900	1	0.976	77.88%		
		NA19312	12	69818961	69819681	720	3	1	72.73%		
		NA19377	12	69819501	69820001	500	1	0.984	43.54%		

Table 4.1 continued

LWK	Luhya individuals	NA19380	12	69819521	69820061	540	1	0.955	41.52%	0.173913043	69
		NA19384	12	69819301	69820001	700	1	0.94	63.74%		
		NA19390	12	69819441	69820001	560	1	0.95	49.60%		
		NA19391	12	69819181	69820001	820	3	0.966	75.86%		
		NA19434	12	69819120	69819681	561	3	0.967	56.67%		
		NA19436	12	69818700	69819681	981	3	0.965	74.65%		
		NA19443	12	69818741	69820001	1260	3	0.992	78.57%		
		NA19457	12	69819541	69820220	679	1	0.95	39.49%		
		NA19463	12	69819141	69819661	520	3	0.893	52.53%		
		NA19467	12	69819041	69820001	960	3	0.971	90.00%		
MXL	HapMap Mexican individuals from LA California	NA19648	12	69819381	69820061	680	1	0.981	55.66%	0.344827586	58
		NA19651	12	69819421	69820201	780	1	0.966	51.62%		
		NA19652	12	69819200	69820061	861	1	0.987	71.94%		
		NA19655	12	69819201	69820441	1240	1	0.974	58.95%		
		NA19657	12	69819341	69820061	720	1	0.981	59.70%		
		NA19660	12	69819240	69820061	821	1	0.968	69.90%		
		NA19661	12	69819261	69820061	800	1	0.984	67.78%		
		NA19669	12	69819341	69820180	839	1	0.977	59.70%		
		NA19670	12	69819321	69820001	680	1	0.98	61.72%		
		NA19675	12	69819281	69820181	900	1	0.973	65.76%		
		NA19682	12	69819321	69819940	619	1	0.961	61.72%		
		NA19685	12	69819161	69820201	1040	1	0.985	74.13%		
		NA19716	12	69819341	69820061	720	1	0.98	59.70%		
		NA19719	12	69819161	69819900	739	1	0.99	74.65%		
		NA19720	12	69819220	69820180	960	1	0.97	71.92%		
		NA19726	12	69819241	69820201	960	1	0.971	69.80%		
		NA19777	12	69819241	69820060	819	1	0.971	69.80%		
		NA19779	12	69819281	69820061	780	1	0.992	65.76%		
		NA19782	12	69819260	69820180	920	1	0.979	67.88%		
		NA19788	12	69819321	69820061	740	1	0.977	61.72%		
PUR	Puerto Rican in Puerto Rico	HG00637	12	69819220	69820021	801	1	0.98	71.92%	0.327868852	61
		HG00638	12	69819261	69819960	699	1	0.948	67.78%		
		HG00734	12	69819141	69820120	979	1	0.986	79.90%		
		HG00739	12	69819320	69820041	721	1	0.97	61.82%		
		HG00740	12	69819501	69820161	660	1	0.953	43.54%		
		HG01051	12	69819181	69820201	1020	1	0.989	73.63%		
		HG01061	12	69819181	69819980	799	1	0.994	75.86%		
		HG01069	12	69819220	69819980	760	1	0.976	71.92%		
		HG01075	12	69819301	69820160	859	1	0.972	63.74%		
		HG01080	12	69819420	69820100	680	1	0.974	51.72%		
		HG01102	12	69819241	69820061	820	1	0.975	69.80%		
		HG01111	12	69819401	69820301	700	1	0.975	53.64%		
		HG01170	12	69819660	69820301	641	1	0.973	27.47%		
		HG01176	12	69819021	69820161	1140	1	0.951	79.91%		
		HG01183	12	69819121	69820200	1079	1	0.991	75.16%		
		HG01187	12	69819241	69819780	539	1	0.997	54.44%		
		HG01197	12	69819341	69820180	939	1	0.976	59.70%		
		HG01241	12	69819480	69820041	561	1	0.976	45.66%		
		HG01242	12	69819420	69820180	760	1	0.974	51.72%		
		HG01248	12	69819081	69819900	819	1	0.978	82.73%		
TSI	Toscan individuals	NA20802	12	69819521	69820120	599	1	0.973	41.52%	0.1	10
YRI	Yoruba individuals	NA18498	12	69819920	69821041	1121	1	0.942	1.07%	0.111111111	45
		NA18516	12	69819560	69820081	521	1	0.927	37.58%		
		NA19197	12	69818760	69819521	761	3	0.997	58.48%		
		NA19223	12	69819000	69819521	521	3	0.96	52.63%		
		NA19248	12	69819560	69820141	581	1	0.963	37.58%		

4.4.2 CNV segmentation

Using the CNV segmentation results from cnvHap and PennCNV (only for NFBC 1966), I was successful in replicating the existence of CNVR5583.1 (Figure 4.5, 4.6, Table 4.1, 4.2). I found a novel CNV containing region overlapping with the 6th and 7th exon of *TSPAN8* located at chr12: 69809401-69812861 region. All the CNV segments outputs from cnvHap and PennCNV had complete overlap with these two regions.

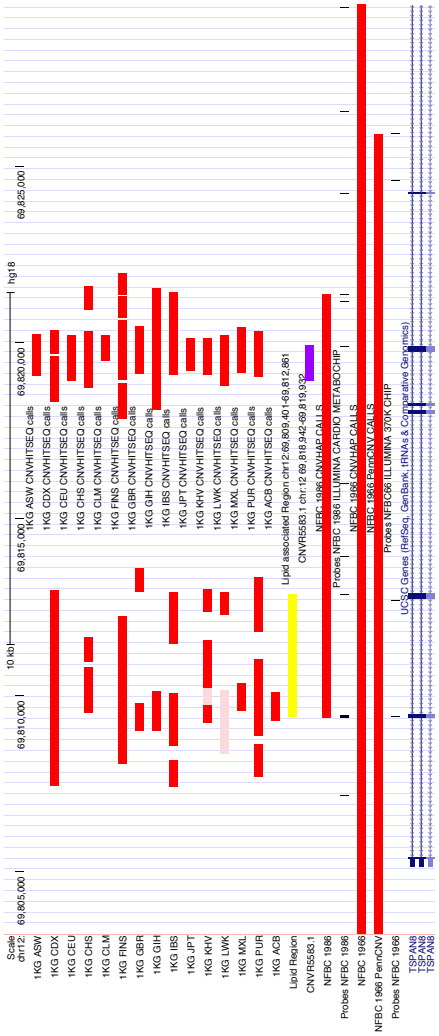


FIGURE 4.5: UCSC genome browser plot

UCSC genome browser plot showing CNV calls (deletion only) in the *TSPAN8* gene for different populations from the 1KG data, including Finnish population (FIN), NFBC 1966, NFBC 1986 and. Probe locations for the various genotyping platforms are indicated by vertical black lines. The region associated with lipids chr12 69809401-69812861 is highlighted in yellow and CNVR5583.1 is indicated by violet.

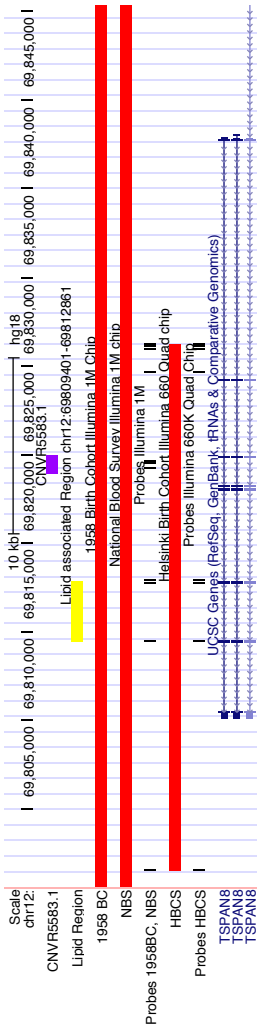


FIGURE 4.6: UCSC genome browser plot for control population

UCSC genome-browser plot for CNVs in the control populations - 1958 birth cohort (BC 1958), National blood service (NBS) and Helsinki Birth Cohort (HBCS).

Table 4.2 Deletion allele-frequency and CNV breakpoint details of all CNVs encompassing the genomic location chr12 69809401, in 11 different populations from the 1000 genomes project NGS data.

Table 4.2: Deletion frequency at chr12 69809401 in 1KG

POPULATION CODE	POPULATION NAME	SAMPLE ID	CHR	START	END	LENGTH	TYPE	CERTAINTY	DELETION FREQUENCY	SAMLES SIZE
ACB	African Caribbean in Barbados	HG02314		69809281	69810061	780	1	0.963	0.019	52
CDX	Chinese Dai in Xishuangbanna, China	HG01796		69809040	69809540	500	1	0.933	0.073	82
		HG02155		69809160	69810481	1321	1	0.905		
		HG02164		69809221	69810580	1359	1	0.984		
		HG02184		69808041	69809981	1940	1	0.961		
		HG02185		69809261	69810681	1420	1	0.978		
		HG02186		69808961	69809541	580	1	0.981		
CLM	Colombian in Medellin, Colombia	HG01377		69809180	69809681	501	1	0.94	0.036	55
		HG01384		69809361	69809921	560	1	0.972		
FIN	HapMap Finnish individuals from Finland	HG00171		69808363	69809444	1081	1	0.986	0.148	61
		HG00180		69808824	69810064	1240	1	0.969		
		HG00182		69808543	69809483	940	1	0.991		
		HG00186		69808544	69809543	999	1	0.99		
		HG00187		69808804	69810044	1240	1	0.975		
		HG00308		69807383	69810544	3161	1	0.997		
		HG00367		69808183	69810064	1881	1	0.988		
		HG00372	12	69808124	69810083	1959	1	0.986		
		HG00373		69808803	69810283	1480	1	0.972		
GBR	British individuals from England and Scotland	HG00242		69808963	69809764	801	1	0.964	0.022	46
GIH	HapMap Gujarati India individuals from Texas	NA21122		69809003	69810123	1120	1	0.984	0.014	74
IBS	Iberian populations in Spain	HG01699		69809200	69810160	960	1	0.973	0.079	76
		HG01709		69809261	69809780	519	1	0.955		
		HG01746		69808841	69809940	1099	1	0.977		
		HG01775		69809201	69810000	799	1	0.95		
		HG01783		69809200	69809701	501	1	0.949		
		HG02235		69808681	69809401	720	1	0.968		
KHV	Kinh in Ho Chi minh City, Vietnam	HG01864		69809101	69809620	519	1	0.969	0.026	76
		HG02046		69809221	69810001	780	1	0.985		
LWK	Luhya individuals	NA19313		69808360	69810140	1780	0	0.966	0.029	69
		NA19313		69808360	69810140	1780	0	0.966		
PUR	Puerto Rican in Puerto Rico	HG00637		69808861	69809641	780	1	0.928	0.033	61
		HG01066		69809360	69809881	521	1	0.821		
YRI	Yoruba individuals	NA18507		69809340	69809960	620	1	0.853	0.022	45

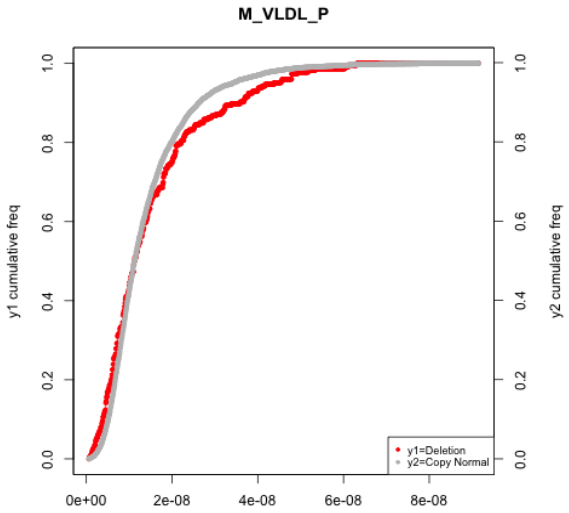


FIGURE 4.7: Cumulative plot for M_VLDL_P in NFBC 1966

Samples having normal copy number are marked in grey and samples with deletion are marked in red

4.4.3 CNV replication and allele frequency

In the above scenario of different CNV breakpoints and difference in probe-coverage in different cohorts, I decided that it was best to leverage cnvHaps probe-by-probe calculation of expected CNV-genotypes (see methods). Briefly, unlike other CNV-calling approaches, cnvHap does CNV-calling at each probe independently to derive the posterior probability, which I refer to as the expected CNV genotype at a particular probe location. For the probes located in the *TSPAN8* coding regions in NFBC 1966 and NFBC 1986, there was no consensus in the probe locations, except for the two probes located at chr12: 69809401 and chr12: 69835363. I could only meta-analyse results for these two probes, and report replication results for other loci by considering the probe located nearest to the region of interest.

Looking into the allele frequency results based on the expected CNV genotypes I found all discovery and control cohorts to carry deletions at chr12: 69809401. Allele frequency results from all the cohorts indicates that the CNV at this location is relatively rare with $MAF < 0.05$. Amongst all cohorts NFBC 1966 had the maximum MAF of

0.041 whereas in Finnish control cohort, HBCS the allele frequency was < 0.01 (Table 4.3). One possible reason for the large difference in allele frequency in Finnish cohorts could be that the Northern Finnish Birth cohorts are geographically different from the Helsinki Birth Cohort, resulting in a different genetic constitution. In the British control populations (BC1958 and NBS) I found an allele frequency around 0.02. However, we should keep in mind the possibility of differences in genotyping platforms and different CNV-calling parameters affecting the allele frequency in the CNV-calling.

Allele frequency of CNVs Allele frequency of CNV calls (deletion only) from the 1KG NGS data and CNV genotype calls (deletion only) in several cohorts genotyped through bead-array platforms. 1a and 1b show the allele frequency of CNVR5583.1 and lipid associated locus chr12 69809401 characterized respectively in the 1KG data. 2a and 2b show the allele frequency for the above regions determined through bead-array based platforms.

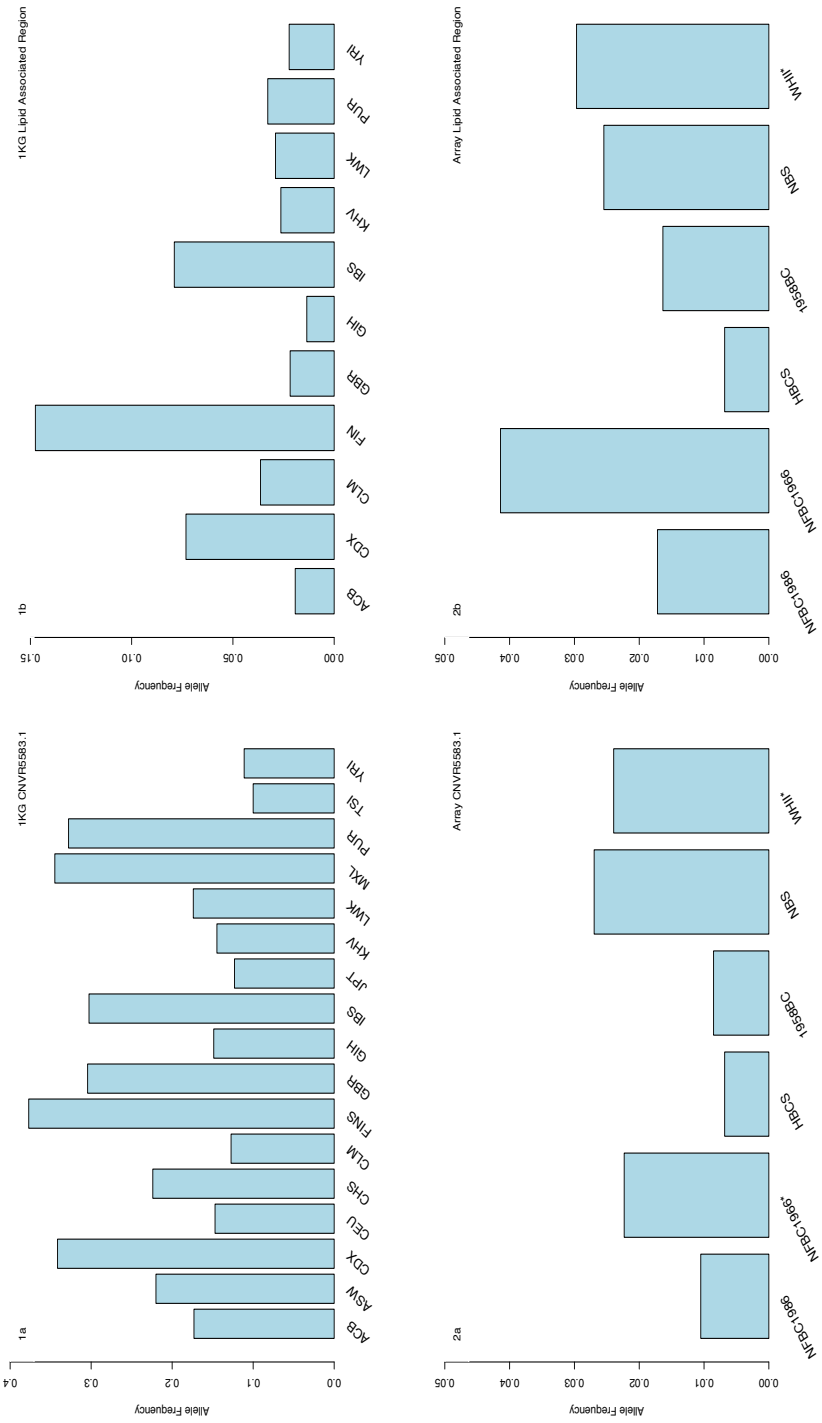


FIGURE 4.8: Allele frequency determined through genotyping array and NGS based platforms

Table 4.3 Deletion allele-frequency of CNV at chr12 69809401 discovered through genotyping-platforms in various cohorts including discovery cohorts (NFBC 1966 and NFBC 1986) and control cohorts (HBCS, BC 1958 and NBS).

TABLE 4.3: Allele frequency of deletion at chr12 69809401

POPULATION	COHORT NAME	GENOTYPING CHIP	SNP NAME	PROBE LOCATION	DELETION FREQUENCY
Finnish	Northern Finnish Birth Cohort 1986	Cardio Metabo chip	chr12:69809401	69809401	0.017
	Northern Finnish Birth Cohort 1966	Illumina 370K	rs1051334		0.041
	Helsinki Birth Cohort	Illumina 610K Quad			0.007
British	1958 British Birth Cohort	Illumina 1M			0.016
	National Blood survey	Illumina 1M			0.025

Table 4.4 CNV deletion frequency for various cohorts characterized through genotyping arrays for probes within or nearest to CNVR5583.1.

TABLE 4.4: CNV frequency from genotyping arrays at *CNVR5583.1*

POPULATION	COHORT NAME	GENOTYPING CHIP	SNP NAME	PROBE LOCATION	DELETION FREQUENCY	SAMPLE SIZE
Finnish	Northern Finnish Birth Cohort 1986	Cardio Metabo chip	hg18_12_69819889	69819889	0.0105	5997
	Northern Finnish Birth Cohort 1966	Illumina 370K	rs1798089	69824577	0.0223	5555
	Helsinki Birth Cohort	Illumina 610K Quad	rs1798089	69824577	0.0068	1614
British	1958 British Birth Cohort	Illumina 1M	rs1798089	69824577	0.0085	2814
	National Blood survey	Illumina 1M	rs2971392	69819189	0.0270	2671

4.4.4 Univariate analysis

4.4.4.1 CNV association with metabonomic and lipid phenotypes in NFBC 1966 and NFBC 1986

Association results for the CNV located at chr12: 69809401 encompassing the 7th exon (LD r^2=0.636 with CNVR5583.1) with high-throughput metabonomic and lipid phenotypes showed a very high proportion of these phenotypes (8/90 (8.8 %) for NFBC 1966 and 56/90 (62.2%) for NFBC 1986 associated at a nominal p-value threshold of < 0.05. On meta-analysing association p-values for 90 common metabonomic lipids in NFBC 1986 and NFBC 1966, I found 11/90 (12.2 %) of the metabonomic lipids to be significantly associated with minimum meta p-value of 7.64E-04 for the phenotype free cholesterol in medium VLDL (M_VLDL_L), after applying a multiple-testing correction threshold p-value of 7.7E-4. On associating CNVs within or nearest to the CNVR5583.1 region in NFBC 1986, NFBC 1966, I found 26/90 (28.8 %) (chr12: 69819889) and 41/90 (45.5 %) (chr12: 69824577) different metabonomic or lipid phenotypes to be nominally associated at a p-value threshold of 0.05. Univariate results for CNVR5583.1 and CNV at the 6th exon are reported in Table 4.5 and Table 4.6.

Table 4.5 This table shows the top CNV-associated metabonomic lipids after meta-analysis of NFBC 1966 and NFBC 1986 Values for probes located in common genomic locations in the coding-regions of the *TSPAN8* gene. This includes the reported lipid-associated region- chr12 69809401. In the association model, the top ten LRR PCs were projected out of the model in order to correct for batch-effects and achieve an inflation-factor close to 1. Only the top two lipids (indicated) at chr12 6980941 survived the multiple-testing threshold (corrected) of 8.05E-4 (Sidak) and 7.8E-4 (Bonferroni) for common metabonomic lipids in NFBC 1966 and NFBC 1986.

TABLE 4.5: Meta-analysis results at chr12 69809401

Meta analysed probe name	Chromosome	Meta analysis probe location	Lipid phenotype	Beta	Meta P-value	Direction	No of samples	Batch-effect correction method
rs1051334 (NFBC 1966); chr12:69809401 (NFBC 1986)	12	69809401	**M_VLDL_L_quantile ****	-0.19	7.64E-04	--	7.84E+03	PC projection
			M_VLDL_PL_quantile **	-0.19	7.65E-04	--		
			M_VLDL_C_quantile	-0.181	1.20E-03	--		
			M_VLDL_CE_quantile	-0.179	1.22E-03	--		
			M_VLDL_FC_quantile	-0.179	1.52E-03	--		
			M_VLDL_TG_quantile	-0.176	1.79E-03	--		
			VLDL_TG_quantile	-0.165	3.39E-03	--		
			IDL_C_eFR_quantile	-0.152	4.43E-03	--		
			XS_VLDL_L_quantile	-0.151	5.17E-03	--		
			XS_VLDL_PL_quantile	-0.151	5.22E-03	--		
rs1614565 (NFBC 1966); chr12:69835363 (NFBC 1986)	12	69835363	Serum_C_quantile	-0.132	3.53E-02	--	7.84E+03	PC projection
			ApoA1_quantile	-0.144	3.68E-02	--		
			HDL_D_quantile	-0.138	3.68E-02	--		
			HDL_C_quantile	-0.136	4.13E-02	--		
			L_HDL_L_quantile	-0.128	5.32E-02	--		

**** Lipids surviving multiple testing threshold**

Table 4.6 Univariate results obtained after meta-analysis of Pvalues at chr12 69809401, derived from association of cnvHap expected CNV calls and lipids in NFBC 1966 and NFBC 1986. Results obtained from using multiphen joint model and multiphen joint model with variable selection, is also reported. Unlike Table 4.5, the batch-effect correction method applied in these results was LRR variance as covariate.

TABLE 4.6: Univariate meta-analysis results at chr12:69809401 using LRR variance as covariate

Association Method	Location	Meta analysis probe	Phenotype(s)	Beta	Meta Pvalue	Batch-effect correction method
Genotype association using univariate assoicaiton model	12_69809401	rs10513334 (NFBC 1966); chr12:69809401 (NFBC 1986)	M_VLDL_L_quantile	-0.19	7.64E-04	LRR variance as covariate
			M_VLDL_PL_quantile	-0.19	7.65E-04	
			M_VLDL_C_quantile	-0.181	1.20E-03	
			M_VLDL_CE_quantile	-0.179	1.22E-03	
			M_VLDL_FC_quantile	-0.179	1.52E-03	
			M_VLDL_TG_quantile	-0.176	1.79E-03	
			VLDL_TG_quantile	-0.165	3.39E-03	
			IDL_C_eFR_quantile	-0.152	4.43E-03	
			XS_VLDL_L_quantile	-0.151	5.17E-03	
			XS_VLDL_PL_quantile	-0.151	5.22E-03	
Genotype association with multipehn joint model using all phenotypes	12_69809401	rs10513334 (NFBC 1966); chr12:69809401 (NFBC 1986)	Joint model for all pheotypes*	NA	2.44E-10	LRR variance as covariate
Genotype associaiton with multipheni joint model with variable selection	12_69809401	rs10513334 (NFBC 1966); chr12:69809401 (NFBC 1986)	Joint model for pheotypes after variable selection	NA	5.70E-13	LRR variance as covariate

Table 4.7 Univariate and meta analysis results for NFBC 1966 and NFBC 1986 for all 90 metabonomic lipids at chr12:69809401.

TABLE 4.7: Univariate meta-analysis results for all metabolomic lipids at chr12:69809401

		NFBC 1966		NFBC 1986		Meta analysis results	
Location	Lipid	beta	pv	beta	pv	beta	pv
	Alb	-0.102	0.0766	0.0249	0.77	-0.0623	1.93E-01
	ApoA1	-0.0187	0.729	0.04	0.636	-0.00174	9.69E-01
	ApoB	-0.0435	0.405	-0.0296	0.728	-0.0397	3.72E-01
	HDL2_C	0.0605	0.27	0.0726	0.356	0.0645	1.52E-01
	HDL3_C	-0.00982	0.865	-0.031	0.7	-0.017	7.17E-01
	HDL_C	0.0291	0.586	0.0537	0.505	0.0367	4.10E-01
	HDL_D	0.00843	0.871	0.102	0.212	0.0353	4.21E-01
	IDL_C_eFR	-0.0269	0.613	-0.0775	0.334	-0.0424	3.40E-01
	IDL_C	0.0296	0.57	-0.0463	0.569	0.00755	8.63E-01
	IDL_FC	0.0523	0.329	-0.0413	0.611	0.0239	5.93E-01
	IDL_L	0.000513	0.992	-0.0871	0.281	-0.0254	5.63E-01
	IDL_PL	-0.0108	0.838	-0.117	0.151	-0.0425	3.40E-01
	IDL_TG	-0.059	0.286	-0.131	0.1	-0.0824	6.96E-02
	LDL_C_eFR	0.0655	0.186	0.00592	0.942	0.0492	2.44E-01
	LDL_C	0.0106	0.832	-0.022	0.781	0.00128	9.76E-01
	LDL_D	0.0485	0.336	-0.0833	0.384	0.0198	6.56E-01
	L_HDL_CE	0.0319	0.56	0.112	0.147	0.0587	1.89E-01
	L_HDL_C	0.0374	0.493	0.112	0.148	0.0621	1.64E-01
	L_HDL_FC	0.0441	0.414	0.123	0.121	0.0691	1.22E-01
	L_HDL_L	0.0232	0.67	0.0732	0.339	0.04	3.67E-01
	L_HDL_PL	0.0218	0.69	0.043	0.57	0.0291	5.11E-01
	L_LDL_CE	0.0204	0.686	-0.0233	0.771	0.00797	8.52E-01
	L_LDL_C	0.022	0.664	-0.0268	0.738	0.00808	8.50E-01
	L_LDL_FC	0.0281	0.582	-0.0491	0.542	0.00592	8.91E-01
	L_LDL_L	0.00594	0.907	-0.0472	0.554	-0.0093	8.28E-01
	L_LDL_PL	-0.0215	0.673	-0.104	0.2	-0.0449	2.98E-01
	L_VLDL_CE	-0.0148	0.8	-0.036	0.646	-0.0224	6.33E-01
	L_VLDL_C	-0.00593	0.919	-0.0532	0.51	-0.022	6.40E-01
	L_VLDL_FC	-0.00223	0.969	-0.0662	0.429	-0.0225	6.33E-01
	L_VLDL_L	-0.0512	0.375	-0.119	0.146	-0.0737	1.18E-01
	L_VLDL_PL	-0.096	0.0985	-0.129	0.114	-0.107	2.38E-02
	L_VLDL_TG	-0.0674	0.244	-0.128	0.115	-0.0878	6.26E-02
	M_HDL_CE	0.0131	0.829	-0.0414	0.588	-0.00795	8.67E-01
	M_HDL_C	0.00958	0.874	-0.0483	0.531	-0.0124	7.94E-01
	M_HDL_FC	0.0106	0.855	-0.0726	0.36	-0.0184	6.94E-01
	M_HDL_L	-0.0548	0.351	-0.12	0.131	-0.0778	9.94E-02
	M_HDL_PL	-0.0823	0.153	-0.162	0.043	-0.11	1.92E-02

Table 4.7 continued

12_69809401	M_LDL_CE	0.0105	0.834	-0.0108	0.892	0.00441	9.17E-01
	M_LDL_C	0.00495	0.921	-0.0154	0.845	-0.000912	9.83E-01
	M_LDL_L	-0.0109	0.827	-0.0364	0.645	-0.0182	6.66E-01
	M_LDL_PL	-0.0433	0.382	-0.0887	0.268	-0.0558	1.85E-01
	M_VLDL_CE	-0.0334	0.54	-0.0802	0.343	-0.0471	3.04E-01
	M_VLDL_C	-0.0351	0.522	-0.108	0.205	-0.0564	2.22E-01
	M_VLDL_FC	-0.0333	0.55	-0.133	0.123	-0.0626	1.80E-01
	M_VLDL_L	-0.086	0.123	-0.164	0.0559	-0.109	1.95E-02
	M_VLDL_PL	-0.0901	0.103	-0.17	0.049	-0.113	1.49E-02
	M_VLDL_TG	-0.12	0.0335	-0.167	0.0467	-0.135	4.06E-03
	Serum_C	0.0088	0.856	-0.0152	0.848	0.00228	9.56E-01
	Serum_TG	-0.0773	0.165	-0.145	0.0891	-0.0975	3.64E-02
	S_HDL_L	-0.0546	0.301	-0.171	0.0426	-0.0873	5.08E-02
	S_HDL_TG	-0.238	1.99E-05	-0.194	0.0181	-0.224	1.19E-06
	S_LDL_C	-0.00152	0.976	-0.018	0.819	-0.00625	8.82E-01
	S_LDL_L	-0.0308	0.534	-0.0534	0.498	-0.0372	3.75E-01
	S_VLDL_C	0.0695	0.176	0.0341	0.677	0.0595	1.72E-01
	S_VLDL_FC	0.0526	0.316	0.00327	0.969	0.0385	3.85E-01
	S_VLDL_L	0.00537	0.919	-0.0408	0.622	-0.00806	8.57E-01
	S_VLDL_PL	0.0233	0.661	-0.0317	0.7	0.00715	8.73E-01
	S_VLDL_TG	-0.0616	0.268	-0.104	0.208	-0.0748	1.05E-01
	VLDL_D	-0.045	0.406	-0.118	0.187	-0.0646	1.63E-01
	VLDL_TG_eFR	-0.0818	0.147	-0.155	0.0706	-0.104	2.74E-02
	VLDL_TG	-0.0882	0.117	-0.147	0.0816	-0.106	2.32E-02
	XL_HDL_CE	0.0262	0.632	0.134	0.107	0.0589	1.98E-01
	XL_HDL_C	0.0134	0.806	0.124	0.135	0.0469	3.04E-01
	XL_HDL_FC	-0.00124	0.982	0.113	0.178	0.0321	4.78E-01
	XL_HDL_L	0.00736	0.89	0.119	0.153	0.0397	3.76E-01
	XL_HDL_PL	0.00482	0.927	0.115	0.162	0.0371	4.05E-01
	XL_HDL_TG	-0.12	0.0384	-0.0258	0.749	-0.0877	6.17E-02
	XL_VLDL_L	-0.00385	0.947	-0.033	0.677	-0.0141	7.64E-01
	XL_VLDL_PL	-0.0457	0.435	-0.0509	0.526	-0.0475	3.15E-01
	XL_VLDL_TG	-0.00975	0.867	-0.0517	0.512	-0.0246	6.00E-01
	XS_VLDL_L	-0.0245	0.652	-0.102	0.208	-0.0485	2.82E-01
	XS_VLDL_PL	-0.00666	0.903	-0.091	0.263	-0.0328	4.69E-01
	XS_VLDL_TG	-0.061	0.281	-0.112	0.157	-0.0783	8.95E-02
	XXL_VLDL_L	-0.123	0.0393	-0.0555	0.473	-0.0978	3.85E-02
	XXL_VLDL_PL	-0.143	0.0192	-0.073	0.316	-0.114	1.48E-02
	XXL_VLDL_TG	-0.12	0.0415	-0.0623	0.435	-0.0997	3.54E-02

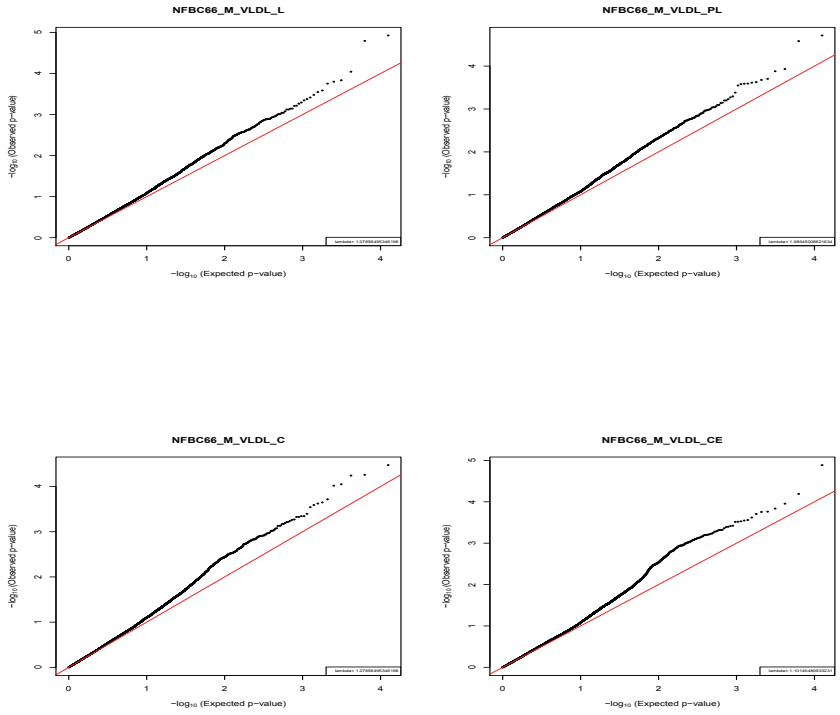


FIGURE 4.9: QQ plot for the top associated lipids

This figure shows the genomic inflation factors for the top four lipids in NFBC 1966.

Table 4.8 Chromosome 12 wide inflation factors observed for the top four CNV-associated lipids (univariate meta-analysis, Table 4.2) in NFBC 1966 and NFBC 1986. Inflation factors using two different methods for correcting for batch effects PC projection approach and LRR variance as covariate, is reported.

TABLE 4.8: Inflation factor for top four lipids

	PC projection		LRR variance as Covariate	
Phenotype	**Inflation Factor NFBC 1966**	**Inflation Factor NFBC 1986**	**Inflation Factor NFBC 1966**	**Inflation Factor NFBC 1986**
M_VLDL_FC_quantile	1.076	1.049	1.254	1.313
M_VLDL_L_quantile	1.076	1.018	1.173	1.36
M_VLDL_PL_quantile	1.086	1.004	1.194	1.33
M_VLDL_C_quantile	1.076	1.101	1.199	1.199

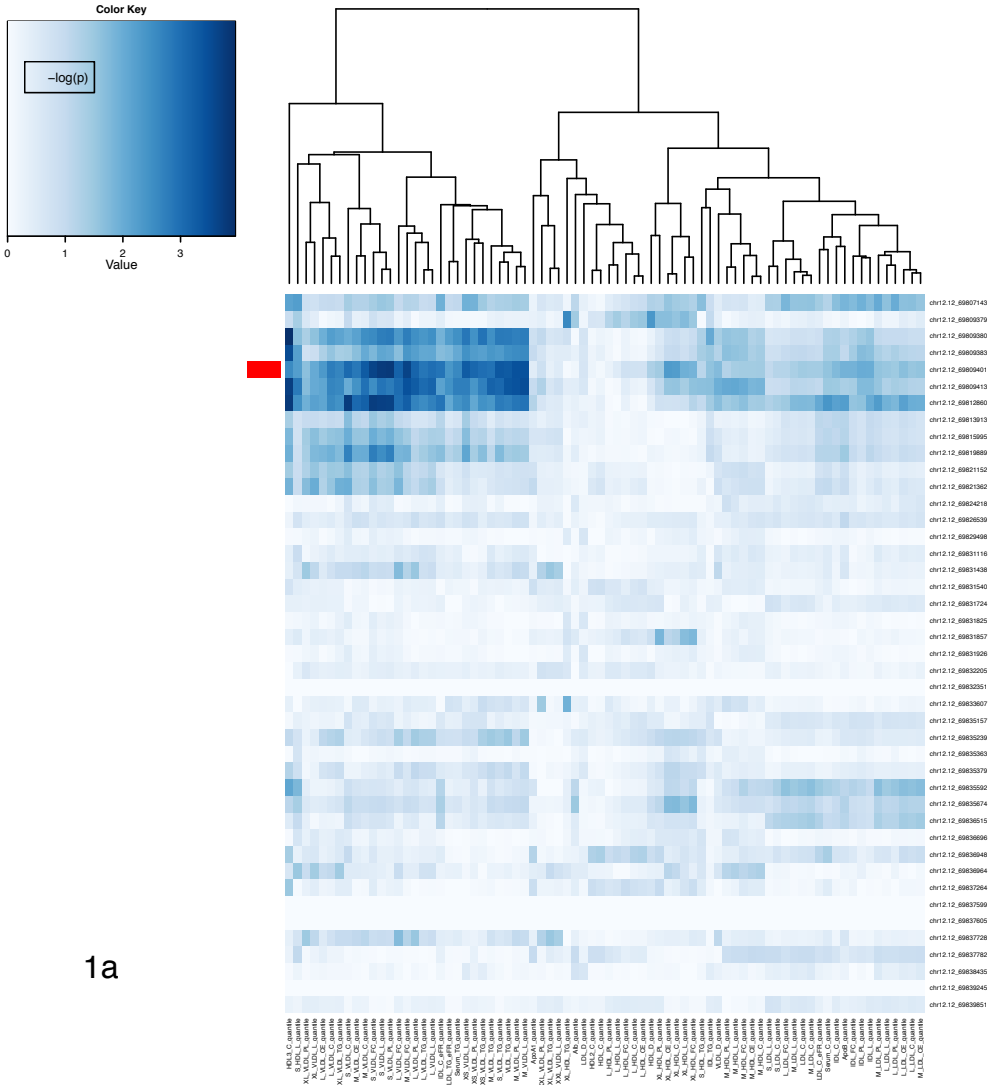


FIGURE 4.10: Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1986

Here, the metabonomic lipids are represented in the X axis and probes within the *TSAPN8* gene are shown in the Y axis. The probe at location chr12: 69809401 is shown in marked in red.

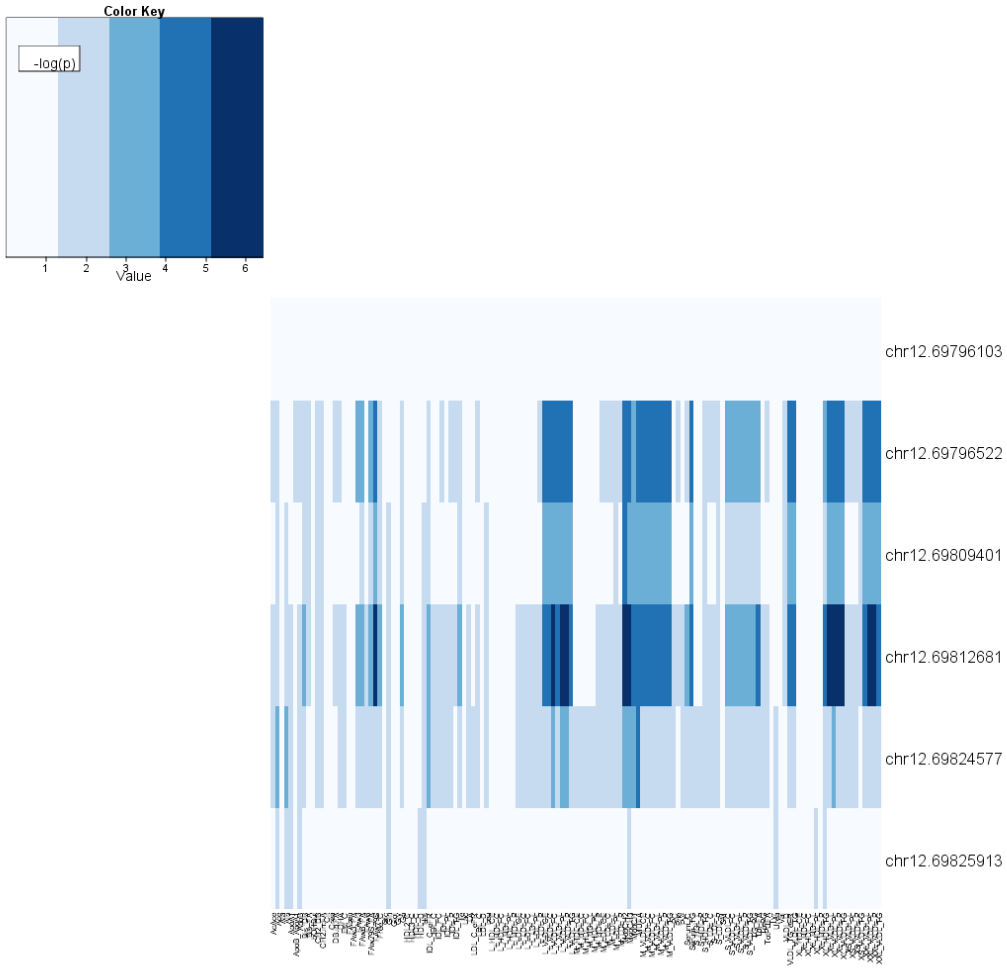


FIGURE 4.11: Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1966

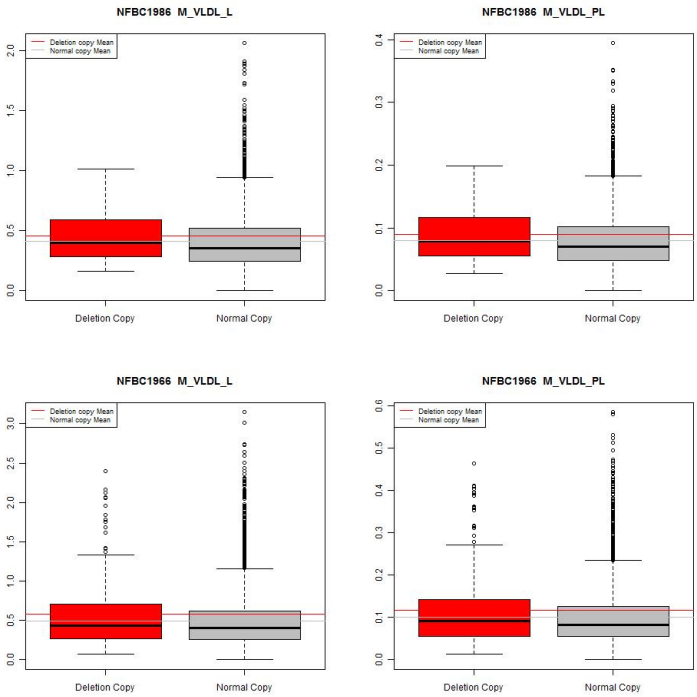


FIGURE 4.12: Box plots for normal and deletion copy

Box plot for the top four metabonomic lipid phenotypes found associated with CNV at chr12: 69809401 after meta-analysis of NFBC 1966 and NFBC 1986. Left and right columns correspond to NFBC 1966 and NFBC 1986 respectively. Box plots for phenotype measurement of deletion copy are marked in red and phenotypes of samples without deletion are marked in grey and a red and grey line marks their respective means of metabonomic phenotypes.

4.4.5 Multivariate analysis

Joint effect of multiple lipid phenotypes I carried out the multiphen based multivariate analysis on the meta-analysis results of NFBC 1966 and NFBC 1986 for 90 common metabonomic phenotypes at chr12: 69809401 (Table 4.9, 4.17). The results from the joint-model with backward selection process indicated that a core set of 4 lipids are significantly associated with the CNV at chr12: 69809401 (joint p-value=1.33E-06), The top three phenotypes identified from the multivariate model M_VLDL_C, M_VLDL_TG and S_HDL_TG were also significant in univariate association results (corrected). The joint-model with no variable selection for all the 90 phenotypes together was 3.8E-06 (Table 4.9, 4.17). In NFBC 1986, for the CNV at chr12: 69812860 (6th exon) I found L_HDL_PL (Phospholipids in large HDL) to be the top associated lipid through multivariate variable selection method (p-value=1.14E-02). In the same or nearest possible probe location in NFBC 1966, I again found strong association patterns for HDL or for its metabonomic sub-categories.

Table 4.9 Results from multiphen association approach with 1) joint-model (all lipid phenotypes) and 2) multiphen joint-model with variable-selection (subset of phenotypes), for common probes in the coding-regions of the TSPAN8 gene, including the lipid-associated region chr12 69809401.

TABLE 4.9: Joint association P-values from multiphen

Location	Lipid Name	MAF	Beta	P-value	N	Batch-effect correction method
12_69809401	M_VLDL_C_quantile M_VLDL_L_quantile M_VLDL_TG_quantile S_HDL_TG_quantile **JointModel (after variable selection)**	0.041	0.12 -0.38 0.26 -0.00796 NA	1.38E-04 1.57E-05 2.03E-05 1.03E-02 1.33E-06	7.34E+03	PC projection
12_69835363	S_HDL_TG_quantile XL_HDL_C_quantile **JointModel (after variable selection)**	0.0261	-0.00965 -0.0013 NA	3.54E-04 6.52E-01 2.17E-03	7.34E+03	PC projection

Table 4.10 Association of CNV at chr12:69812681 in NFBC 1966 using multi-variate, multi-variate with variable selection and LRR based methods.

TABLE 4.10: Association of CNV at chr12:69812681 in NFBC 1966

Association Method	Location	Cohort Name	Phenotype(s)	Beta	P-value	Batch-effect correction method
Genotype association using univariate assoicaiton model	12_69812681	NFBC 1966	L_HDL_PL_quantile	-0.23	1.14E-02	PC projection
			L_HDL_L_quantile	-0.223	1.38E-02	
			L_HDL_FC_quantile	-0.212	1.82E-02	
			M_HDL_FC_quantile	-0.219	2.28E-02	
			L_HDL_C_quantile	-0.198	2.86E-02	
			HDL_C_quantile	-0.188	3.28E-02	
			M_HDL_PL_quantile	-0.204	3.28E-02	
			L_HDL_CE_quantile	-0.19	3.59E-02	
			M_HDL_L_quantile	-0.197	4.26E-02	
			HDL2_C_quantile	-0.184	4.32E-02	
Genotype association with multipehn joint model using all phenotypes	12_69812681	NFBC 1966	Joint model for all pheotypes *	NA	0.503	PC projection
Genotype associaiton with multiphen joint model with variable selection	12_69812681	NFBC 1966	Joint model for pheotypes after variable selection	NA	NA	PC projection
Genotype association with multipehn joint model using all phenotypes	12_69812681	NFBC 1966	Joint model for all pheotypes	NA	1.78E-03	LRR variance as covariate
Genotype associaiton with multiphen joint model with variable selection	12_69812681	NFBC 1966	Joint model for pheotypes after variable selection	NA	6.07E-05	LRR variance as covariate
LRR association using univariate method	12_69812681	NFBC 1966	XXL_VLDL_PL_quantile	-0.164	1.65E-02	PC projection
			S_HDL_TG_quantile	-0.13	3.67E-02	
			Alb_quantile	-0.128	4.77E-02	
LRR association with multipehn joint model using all phenotypes	12_69812681	NFBC 1966	joint model for all pheotypes	NA	0.242	PC projection
LRR associaiton with multiphen joint model with Variable selection	12_69812681	NFBC 1966	Joint model for pheotypes after variable selection	NA	NA	PC projection
LRR association with multipehn joint model using all phenotypes	12_69812681	NFBC 1966	Joint model for all pheotypes	NA	4.39E-09	LRR variance as covariate
LRR associaiton with multiphen joint model with Variable selection	12_69812681	NFBC 1966	Joint model for pheotypes after variable selection	NA	1.39E-13	LRR variance as covariate

* all phenotypes refers to the 90 common metabonomic lipid phenotypes between NFBC 1966 and NFBC 1986

4.4.6 Validation using raw intensity

T2D association In the DESIR T2D cohort, the probe nearest to CNVR5583.1 located at chr12: 69824577, the probe rs1798089, was found to be nominally associated with T2D with p-value of 5.55E-06 (Table 4.11, 4.12). Probes at chr12: 69809401 (rs1051334) and chr12: 69812681 (rs2270588) were also found to be nominally associated with T2D with p-values of 3.32e-07 and 1.24e-05 respectively. I further highlight that the WTCCC CNV study design did not cover the lipid-associated region I analysed in details here (chr12: 69809401-69812861). Hence, this finding also promises to be a novel T2D-associated locus which is also further indirectly confirmed by the LD results.

Lipid association I have further replicated some of the association of CNVs with metabonomic and lipid phenotypes. At chr12: 69809401 location, on meta-analysing results from NFBC 1966 and NFBC 1986 I found 10/90 lipids to be nominally associated at a p-value threshold of 0.05 and further three lipids replicating from the univariate meta-results (M_VLDL_TG, M_VLDL_PL and M_VLDL_L) and two lipid phenotypes (M_VLDL_TG and M_VLDL_L) from meta-multivariate results of NFBC 1966 and NFBC 1986 (Table 4.11). These results indicated some degree of consistency in the association results from univariate, multivariate and LRR based methods. This gives more confidence to the results presented here.

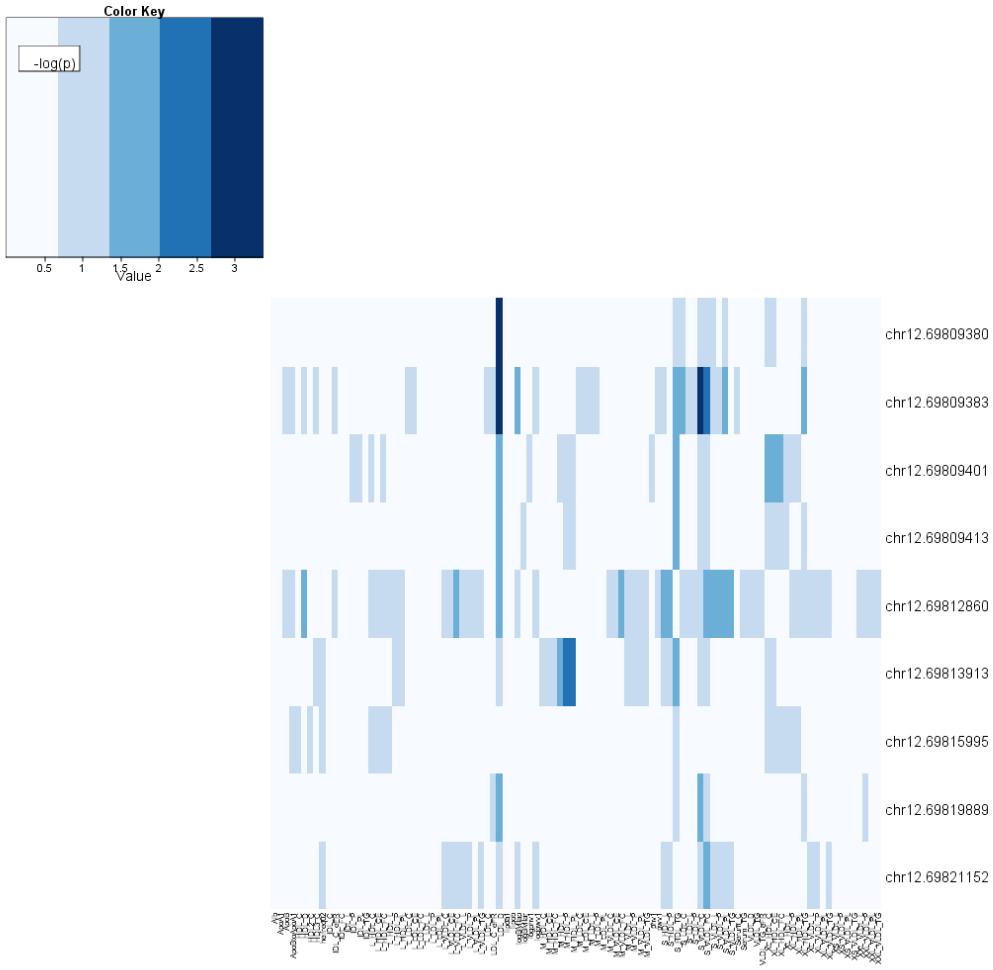


FIGURE 4.13: Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1986 using raw intensity

Table 4.11 Lipids found associated with CNV at chr12 69809401 are reported after meta-analysing LRR intensity based association Pvalues from NFBC 1966 and NFBC 1986. Multiphen association approach with 1) joint-model (all lipid phenotypes) and 2) multiphen joint-model with variable-selection (subset of phenotypes) using two different batch-effect based correction methods (PC projection approach and LRR variance as covariate) is reported.

TABLE 4.11: Meta-analysis results using LRR intensity values at chr12:69809401

Association Method	Location	Meta analysis probe	Phenotype(s)	Beta	P-value	Batch-effect correction method
LRR association using univariate assoicaiton model	12_69809401	rs1051334 (NFBC 1966); chr12:69809401 (NFBC 1986)	S_HDL_TG_quantile	-0.224	1.19E-06	PC projection
			M_VLDL_TG_quantile	-0.135	4.06E-03	
			XXL_VLDL_PL_quantile	-0.114	1.48E-02	
			M_VLDL_PL_quantile	-0.113	1.49E-02	
			M_HDL_PL_quantile	-0.11	1.92E-02	
			M_VLDL_L_quantile	-0.109	1.95E-02	
			VLDL_TG_quantile	-0.106	2.32E-02	
			L_VLDL_PL_quantile	-0.107	2.38E-02	
			VLDL_TG_eFR_quantile	-0.104	2.74E-02	
			XXL_VLDL_TG_quantile	-0.0997	3.54E-02	
LRR association using univariate assoicaiton model	12_69809401	rs1051334 (NFBC 1966); chr12:69809401 (NFBC 1986)	S_HDL_TG_quantile	-0.125	0.0025	LRR variance as covariate
			XL_HDL_TG_quantile	-0.0869	0.0365	
			XXL_VLDL_PL_quantile	-0.0733	0.0666	
			M_VLDL_TG_quantile	-0.0752	0.0751	
LRR association with multiphen joint model using all phenotypes	12_69809401	rs1051334 (NFBC 1966); chr12:69809401 (NFBC 1986)	Joint model for all phenotypes	NA	3.80E-06	PC projection
LRR association with multiphen joint model using all phenotypes	12_69809401	rs1051334 (NFBC 1966); chr12:69809401 (NFBC 1986)	Joint model for all phenotypes	NA	3.86E-04	LRR variance as covariate

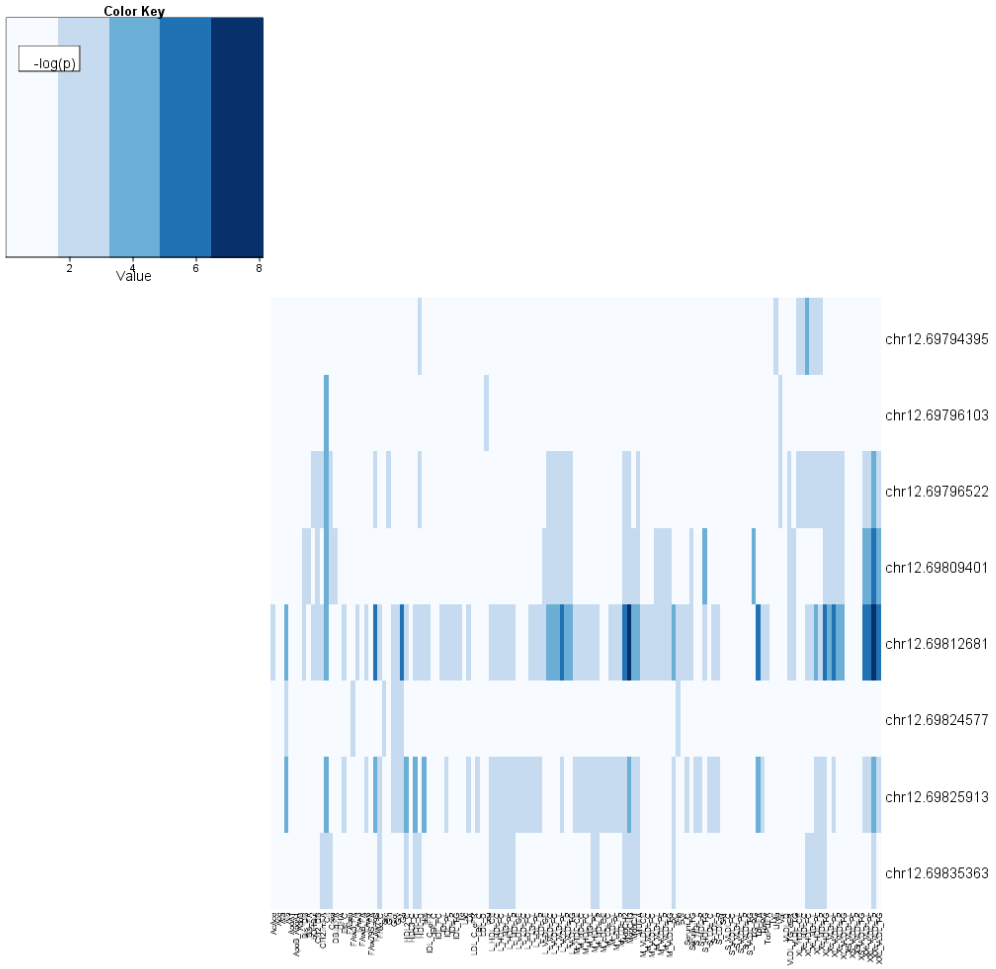


FIGURE 4.14: Heatmap for association with lipids in the *TSPAN8* gene in NFBC 1966 using raw intensity

Table 4.12 Association of LRR intensity with T2D disease outcome. The association results were adjusted for gender and batch effects using LRR PC based method.

TABLE 4.12: Association with T2D outcome using LRR intensity values

Location	SNP Name	Phenotype	MAF	Beta	P-value	No of Samples	Batch effect correction method
12_69796522	12_69796522	T2D	0.0182	-0.225	3.35E-01	1.35E+03	PC projection
12_69809401	12_69809401		0.0143	-1.08	3.32E-07		
12_69812681	12_69812681		-0.0453	-1.09	1.24E-05		
12_69824577	12_69824577		-0.014	-1.89	5.55E-06		
12_69825913	12_69825913		-0.0118	-1.82	2.27E-10		
12_69872569	12_69872569		-0.0592	-1.44	9.87E-08		
12_69875513	12_69875513		-0.00854	1.28	1.86E-02		
12_69899139	12_69899139		-0.0416	0.9	3.44E-03		
12_69899543	12_69899543		-0.0243	0.089	8.04E-01		

4.4.7 LD results for *TSPAN8*

The Wellcome Trust CNV study reported the best tagging SNP for CNVR55831 to be rs1705261 with $r^2 = 0.998$. In NFBC 1986, I have reported a similar observation where the same SNP rs1705261was found to be in high LD with $r^2 = 0.958$ for a probe within CNVR5583.1 (probe hg1869819889) (Table 4.13). Further, in NFBC 1986 the best tagging SNP for the CNV at chr12: 69809401 was found to be the probe hg181269809383 with r^2=0.939 located at chr12: 69809383. In NFBC 1986, the probe at chr12: 69809401 was in LD with CNVR55831 (probe hg181269819889) with r^2=0.623. I observed that within a 12 kb window there were several CNVs, which were found to be in high LD with CNVR5583.1, suggesting a LD block in that region. However, LD results for NFBC 1966 was not significant with no probes having r^2 (between CNV-SNP) greater than 0.1. This negative result could be due to the lack of probe coverage within and around CNVR5583.1.

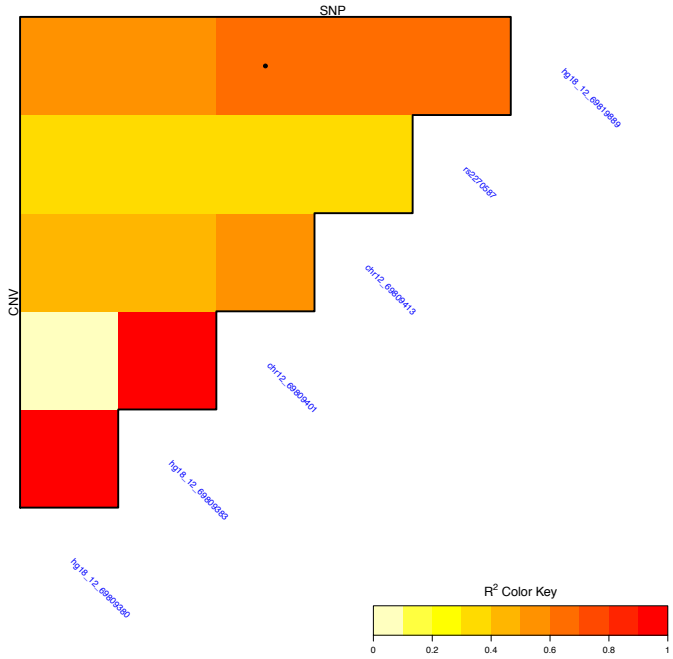


FIGURE 4.15: *TSPAN8* LD plot in NFBC 1986

LD heatmap showing the r^2 values for the NFBC 1986 cohort, where r^2 was calculated between total number of copies of the CNV and the B allele from the SNP genotype calculated by cnvHap. LD r^2 for the CNV at chr12: 69809401 and the best tagging SNP probe within CNVR5583.1 is indicated by a black circle

Table 4.13 LD results in NFBC 1986 (genotyped on the Illumina Cardio Metabochip). LD was calculated between the number of copy number from HMM model and no of B allele for all probes within or around CNVR5581 using a regression model. All probes where LD $r^2 > 0.5$ is reported.

TABLE 4.13: LD results in NFBC 1986 at CNVR5583.1

LD between		Location A	Location B	Distance between probes	rsid_A	rsid_B	r^2
No of copies	B allele	69809380	69809383	3	hg18_12_69809380	hg18_12_69809383	1
		69815995	69819889	3894	rs1495381	hg18_12_69819889	1
		69813913	69819889	5976	rs1705261	hg18_12_69819889	0.974
		69809401	69809383	18	chr12:69809401	hg18_12_69809383	0.939
		69809413	69809383	30	chr12:69809413	hg18_12_69809383	0.924
		69821152	69819889	1263	rs3794217	hg18_12_69819889	0.914
		69821362	69819889	1473	rs1798090	hg18_12_69819889	0.914
		69812860	69809383	3477	rs2270587	hg18_12_69809383	0.893
		69812860	69819889	7029	rs2270587	hg18_12_69819889	0.66
		69809413	69819889	10476	chr12:69809413	hg18_12_69819889	0.636
		69809401	69819889	10488	chr12:69809401	hg18_12_69819889	0.623
		69813913	69809383	4530	rs1705261	hg18_12_69809383	0.597
		69809380	69819889	10509	hg18_12_69809380	hg18_12_69819889	0.585
		69809383	69819889	10506	hg18_12_69809383	hg18_12_69819889	0.585
		69815995	69809383	6612	rs1495381	hg18_12_69809383	0.585
		69819889	69809383	10506	hg18_12_69819889	hg18_12_69809383	0.585
		69809401	69809413	12	chr12:69809401	chr12:69809413	0.531
		69821152	69809383	11769	rs3794217	hg18_12_69809383	0.53
		69821362	69809383	11979	rs1798090	hg18_12_69809383	0.53
		69812860	69809413	3447	rs2270587	chr12:69809413	0.522

4.4.8 Association with glucose stratification

In accordance with the earlier published classification, I divided the NFBC cohorts into three categories based on fasting blood glucose levels. This included non-diabetics (glucose concentration of 5.6 mmol/L or less), pre-diabetics (glucose concentration between 5.6mmol/L and 6.9mmol/L) and diabetics (glucose levels $>$ 6.9 mmol/L). Based on this classification there were 7 diabetics and 31 diabetics in NFBC 1966 (Table 4.14). To increase power in association analysis (especially in NFBC 1986 for the diabetics cohort), I pooled the data from NFBC 1966 and NFBC 1986 together and then carried out my association analysis. The stratified results show strong association patterns in the non-diabetics population (N=7230).

The top associated lipid I found is a sub-category of VLDL (S_VLDL_L, p-value=7.65E-09. The earlier reported top lipid M_VLDL_FC was found to be associated with p-value=2.48E-07. In the pre-diabetics cohort, there was a strong trend for association in sub-categories of VLDL lipids, with 5/7 VLDL lipids associated with p-value < 0.05 (N=1062), but the association p-values were much weaker. In the diabetics cohort though the sample size was very small (N=38), I still found association for at least 3 VLDL sub-categories at a p-value threshold of 0.05. These results indicate the presence of VLDL association in all three categories (diabetics, pre-diabetics and non-diabetics) and further suggests the possibility that *TSPAN8* CNVs are indeed a true susceptibility variant for T2D. But these results need to be further explored using larger samples sizes, especially for the pre-diabetics and diabetics cohort.

Table 4.14 Stratified association of CNVs at chr12: 69809401 with metabonomic lipid phenotypes. Here, the phenotype stratification was based on the fasting glucose levels. Classification was done using the following criteria: Non-diabetics (glucose concentration of 5.6 mmol/L or less), pre-diabetics (glucose concentration between 5.6mmol/L and 6.9mmol/L) and diabetics (glucose levels $>$ 6.9 mmol/L)

TABLE 4.14: Univariate meta-analysis results at chr12:69809401 based on sample stratification based on blood glucose levels. LRR variance is used as a covariate

Phenotype	Location	Phenotype	Beta	Pvalue	N
		M_VLDL_PL_quantile	-0.235	1.20E-04	7230
Threshold 0		M_VLDL_L_quantile	-0.234	1.36E-04	7230
		M_VLDL_TG_quantile	-0.223	2.92E-04	7230
		XL_HDL_C_quantile	0.114	0.402	1060
Threshold 1		XL_HDL_CE_quantile	0.113	0.407	1060
	12_69809401	S_VLDL_C_quantile	-0.0744	0.577	1060
		ApoA1_quantile	-1.31	2.75E-02	36
Threshold 2		LDL_D_quantile	-1.41	3.08E-02	38
		M_HDL_FC_quantile	-1.05	4.12E-02	38

4.4.9 Association with PC projection

Confounding factors like batch effects are a major problem in genetic association studies. To address this problem, I implemented a new approach to account for batch effects by first calculating genome wide LRR PCs using a fast iterative procedure (described in chapter 3) and then projecting out these top PCs from our CNV association model. The main difference here is that in this association model, the PCs were projected out of the outcome variable not used as covariates, like in the epilepsy analysis (Table 4,5, 4.9, 4.17). Using this approach, I was successful in replicating the CNV associations I reported earlier, and was further able to reduce genomic inflation factor. Using this approach seems to be promising and I obtained good concordance with the LRR-variance method (when used a covariate). I still further need to do a genome-wide analysis for determining the proper number of PCs to use in the association model.

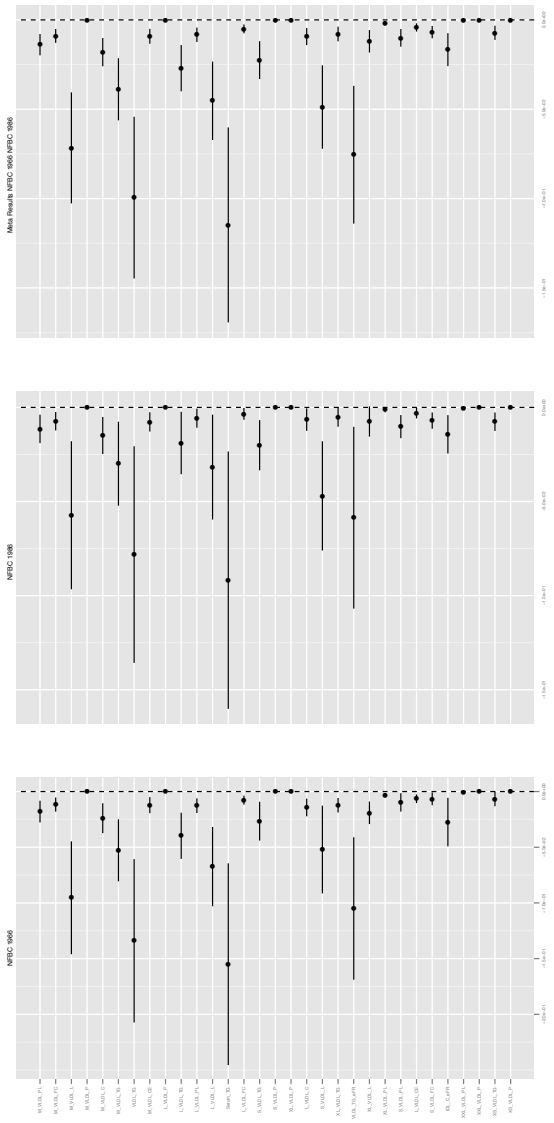


FIGURE 4.16: Forest plot of top meta p-values

Forest plot showing the beta and standard errors in NFBC 1966, NFBC 1986 and the meta analysis results.

4.4.10 Association with lipids at chr12: 69812860

I further discovered and report here a CNV in the 6th exon of the *TSPAN8* gene in NFBC 1986 (rs2270587) with an allele frequency of 0.013, which also showed a strong trend in association with lipids. The top lipids associated for this CNV was found to be M_VLDL_FC (Free cholesterol in medium VLDL) (P=6.54E-04) (Table 4.10). In NFBC 1986, the multivariate model with variable selection a joint p-value= 2.19E-04. The top lipid associated using the LRR PCs projection approach method was HDL, but this time a different sub-category was obtained (S_HDL_L, p-value= 1.14E-04).

In NFBC 1966, there was a trend in association with HDL sub-categories at chr12: 69812681 using all the association approaches described here (univariate p-value=1.14E-02 for L_HDL_PL), multivariate (p-value= 6.07E-05) (Table 4.10). Despite the lack of consistency for all probes in the three cohorts analysed, there seems to be an indication of HDL association near the 6th exon of *TSPAN8* gene using univariate, multivariate and LRR PC projection association methods. I also underline that the British and NFBC cohorts are different populations and this might be a real possible reason for why the association level do not match completely at the same genomic location.

Table 4.15 Association of CNV with metabonomic lipids in the 6th exon of *TSPAN8* gene in NFBC 1986. Replication in NFBC 1966 is reported for same or nearest probe possible. Results from multiphen association approach with 1) joint-model (all lipid phenotypes) and 2) multiphen joint-model with variable-selection (subset of phenotypes) are reported for CNVs characterized through a) cnvHap derived expected genotypes b) LRR raw intensity.

TABLE 4.15: Association of CNV at 6th exon of *TSPAN8* with metabonomic lipids in NFBC 1966 and NFBC 1986

Association Method	Location	Cohort Name	Lipid Name	Beta	P value	Batch-effect correction method
Genotype association using univariate assoicaiton model	12_69812860	NFBC 1986	M_VLDL_FC_quantile	-0.4	6.50E-04	PC projection
			S_VLDL_L_quantile	-0.38	7.70E-04	
			M_VLDL_C_quantile	-0.388	8.81E-04	
			S_VLDL_PL_quantile	-0.372	9.46E-04	
			S_VLDL_FC_quantile	-0.373	9.78E-04	
			S_VLDL_C_quantile	-0.361	1.24E-03	
			L_VLDL_FC_quantile	-0.368	1.34E-03	
			M_VLDL_L_quantile	-0.375	1.36E-03	
			M_VLDL_CE_quantile	-0.366	1.54E-03	
			M_VLDL_PL_quantile	-0.368	1.83E-03	
			L_VLDL_L_quantile	-0.345	2.09E-03	
			S_VLDL_TG_quantile	-0.34	2.60E-03	
			XL_VLDL_TG_quantile	-0.324	2.63E-03	
			L_VLDL_TG_quantile	-0.335	2.66E-03	
			VLDL_TG_quantile	-0.346	2.69E-03	
			XL_VLDL_L_quantile	-0.319	3.29E-03	
			L_VLDL_C_quantile	-0.322	3.67E-03	
			L_VLDL_PL_quantile	-0.325	3.79E-03	
			M_VLDL_TG_quantile	-0.327	4.54E-03	
			IDL_C_eFR_quantile	-0.301	6.21E-03	
			XL_VLDL_PL_quantile	-0.299	6.23E-03	
			HDL3_C_quantile	-0.297	7.07E-03	
			VLDL_D_quantile	-0.327	7.22E-03	
			Serum_TG_quantile	-0.312	7.31E-03	
			L_VLDL_CE_quantile	-0.274	1.07E-02	
			VLDL_TG_eFR_quantile	-0.298	1.10E-02	
			S_HDL_L_quantile	-0.284	1.38E-02	
			XS_VLDL_L_quantile	-0.266	1.61E-02	
			XS_VLDL_TG_quantile	-0.261	1.67E-02	
			ApoB_quantile	-0.266	2.19E-02	
			M_LDL_PL_quantile	-0.239	2.90E-02	
			Serum_C_quantile	-0.233	3.25E-02	
			IDL_C_quantile	-0.235	3.46E-02	
			XL_HDL_FC_quantile	0.241	3.56E-02	
			L_LDL_CE_quantile	-0.23	3.59E-02	
			XL_HDL_PL_quantile	0.235	3.77E-02	
			LDL_C_eFR_quantile	-0.217	4.91E-02	
Genotype association with multipehn joint model using all phenotypes	12_69812860	NFBC 1986	Joint model for all pheotypes	NA	0.0438	PC projection
Genotype associaiton with multiphen joint model with Variable selection	12_69812860	NFBC 1986	Joint model for pheotypes after variable selection	NA	NA	PC projection
Genotype association with multipehn joint model using all phenotypes	12_69812860	NFBC 1986	Joint model for all pheotypes	NA	2.19E-04	LRR variance as covariate
Genotype associaiton with multiphen joint Model with variable selection	12_69812860	NFBC 1986	Joint model for pheotypes after variable selection	NA	3.40E+03	LRR variance as covariate
LRR association using univariate method	12_69812860	NFBC 1986	S_HDL_L_quantile	0.228	1.14E-03	PC projection
			XL_HDL_CE_quantile	-0.223	1.26E-03	
			XL_HDL_C_quantile	-0.221	1.38E-03	
			XL_HDL_FC_quantile	-0.195	5.08E-03	
			XL_HDL_L_quantile	-0.184	8.03E-03	
			M_HDL_CE_quantile	0.16	1.20E-02	
			M_HDL_C_quantile	0.154	1.62E-02	
			M_HDL_L_quantile	0.155	1.86E-02	
			XL_HDL_TG_quantile	-0.152	2.33E-02	
			M_HDL_PL_quantile	0.139	3.78E-02	
			HDL_D_quantile	-0.135	4.79E-02	
LRR association with multipehn joint model using all phenotypes	12_69812860	NFBC 1986	Joint model for all pheotypes	NA	0.124	PC projection
LRR associaiton with multiphen joint model with variable selection	12_69812860	NFBC 1986	Joint model for pheotypes after variable selection	NA	NA	PC projection
LRR association with multipehn joint model using all phenotypes	12_69812860	NFBC 1986	Joint model for all pheotypes	NA	0.13	LRR Var as covariate
LRR associaiton with multiphen joint model with variable selection	12_69812860	NFBC 1986	Joint Model for pheotypes after variable selection	NA	NA	LRR Var as covariate

4.4.11 Association with lipids at CNVR5583.1

Only NFBC 1986 had a probe within CNVR5583.1 located at chr12 69819889 (probe hg181269819889). The top lipid associated at this locus was S_VLDL_PL (free cholesterol in small VLDL, p-value= 5.39E-03) and 25 lipids were associated at a p-value < 0.05 (Table 4.16). However, none of these univariate results survived multiple-testing threshold. I further report that the top lipid associated in this case is similar to the association of top lipid at chr12: 69809401, with weaker association p-values.

Table 4.16 Association of CNV (deletion) with lipid phenotypes for probes within or nearest to CNVR5583.1 in NFBC 1966 and NFBC 1986.

TABLE 4.16: Association of deletion CNV at or near CNVR5583.1

				Batch effect correction method	
chr_start	Cohort Name	Phenotype	beta	Pvalue with PC projection	Pvalue with LRR variance as covariate
12_69819889	NFBC 1986	S_VLDL_PL_quantile	-0.388	5.39E-03	4.07E-03
		S_VLDL_FC_quantile	-0.381	6.63E-03	4.36E-03
		S_VLDL_L_quantile	-0.378	6.95E-03	5.97E-03
		M_VLDL_C_quantile	-0.384	7.90E-03	1.40E-02
		S_VLDL_C_quantile	-0.364	8.74E-03	5.61E-03
		M_VLDL_CE_quantile	-0.374	8.99E-03	1.55E-02
		M_VLDL_FC_quantile	-0.376	9.81E-03	1.69E-02
		L_VLDL_C_quantile	-0.351	1.04E-02	2.51E-02
		L_VLDL_CE_quantile	-0.339	1.10E-02	2.97E-02
		L_VLDL_FC_quantile	-0.355	1.24E-02	2.58E-02
		S_VLDL_TG_quantile	-0.329	1.86E-02	2.21E-02
		XL_VLDL_L_quantile	-0.313	1.98E-02	4.30E-02
		M_VLDL_PL_quantile	-0.338	2.08E-02	3.51E-02
		L_VLDL_L_quantile	-0.321	2.11E-02	4.30E-02
		XL_VLDL_TG_quantile	-0.306	2.18E-02	4.40E-02
		M_VLDL_L_quantile	-0.331	2.29E-02	3.83E-02
		XS_VLDL_L_quantile	-0.305	2.62E-02	1.33E-02
		VLDL_TG_quantile	-0.31	3.01E-02	4.45E-02
		L_VLDL_TG_quantile	-0.295	3.28E-02	6.02E-02
		XL_VLDL_PL_quantile	-0.287	3.43E-02	7.70E-02
		L_VLDL_PL_quantile	-0.289	3.81E-02	7.74E-02
		M_VLDL_TG_quantile	-0.289	4.34E-02	7.67E-02
		HDL3_C_quantile	-0.274	4.47E-02	1.56E-02
		XS_VLDL_TG_quantile	-0.266	4.89E-02	2.63E-02
		ApoB_quantile	-0.281	5.05E-02	4.76E-02

4.4.12 Multivariate analysis joint association and joint association with variable selection

Table 4.9 and Table 4.17 gives the multiphen based p-values for joint-model and for joint-model with variable selection for all probes within the *TSPAN8* coding regions in NFBC 1966 and NFBC 1986. The joint-model p-values and joint-model p-values with variable selection results are fairly consistent in showing the positive associations but it can be observed that the p-values obtained with variable selection are usually much smaller than the joint-model p-values when all the variables are considered in the model. One probable reason for this could be that the non correlated and non associated phenotypes considered in the model can decrease the overall parameter fit in the model when in reality only a subset of the phenotypes are actually associated with the outcome. In these cases, results from the joint-model with variable selection can be closer to reality and give more accurate results. One such possible scenario is at chr12: 6983529 locus lying in the start codon of *TSPAN8* in NFBC 1966 and NFBC 1986 where the overall joint-model p-value was found to be 1.99E-41, but on variable selection the overall p-value and joint p-value increased giving a joint p-value= 1.04E-50 (Table 4.9, 4.17).

The other advantage of the backward variable selection process I implemented is that it can filter out only the most relevant variables from all other correlated variables and thus giving only a small and best possible subset of phenotypes driving the association. This can be quite advantageous when experimental follow up of phenotypes is expensive and limited by the available resources. However, when there is a prior knowledge that all phenotypes have some effect on the outcome then the joint-model might prove to be more accurate than the joint-model with variable selection. But the variable selection process is a time consuming process and performing this step for all the probes genome wide can take a very long time. In such cases the joint-model with all the phenotypes, which is much faster to run, can be a proxy analysis approach for very large regions or genome wide coverage. This step can then pick up regions of interest before applying a variable selection process.

Table 4.17 Results from multiphen association approach for cnvHap derived probabilistic genotypes with 1) joint-model (all lipid phenotypes) and 2) multiphen joint-model with variable-selection (subset of phenotypes) are reported for CNVs in the *TSPAN8* coding region for NFBC 1966 and NFBC1986 separately. Two different approaches for correcting batch-effects using LRR PC projection method and LRR variance as covariate is reported.

TABLE 4.17: Multiphen based association results using joint model in *TSPAN8* coding regions in NFBC 1986 and NFBC 1966

		Batch effect correction method			
		LRR Pcs projected out		LRR Variance as covariate	
Location	Cohort Name	Joint Model P-value	Joint Model with Variable Selection	Joint Model P-value	Joint Model with Variable Selection
12_69807143	NFBC 1986	9.34E-01	1.89E-02	2.83E-01	1.58E-03
12_69809379		9.95E-01	3.06E-05	8.29E-03	3.13E-05
12_69809380		5.11E-02	2.79E-03	4.93E-02	4.52E-04
12_69809383		1.11E-01	3.83E-02	1.55E-02	4.99E-07
12_69809401		2.06E-03	3.80E-03	7.58E-02	2.35E-07
12_69809413		7.52E-03	4.27E-05	8.68E-02	1.03E-05
12_69812681		NA	NA	NA	NA
12_69812860		4.38E-02	9.79E-03	1.30E-01	3.69E-04
12_69813913		5.88E-01		1.42E-01	7.31E-02
12_69815995		7.56E-01	2.60E-02	8.63E-01	2.45E-02
12_69819889		6.89E-01	6.74E-03	6.63E-01	2.31E-02
12_69821152		1.48E-01	8.41E-04	1.87E-02	5.35E-02
12_69821362		1.39E-01	4.60E-02	2.89E-02	NA
12_69824218		4.65E-01	NA	7.77E-04	NA
12_69824577		NA	NA	NA	NA
12_69825913		NA	NA	NA	NA
12_69826539		7.75E-01	NA	6.57E-01	NA
12_69829498		6.01E-02	NA	6.58E-03	NA
12_69831116		7.71E-02	NA	1.76E-01	NA
12_69831438		8.14E-02	1.63E-13	7.79E-01	9.33E-14
12_69831540		1.06E-01	NA	1.13E-01	NA
12_69831724		5.66E-06	4.36E-03	2.99E-01	1.01E-02
12_69831738		5.07E-09	NA	2.60E-02	NA
12_69831825		1.20E-07	NA	1.33E-02	NA
12_69831857		4.29E-08	3.59E-02	5.71E-04	1.49E-02
12_69831926		6.05E-08	NA	2.57E-01	NA
12_69832205		2.27E-06	NA	8.34E-01	NA
12_69832274		7.87E-74	NA	1.74E-01	NA
12_69832351		6.45E-01	NA	2.46E-02	NA
12_69832529		9.87E-01	NA	3.93E-02	NA
12_69833589		1.00E+00	NA	3.40E-02	NA
12_69833607		1.00E+00	2.08E-05	4.29E-02	2.06E-05
12_69833786		NA	NA	1.30E-02	NA
12_69834401		9.97E-01	NA	4.56E-04	NA
12_69835157		7.44E-01	NA	4.19E-01	NA
12_69835239		8.35E-01	1.99E-41	1.71E-01	1.04E-50
12_69835363		8.31E-01	3.84E-02	3.12E-02	NA

4.4.13 NFBC1966 PennCNV results

PennCNV calls for NFBC 1966 was generated using recommended settings. Six samples showed heterozygous deletions encompassing chr12: 69809401 CNV and the CNVR5583.1 region.

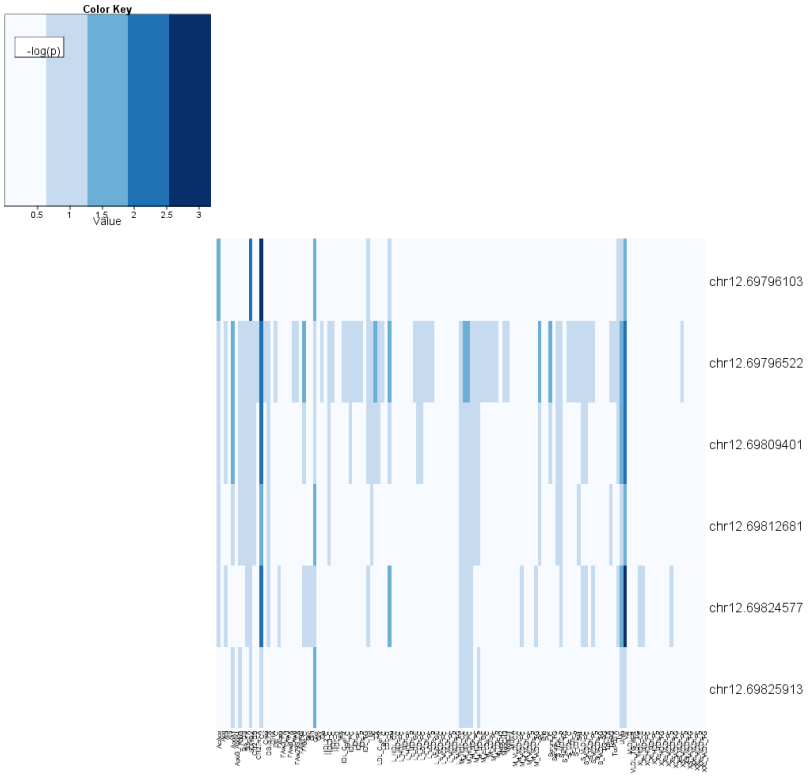


FIGURE 4.17: Heatmap results for association using PennCNV

4.4.14 LD results for NFBC 1966

No probes were found to be in LD with chr12: 69809401 CNV or CNVR5583.1 with $r^2 > 0.5$, hence LD results for NFBC 1966 is not reported.

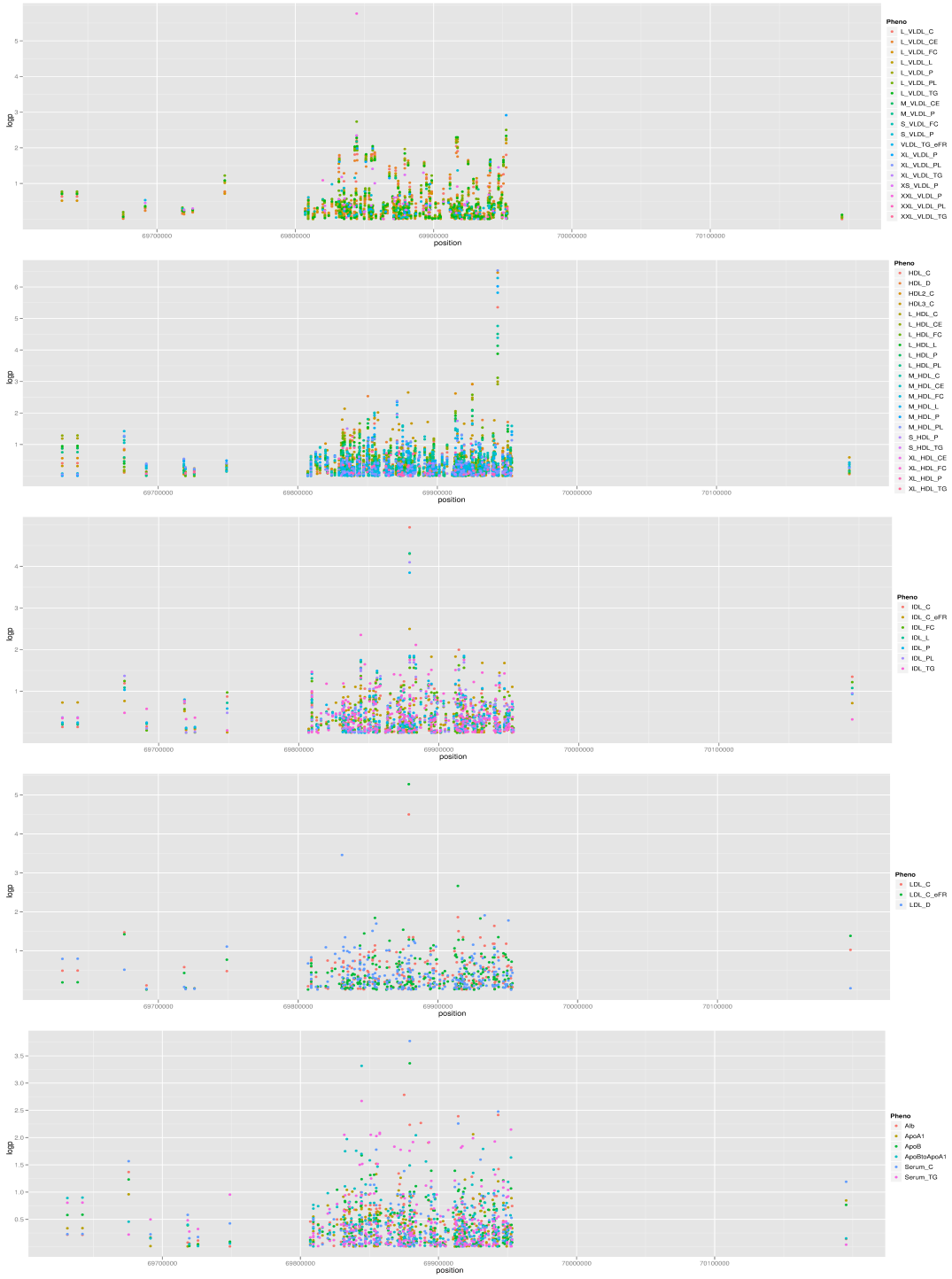


FIGURE 4.18: Manhattan plot of SNPs within the *TSPAN8* gene



FIGURE 4.19: Manhattan plot of SNPs within the *TSPAN8* gene in NFBC 1966

4.5 Missing data

While doing normal regression rows containing missing values are usually removed. This can limit the amount of information and if significant number of data points are missing and this could lead to a loss of power. There are usually four accepted mechanisms which can theoretically lead to missingness in a given dataset.

- Missing completely at random
- Missing at random
- Missingness that depends on the missing value
- Missingness that depends on unobserved predictors

In the first case, the probability of missingness is the same for all units and removing data may not affect inferences, unless power is affected to a large extent. In the second case, the probability of missingness of a variable depends only on the available information. If the variable of interest is missing at random, then the probability of missingness depends only on the other fully recorded variables. For the last case the missingness depends on the unobserved predictors. In this case the missingness may not be random anymore and this can predict missing values. Another interesting instance is when the missingness may depend on the missing value itself. For example if a variable predictor's probability of missingness increases with its value or vice versa. In general, it is impossible to predict that the data is missing at random so it may lead to estimates with larger standard errors and biases. There are two approaches for modelling missing data namely joint-modelling (JM) [243] and fully conditional specification (FCS) [244]. Joint-modelling specifies a multivariate distribution for missing data and then draws imputations from conditional distributions. Fully conditional specification specifies multivariate imputation model on variable by variable basis by a set of conditional densities for each incomplete variable. FCS starts with initial imputation and then draws imputations by iterating over conditional densities, usually 10-20 is sufficient. FCS is used to JM when no suitable distribution can be found. In my analysis, I used chained equations for imputing data. It is chained because it can be used as a concatenation of univariate procedures. The method executes m-streams in parallel and generates one imputed data set for each of them. In this multivariate method instead of filling each missing value, this approach fills each of the missing-values with a set of plausible values.

These are then combined together for the main analysis and it is probably closer to the actual uncertainty in the missing values.

TABLE 4.18: Association after imputation of missing values

Covariate	Location	Lipid	Beta	Meta P value
Missing value imputed and PC projection+Gender	12_69809401	M_VLDL_L_quantile	-1.90E-01	7.42E-04
	12_69809401	M_VLDL_PL_quantile	-1.90E-01	7.70E-04
	12_69809401	M_VLDL_C_quantile	-1.81E-01	1.17E-03
	12_69809401	M_VLDL_CE_quantile	-1.79E-01	1.22E-03
	12_69809401	M_VLDL_FC_quantile	-1.79E-01	1.54E-03
	12_69809401	M_VLDL_TG_quantile	-1.76E-01	1.78E-03
	12_69809401	IDL_C_eFR_quantile	-1.57E-01	3.20E-03
	12_69809401	VLDL_TG_quantile	-1.64E-01	3.40E-03
	12_69809401	S_VLDL_L_quantile	-1.51E-01	5.33E-03
	12_69809401	XS_VLDL_L_quantile	-1.50E-01	5.33E-03

4.6 CNV analysis in *LCE3C-LCE3B*

Background : Psoriasis occurs in a person when the body starts making skin cells in an accelerated way. Skin cells are usually replenished every three to four months, but in psoriasis it happens every three to four days. This results in build-up of patches which is a common symptom of psoriasis. Although the exact mechanism is unknown, it is related to the body immune system. The immune system in this case attacks the healthy skin cells by mistake. Psoriasis also has a known genetic element, however the exact role of genetics remains unclear upto now. Some recent studies have indicated that the late cornified envelope (*LCE*) gene within the epidermal differentiation complex on chromosome 1 is associated with psoriasis in European populations [245]. This study replicated associations of deletions in the *LCE3C LCE3B* gene with psoriasis and reported strong LD for three SNPs with the CNV. They further characterized a GGA deletion haplotype strongly associated with psoriasis [246].

***LCE3C-LCE3B* association with metabolomics** In my preliminary analysis, I found that, similarly to the *TSPAN8* gene there was a strong trend in association with lipids near the *LCE3C-LCE3B* region (Figure 4.22, 4.23, 4.24). In particular, I found association between urea and CNVs in the *LCE3C-LCE3B* region in NFBC 1966. Interestingly, urea is also a known treatment for psoriasis. Another interesting fact is that there is also a known link between psoriasis and Crohns disease, and there have been earlier reports of elevated levels of urea in patients with Crohns disease [247]. The association of CNVs in the general population with urea in *LCE3C-LCE3B* region is thus an exciting observation, which can perhaps relate urea regulation and susceptibility to psoriasis and Crohn's disease. These results are similar to the results I reported for the *TSPAN8* gene in this chapter and will require further analysis for confirmation. In addition. These results again underline the power of characterizing genomics (or other omics a well) for normal populations, which can reveal important information for diseases susceptibility. Replication of this association in other cohorts and functional follow-up studies remain to be done.

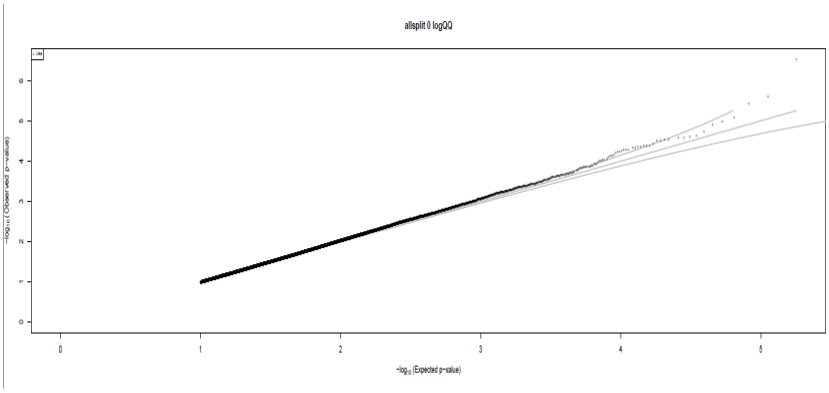


FIGURE 4.20: QQ plot for urea in NFBC 1986

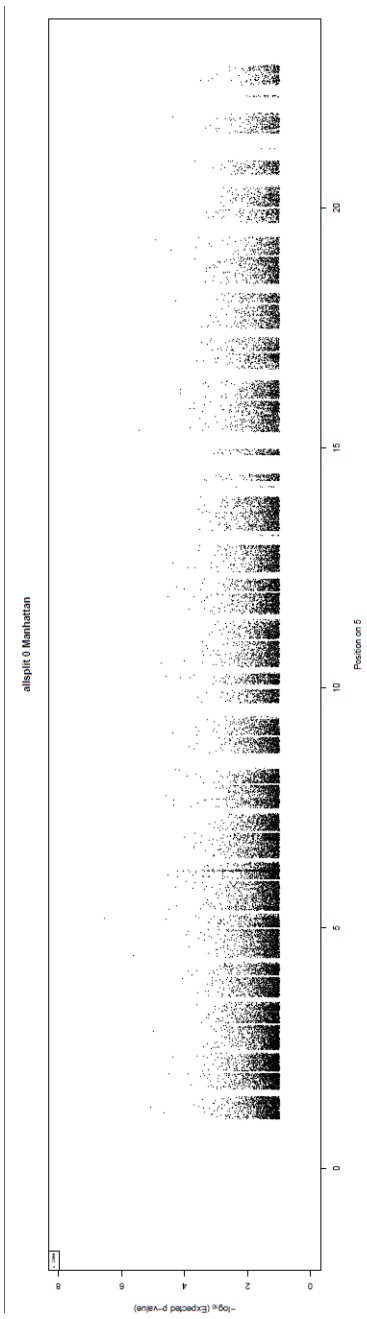


FIGURE 4.21: Manhattan plot for urea in NFBC 1986

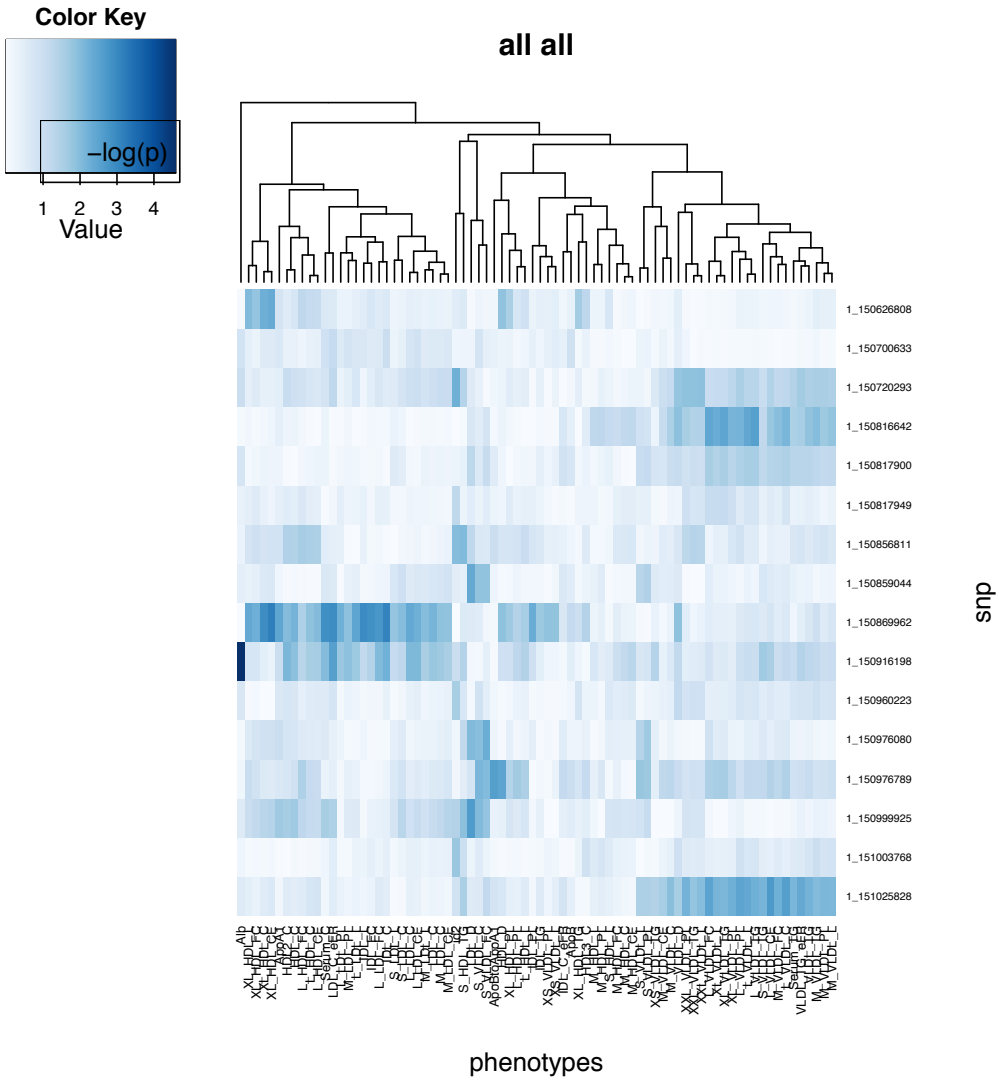


FIGURE 4.22: Heatmap for association with lipids near the *LCE3C* gene in NFBC 1986

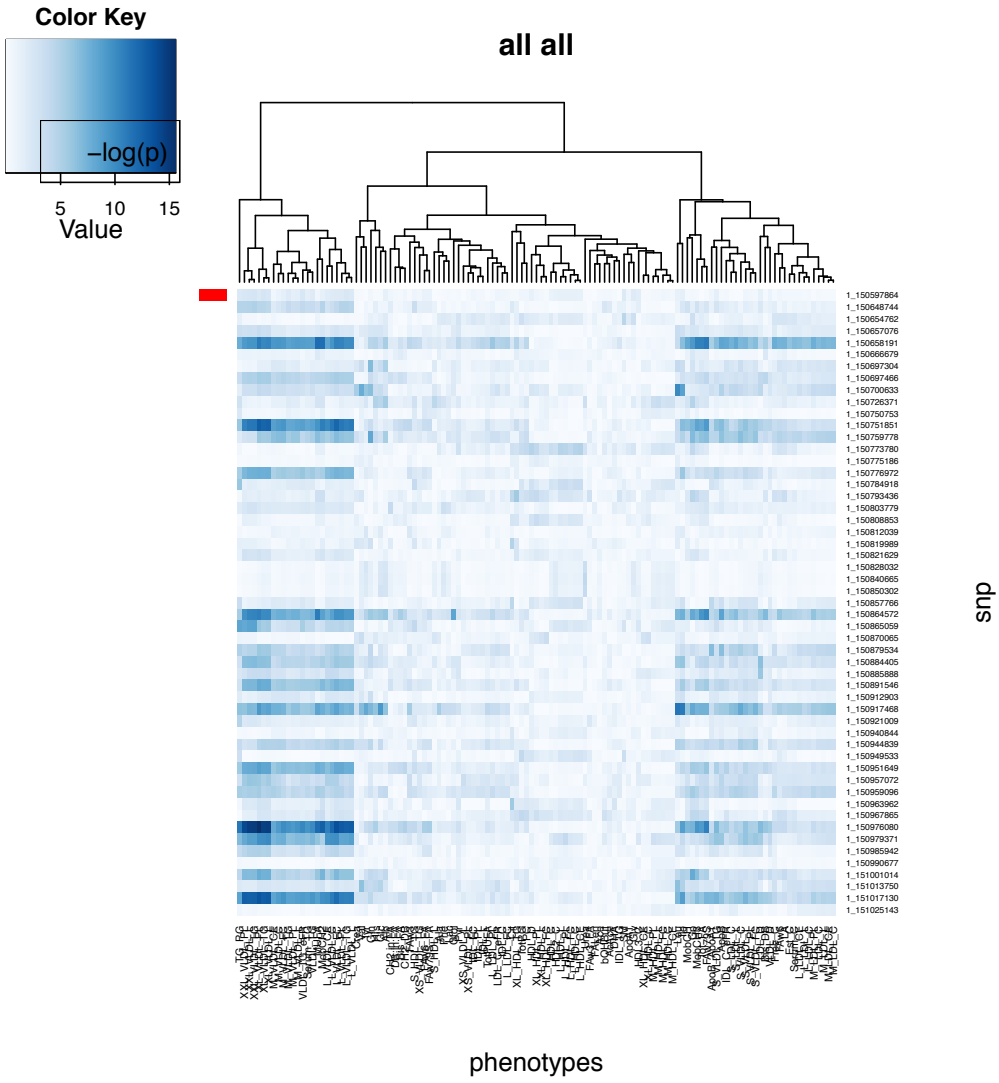


FIGURE 4.23: Heatmap for association with lipids near the *LCE3C* gene in NFBC 1966

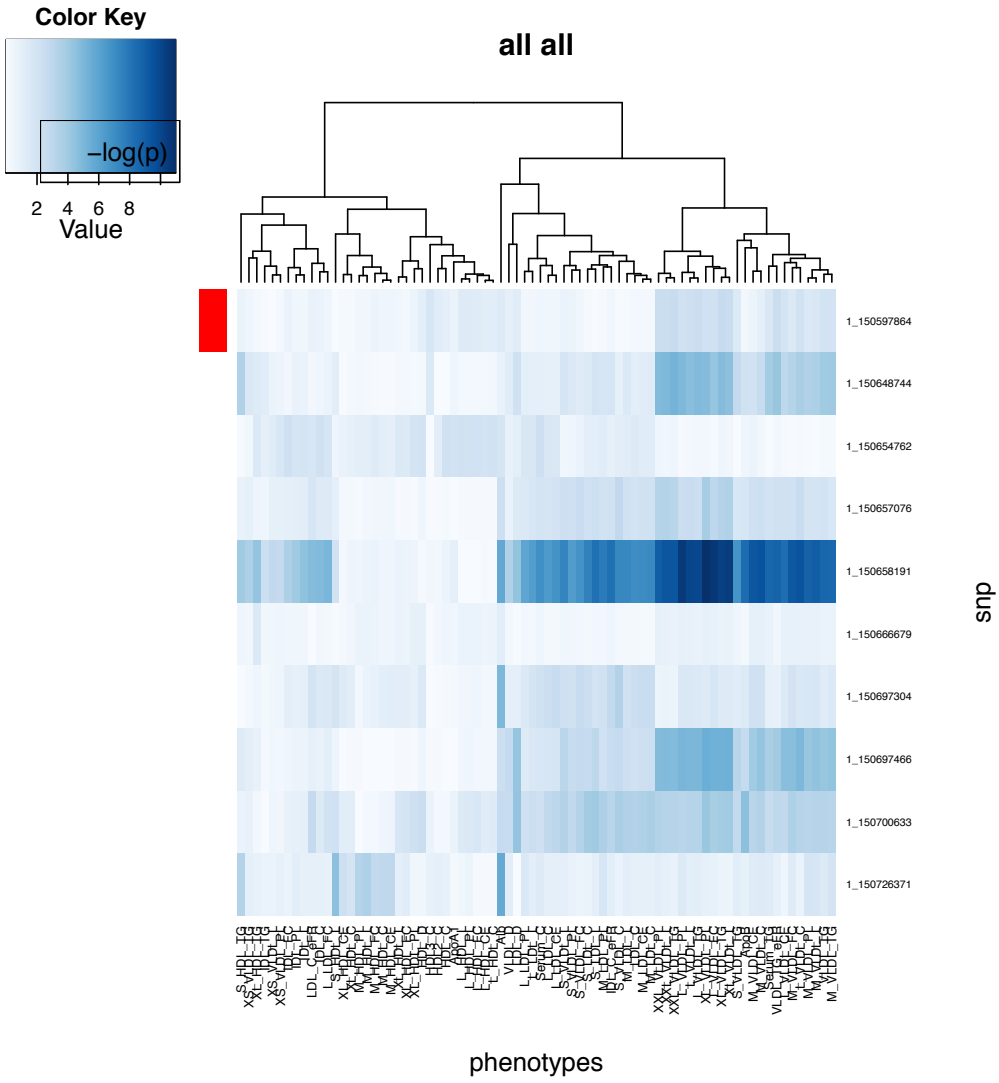


FIGURE 4.24: Heatmap for association with lipids near the *LCE3C* gene in joint analysis of NFBC 1966 and NFBC 1986

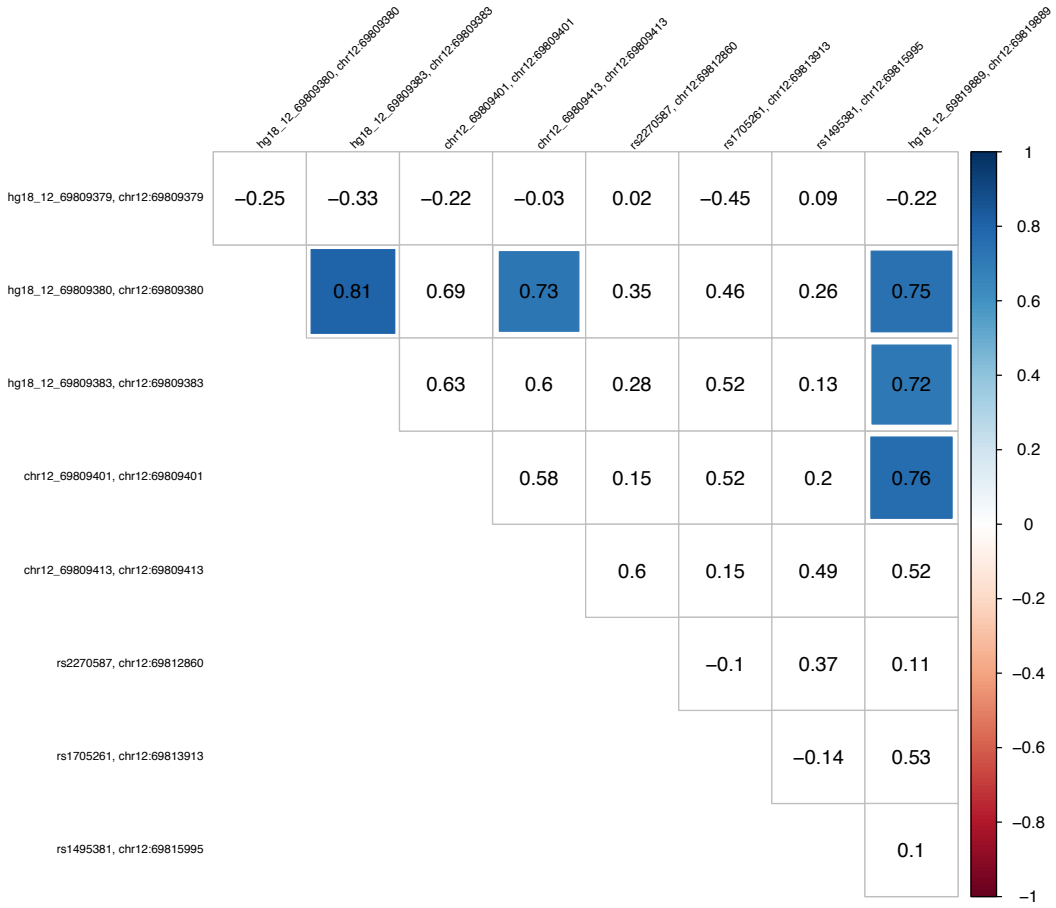


FIGURE 4.25: Correlation matrix plot showing the LD between the lipid associated region chr12:69809401-69812861 and CNVR5583.1 in NFBC 1986

Chapter 5

CNV Association with Gene Expression

5.1 Introduction

In this chapter, I have analysed the effect of CNVs on gene expression. This relationship is known to play an important role in regulating gene products and phenotypes. Such association approaches try to find the regulation of gene expression by SNPs and in our case CNVs. It does not look directly into the association with any of the disease traits or phenotypes but nonetheless results from such an analysis can indirectly reveal important information on diseases [248]. Most GWAS studies have found that the SNPs or CNVs associated with diseases do not seem to act through changes in the protein levels hence the mediation through regulation of gene expression remains a strong possibility. To test and further prove such an effect it will require further reanalysis of earlier GWAS candidate SNPs in light of eQTLs or CNV-eQTLs. To enable such an approach, underlines the rationale of my analysis presented here. In addition disease susceptibility has also been found to be tissue specific eg. variants associated with obesity where expression has been found to be mainly present in the adipose tissue [249, 250]. This observation mandates studies like ours where we aimed to find the association of CNVs with gene expression simultaneously in multiple tissues.

I have divided this chapter into two sections. In the first section, I have presented results from an association analyses of CNVs with gene expression in different brain regions from the UK brain expression consortium (UKBEC) project [251]. In the first few sections of this part, I have reviewed some of the background information and recent developments in the field of transcriptomics and highlight their important role in shaping human phenotypes and disease outcomes. These include some of the studies and findings from tissue and population specific analysis reported earlier. I have also discussed some of the current NGS approaches for transcriptomics and further discuss methods for association of gene expression with CNVs. In the second part of this chapter, I have presented some of the methods for carrying out association using simple linear models in Java. Such models can be generalized and can also include approaches for CNV-eQTL analysis. In addition I have performed some test analysis on the graphics processing unit (GPU) through the Java interface (jcuda) and further discuss how this can help in developing linear models through the GPU technology. This kind of framework if successfully built and perfected can perhaps be deployed in the future for analysis of large-scale omics data in a time efficient manner.

5.2 Section 1: CNV association with gene expression

5.2.1 Gene expression and phenotypes

All cells in an organism have the same DNA code but these cells are themselves quite different from each other. One of the main reasons behind this observed phenomenon is due to the expression of different RNA molecules in different cells. RNA molecules vary in quantity as well as quality (type), which gives rise to various different cellular characteristics. This variation has also been associated with various diseases outcomes like cancer [252]. An in-depth analysis and characterization of such variation and differences between different cells remain a crucial aspect to be addressed for the realisation of personalised medicine and for curing diseases. Most GWAS studies rely on the static nature of DNA and try to find its association with various phenotypes. However, this static DNA translates into lots of intermediate molecular steps which can vary with time, environmental and cellular factors. Usually, the link between the DNA and disease phenotype takes place through such molecular mechanisms but the current analysis methods are usually agnostic to such factors. In addition, complex disease phenotypes are often

associated with multiple variants of modest effect size detected indirectly through LD. One way to address this issue is to disentangle the true causal variant by finding out its effect on gene expression [248, 253]. This will allow finding how the variation in the regulatory regions is affecting the phenotypes and perhaps also find other associated traits [254, 255]. This will also allow taking into consideration the way regulatory elements act in a tissue specific manner, which can help in characterizing the disease association traits in greater details. In the future, these details need to be unravelled in order to pin point the true causal factors for diseases and developing targeted cure for them. Further, efforts for large-scale integrated analysis of biomarkers and gene expression will also help in identifying in greater details the molecular mechanisms and pathways through which the disease manifestation takes place.

5.2.2 eQTL studies

In genomics, eQTL studies try to find the effect of genomic variants like SNPs on gene expression and are sometimes referred to as genetics of gene expression [256]. Systematic analysis through these studies has led to the discovery of many regulatory elements and their location, tissue specificity of expression and further contribution of *cis* and *trans* elements. *cis*-elements are currently defined as SNPs associated with expression (refers to CNVs for this analysis) which are located near the gene and are most likely to be located within 100 kb of transcription start site (TSS) of their respective genes. While *trans*-elements are distal SNPs or CNVs which are located farther away and sometimes can also be located on a different chromosome. Both *cis* and *trans* elements have their own characteristic features on expression and can manifest influence in a complex manner [257]. Some studies have shown through expression profiling of *cis* elements that proximal TSS which are involved in RNA-polymerase-II recruitment can lead to subsequent transcription regulation which has greater influence on expression than the distal elements. On the other hand, *trans*-elements have weaker effect compared to *cis*-elements but they can cumulatively explain greater heritability [258]. Thus, we see that though the *cis* and *trans* elements can confer different levels of influence they can be equally important from a functional point of view.

RNA molecules are much harder to characterize than DNA molecules due their dynamic nature and diversity. This complicates eQTL analysis. To address this, current

studies have increasingly tried to analyse eQTLs based on condition [259] and tissue specificity [260] in the first instance. One of the advantages of looking into the degree of sharing of eQTLs across various conditions and tissues to allow the determination of regulatory elements. This analysis often applies network-based architectures [261] and several results through such methods have shown that there can be up to 50% sharing of eQTLs between different tissues [262]. Another study found that even if the eQTLs are shared at a statistically significant level in similar tissues their mean expression varies, hence indicating the importance of integration of results across various tissues as well [263]. Some studies have also shown the condition specific eQTLs where the degree of sharing of eQTLs across different conditions allows the estimation of regulatory elements and also further build various disease models from it. The ENCODE study concluded that about 69-80 % of the cis-eQTLs can be cell specific, even amongst closely related cells [264]. They also indicated that enhancers are likely to be tissue-specific but similar across different lineages. To this end, several studies have found 50% sharing of cis-eQTLs between adipose and blood tissues [265]. These findings show the complex nature of RNA biology and the importance of taking into consideration tissue type, conditions and related features during analysis.

5.2.3 CNV eQTL studies

CNVs have known implications in Mendelian disorders as well as for complex diseases outcomes with polygenic causes. They are also known to affect gene expression and metabolic traits in mice [266] and humans. The landmark study for CNVs by the Wellcome Trust found several CNVs associated with T2D and Crohns disease but their overall conclusion was that common CNVs are unlikely to have a causal role for common diseases. They also reported that since there is strong LD between common CNVs and SNPs, their association with disease outcomes have already been explored indirectly through earlier GWAS analysis [76]. Some examples include a 20 kbps deletion with $r^2 = 1$ with a SNP which was shown to be associated with Chrons disease [267]. The deletion in this case was found to be the actual causal-variant. Another such example is the *NEGR1* gene where SNPs as well as a 43-kb deletion were found to be associated with severe obesity [268]. Despite the WTCCC results there have been several other studies which have pointed out that the role of common CNVs should not be overlooked. CNVs have known effect on gene expression through dosage effect and can manifest as complex disease phenotypes at the molecular level, as I describe earlier. This is further

corroborated by the fact that most CNVs lie in the non-coding part of the genome and their possible effect on phenotypes through gene expression remains plausible. One particular study looked specifically into the overlap of common CNVs and eQTLs and reported several important conclusions and observations for CNV-eQTLs. This includes results showing that CNVs are enriched in the eQTL regions in a significant way and further found that out of the 3600 CNVs reported by the WTCCC study, about 1306 of them had at least one eQTL SNP in that region [269]. They report that CNVs which are less well tagged by SNPs ($r^2 < 0.3$)) contain about 44% of the SNP-eQTLs and are further enriched for reproducible trait associations. They further add that CNV which have functional effects are enriched for cis-eQTLs and are not usually well tagged by SNPs. Lastly, we should keep in mind that, of all the CNVs discovered by the WTCCC study, about 60 % of them were discarded since they could not be assigned to any specific copy-number class. These facts and results suggest that a deep characterization of the relationship of CNVs and eQTL regions still remain unknown.

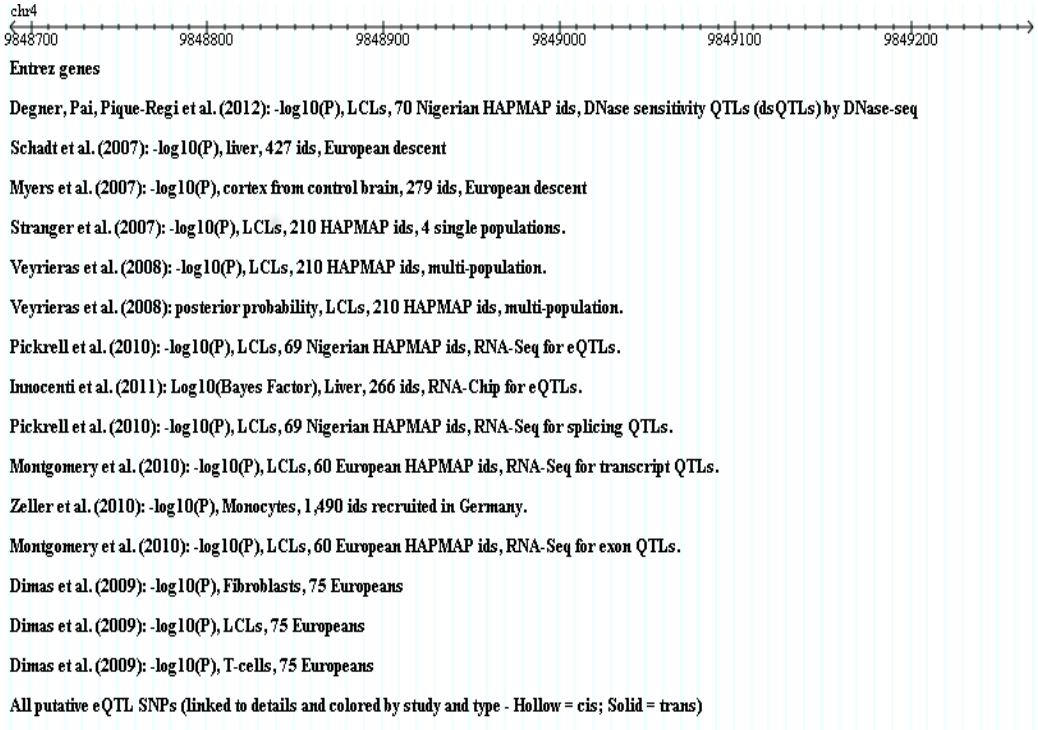


FIGURE 5.1: Previous reports of SNP eQTLs on chromosome 4

The region shown above contains a chromosome 4 duplication which is also one the top CNV-eQTLs reported later

5.2.4 Population specific gene expression

Current genomic findings for diseases and other human traits are mainly based on European populations. However, for full functional characterization of such variants they also need to be analysed in other populations. Population specific LD patterns, physical traits and susceptibility to diseases are well known problems to be addressed in the current genomic community. Many eQTL studies and other methods for characterizing gene expression have been helpful in characterizing such inter-population differences [270]. Though earlier studies have been mainly restricted to reference samples from the Hapmap study in the past, but at present it is now increasingly becoming more important to take into consideration the genomic variations existing across other populations as well. Data and results from the 1KG project which currently has sequenced samples from about 21 different populations can prove to be a useful resource in this purpose [105].

5.2.5 RNA-Seq

All DNA in our cells have two different alleles, one from each parent. Mutations or difference in any one of these alleles can directly lead to difference in the RNA expression. Some of the current research has shown that allele-specific transcripts are abundant in the cell, hence suggesting the possibility that some of this allelic-imbalance can have effect on disease outcomes as well [271]. In addition, integration of allele-specific expression from RNA-Seq and linkage-analysis has shown that about 20% of the *trans* regulators have allelic imbalance on target genes [263]. The earlier technology based on micro-arrays was not able to disentangle this allelic imbalance and can only quantify the total expression, irrespective of the allelic-differences. The development of the RNA-Seq technology can now accurately quantify RNA at a higher resolution as well as differentiate the allele-specific effects. In addition, haplotype-specific isoform information can be further derived from it. Some of the applications of such allele-specific expression aims to test whether one of these alleles is associated with higher or lower levels of gene expression or in analogy with SNPs, to characterize the heterozygous vs homozygous allele-specific expression. Sequencing approach through RNA-Seq works in a similar way as other DNA sequencing methods. RNA is first extracted and pre-processed and then shredded into millions of fragments. These fragments are then amplified and mapped back to a reference genome. There are usually two options of mapping the RNA-Seq.

This can either be the phased personal genome of the person being analysed or the publicly available reference genome. Both these approaches consist of two main areas: transcript reconstruction and mapping the transcripts back to the reference. This step poses several big statistical and computational challenges and at present is confounded by several factors. Some of these include difference in the source of the RNA, noise due to incomplete processed RNA and problems in PCR due to high GC content. Counter-intuitive to the notion that deep sequencing might be better, it should be kept in mind that many of these problems increase with the depth-coverage, hence care should be taken in designing the experiments and analysis [272].

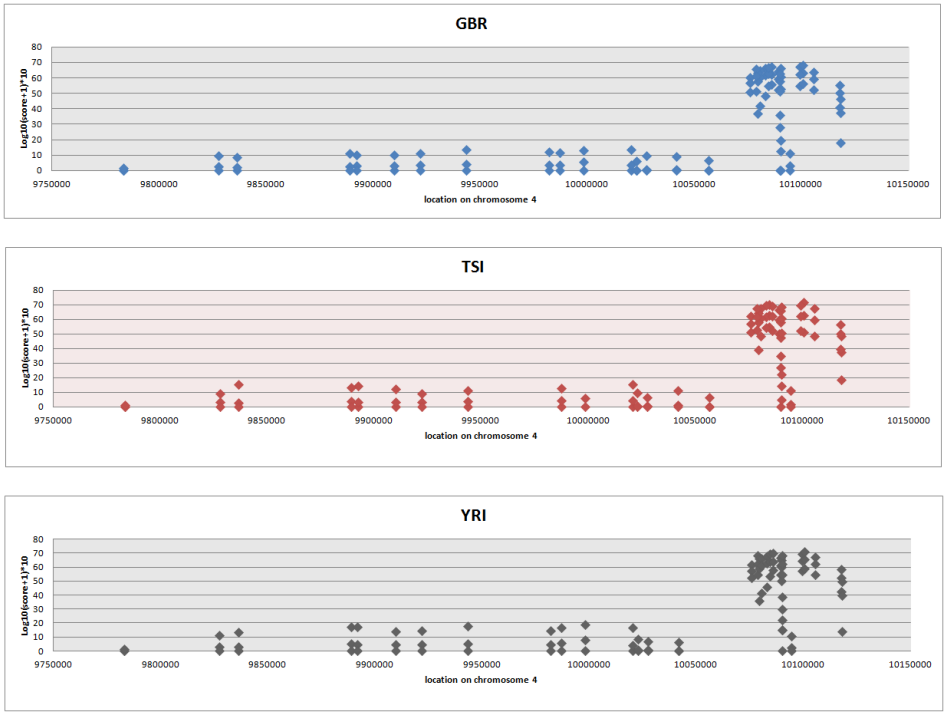


FIGURE 5.2: RNA-Seq results for expression around *DRD5* in 3 different populations

A strong CNV-eQTL for the *DRD5* is reported in the results later. Here the X axis is genomic location and the Y axis is the expression score (log scaled) for British (GBR), Yoruba in Ibadan, Nigera (YRI) and Toscani in Italia (TSI)

5.2.6 Methods for eQTL analysis

Present methods for associating gene expression with genotypes can involve billions of association comparisons, which can make such analysis for large datasets computationally prohibitive [8]. Due to this limitation, large-scale eQTL analysis has remained limited till now. At present many statistical methods are available for calculating the association metrics and some of these include linear models, ANOVA, GLMs, mixed models and also Bayesian methods [273]. The possibility of applying multivariate methods to determine simultaneous effect of all SNPs on expression profile of an individual also remains to be explored. Some of the current methods used for eQTL analysis are plink [274], R/qtl [275] and fastMap [276]. But most of these methods face the problem of computational bottlenecks, and can sometimes take enormous amount of time to run for large-scale studies.

Matrix eQTL software The matrix eQTL software was initially published as an ultra-fast method for doing eQTL analysis [8]. Some of the recent versions have also been developed for multi-tissue comparisons and these methods are planned to be used for several large-scale consortium. The general approach implemented in the eQTL software is to first test for each gene-SNP pair and then correct for multiple testing. For association, first a simple linear model is solved and then a test statistics is calculated. Final results are reported after checking if the test statistics exceeds a particular significance level. Matrix eQTL gains its speed by doing the computationally most intense part of the algorithm in terms of fast matrix-multiplications. It first splits the main large data matrix consisting of all SNPs and gene expression into chunks of $10k * 10k$ sub-matrices and then carries out the association analysis on these sub-matrices. Test statistic is calculated only for values which are above a certain threshold level, thus further saving running time. The matrix multiplication routine is implemented using BLAS (Basic Linear Algebra Sub programs), which has a running time of $O(n^3)$. For the matrix multiplications routine, several alternative methods have been developed which have lower running time complexity $O(n^{2.76})$ [277], but for practical issues these are not widely used at present.

5.3 UKBEC Analysis

The UK Brain Expression consortium was set up to study the regulation and alternative splicing of gene expression in the human brain [251]. This consortium initially collected brains from 134 subjects who were free from any kind of neuro-degenerative disease. For each individual, twelve different brain regions were extracted for further detailed analysis using expression-profiling approaches. These included array and sequencing based mRNA quantification methods. Array-based analysis has already been done and in the future RNA-Seq based profiling is also planned. Genotyping of the same individuals was done on two Illumina platforms (described later). Using the matrix eQTL software, a SNP and CNV-eQTL analysis was done. In this chapter, I have focussed on the CNV-eQTL analysis. An overall detailed analysis plan is shown in Figure 5.3.

5.3.1 Materials and Methods

For all study subjects, tissues samples from the central nervous system were collected by the Medical Research Council (MRC) sudden death tissue bank, Edinburgh, UK and the Sun health corporation, USA. Fresh tissues samples were removed from the brains and were extracted during autopsy. These were next immediately flash frozen. After a detailed examination by a doctor, all samples had been declared as neuropathologically normal. A detailed review of the samples and procedures has been made in an earlier review [251]. For each sample ten different brain regions were extracted. These included cerebellar cortex (CRBL), frontal cortex (FCTX), hippocampus (HIPP), medulla (inferior olivary nucleus), occipital cortex (OCTX), putamen (PUTM), substantia nigra (SNIG), thalamus (THAL), temporal cortex (TCTX) and intralobular white matter (WHMT). Expression-profiling and association analysis were done for all regions separately as well as by pooling the data together.

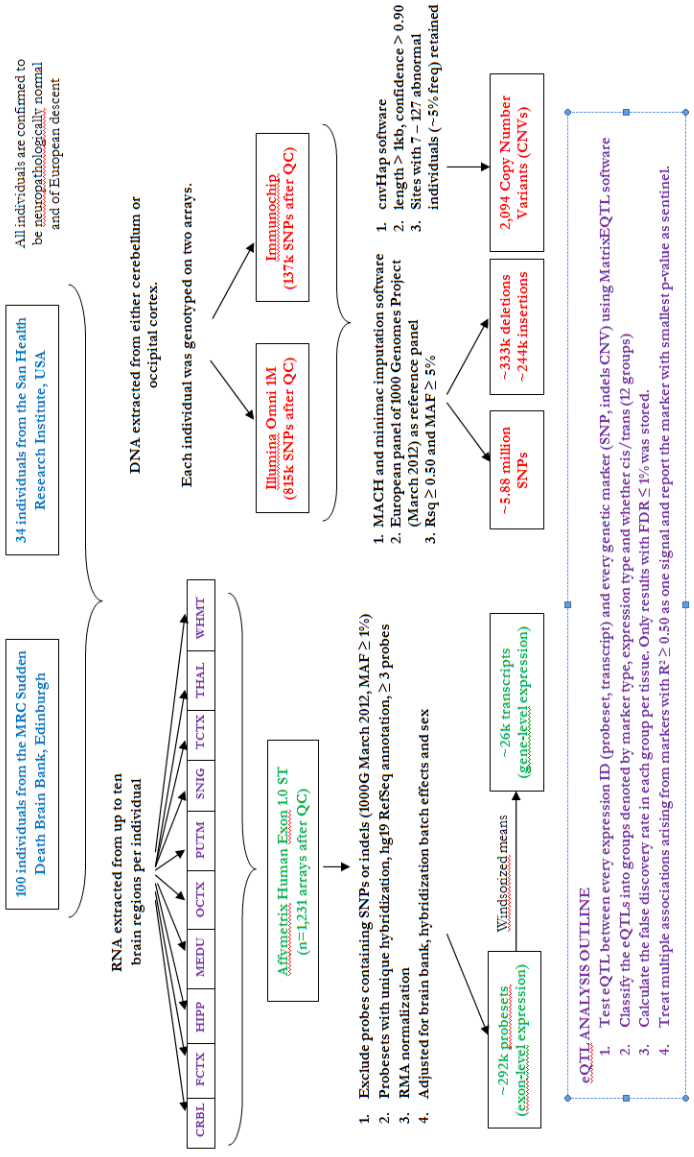


FIGURE 5.3: Overview of UKBEC association analysis

Source: http://ukbec.wordpress.com/

RNA extraction and processing A brief review of the RNA processing and genotyping is presented here. RNA was first extracted from the different post-mortem human brain tissues. This was based on a single-step RNA extraction procedure using custom miRNeasy kit. Quality check for the RNA was done using 2100-bio-analyzer and RNA 6000-Nano-kit. Next, hybridization was done using Affymetrix Exon 1ST-arrays following the manufacturers instructions. After hybridization, arrays were scanned using custom Affymetrix GeneChip-scanner-3000-7G and visually inspected. Pre-processing was done using the robust multi-array average (RMA) normalisation and log-transformed using the Affymetrix power-tools. After applying metrics like detection above background (DABG), the probes were re-mapped to the reference human genome (hg19). The analysis was restricted to 292k probes having proper gene names. In addition, probes which were not catalogued as common European SNPs ($MAF > 0.01$) or indels (1KG interim-phase-v3) were also excluded. After this, expression levels for 26k genes were calculated using the winzorized mean signal (below 10 % and above 90 %) for all probes sets corresponding to each gene. The final expression data were adjusted for gender, batch effects and source. DNA samples were also extracted from the brain tissues for genotyping. This was done on two chips which included the Illumina Infinium Omni1M quad custom chip and the Illumina Immunochip. After this step, scanning was done using iScan-Illumina and the generated data was imported and analysed through the Illumina genome studio software. In genome studio, further QC measures based on European ancestry and call-rate ≤ 95 % were applied. Finally, checks on gender, cryptic relatedness, autosomal heterozygosity rate and monomorphic SNPs were applied. The HWE disequilibrium p-value threshold was set to 1E-4.

5.3.2 Results

In this section I present some of the preliminary results from a pilot CNV-eQTL analysis done on all UKBEC samples. This analysis was done jointly with my colleagues in Kings College London. At present I am in the process of independently replicating some of these initial hits in other cohorts and in addition trying to implement an alternative and more robust method for performing CNV-eQTL association. This approach is based on raw intensity signals. In the following section, I have presented some of the CNV-eQTL hits from our initial analysis which survived initial filtering criteria based on 1 % FDR, and were further deemed as potential candidates for validation.

CNV segment at chr1:54557270-54560027 This CNV lies within the *TCEANC2* gene (transcription elongation factor). There are lots of other genes flanking this region. This CNV also has a cis-eQTL with TCTX region with the *DHCR24* gene which is 800 kb away. The *DHCR24* gene has been found to have reduced expression in TCTX of patients with Alzheimers disease. Further, this CNV has trans-eQTL with 21 other genes but mostly for TCTX. Functional analysis using DAVID tool (http://david.abcc.ncifcrf.gov/) did not reveal anything interesting.

Table 5.1: Association results from UKBEC CNV-eQTL analysis

CNVR	exprID	expr.chr:TSS	gene	dist2TSS	tissue	Z	pvalue	FDR
1:54557270-54560027	2413917	1:55352906	DHCR24	-795636	TCTX	6.379	1.80E-10	9.40E-05
	2413913	1:55352906	DHCR24	-795636	TCTX	6.109	1.00E-09	2.60E-04
	t2413907	1:55352906	DHCR24	-795636	TCTX	5.466	4.60E-08	2.30E-03
	2413918	1:55352906	DHCR24	-795636	TCTX	5.617	1.90E-08	3.40E-03
4:10239590-10240170	t2718259	4:9782216	DRD5	457374	aveALL	5.728	1.00E-08	1.70E-04
	2760720	4:10486017	ZNF518B	-246427	aveALL	5.586	2.30E-08	3.70E-03
	t2760626	4:10446650	ZNF518B	-207060	aveALL	-6.051	1.40E-09	7.10E-05
	2760627	4:10446650	ZNF518B	-207060	aveALL	-6.051	1.40E-09	7.50E-04
	t2760628	4:10447946	ZNF518B	-208356	aveALL	-5.729	1.00E-08	1.70E-04
	2760629	4:10447946	ZNF518B	-208356	aveALL	-5.729	1.00E-08	2.70E-03
	2760658	4:10686504	CLNK	-446914	aveALL	5.552	2.80E-08	3.70E-03
	t2760632	4:10686504	CLNK	-446914	aveALL	4.952	7.30E-07	9.10E-03
6:29275507-29286670	2947783	6:29390839	OR5V1	-115332	HIPP	5.65	4.70E-08	8.20E-03
	2947783	6:29390839	OR5V1	-104169	HIPP	5.65	4.70E-08	8.20E-03
	2947783	6:29390839	OR5V1	-51568	HIPP	5.65	4.70E-08	8.20E-03
18:41978081	3786614	18:42179671	SETBP1	-201590	FCTX	5.583	2.40E-08	6.20E-03

CNV segment chr4:10239590-10240170 This CNV lies on chromosome 4 intergenic-region. There are five genes around this CNV out of which three of them have significant cis-eQTLs namely *ZNF518B*, *DRD5* and *CLNK*. Results from Uniprot show that *ZNF518B* has known involvement in transcriptional regulation. *DRD5* (Dopamine receptor D5) is a strong candidate, expressed in the neurons of the limbic region of the brain. *CLNK* (cytokine-dependent hematopoietic cell linker) has known role in immunoreceptor signalling.

The chromosome 4 CNV was found to have large number of trans-eQTLs. There were 67,882 trans-eQTLs out of which 94 % had FDR $< 1\%$. The best trans-eQTL with this CNV was found to be *CACNA1A* (calcium channel, voltage-dependent, P/Q type, alpha 1A subunit) in Substantia Nigra and average-all category with FDR $< 2.04e-11$. *CACNA1A* is involved in a variety of calcium-dependent processes, including muscle contraction and functions as a neurotransmitter. The next best trans-eQTLs were found to be *MMP16* and *GRIA4*. *GRIA4* is a glutamate-receptor excitatory neurotransmitter in the brain and has been linked to schizophrenia.

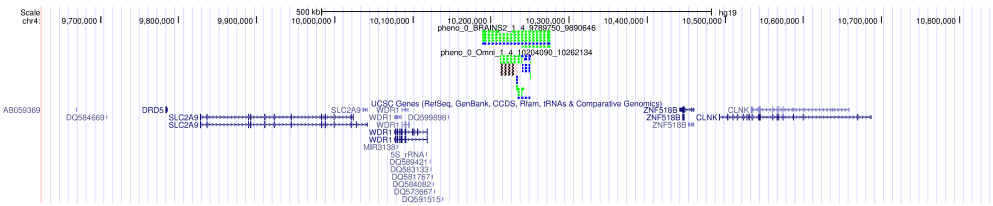


Figure 5.4: Chromosome 4 duplication region in UKBEC and NABEC

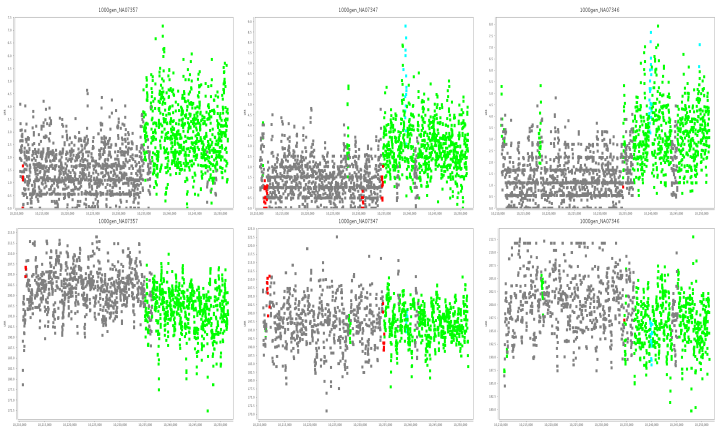


FIGURE 5.5: Chromosome 4 duplication in *Thousand Genomes* CEU samples

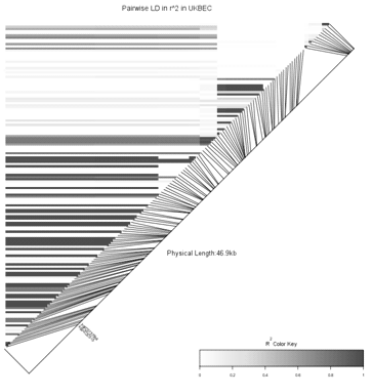


FIGURE 5.6: LD plot in UKBEC around *DRD5*

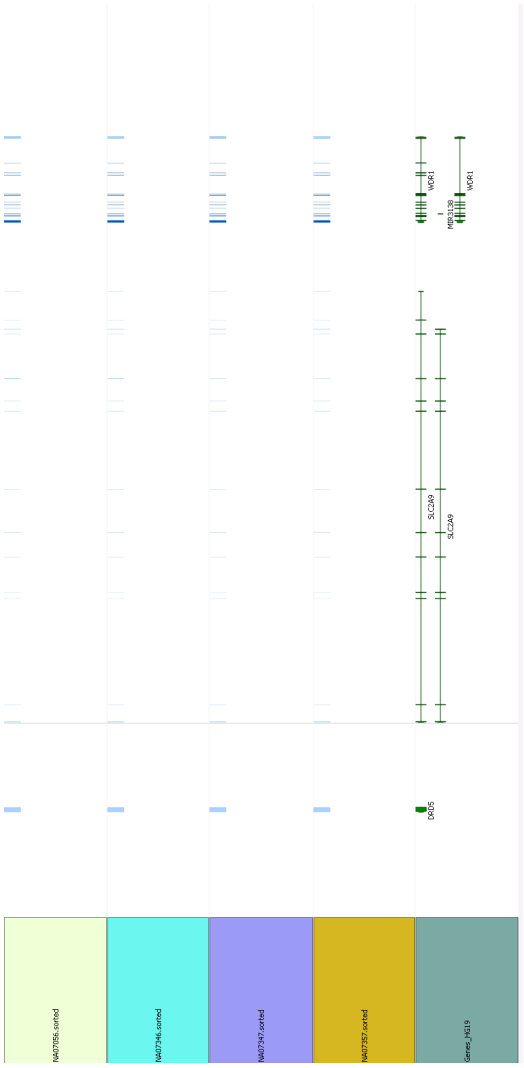


FIGURE 5.7: RNA-Seq results from *The 1000 Genomes CEU* population for samples having duplication at *DRD5*

The intensity of the blue colour above the genomic location on the X axis is directly correlated to the total number of reads mapped to that location.

CNV at chr6:29275507-29341623 We found three different CNVR spanning this region. This CNV has a cis-eQTL with *OR5V1* gene in hippocampus only. Also, no trans-eQTLs was found.

CNV at chr18:41978081-41981339 This CNV lies in an intergenic region. The nearest gene was found to be *SETBP1* (SET binding protein 1) located about 1.6 Mb away and this CNV has a cis-eQTL associated with frontal cortex. This CNV also has a cis-eQTL with *SYT4* (synaptotagmin IV) which has a cis-eQTL in frontal cortex and is known to be involved in calcium ion exocytosis. There were 213 trans-eQTLs associated with the frontal cortex and some with average-all category. A DAVID analysis showed that the best p-value for enrichment of 0.0212 with phosphoprotein ($p < 1E-8$) and alternate splicing ($p < 1E-6$)

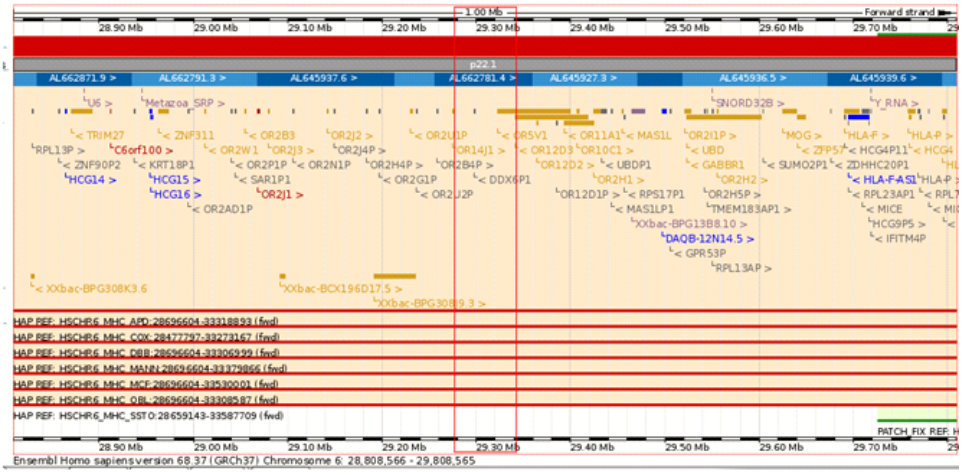


FIGURE 5.8: Ensembl plot of the region around chromosome 4 duplication

Source: http://www.ensembl.org/index.html

5.4 Section 2: CNV-eQTL methodology development

In the following sections, I have explored some preliminary background theory and results for linear models currently used in genomic studies. I have specifically discussed the association framework in such studies in light of matrix algebra. I have further explored some of the software implementation of such matrix routines which includes the GPU framework for computation and some specific highly optimized scientific libraries in Java.

5.5 Graphics processing unit (GPU)

Graphics processing units (GPU) are a recent parallel computing architecture which can deliver up to 1 teraflops (TFLOPs) of computing power on a single desktop computer. The Compute Unified Device Architecture (CUDA) [278] developed by Nvidia is a general data programming model which helps programmers to use the GPU architecture without delving into the instruction set details for GPUs. Alternately, the recently developed openCL provides a standard for common computing model not only for GPUs but also for other platforms like multicore and many core platforms from Intel.

The commonly used CUDA framework presents a heterogeneous programming model, where the GPU parallel hardware is used in conjunction with the CPU. In this case, the CPU decides the barrier of synchronization and requires no direct involvement from the user. The current GPU applications using frameworks like CUDA are designed in such a way that the compiler and the run-time can control about the parallelism. This framework allows for parallelism at an unprecedented scale and when applied for scientific computations, this can lead to several orders of improvement in the computational time. One possible application of this method lies in eQTL analysis, where current computation time for even modest datasets can take up to hours or days.

The matrix eQTL software developed is a powerful method which has been able to decrease this computation time by several orders of magnitudes. The main performance improvement for this software comes from the speed of matrix multiplication operation by using the BLAS routine. In the following sections,my aim is to use the GPU architecture for solving normal equations through matrix algebra. The use of normal equations can prove to be a powerful method for calculating association between two variables

of interest and can be further extended to machine learning problems like predictions. Similar to matrix eQTL approach, I plan to use the BLAS library. In this work I have used a version of a customized BLAS library for GPU known as *CUBLAS* provided by Nvidia. Jcuda is a Java wrapper made for the Nvidia CUDA and CUBLAS libraries, and I use this *jcuda* library [279] for obtaining the results presented in the following sections.

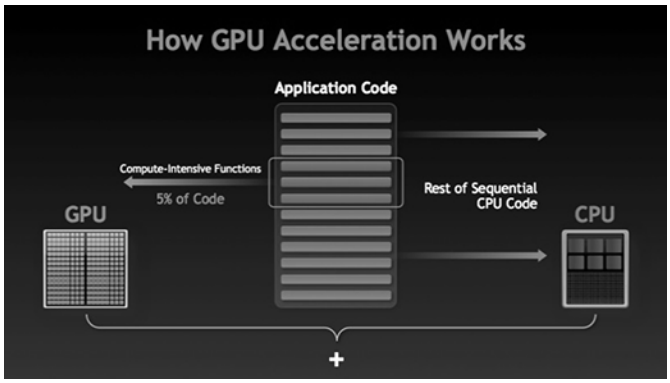


FIGURE 5.9: Overview of GPU acceleration framework

Source: http://www.nvidia.com/

5.6 Linear systems and matrix algebra

Huge datasets can take up lots of space can take up large amount of time for processing. Nowadays, this phenomenon has been popularly referred to as the *big data* problem. The big data issue is also currently prevalent in the current genomic studies and is likely to increase with the advent of NGS data and deep phenotyping through several omics approaches.

However, for analysing such kind of datasets in general one of the most common problem is to find a relationship between two variables X and Y and this can be interpreted as fitting a line through a set of data points. One simple form of this relationship is also known as the linear equation. This linear relationship idea can further easily be

interpolated to prediction problems in machine learning. In machine learning, there is usually a prior training step for determining the parameters of a model (which can be linear) and this is used to predict unknown observations. In the following sections, I have described some of the simple basic steps involved in finding parameters of a linear model. The linear model can be understood and can be solved through several domains in mathematics including vectors and 3D geometry, numerical analysis and matrix algebra. I have mainly described the the matrix algebra aspects of the linear model. This also happens to be the most common routine in current software implementations.

In linear algebra and statistics, one of the most important operations is matrix factorization which decomposes a given matrix into its canonical form. This process of producing the right-hand side product of canonical matrices is also called "matrix factorization". This concept has very important applications in linear algebra, applied statistics and also engineering. The objective of matrix-decomposition involve two main purposes 'computational convenience' and 'analytical simplicity'. One common current scientific application is that it is not possible to create a pragmatic solution in an explicit way for problems involving matrix operations such as matrix inversions, solving linear systems, finding matrix determinants and least square fitting. Thus, one approach is to convert the difficult matrix computation problem into several easier steps. Some of the common issues arising from this are solving triangular systems or diagonal systems. One method is to convert the matrix into some kind of lower order or lower rank canonical form, which then can reveal the inherent nature of the matrix. This can help in interpreting their meaning. In the next section, I have reviewed some of the common methods for matrix decomposition.

5.6.1 Methods for matrix decomposition

In the following section, I have described some of the basic theory behind matrix decomposition. This background helps in understanding their algorithmic implementation.

Singular value decomposition For a matrix $A\ \epsilon\mathbb{R}^{m\times n}$, there exists orthogonal matrices $U = (u_1, ...u_m)\epsilon\mathbb{R}^{m\times m}, V = (v_1, ...v_n)\epsilon\mathbb{R}^{n\times n}$ and a diagonal matrix $\sum = diag(\sigma_1, ...\sigma_p)\epsilon\mathbb{R}^{m\times n}$ with $p = min(m, n)$, such that $A = U\sum V^T$, and $\sigma_1 \geqslant \sigma_2 ... \geqslant \sigma_p \geqslant 0$

SVD decomposition helps in revealing many intrinsic properties of a given matrix in numerically stable manner.

LU and Cholesky decomposition The basic problem in linear algebra is often to find the following solution $x = A^{-1}b$. One of the easier method of solving this system is through LU decomposition. LU decomposition is defined as follows. Let there be a matrix $A\epsilon\mathbb{R}^{n\times n}$ and all the leading principal minors be $det(A(1:k,1:k)) \neq 0, k = 1,...n-1$, then there exists a lower triangular matrix L with diagonal elements all equal to one and a unique upper triangular matrix U such that $A = LU$ and $det(A) = u_{11}u_{22}...u_{nn}$. If matrix A has additional properties the LU decomposition will have particular forms. Cholesky decomposition is referred to as follows. If matrix $A\epsilon\mathbb{R}^{n\times n}$ is symmetric positive definite, then there exists a unique lower triangular $G\epsilon\mathbb{R}^{n\times n}$ with positive diagonal entries such that $A = GG^T$. Here G is referred to as Cholesky triangle.

Householder reflections If $v\epsilon\mathbb{R}^n$ is a nonzero vector and a n-by-n matrix P can be expanded as

$$P = I - 2vv^T/v^Tv$$

then this expansion is called the Householder reflection or Householder matrix or Householder transformation. The vector v is called the Householder vector.

Gram-Schmidt orthogonalization For matrix $A = (a_1,...a_n)\epsilon\mathbb{R}^{m\times n}$, there exits a linear independent set of vectors obtained by subtracting from the projections of a_k onto $a_i(i<k)$. Using an adequate normalization, we can gradually orthogonalize A to an orthonormal set $Q = (q_1,...q_m)$ as follows

$$\begin{array}{ll} q_1 = a_1/r_{11} & r_{11} = ||a_1|| \\ q_2 = (a_2 - r_{21}q_1)/r_{22} & r_{21} = x_2^T\ r_{22} = ||a2 - r_{21}q_1||_2 \\ ... = ... & .. \\ q_k = (a_k - \sum_{i=1}^{k-1} r_{ik}q_i/r_{kk}) & r_{ik} = q_i^T a_k, r_{kk} = ||a_k - \sum_{i=1}^{k-1} r_{ik}q_i||_2 \end{array}$$

Here $q_k = \sum_{i=1}^{k}$ and A is the product of Q and upper triangular $R = (r_{ij})$. This process is called the Gram-Schmidt orthogonalization process.

QR decomposition For a matrix $A\epsilon\mathbb{R}^{m\times n}$, there exists an orthogonal matrix $Q\epsilon\mathbb{R}^{m\times m}$ and an upper triangular matrix $R\epsilon\mathbb{R}^{m\times n}$ such that

$$A = QR$$

Some of the above theorems can be used to derive the proofs for QR decomposition including Householder reflection, Givens rotations, and Gram-Schimdt orthogonalization. I avoid going into details for these steps.

5.6.2 Least squares solution

In overdetermined systems, for the equation $Ax = b$ with data matrix $A\epsilon\mathbb{R}^{m\times n}$ and vector $b\epsilon\mathbb{R}^{m}$ with $m > n$, there is no exact solution, if b is not an element of range(A). The goal of least squares (LS) fitting problem is to find $x\epsilon\mathbb{R}^{n}$ minimizing $J = ||Ax - b||$. If A does not have full column rank then the solution to the LS fitting is not unique, i.e if x minimizes J, then $x + z$, $z\epsilon null(A)$ is also a solution. If A has full column rank then the unique least squares solution x_{LS} is given by the pseudo inverse $x_{LS} = (A^T A)^{-1}A^T b$. As I mentioned before, inverse of a matrix is not easy to compute and is more conveniently solved by QR, LU and SVD decomposition.

5.6.3 Linear regression in matrix form

I this section, I review the linear equations in matrix form. Equation 5.1 is the general linear equation commonly used

$$Y_i = \beta_0 + \beta_1 X_1 + \epsilon_1 \tag{5.1}$$

where

$$\epsilon_i \sim^{iid} N(0, \sigma^2) \tag{5.2}$$

Equation 5.1 can be rewritten in matrix form as follows

$$\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} = \begin{bmatrix} \beta_0 + \beta_1 X_1 + \epsilon_1 \\ \beta_0 + \beta_1 X_2 + \epsilon_2 \\ ... \\ \beta_0 + \beta_1 X_n + \epsilon_n \end{bmatrix} \tag{5.3}$$

$$\begin{bmatrix} Y_1 \\ Y_2 \\ ... \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ ... & \\ 1 & X_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{bmatrix} \tag{5.4}$$

In equation 5.3, X_{nx2} is the design matrix, β_{nx1} is the vector of parameters, ϵ_{nx1} is the error term and Y_{nx1} is the response vector.

Parameter estimation One popular method for estimating parameters in equation 5.3 is the least squares estimate. For equation 5.3, we want to minimize the squared residuals $\epsilon = Y - X\beta$:

$$\sigma\epsilon^2 = [\epsilon_1 \epsilon 2 ... \epsilon n] \begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ ... \\ \epsilon_n \end{bmatrix} = e^T e \tag{5.5}$$

Minimizing $\epsilon^T \epsilon = (Y - X\beta)^T (Y - X\beta)$ requires differentiating and setting the derivatives equal to 0, this provides the solution

$$b = (X^T X)^{-1} X^T Y \tag{5.6}$$

In equation 5.5, $(X^T X)^{-1} X^T$ is also called the hat matrix. The hat matrix helps convert the response values to the predicted values of Y

$$\hat{Y} = Xb$$
$$\hat{Y} = X(X'X)^{-1}X'Y$$

$$\hat{Y} = HY \tag{5.7}$$

Using multivariate theorem it can be shown that the above framework can also be extended to multiple regression problems. Once a least squares estimate is determined, a statistical test, e.g. F-test, can be used to calculate p-values and to determine if any of the explanatory variables can be used to model the response variable in a significant way. Another commonly used metric is the R^2, which is used to determine the proportion of variation in the response variable explained by the explanatory variables.

Variance-covariance matrix In general for any set of variables $U_1, U_2, ..., U_n$ the variance-covariance matrix is defined as

$$\sigma^2\{U\} = \begin{bmatrix} \sigma^2\{U_1\} & \sigma\{U_1, U_2\} & ... & \sigma\{U_1, U_n\} \\ \sigma\{U_2, U_1\} & \sigma^2\{U_2\} & ... & ... \\ ... & ... & ... & \sigma\{U_{n-1}, U_n\} \\ \sigma\{U_n, U_1\} & ... & \sigma\{U_n, U_{n-1}\} & \sigma^2\{U_n\} \end{bmatrix} \tag{5.8}$$

Here $\sigma^2\{U_i\}$ is the variance of U_i and $\sigma\{U_i, U_j\}$ is the covariance of U_i and U_j. This matrix has an important role in determining the correlation structure between variables. When variables are uncorrelated, the covariance is 0. The variance-covariance matrix of uncorrelated variables will be a diagonal matrix since all covariances are 0. When variables are independent they will also be automatically uncorrelated. But it should be kept in mind that it is possible to have variables which are dependent but uncorrelated since correlation only measures linear dependence.

5.7 Results

Matrix multiplication is an important step for doing correlation analysis for any given two variables and for calculating the hat matrix in the normal equations. In this section, I have presented some of the preliminary results comparing the GPU and CPU operations. Figure 5.11 shows a plot comparing GPU vs CPU time for matrix multiplication. It can be seen that for matrix size of up to 600x600, the CPU time is better than GPU. The reason for this is though the actual computation time for the GPU is faster than the CPU by several magnitudes, there is an overhead for copying the data from CPU to the GPU chip. Due to this overhead, for smaller matrix sizes the computation time is better for CPU than GPU. But once the matrix size gets bigger ($> 700x700$), then the GPU outperforms the CPU by up to 9 times. This speed up is comparable to the results reported by the matrix eQTL paper. Since the time for matrix multiplication almost scales exponentially for the CPU, as the matrix size increases, the time for GPU will also get better. This makes an ideal solution for analysis of large-scale datasets.

Matrix inversion is an important step for various scientific computations, including for regression and matrix decomposition analysis. In this section, I have compared the matrix-decomposition times for CPU and GPU for various matrix sizes. Figure 5.12 shows a plot comparing CPU vs GPU time for matrix inversion. Unlike matrix multiplication the time for CPU is consistently better than GPU in this case for matrix size up to 4096. However, on closer examination reveals that for matrix size 500x500 the CPU is better than GPU by 3.73 times, but for matrix size 4096x4096, the gap is only 2.64 times. From the graph we can see though CPU is performing better but the gap is slowly decreasing. Thus it can be reasonably argued that for higher matrix sizes, GPU speed will eventually be better than CPU, but for lower matrix sizes the computational overhead for copying data from CPU to GPU hinders the speed up. An online research through discussion forums confirms this and it is reported that the real speed up of GPU for matrix inversion is only observed for matrices greater than $10Kx10K$. A test-harness and improvement of the codes and algorithms remain to be done in the future. I also report some benchmarking results for various matrix operations done purely in Java and implemented through a modified JAMA Java library packages. This includes various results for linear solvers (Table 5.3), matrix decomposition (Table 5.4) and times for matrix inversions (Table 5.5). These results can help for developing various routines and algorithms GPU in conjunction with various other Java libraries.

```
Found 1 devices
Device 0: Quadro FX 380 with Compute Capability 1.1
    Maximum number of threads per block                 : 512
    Maximum x-dimension of a block                      : 512
    Maximum y-dimension of a block                      : 512
    Maximum z-dimension of a block                      : 64
    Maximum x-dimension of a grid                       : 65535
    Maximum y-dimension of a grid                       : 65535
    Maximum z-dimension of a grid                       : 1
    Maximum shared memory per thread block in bytes     : 16384
    Total constant memory on the device in bytes        : 65536
    Warp size in threads                                : 32
    Maximum pitch in bytes allowed for memory copies    : 2147483647
    Maximum number of 32-bit registers per thread block : 8192
    Clock frequency in kilohertz                        : 1100000
    Alignment requirement                               : 256
    Number of multiprocessors on the device             : 2
    Whether there is a run time limit on kernels        : 1
    Device is integrated with host memory               : 0
    Device can map host memory into CUDA address space  : 1
    Compute mode                                        : 0
    Maximum 1D texture width                            : 8192
    Maximum 2D texture width                            : 65536
    aximum 2D texture height                            : 32768
    Maximum 3D texture width                            : 2048
    aximum 3D texture height                            : 2048
    Maximum 3D texture depth                            : 2048
    Maximum 2D layered texture width                    : 8192
    Maximum 2D layered texture height                   : 8192
    Maximum layers in a 2D layered texture              : 512
    Alignment requirement for surfaces                  : 256
    Device can execute multiple kernels concurrently    : 0
    Device has ECC support enabled                      : 0
    PCI bus ID of the device                            : 15
    PCI device ID of the device                         : 0
    Device is using TCC driver model                    : 0
    Peak memory clock frequency in kilohertz            : 700000
    Global memory bus width in bits                     : 128
    Size of L2 cache in bytes                           : 0
    Maximum resident threads per multiprocessor         : 768
    Number of asynchronous engines                      : 1
    Device shares a unified address space with the host : 0
    Maximum 1D layered texture width                    : 8192
    Maximum layers in a 1D layered texture              : 512
    PCI domain ID of the device                         : 0
```

FIGURE 5.10: GPU configuration

Device query results for the GPU chip used to obtain the results for the following sections

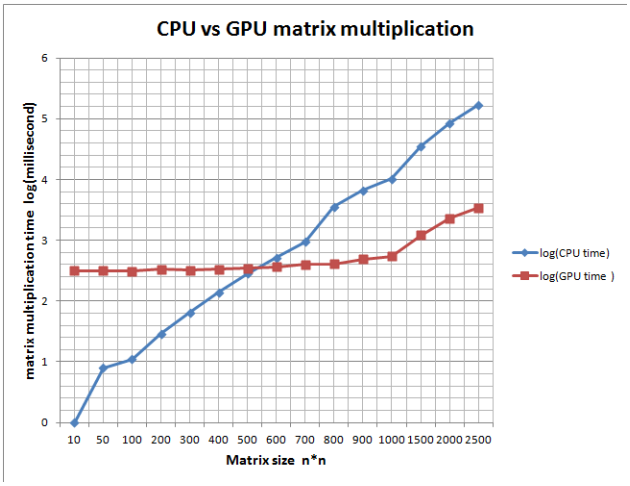


FIGURE 5.11: CPU vs GPU matrix multiplication time

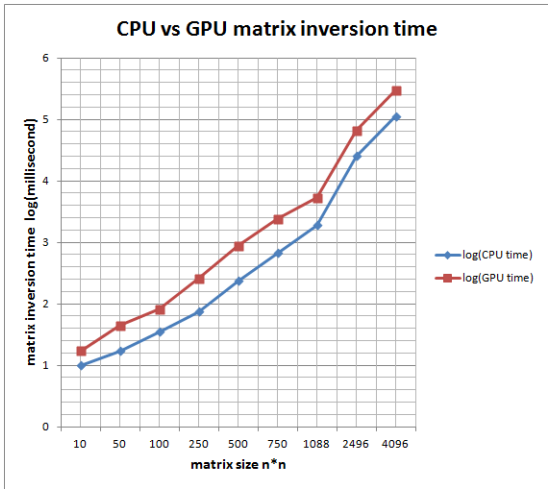


FIGURE 5.12: CPU vs GPU matrix inversion time

TABLE 5.2: Comparison of linear solvers

Matrix Size	**No of trials**	**Solving method for matrix A**	**Time**	**No of trials**	**Solving method for matrix B**	**Time**
		LU A	844		LU A	608
		LU B	1320		LU B	589
10	1000000	QR house	1017	200000	QR house	807
		QR house	2280		QR house	2140
		PInv	342		PInv	377
		LU A	88		LU A	2181
		LU B	192		LU B	1560
100	2000	QR house	117	70000	QR house	2866
		QR house	307		QR house	7522
		PInv	55		PInv	999
		LU A	34		LU A	1234
		LU B	102		LU B	910
1000	8	QR house	86	4000	QR house	1659
		QR house	127		QR house	4278
		PInv	22		PInv	549

Table 5.3: Comparison of LU decomposition

Matrix Size	No of trials	LU method	Time
10	1000000	LUDecompositionAlt	1757
		LUDecompositionNR	2799
100	2000	LUDecompositionAlt	1345
		LUDecompositionNR	2754
1000	3	LUDecompositionAlt	1774
		LUDecompositionNR	3556

Table 5.4: Comparison of QR decomposition

Decomposing size 40X10		
100000	column	1271
	tran	1306
	pivot	1381
	block	1886
Decomposing size 400X100		
400	column	2076
	tran	2208
	pivot	2097
	block	7069
Decomposing size 4000 X1000		
1	column	5391
	tran	5623
	pivot	5385
	block	6524

Table 5.5: Comparison of time for matrix inversion

Inverting size 10 for 1000000 trials	**Time**
LU	3329
LU NR	3343
Ops	5358
Inverting size 100 for 1000 trials	
LU	1425
LU NR	1436
Ops	2087
Inverting size 1000 for 3 trials	
LU	4060
LU NR	3950
Ops	5068

5.8 Future work

My main objective in the long run is to create a highly optimized linear model for association in large-scale omics studies, such as sequencing studies. To do that, I will need to address the main computational bottlenecks which usually narrow down to multiplication and inversion of large matrices. Once I addressed these problems, I will further need to create additional methods suited for application, but this is not discussed here and left as future work.

Chapter 6

Conclusions and future work

In this thesis I have analysed approaches for CNV association with a wide spectrum of phenotypes encompassing several omics domains. This analysis has led to the discovery of several novel CNV associations with phenotypes. These findings thus confirm the underlying importance of CNVs not only for investigating disease outcomes but also for their regulatory role at the molecular level. I have also explored several novel statistical approaches for improved CNV detection, correcting for batch effects and variable selection for metabonomic phenotypes. Due to time and other limitations, the broad scope of this thesis has been analytical in nature in order to prioritize best candidate CNVs associated with phenotypes which can be taken forward for validation.

For methods like the clustering based CNV detection and the variable selection approach applied for the metabonomic data, an empirical proof for the justification of using such approaches and further comparison to gold standard data and other methods still need to be done in the near future.

6.1 Discussion for CNV analysis in Epilepsy

This chapter explores the role of CNVs and its association with epilepsy susceptibility and prognosis using several novel analytical frameworks implemented in cnvHap. Some of these approaches included multi-platform integration for finding phenotype associations and adjusting for genotyping platform-based batch effects for reducing inflation. The multi-platform capability for finding associations has important implications. On

testing this approach in the WTCCC CNV datasets I was successful in reproducing the earlier reported CNV-phenotype associations. This not only confirmed that such an approach can work but also opens up the possibility to re-analyse many of the previous datasets. This implies we can re-analyse previously genotyped datasets for numerous disease-phenotype outcomes to discover new associations. One such possibility may include associating two diseases with known overlapping phenotypes, which might reveal common genomic variations between the two, as already undertaken for single-SNP analyses [280]. Further, a PCA analysis of associated CNVs may also reveal some interesting visualizations. In our study the application of these approaches has suggested several novel CNVs associations with epilepsy susceptibility and treatment response. The top CNVs associated with susceptibility represent biologically important candidates like *SLC31A1* and *NLGN1*. We postulate that in light of some of the previous findings, that these CNVs may have important functional role for epilepsy. However, more research is needed to untangle these overlapping causal factors and possible traits between these two diseases. The CNV in the *NLGN1* gene is another interesting candidate from our susceptibility analysis. While *NLGN1* is a known gene for autism and given that about 30% of autism patients have seizures, this corroborates the hypothesis that these two diseases may share common aetiology. There have been several other recent studies which have reported important functional mechanism of the *NLGN1* and how this gene can cause seizures in rats. Results from our drug-response analysis also show a promising candidate CNV in the *WWOX* gene.

Again in this case, I found several previous reports where a deletion in the exonic region of this gene in rats was shown to cause seizures. Further the *WWOX* CNV has strong validation through sequencing data. The presence of consistent CNV-breakpoints and good separation in the clusters with greater than 50% reciprocal overlap with the sequencing CNV-breakpoints, lends a degree of confidence that these results are real and worthy of laboratory validation. In the future, one other important step would be to characterize the *WWOX* CNV in different populations from the 1KG project. Expression profiling of this gene in multiple tissues and multiple individuals may shed more light into its population specific functional characteristics and effects on phenotypes. Further, a global CNV-map, specific for epilepsy patients can also be a useful resource for the scientific community. Several studies have highlighted that for several neuro-psychiatric diseases like autism, there is an increased global burden of CNVs in the patients. It will be interesting to explore for epilepsy whether such a burden of

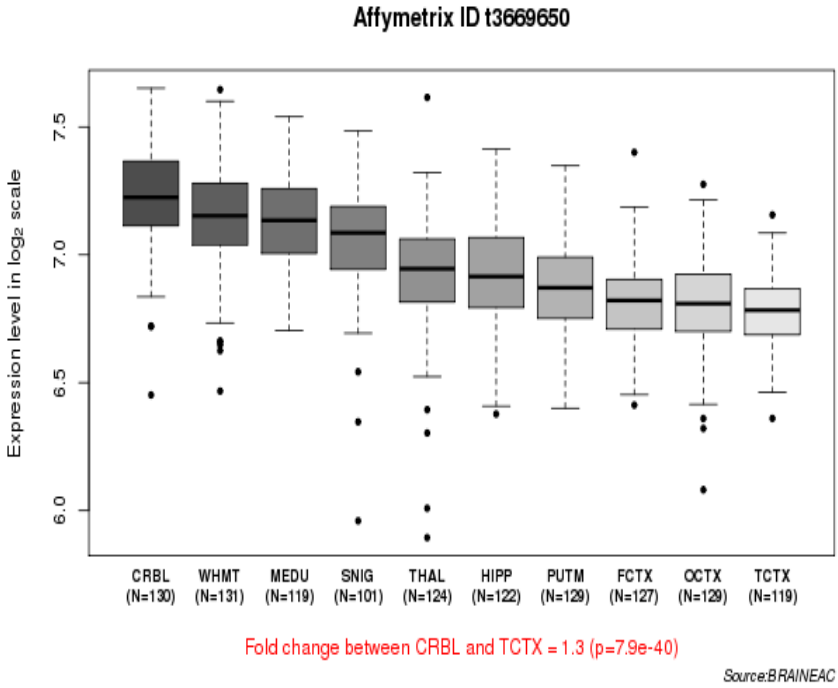


FIGURE 6.1: RNA expression values for the *WWOX* gene

These values are from the ten different regions of the brain obtained from the UKBEC. Some regions of brain have higher expression profiles than the other. This included the hippocampus region which lies in the middle range of expression boxplots. These differences can reveal tissue specificity and further functional role for *WWOX*.

CNVs exists. In addition, approaches like a global CNV analysis through network-based approaches can reveal the pathway dynamics of these variants. An integrated analysis of CNVs, expression and phenotype remains a promising avenue to be explored in the future. I further highlight that future studies like ours can open up new avenues for better prognosis and medication for epilepsy and ultimately leading to better medication and lifestyles for the patients.

6.1.1 Future directions

In the latter sections of chapter 3, I have explored some of these initial efforts for applying clustering for CNV detection. In future, I indicate that the best way forward will be to leverage the power of multiple clustering algorithms to derive the final partition of high quality clusters. Some of the approaches to do this will be to use different partition algorithms and combine them using several approaches based on like median partition

based methods, voting-based method, co-association based methods and graph-based methods. All these ideas are related to the concept of ensemble clustering.

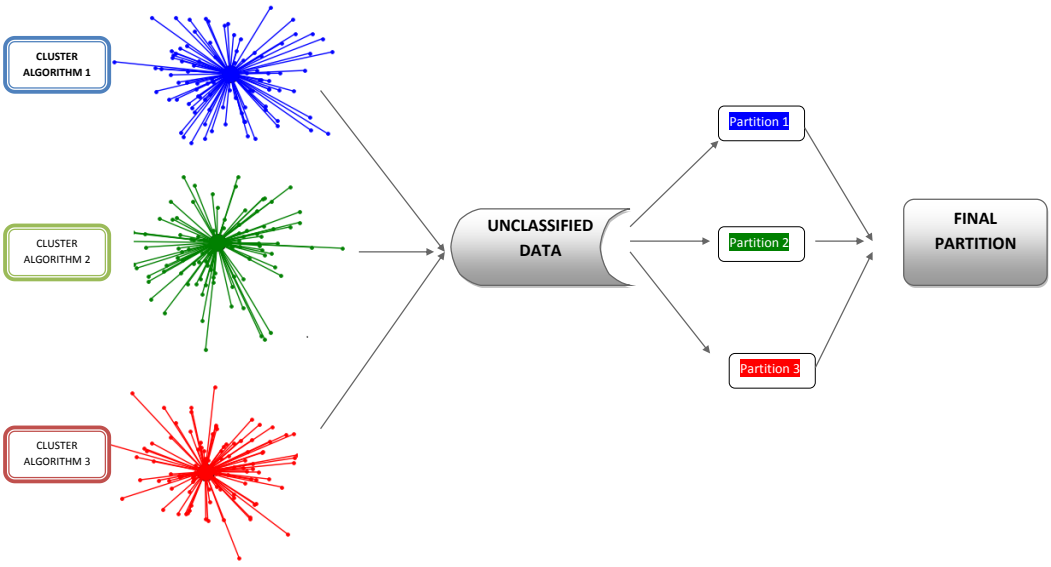


FIGURE 6.2: Merging clustering algorithms

Here the same clustering approach can be run multiple times to obtain different partitions or different clustering methods can be merged to obtain the final partition of the data. Special methods are needed to derive the final partition from the different results. Such an approach has the potential to show improved results.

6.2 Discussion for CNV analysis in metabolomics

The successful association, replication and validation of CNVs in the exonic regions of the *TSPAN8* gene with high-throughput metabonomic and lipid phenotypes in Finnish and British populations is a novel finding. These findings add new information to the current understanding of the role of lipids in T2D. Further, the strong LD structure of CNVs around CNVR5583.1 and the overlap of these CNVs with exonic regions suggest a causal role of these lipid-regulating CNVs in the pathogenesis of T2D.

From my association analysis, I found that multivariate approaches using multiple phenotypes showed stronger signals for associations when compared to univariate approaches. This confirms multiphen original finding that multivariate approaches can prove to be more powerful than univariate approaches for association studies. Also, I found that increasing the sample size by pooling NFBC 1966 and NFBC 1986 data together, the association signals increased 6 folds. However, I do caution that multivariate analysis variable selection process is a crucial step since having nonrelevant predictors in the model can dilute association results to some extent. I also conclude from my association results that for common diseases like T2D, the genetic risk for susceptibility can be accounted by two or more genomic structural variants like CNVs present in a single gene, regulating multiple phenotypes simultaneously. This indicates the presence of pleiotropy at the molecular metabolic level. These lipid findings also confirm the already well-established link between dyslipidemia and T2D.

In accordance with the ENCODE results [264], I suggest that phenotypic characterization should also be done at the molecular level like gene expression or metabonomics. This can help in ascertaining the role of genetic variants for complex diseases like T2D at deeper levels. These results also indicate the necessity of phenotypic characterization of different cohorts being done comprehensively in order to make use of the increased power of multivariate association approaches for discovering causal variants.

I observed that results from NGS projects like the 1KG can provide rich resources for insilico validation of genomic structural variations like CNVs and a reference for comparing allele frequency inferred though genotyping platforms in different population cohorts. Analysing the sequencing results, I found CNVR5583.1 to be present in all 17

different populations from the 1KG project, but the strong lipid regulating CNV located at chr12:69809401 was present in the form of heterozygous deletion in 11 populations and as homozygous deletion in 2 other populations. These findings indicate that, in addition to the association of these CNVs with T2D disease outcome and lipids, there might be population-specific susceptibility as well. For CNVR5583.1, I found a trend in association with lipids but replication and meta-analysis was not possible due to the lack of consensus in the of probe locations across the different genotyping chips. Hence, the effect of CNVR5583.1 on lipids needs confirmation.

I also highlight that some important issues related to CNV segmentation results in matured and widely used CNV calling algorithms like PennCNV and cnvHap. Although both these algorithms detected CNVs at chr12: 69809401 and CNVR5583.1, the CNV breakpoints from these algorithms when compared to the NGS results show that several smaller CNVs are being merged as one long CNV segment. The main reason for this I believe is variation or sparsity in the probe density across different genotyping platforms. This affects the breakpoint annotation for CNV calling algorithms. In this situation the ascertainment of allele frequency in the cohorts probably becomes inaccurate, since one or more CNVs having different allele frequencies might get merged together. I believe that compared to other CNV calling algorithm, cnvHap probe-by-probe calling approach (where each probe is treated independently) for reclustering intensity data, while calling CNVs and using expected CNV genotypes for association, might circumvent this problem to some extent. However, I believe that the best solution for this problem could be the future possibility of imputing CNV breakpoint information from NGS platforms onto genotyping arrays for more accurate determination of CNV segments and allele frequency.

Results from the WTCCC results concluded that the role of common CNVs is unlikely to contribute greatly to the susceptibility of common diseases, but I would like to point out that the design of the genotyping arrays used by the WTCCC study has no probes covering the region chr12: 69164368- 69819067. Thus the WTCCC study did not cover the region from 5th - 8th exon of the *TSPAN8* gene and hence missed the lipid regulating CNVs, which I have reported here. I believe that disease susceptibility can be manifested in a complex manner (at several levels like molecular, expression or methylation patterns) and hence there is a need for characterizing the role of common CNVs in light of greater

coverage through high-density genotyping platforms, NGS, and multiple phenotypes. The variation in disease susceptibility across different populations also needs to be taken in consideration. This can perhaps help in the better ascertainment of the role of genetics for complex diseases like T2D.

Lastly, I echo to the current opinion that, apart from analysing omics of disease specific cohorts, there is also a need for characterizing the normal control population in details. Comparing omics results from the normal population can provide novel insights into the genomics, transcriptomics or metabonomic signatures, that confer healthiness in an individual or have protective features against diseases like T2D. In light of these findings, I suggest my results possibly indicate some metabonomic signatures for T2D. I further add that there is potential of integrating baseline characteristics of metabonomic with other omics data, which in the future has the potential to accurately diagnose and treat disease. This also holds the promise of creation of new avenues of gene therapy and molecular diagnosis for T2D.

6.3 Discussion for CNV association with gene expression

The analysis results presented in chapter 5 are preliminary. I will need to replicate these findings. Some possibilities for replication are several other cohorts where expression profiling and genotyping have been done with the same samples. These include a study from [281] which used the similar expression-array for a smaller sample size (n=84) and for tissues from frontal cortex (only one brain region analysed). The NABEC study [282] used a different expression array but larger sample size (n=304) and two brain regions (cerebellum and frontal cortex) for each individual. Lastly, another study with large sample size (n=1490) by [283] looked for eQTLs in monocytes. Monocytes are different cell types from the brain and it will only count as a partial replication.

At present the duplication at chromosome 4 which I described earlier in this chapter, seems to be a strong hit. I successfully validated this CNV in the 1Kg data as well. Given that this CNV also has lots of trans-eQTLs, it remains a potential candidate for follow up-studies. We could also potentially carry out an additional analysis to find

regions of the genome where eQTLs are shared not just in different brain tissues but with other tissues in the body as well. This might help us determine a new metric to measure tissue similarity which can prove to be a useful resource in the field of transcriptomics. At present, I am processing the data from NABEC and hope to replicate some of our preliminary results. NABEC has data for only two tissues and hence tissue-specific replication may not be possible. In this scenario, I might follow an alternate strategy and use the NABEC as the primary discovery cohort and then replicate the top hits in UKBEC.

There are several open questions in eQTL analysis which I aim to address in the future. One possibility is a multi tissue network analysis of eQTLs for UKBEC and NABEC datasets. This can perhaps show the enrichments of eQTLs in the CNV regions and perhaps I can replicate some of the other reports that the presence of eQTLs in CNV regions is non random. Further, more such an analysis in the light of multiple tissues and pathway analysis can shed more light on the underlying RNA biology. Another possibility to be explored is to analyse regions of homozygosity (ROH) and the nature of expression in them.

In the end, I will like to conclude that these analyses in the light of NGS results can further help in better understanding and characterizing the true nature of relationship between RNA expression and genomic variations like CNVs and SNPs.

6.4 Conclusions

This work also lays an important background work for some novel future work in the area of CNVs applicable to the datasets analysed in this thesis. As we saw in the earlier chapters for epilepsy and psoriasis, there is an overlap of associated genomic variants for more than one disease. In addition, I have indicated that there was a possibility of association of CNVs in the *TSPAN8* gene with obesity. These results indicate that there might be an inherent common aetiology for diseases like epilepsy, psoriasis and diabetes. In this scenario, a pathway-based analysis or other multivariate approaches which can incorporate several variants into a single model might be ideal and may shed more light into the true mechanistic causes for the disease.

Common disease common variant hypothesis and CNVs The Wellcome Trust study for common CNVs and common diseases concluded that common CNVs are unlikely to play a major role in the pathogenesis of common diseases, except for T2D and Crohn's disease [76]. They reported two common CNVs - CNVR5583.1 and CNVR2647.1 to be strongly associated with T2D and Crohn's disease, respectively.

Based on the results from this thesis, I concur with the Wellcome Trust CNV analysis, that for T2D, a common CNV in the $3rd$ exon of the *TSPAN8* gene (CNVR5583.1) is strongly associated with the disease outcome. For CNVR5583.1, I further report its association with lipids in two Finnish populations. In addition to this confirmation, I have discovered and reported a new low frequency ($< 5\%$) CNV variant to be associated strongly with T2D and also associated this CNV with large proportion of lipid phenotypes. These results highlights the role of common CNVs in regulating lipids levels in humans also affecting disease outcome at the same time. In chapter 3 and 4, I again discovered and reported two examples of common CNVs, *WWOX* and *DRD5* which were not associated with disease outcome like in the normal GWAS setting, but were found to be associated with other phenotypes such as epilepsy drug response and gene expression. These results indicate that common CNVs have some influence (modest effect size) in the pathogenesis of common disease but may have greater influence at the molecular level such as expression and drug regulation. In this scenario, their role cannot be completely ignored and should be re-characterized in light of improved methods

for detecting small and common CNVs and associating them high resolution molecular phenotypes such metabolomics and gene expression.

Current focus in genomic studies is further shifting from common variants towards rare variants, due to the limited success of common disease common variant hypothesis. Pathway models become more attractive in this setting since pathway models can capture multiple rare variants acting through common pathways leading to disease manifestation. Analysis of rare variants alone without taking the pathway into consideration may not reveal the underlying biology behind the disease, hence making therapeutic approaches more difficult. Pathway models for CNVs and SNPs together and their dosage effects are still limited in the current genomic studies and need to be explored in the future.

We also saw with example of Crohns disease where the true causal variant was a CNV and it was in perfect LD with surrounding SNP. A haplotype analysis taking both the SNPs and CNVs into a single model showed stronger association patterns. Now with the advent of high density genotyping chips there has been development of matured methods for calling SNPs and CNVs from these platforms. However, a joint association framework for combined SNPs and CNVs on a haplotype level is limited. This will require a calibrated metric for characterizing SNPs and CNVs on a genome wide level and perhaps my proposed clustering based method can prove to be useful to this end.

Most of the GWAS analysis done in the last decade has centred on European populations. An in depth analysis to find out how these findings translate to other populations still remains to be done. Now as sequencing and genotyping become cheaper and more widely used around the world a multi-population disease model can emerge. Medical diagnosis and treatment can be based not only on the personal level but also at the population level, where targeted medication can be developed for different ethnic groups. At present, the cost of sequencing is decreasing faster than predicted by Moores law. In addition, several other high throughput technologies are catching up fast in terms of fidelity and cost, thus heralding a data surge in the omics field. This will not only allow better characterization of biomarkers and bio-molecules associated with diseases but will perhaps allow building models for predicting disease susceptibility for individuals. Integrated model for genomic variants with multiple omics like metabonomics, transcriptomics and proteomics can be built using deep learning methods. These approaches can make the concept of personalized medicine closer to reality.

In the immediate future much work needs to be done as a follow-up to the different analyses reported in this thesis. This includes replication of the *WWOX* CNV with increased power in the Australian cohort and perhaps a sequencing based validation approach for some of the other CNVs like *NLGN1*, which were found associated with epilepsy susceptibility. In the metabonomics chapter, I focussed on the coding regions in the first instance. Analysis in the *TSPAN8* gene region was also hindered because of the lack of common probes between NFBC 1966 and NFBC 1986.

Interestingly, some CNVs present in the intronic regions also showed strong association signals and one of these CNVs was also found to be nominally associated with obesity. A targeted genotyping or sequence based analysis for the *TSPAN8* gene in NFBC cohorts and other diabetes cohort has the potential to reveal novel associations of CNVs and SNPs with lipids. The fact that SNPs in the *TSPAN8* gene has also been found to be associated with diabetes in earlier GWAS, this corroborates the possible causal role of our reported CNVs in the pathogenesis of type 2 diabetes. The association of the Psoriasis CNV in the *LCE3B-LCE3C* gene with urea is most interesting and warrants further investigation. This association with urea adds further proof to the already known relationship between psoriasis and Crohns disease. First, that urea levels have been found to be high in patients with Crohns disease and ,second that urea is used as a treatment for psoriasis. This association shows preliminary molecular basis for these two diseases but needs stronger experimental validation.

The results from the UKBEC analysis show much promise for the chromosome 4 duplication and its association with gene expression. I have already made preliminary efforts in order to further characterize this CNV and the expression profile from the 1000 genomes project data. This CNV was successfully validated in the 1KG data and further replicated in the separate North American Brain Expression Consortium dataset (NABEC). However, replication of the CNV-eQTL still needs to be done in NABEC and the 1KG data. Replication in several other cohorts as well as an analysis using the NABEC study as the discovery cohort is needed. I also plan to carry out a loss of heterozygosity analysis for these datasets which can shed light on whether CNV-eQTL has been conserved through selection. I have plans to further develop methods for CNV-eQTL analysis for sequencing based data. To this end, I have already successfully tried out some fast approaches through the Java interface to the GPU and got some positive results. As we saw, for performing eQTL analysis can be computationally prohibitive and often infeasible to carry out on a desktop computer. GPU based approaches remain a promising

solution to tackle this problem.

Amongst the future work discussed above, I would prioritize the follow up of the *WWOX* CNV most. This would include replication of this CNV in the Australian cohort with more power to replicate the association in the SANAD cohort. My second priority would be to replicate the *DRD5* association with gene expression. Given the uniqueness of these two analysis, when complete it would enable some possibility of publication, indicating the positive or negative findings.

6.5 Original contributions

Following are the novel contributions made for CNV association approaches in the field of omics data in humans:

- Novel candidate genes identified for epilepsy susceptibility and drug remission.
- First application and real GWAS demonstration of cnvHap multiplatform integration capability for the epilepsy and the WTCCC aCGH datasets.
- Parameter optimisation of cnvHap, implemented improvements in the cnvHap algorithm - addition of new features added like removal of missing data, modification in the clustering model for chromosome X results.
- Novel idea of possible application of clustering based approach for detecting CNVs on a probe by probe basis from high-throughput genomic data.
- Application of a novel iterative PCA based approach for calculating PCs for genotyping intensity values at genome wide scale and further using these PCs in the association model to reduce inflation.
- Definition of new phenotype for T2D based on metabonomic signatures.
- Development and application of a novel backward-selection based approach for variable selection which selects a minimum set of uncorrelated predictors.
- Identification of the role of CNVs in T2D and psoriasis through metabonomics data.

- Identification of novel CNV-eQTL loci in the UKBEC study and their characterization through NGS data for DNA and RNA.
- Revision of current aspects related to methods for CNV-eQTL work.
- Display of preliminary work for creating association models through the GPU framework.

Bibliography

[1] Andrea S Foulkes. *Applied statistical genetics with R*. Springer, 2009.

[2] Jay Shendure and Hanlee Ji. Next-generation dna sequencing. *Nature biotechnology*, 26(10):1135–1145, 2008.

[3] Cristen J Willer, Elizabeth K Speliotes, Ruth JF Loos, Shengxu Li, Cecilia M Lindgren, Iris M Heid, Sonja I Berndt, Amanda L Elliott, Anne U Jackson, Claudia Lamina, et al. Six new loci associated with body mass index highlight a neuronal influence on body weight regulation. *Nature genetics*, 41(1):25–34, 2008.

[4] Beben Benyamin, Allan F McRae, Gu Zhu, Scott Gordon, Anjali K Henders, Aarno Palotie, Leena Peltonen, Nicholas G Martin, Grant W Montgomery, John B Whitfield, et al. Variants in tf and hfe explain 40% of genetic variation in serum-transferrin levels. *The American Journal of Human Genetics*, 84(1):60–65, 2009.

[5] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[6] William S Bush and Jason H Moore. Genome-wide association studies. *PLoS computational biology*, 8(12):e1002822, 2012.

[7] Karol Estrada, Anis Abuseiris, Frank G Grosveld, André G Uitterlinden, Tobias A Knoch, and Fernando Rivadeneira. Grimp: a web-and grid-based tool for high-speed analysis of large-scale genome-wide association using imputed data. *Bioinformatics*, 25(20):2750–2752, 2009.

[8] Andrey A Shabalin. Matrix eqtl: ultra fast eqtl analysis via large matrix operations. *Bioinformatics*, 28(10):1353–1358, 2012.

[9] Clive J Hoggart, John C Whittaker, Maria De Iorio, and David J Balding. Simultaneous analysis of all snps in genome-wide and re-sequencing association studies. *PLoS genetics*, 4(7):e1000130, 2008.

[10] Lachlan JM Coin, Julian E Asher, Robin G Walters, Julia S El-Sayed Moustafa, Adam J de Smith, Rob Sladek, David J Balding, Philippe Froguel, and Alexandra IF Blakemore. cnvhap: an integrative population and haplotype–based multiplatform model of snps and cnvs. *Nature methods*, 7(7):541–546, 2010.

[11] Jeremy K Nicholson, John C Lindon, and Elaine Holmes. 'metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological nmr spectroscopic data. *Xenobiotica*, 29(11):1181–1189, 1999.

[12] Peter J Turnbaugh, Micah Hamady, Tanya Yatsunenko, Brandi L Cantarel, Alexis Duncan, Ruth E Ley, Mitchell L Sogin, William J Jones, Bruce A Roe, Jason P Affourtit, et al. A core gut microbiome in obese and lean twins. *nature*, 457(7228): 480–484, 2008.

[13] Susan C Connor, Michael K Hansen, Adam Corner, Randall F Smith, and Terence E Ryan. Integration of metabolomics and transcriptomics data to aid biomarker discovery in type 2 diabetes. *Molecular BioSystems*, 6(5):909–921, 2010.

[14] Arun Sreekumar, Laila M Poisson, Thekkelnaycke M Rajendiran, Amjad P Khan, Qi Cao, Jindan Yu, Bharathi Laxman, Rohit Mehra, Robert J Lonigro, Yong Li, et al. Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, 457(7231):910–914, 2009.

[15] Gregory D Lewis, Aarti Asnani, and Robert E Gerszten. Application of metabolomics to cardiovascular biomarker and pathway discovery. *Journal of the American College of Cardiology*, 52(2):117–123, 2008.

[16] David S Wishart, Dan Tzur, Craig Knox, Roman Eisner, An Chi Guo, Nelson Young, Dean Cheng, Kevin Jewell, David Arndt, Summit Sawhney, et al. Hmdb: the human metabolome database. *Nucleic acids research*, 35(suppl 1):D521–D526, 2007.

[17] Zhong Wang, Mark Gerstein, and Michael Snyder. Rna-seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 10(1):57–63, 2009.

[18] Mark Schena, Dari Shalon, Ronald W Davis, and Patrick O Brown. Quantitative monitoring of gene expression patterns with a complementary dna microarray. *Science*, 270(5235):467–470, 1995.

[19] Pieter Mestdagh, Tom Feys, Nathalie Bernard, Simone Guenther, Caifu Chen, Frank Speleman, and Jo Vandesompele. High-throughput stem-loop rt-qpcr mirna expression profiling using minute amounts of input rna. *Nucleic acids research*, 36 (21):e143–e143, 2008.

[20] Wolfgang Huber, Anja von Heydebreck, Holger Sueltmann, Annemarie Poustka, and Martin Vingron. Parameter estimation for the calibration and variance stabilization of microarray data. *Statistical applications in genetics and molecular biology*, 2(1), 2003.

[21] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[22] Baolin Wu. Differential gene expression detection using penalized linear regression models: the improved sam statistics. *Bioinformatics*, 21(8):1565–1571, 2005.

[23] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli Yu, Huanming Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. *Nature*, 426(6968):789–796, 2003.

[24] Manuel A Rivas, Mélissa Beaudoin, Agnes Gardet, Christine Stevens, Yashoda Sharma, Clarence K Zhang, Gabrielle Boucher, Stephan Ripke, David Ellinghaus, Noel Burtt, et al. Deep resequencing of gwas loci identifies independent rare variants associated with inflammatory bowel disease. *Nature genetics*, 43(11): 1066–1073, 2011.

[25] 1000 Genomes Project Consortium et al. A map of human genome variation from population-scale sequencing. *Nature*, 467(7319):1061–1073, 2010.

[26] Jiannis Ragoussis. Genotyping technologies for genetic research. *Annual review of genomics and human genetics*, 10:117–133, 2009.

[27] Chris Barnes, Vincent Plagnol, Tomas Fitzgerald, Richard Redon, Jonathan Marchini, David Clayton, and Matthew E Hurles. A robust statistical method for case-control association testing with copy number variation. *Nature genetics*, 40 (10):1245–1252, 2008.

[28] Richard Redon, Shumpei Ishikawa, Karen R Fitch, Lars Feuk, George H Perry, T Daniel Andrews, Heike Fiegler, Michael H Shapero, Andrew R Carson, Wenwei Chen, et al. Global variation in copy number in the human genome. *nature*, 444 (7118):444–454, 2006.

[29] Donald F Conrad, Dalila Pinto, Richard Redon, Lars Feuk, Omer Gokcumen, Yujun Zhang, Jan Aerts, T Daniel Andrews, Chris Barnes, Peter Campbell, et al. Origins and functional impact of copy number variation in the human genome. *Nature*, 464(7289):704–712, 2010.

[30] Steven A McCarroll and David M Altshuler. Copy-number variation and association studies of human disease. *Nature genetics*, 39:S37–S42, 2007.

[31] Manuela Fanciulli, Penny J Norsworthy, Enrico Petretto, Rong Dong, Lorraine Harper, Lavanya Kamesh, Joanne M Heward, Stephen CL Gough, Adam de Smith, Alexandra IF Blakemore, et al. Fcgr3b copy number variation is associated with susceptibility to systemic, but not organ-specific, autoimmunity. *Nature genetics*, 39(6):721–723, 2007.

[32] James R Lupski and Pawel Stankiewicz. Genomic disorders: molecular mechanisms for rearrangements and conveyed phenotypes. *PLoS genetics*, 1(6):e49, 2005.

[33] Christine J Shaw and James R Lupski. Implications of human genome architecture for rearrangement-based disorders: the genomic basis of disease. *Human molecular genetics*, 13(suppl 1):R57–R64, 2004.

[34] Aikaterini Barbouti, Pawel Stankiewicz, Chad Nusbaum, Christina Cuomo, April Cook, Mattias Höglund, Bertil Johansson, Anne Hagemeijer, Sung-Sup Park, Felix Mitelman, et al. The breakpoint region of the most common isochromosome, i (17q), in human neoplasia is characterized by a complex genomic architecture with large, palindromic, low-copy repeats. *The American Journal of Human Genetics*, 74(1):1–10, 2004.

[35] Jeffrey A Bailey and Evan E Eichler. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nature Reviews Genetics*, 7(7):552–564, 2006.

[36] James R Lupski. Genomic disorders: structural features of the genome can lead to dna rearrangements and human disease traits. *Trends in genetics*, 14(10):417–422, 1998.

[37] Lawrence T Reiter, Tatsufumi Murakami, Thearith Koeuth, Liu Pentao, Donna M Muzny, Richard A Gibbs, James R Lupski, et al. A recombination hotspot responsible for two inherited peripheral neuropathies is located near a mariner transposon-like element. *Nature genetics*, 12(3):288–297, 1996.

[38] Weimin Bi, Sung-Sup Park, Christine J Shaw, Marjorie A Withers, Pragna I Patel, and James R Lupski. Reciprocal crossovers and a positional preference for strand exchange in recombination events resulting in deletion or duplication of chromosome 17p11. 2. *The American Journal of Human Genetics*, 73(6):1302–1315, 2003.

[39] Kwan-Wood G Lam and Alec J Jeffreys. Processes of copy-number change in human dna: the dynamics of α-globin gene deletion. *Proceedings of the National Academy of Sciences*, 103(24):8921–8927, 2006.

[40] Katharina Steinmann, David N Cooper, Lan Kluwe, Nadia A Chuzhanova, Cornelia Senger, Eduard Serra, Conxi Lazaro, Montserrat Gilaberte, Katharina Wimmer, Viktor-Felix Mautner, et al. Type 2 nf1 deletions are highly unusual by virtue of the absence of nonallelic homologous recombination hotspots and an apparent preference for female mitotic recombination. *The American Journal of Human Genetics*, 81(6):1201–1220, 2007.

[41] Daniel J Turner, Marcos Miretti, Diana Rajan, Heike Fiegler, Nigel P Carter, Martyn L Blayney, Stephan Beck, and Matthew E Hurles. Germline rates of de novo meiotic deletions and duplications causing several genomic disorders. *Nature genetics*, 40(1):90–95, 2007.

[42] Lorraine Potocki, Weimin Bi, Diane Treadwell-Deering, Claudia Carvalho, Anna Eifert, Ellen M Friedman, Daniel Glaze, Kevin Krull, Jennifer A Lee, Richard Alan Lewis, et al. Characterization of potocki-lupski syndrome (dup (17)(p11. 2p11. 2)) and delineation of a dosage-sensitive critical interval that can convey an autism phenotype. *The American Journal of Human Genetics*, 80(4):633–649, 2007.

[43] Conxi Lázaro, Antonia Gaona, Peter Ainsworth, Romano Tenconi, Dominique Vidaud, Helena Kruyer, Elisabet Ars, Víctor Volpini, and Xavier Estivill. Sex differences in mutational rate and mutational mechanism in the nf1 gene in neurofibromatosis type 1 patients. *Human genetics*, 98(6):696–699, 1996.

[44] Mariko Sasaki, Julian Lange, and Scott Keeney. Genome destabilization by homologous recombination in the germ line. *Nature Reviews Molecular Cell Biology*, 11(3):182–195, 2010.

[45] Michael R Lieber, Yunmei Ma, Ulrich Pannicke, and Klaus Schwarz. Mechanism and regulation of human non-homologous dna end-joining. *Nature reviews Molecular cell biology*, 4(9):712–720, 2003.

[46] Eric Weterings and Dik C van Gent. The mechanism of non-homologous end-joining: a synopsis of synapsis. *DNA repair*, 3(11):1425–1435, 2004.

[47] Michael R Lieber. The mechanism of human nonhomologous dna end joining. *Journal of Biological Chemistry*, 283(1):1–5, 2008.

[48] Klaus Schwarz, Yunmei Ma, Ulrich Pannicke, and Michael R Lieber. Human severe combined immune deficiency and dna repair. *Bioessays*, 25(11):1061–1070, 2003.

[49] Irene Tiemann-Boege, Peter Calabrese, David M Cochran, Rebecca Sokol, and Norman Arnheim. High-resolution recombination patterns in a region of human chromosome 21 measured by sperm typing. *PLoS genetics*, 2(5):e70, 2006.

[50] Jennifer A Lee, Claudia Carvalho, and James R Lupski. A dna replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *cell*, 131(7):1235–1247, 2007.

[51] Zhujun Zhang, Yasuhiro Takeshima, Hiroyuki Awano, Atsushi Nishiyama, Yo Okizuka, Mariko Yagi, and Masafumi Matsuo. Tandem duplications of two separate fragments of the dystrophin gene in a patient with duchenne muscular dystrophy. *Journal of human genetics*, 53(3):215–219, 2008.

[52] LELM Vissers, P Stankiewicz, SA Yatsenko, E Crawford, H Creswick, VK Proud, BBA de Vries, R Pfundt, CLM Marcelis, J Zackowski, et al. Complex chromosome 17p rearrangements associated with low-copy repeats in two patients with congenital anomalies. *Human genetics*, 121(6):697–709, 2007.

[53] Ryan E Mills, Christopher T Luttig, Christine E Larkins, Adam Beauchamp, Circe Tsui, W Stephen Pittard, and Scott E Devine. An initial map of insertion and deletion (indel) variation in the human genome. *Genome research*, 16(9):1182–1190, 2006.

[54] Curt Stern. The hardy-weinberg law. *Science*, 97(2510):137–138, 1943.

[55] Mark P Sawicki, Ghassan Samara, Michael Hurwitz, and Edward Passaro Jr. Human genome project. *The American journal of surgery*, 165(2):258–264, 1993.

[56] Teri A Manolio, Lisa D Brooks, and Francis S Collins. A hapmap harvest of insights into the genetics of common disease. *The Journal of clinical investigation*, 118(5):1590, 2008.

[57] Jonathan K Pritchard and Nancy J Cox. The allelic architecture of human disease genes: common disease–common variant or not? *Human molecular genetics*, 11 (20):2417–2423, 2002.

[58] Timothy M Frayling, Nicholas J Timpson, Michael N Weedon, Eleftheria Zeggini, Rachel M Freathy, Cecilia M Lindgren, John RB Perry, Katherine S Elliott, Hana Lango, Nigel W Rayner, et al. A common variant in the fto gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316(5826):889–894, 2007.

[59] Paul R Burton, David G Clayton, Lon R Cardon, Nick Craddock, Panos Deloukas, Audrey Duncanson, Dominic P Kwiatkowski, Mark I McCarthy, Willem H Ouwehand, Nilesh J Samani, et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–678, 2007.

[60] Robert Sladek, Ghislain Rocheleau, Johan Rung, Christian Dina, Lishuang Shen, David Serre, Philippe Boutin, Daniel Vincent, Alexandre Belisle, Samy Hadjadj, et al. A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, 445(7130):881–885, 2007.

[61] Jeffrey C Barrett and Lon R Cardon. Evaluating coverage of genome-wide association studies. *Nature genetics*, 38(6):659–662, 2006.

[62] M Brion, I Quintela, B Sobrino, M Torres, C Allegue, and A Carracedo. New technologies in the genetic approach to sudden cardiac death in the young. *Forensic science international*, 203(1):15–24, 2010.

[63] Vijayaprakash Suppiah, Max Moldovan, Golo Ahlenstiel, Thomas Berg, Martin Weltman, Maria Lorena Abate, Margaret Bassendine, Ulrich Spengler, Gregory J Dore, Elizabeth Powell, et al. Il28b is associated with response to chronic hepatitis c interferon-α and ribavirin therapy. *Nature genetics*, 41(10):1100–1104, 2009.

[64] Yasuhito Tanaka, Nao Nishida, Masaya Sugiyama, Masayuki Kurosaki, Kentaro Matsuura, Naoya Sakamoto, Mina Nakagawa, Masaaki Korenaga, Keisuke Hino, Shuhei Hige, et al. Genome-wide association of il28b with response to pegylated interferon-α and ribavirin therapy for chronic hepatitis c. *Nature genetics*, 41(10): 1105–1109, 2009.

[65] CT Falk and P Rubinstein. Haplotype relative risks: an easy reliable way to construct a proper control sample for risk calculations. *Annals of Human Genetics*, 51(3):227–233, 1987.

[66] Richard S Spielman, Ralph E McGinnis, and Warren J Ewens. Transmission test for linkage disequilibrium: the insulin gene region and insulin-dependent diabetes mellitus (iddm). *American journal of human genetics*, 52(3):506, 1993.

[67] Jurg Ott. Statistical properties of the haplotype relative risk. *Genetic Epidemiology*, 6(1):127–130, 1989.

[68] Joseph D Terwilliger and J Ott. A haplotype-based haplotype relative riskapproach to detecting allelic associations. *Human Heredity*, 42(6):337–346, 1992.

[69] William YS Wang, Bryan J Barratt, David G Clayton, and John A Todd. Genome-wide association studies: theoretical and practical concerns. *Nature Reviews Genetics*, 6(2):109–118, 2005.

[70] Nan M Laird and Christoph Lange. Family-based designs in the age of large-scale gene-association studies. *Nature Reviews Genetics*, 7(5):385–394, 2006.

[71] Hariklia Eleftherohorinou, Victoria Wright, Clive Hoggart, Anna-Liisa Hartikainen, Marjo-Riitta Jarvelin, David Balding, Lachlan Coin, and Michael Levin. Pathway analysis of gwas provides new insights into genetic susceptibility to 3 inflammatory diseases. *PloS one*, 4(11):e8068, 2009.

[72] Andrew S Allen and Glen A Satten. Genome-wide association analysis of rheumatoid arthritis data via haplotype sharing. In *BMC proceedings*, volume 3, page S30. BioMed Central Ltd, 2009.

[73] Joel N Hirschhorn and Mark J Daly. Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, 6(2):95–108, 2005.

[74] Richard R Hudson, Kevin Bailey, Douglas Skarecky, Jan Kwiatowski, and Francisco J Ayala. Evidence for positive selection in the superoxide dismutase (sod) region of drosophila melanogaster. *Genetics*, 136(4):1329–1340, 1994.

[75] Pardis C Sabeti, David E Reich, John M Higgins, Haninah ZP Levine, Daniel J Richter, Stephen F Schaffner, Stacey B Gabriel, Jill V Platko, Nick J Patterson, Gavin J McDonald, et al. Detecting recent positive selection in the human genome from haplotype structure. *Nature*, 419(6909):832–837, 2002.

[76] Nick Craddock, Matthew E Hurles, Niall Cardin, Richard D Pearson, Vincent Plagnol, Samuel Robson, Damjan Vukcevic, Chris Barnes, Donald F Conrad, Eleni Giannoulatou, et al. Genome-wide association study of cnvs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature*, 464(7289):713–720, 2010.

[77] Nathan E Wineinger, Nicholas M Pajewski, Richard E Kennedy, Mary K Wojczynski, Laura K Vaughan, Steven C Hunt, C Charles Gu, Dabeeru C Rao, Rachel Lorier, Ulrich Broeckel, et al. Characterization of autosomal copy-number variation in african americans: the hypergen study. *European Journal of Human Genetics*, 19(12):1271–1275, 2011.

[78] Kui Zhang, Peter Calabrese, Magnus Nordborg, and Fengzhu Sun. Haplotype block structure and its applications to association studies: power and study designs. *The American Journal of Human Genetics*, 71(6):1386–1394, 2002.

[79] Shu-Yi Su, David J Balding, and Lachlan JM Coin. Disease association tests by inferring ancestral haplotypes using a hidden markov model. *Bioinformatics*, 24 (7):972–978, 2008.

[80] Ryan Tewhey, Vikas Bansal, Ali Torkamani, Eric J Topol, and Nicholas J Schork. The importance of phase information for human genomics. *Nature Reviews Genetics*, 12(3):215–223, 2011.

[81] Timothy J Aitman, Rong Dong, Timothy J Vyse, Penny J Norsworthy, Michelle D Johnson, Jennifer Smith, Jonathan Mangion, Cheri Roberton-Lowe, Amy J Marshall, Enrico Petretto, et al. Copy number polymorphism in fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature*, 439(7078):851–855, 2006.

[82] Kai Wang, Mingyao Li, Dexter Hadley, Rui Liu, Joseph Glessner, Struan FA Grant, Hakon Hakonarson, and Maja Bucan. 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, 2007.

[83] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl 1):D668–D672, 2006.

[84] David E Reich and Eric S Lander. On the allelic spectrum of human disease. *TRENDS in Genetics*, 17(9):502–510, 2001.

[85] Jonathan K Pritchard. Are rare variants responsible for susceptibility to complex diseases? *The American Journal of Human Genetics*, 69(1):124–137, 2001.

[86] Lucia A Hindorff, Praveen Sethupathy, Heather A Junkins, Erin M Ramos, Jayashri P Mehta, Francis S Collins, and Teri A Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences*, 106(23):9362–9367, 2009.

[87] Peter M Visscher. Sizing up human height variation. *Nature genetics*, 40(5): 489–490, 2008.

[88] Francis S Collins. No longer just looking under the lamppost. *American journal of human genetics*, 79(3):421, 2006.

[89] Brendan Maher. The case of the missing heritability. *Nature*, 456(7218):18–21, 2008.

[90] Johanna Jakobsdottir, Michael B Gorin, Yvette P Conley, Robert E Ferrell, and Daniel E Weeks. Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. *PLoS genetics*, 5(2): e1000337, 2009.

[91] Matthew C Keller and Geoffrey Miller. Resolving the paradox of common, harmful, heritable mental disorders: which evolutionary genetic models work best? *Behavioral and Brain Sciences*, 29(04):385–404, 2006.

[92] Greg Gibson and Günter Wagner. Canalization in evolutionary genetics: a stabilizing theory? *Bioessays*, 22(4):372–380, 2000.

[93] Sekar Kathiresan, Cristen J Willer, Gina M Peloso, Serkalem Demissie, Kiran Musunuru, Eric E Schadt, Lee Kaplan, Derrick Bennett, Yun Li, Toshiko Tanaka, et al. Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature genetics*, 41(1):56–65, 2008.

[94] Aldons J Lusis and Päivi Pajukanta. A treasure trove for lipoprotein biology. *Nature genetics*, 40(2):129–130, 2008.

[95] Krishna R Veeramah and Michael F Hammer. The impact of whole-genome sequencing on the reconstruction of human population history. *Nature Reviews Genetics*, 15(3):149–162, 2014.

[96] Kelly A Frazer, Dennis G Ballinger, David R Cox, David A Hinds, Laura L Stuve, Richard A Gibbs, John W Belmont, Andrew Boudreau, Paul Hardenbol, Suzanne M Leal, et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851–861, 2007.

[97] Or Zuk, Stephen F Schaffner, Kaitlin Samocha, Ron Do, Eliana Hechter, Sekar Kathiresan, Mark J Daly, Benjamin M Neale, Shamil R Sunyaev, and Eric S Lander. Searching for missing heritability: Designing rare variant association studies. *Proceedings of the National Academy of Sciences*, 111(4):E455–E464, 2014.

[98] Mark I McCarthy, Gonçalo R Abecasis, Lon R Cardon, David B Goldstein, Julian Little, John PA Ioannidis, and Joel N Hirschhorn. Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Reviews Genetics*, 9(5):356–369, 2008.

[99] RG Walters, S Jacquemont, A Valsesia, AJ De Smith, D Martinet, J Andersson, M Falchi, F Chen, J Andrieux, S Lobbens, et al. A new highly penetrant form of obesity due to deletions on chromosome 16p11. 2. *Nature*, 463(7281):671–675, 2010.

[100] M Fanciulli, E Petretto, and TJ Aitman. Gene copy number variation and common human disease. *Clinical genetics*, 77(3):201–213, 2010.

[101] Barbara E Stranger, Alexandra C Nica, Matthew S Forrest, Antigone Dimas, Christine P Bird, Claude Beazley, Catherine E Ingle, Mark Dunning, Paul Flicek, Daphne Koller, et al. Population genomics of human gene expression. *Nature genetics*, 39(10):1217–1224, 2007.

[102] Johan Staaf, Johan Vallon-Christersson, David Lindgren, Gunnar Juliusson, Richard Rosenquist, Mattias Höglund, Åke Borg, and Markus Ringnér. Normalization of illumina infinium whole-genome snp data improves copy number estimates and allelic intensity ratios. *BMC bioinformatics*, 9(1):409, 2008.

[103] C Yau and CC Holmes. Cnv discovery using snp genotyping arrays. *Cytogenetic and genome research*, 123(1-4):307–312, 2009.

[104] John C Marioni, Natalie P Thorne, Armand Valsesia, Tomas Fitzgerald, Richard Redon, Heike Fiegler, T Daniel Andrews, Barbara E Stranger, Andrew G Lynch, Emmanouil T Dermitzakis, et al. Breaking the waves: improved detection of copy number variation from microarray-based comparative genomic hybridization. *Genome Biol*, 8(10):R228, 2007.

[105] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 2012.

[106] Nicolas Wicker, Annaïck Carles, Ian G Mills, Maija Wolf, Abhi Veerakumarasivam, Henrik Edgren, Fabrice Boileau, Bohdan Wasylyk, Jack A Schalken, David E Neal, et al. A new look towards bac-based array cgh through a comprehensive comparison with oligo-based array cgh. *BMC genomics*, 8(1):84, 2007.

[107] Femke D Hannes, Andrew J Sharp, Heather C Mefford, Thomy de Ravel, Claudia A Ruivenkamp, Martijn H Breuning, Jean-Pierre Fryns, Koen Devriendt, Griet Van Buggenhout, Annick Vogels, et al. Recurrent reciprocal deletions and duplications of 16p13. 11: the deletion is a risk factor for mr/mca while the duplication may be a rare benign variant. *Journal of medical genetics*, 46(4):223–232, 2009.

[108] Cornelis A Albers, Gerton Lunter, Daniel G MacArthur, Gilean McVean, Willem H Ouwehand, and Richard Durbin. Dindel: accurate indel calls from short-read data. *Genome research*, 21(6):961–973, 2011.

[109] Razi Khaja, Junjun Zhang, Jeffrey R MacDonald, Yongshu He, Ann M Joseph-George, John Wei, Muhammad A Rafiq, Cheng Qian, Mary Shago, Lorena Pantano, et al. Genome assembly comparison identifies structural variants in the human genome. *Nature genetics*, 38(12):1413–1418, 2006.

[110] Samuel Levy, Granger Sutton, Pauline C Ng, Lars Feuk, Aaron L Halpern, Brian P Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F Kirkness, Gennady Denisov, et al.

The diploid genome sequence of an individual human. *PLoS biology*, 5(10):e254, 2007.

[111] David A Wheeler, Maithreyan Srinivasan, Michael Egholm, Yufeng Shen, Lei Chen, Amy McGuire, Wen He, Yi-Ju Chen, Vinod Makhijani, G Thomas Roth, et al. The complete genome of an individual by massively parallel dna sequencing. *nature*, 452(7189):872–876, 2008.

[112] Eray Tuzun, Andrew J Sharp, Jeffrey A Bailey, Rajinder Kaul, V Anne Morrison, Lisa M Pertz, Eric Haugen, Hillary Hayden, Donna Albertson, Daniel Pinkel, et al. Fine-scale structural variation of the human genome. *Nature genetics*, 37 (7):727–732, 2005.

[113] Evangelos Bellos, Michael R Johnson, and Lachlan JM Coin. cnvhitseq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome biology*, 13(12):R120, 2012.

[114] Tobias Rausch, Thomas Zichner, Andreas Schlattl, Adrian M Stütz, Vladimir Benes, and Jan O Korbel. Delly: structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, 28(18):i333–i339, 2012.

[115] Loryn N Sellner and Graham R Taylor. Mlpa and maph: new techniques for detection of gene deletions. *Human mutation*, 23(5):413–419, 2004.

[116] John AL Armour, Raquel Palla, Patrick LJM Zeeuwen, Martin den Heijer, Joost Schalkwijk, and Edward J Hollox. Accurate, high-throughput typing of copy number variation using paralogue ratios from dispersed repeats. *Nucleic acids research*, 35(3):e19–e19, 2007.

[117] Edward J Hollox, Ulrike Huffmeier, Patrick LJM Zeeuwen, Raquel Palla, Jesús Lascorz, Diana Rodijk-Olthuis, Peter CM van de Kerkhof, Heiko Traupe, Gys de Jongh, Martin den Heijer, et al. Psoriasis is associated with increased beta-defensin genomic copy number. *Nature genetics*, 40(1):23, 2008.

[118] Donald F Conrad, T Daniel Andrews, Nigel P Carter, Matthew E Hurles, and Jonathan K Pritchard. A high-resolution survey of deletion polymorphism in the human genome. *Nature genetics*, 38(1):75–81, 2006.

[119] Steven A McCarroll, Tracy N Hadnott, George H Perry, Pardis C Sabeti, Michael C Zody, Jeffrey C Barrett, Stephanie Dallaire, Stacey B Gabriel, Charles Lee, Mark J

Daly, et al. Common deletion polymorphisms in the human genome. *Nature genetics*, 38(1):86–92, 2006.

[120] ES Venkatraman and Adam B Olshen. A faster circular binary segmentation algorithm for the analysis of array cgh data. *Bioinformatics*, 23(6):657–663, 2007.

[121] Thomas S Price, Regina Regan, Richard Mott, Åsa Hedman, Ben Honey, Rachael J Daniels, Lee Smith, Andy Greenfield, Ana Tiganescu, Veronica Buckle, et al. Sw-array: a dynamic programming solution for the identification of copy-number changes in genomic dna using array comparative genome hybridization data. *Nucleic acids research*, 33(11):3455–3464, 2005.

[122] Thomas LaFramboise, Barbara A Weir, Xiaojun Zhao, Rameen Beroukhim, Cheng Li, David Harrington, William R Sellers, and Matthew Meyerson. Allele-specific amplification in cancer revealed by snp array analysis. *PLoS computational biology*, 1(6):e65, 2005.

[123] Daisuke Komura, Fan Shen, Shumpei Ishikawa, Karen R Fitch, Wenwei Chen, Jane Zhang, Guoying Liu, Sigeo Ihara, Hiroshi Nakamura, Matthew E Hurles, et al. Genome-wide detection of human copy number variations using high-density dna oligonucleotide arrays. *Genome research*, 16(12):1575–1584, 2006.

[124] Cheng Li and Wing Hung Wong. Dna-chip analyzer (dchip). In *The Analysis of Gene Expression Data*, pages 120–141. Springer, 2003.

[125] Yasuhito Nannya, Masashi Sanada, Kumi Nakazaki, Noriko Hosoya, Lili Wang, Akira Hangaishi, Mineo Kurokawa, Shigeru Chiba, Dione K Bailey, Giulia C Kennedy, et al. A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer research*, 65(14):6071–6079, 2005.

[126] Xuedong D Huang, Yasuo Ariki, and Mervyn A Jack. *Hidden Markov models for speech recognition*, volume 2004. Edinburgh university press Edinburgh, 1990.

[127] Richard Durbin. *Biological sequence analysis: probabilistic models of proteins and nucleic acids*. Cambridge university press, 1998.

[128] Robert J Elliott, Lakhdar Aggoun, and John B Moore. *Hidden Markov Models*. Springer, 1994.

[129] Lawrence Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. *Proceedings of the IEEE*, 77(2):257–286, 1989.

[130] Mark Stamp. A revealing introduction to hidden markov models. *Department of Computer Science San Jose State University*, 2004.

[131] G David Forney Jr. The viterbi algorithm. *Proceedings of the IEEE*, 61(3):268–278, 1973.

[132] Geoffrey McLachlan and Thriyambakam Krishnan. *The EM algorithm and extensions*, volume 382. John Wiley & Sons, 2007.

[133] Stefano Colella, Christopher Yau, Jennifer M Taylor, Ghazala Mirza, Helen Butler, Penny Clouston, Anne S Bassett, Anneke Seller, Christopher C Holmes, and Jiannis Ragoussis. Quantisnp: an objective bayes hidden-markov model to detect and accurately map copy number variation using snp genotyping data. *Nucleic acids research*, 35(6):2013–2025, 2007.

[134] Laura Winchester, Christopher Yau, and Jiannis Ragoussis. Comparing cnv detection methods for snp arrays. *Briefings in functional genomics & proteomics*, 8 (5):353–366, 2009.

[135] Dan Levy, Michael Ronemus, Boris Yamrom, Yoon-ha Lee, Anthony Leotta, Jude Kendall, Steven Marks, B Lakshmi, Deepa Pai, Kenny Ye, et al. Rare de novo and transmitted copy-number variation in autistic spectrum disorders. *Neuron*, 70(5): 886–897, 2011.

[136] Sujit Maiti, Kiran Halagur Bhoge Gowda Kumar, Christina A Castellani, Richard O'Reilly, and Shiva M Singh. Ontogenetic de novo copy number variations (cnvs) as a source of genetic individuality: studies on two families with mzd twins for schizophrenia. *PLoS One*, 6(3):e17125, 2011.

[137] Robert B Scharpf, Terri H Beaty, Holger Schwender, Samuel G Younkin, Alan F Scott, and Ingo Ruczinski. Fast detection of de novo copy number variants from snp arrays for case-parent trios. *BMC bioinformatics*, 13(1):330, 2012.

[138] Thomas LaFramboise. Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances. *Nucleic acids research*, page gkp552, 2009.

[139] Roger Pique-Regi, Antonio Ortega, and Shahab Asgharzadeh. Joint estimation of copy number variation and reference intensities on multiple dna arrays using gada. *Bioinformatics*, 25(10):1223–1230, 2009.

[140] Guillem Rigaill, Philippe Hupé, Anna Almeida, Philippe La Rosa, Jean-Philippe Meyniel, Charles Decraene, and Emmanuel Barillot. Italics: an algorithm for normalization and dna copy number calling for affymetrix snp arrays. *Bioinformatics*, 24(6):768–774, 2008.

[141] Philippe Hupé, Nicolas Stransky, Jean-Paul Thiery, François Radvanyi, and Emmanuel Barillot. Analysis of array cgh data: from signal ratio to gain and loss of dna regions. *Bioinformatics*, 20(18):3413–3422, 2004.

[142] David J Balding. A tutorial on statistical methods for population association studies. *Nature Reviews Genetics*, 7(10):781–791, 2006.

[143] Mark J Daly, John D Rioux, Stephen F Schaffner, Thomas J Hudson, and Eric S Lander. High-resolution haplotype structure in the human genome. *Nature genetics*, 29(2):229–232, 2001.

[144] Jacob Cohen, Patricia Cohen, Stephen G West, and Leona S Aiken. *Applied multiple regression/correlation analysis for the behavioral sciences*. Routledge, 2013.

[145] John Neter, Michael H Kutner, Christopher J Nachtsheim, and William Wasserman. *Applied linear statistical models*, volume 4. Irwin Chicago, 1996.

[146] Rita M Cantor, Kenneth Lange, and Janet S Sinsheimer. Prioritizing gwas results: a review of statistical methods and recommendations for their application. *The American Journal of Human Genetics*, 86(1):6–22, 2010.

[147] Herman O Hartley. The modified gauss-newton method for the fitting of non-linear regression functions by least squares. *Technometrics*, 3(2):269–280, 1961.

[148] David P MacKinnon, Chondra M Lockwood, C Hendricks Brown, Wei Wang, and Jeanne M Hoffman. The intermediate endpoint effect in logistic and probit regression. *Clinical Trials*, 4(5):499–513, 2007.

[149] Evangelos Evangelou, Jacques Fellay, Sara Colombo, Javier Martinez-Picado, Niels Obel, David B Goldstein, Amalio Telenti, and John PA Ioannidis. Impact of phenotype definition on genome-wide association signals: empirical evaluation in

human immunodeficiency virus type 1 infection. *American journal of epidemiology*, 173(11):1336–1342, 2011.

[150] Harris Cooper, Larry V Hedges, and Jeffrey C Valentine. *The handbook of research synthesis and meta-analysis*. Russell Sage Foundation, 2009.

[151] Robert Lawrence, Aaron G Day-Williams, Katherine S Elliott, Andrew P Morris, and Eleftheria Zeggini. Ccravat and qutie-enabling analysis of rare variants in large-scale case control and quantitative trait association studies. *BMC bioinformatics*, 11(1):527, 2010.

[152] Tiago V Pereira, Nikolaos A Patsopoulos, Georgia Salanti, and John PA Ioannidis. Discovery properties of genome-wide association signals from cumulatively combined data sets. *American journal of epidemiology*, page kwp262, 2009.

[153] Buhm Han and Eleazar Eskin. Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *The American Journal of Human Genetics*, 88(5):586–598, 2011.

[154] David Altshuler, Mark J Daly, and Eric S Lander. Genetic mapping in human disease. *science*, 322(5903):881–888, 2008.

[155] William Astle and David J Balding. Population structure and cryptic relatedness in genetic association studies. *Statistical Science*, pages 451–471, 2009.

[156] Sholom Wacholder, Nathaniel Rothman, and Neil Caporaso. Counterpoint: bias from population stratification is not a major threat to the validity of conclusions from epidemiological studies of common polymorphisms and cancer. *Cancer Epidemiology Biomarkers & Prevention*, 11(6):513–520, 2002.

[157] Lori J Lawson Handley, Andrea Manica, Jérôme Goudet, and Francois Balloux. Going the distance: human population genetics in a clinal world. *TRENDS in Genetics*, 23(9):432–439, 2007.

[158] Alkes L Price, Noah A Zaitlen, David Reich, and Nick Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11(7):459–463, 2010.

[159] Jonathan Marchini, Lon R Cardon, Michael S Phillips, and Peter Donnelly. The effects of human population structure on large genetic association studies. *Nature genetics*, 36(5):512–517, 2004.

[160] B Devlin, Silviu-Alin Bacanu, and Kathryn Roeder. Genomic control to the extreme. *Nature genetics*, 36(11):1129–1130, 2004.

[161] Gang Zheng, Boris Freidlin, and Joseph L Gastwirth. Robust genomic control for association studies. *The American Journal of Human Genetics*, 78(2):350–356, 2006.

[162] David E Reich and David B Goldstein. Detecting association in a case-control study while correcting for population stratification. *Genetic epidemiology*, 20(1): 4–16, 2001.

[163] Giovanni Montana and Clive Hoggart. Statistical software for gene mapping by admixture linkage disequilibrium. *Briefings in bioinformatics*, 8(6):393–395, 2007.

[164] Guillaume Evanno, Sebastien Regnaut, and Jérôme Goudet. Detecting the number of clusters of individuals using the software structure: a simulation study. *Molecular ecology*, 14(8):2611–2620, 2005.

[165] Efrosini Setakis, Heide Stirnadel, and David J Balding. Logistic regression protects against population structure in genetic association studies. *Genome research*, 16 (2):290–296, 2006.

[166] Alkes L Price, Nick J Patterson, Robert M Plenge, Michael E Weinblatt, Nancy A Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nature genetics*, 38(8):904–909, 2006.

[167] Martin Senne, Benjamin Trendelkamp-Schroer, Antonia SJS Mey, Christof Schutte, and Frank Noe. Emma: a software package for markov model building and analysis. *Journal of Chemical Theory and Computation*, 8(7):2223–2238, 2012.

[168] Yurii S Aulchenko, Dirk-Jan De Koning, and Chris Haley. Genomewide rapid association using mixed model and regression: a fast and simple method for genomewide pedigree-based quantitative trait loci association analysis. *Genetics*, 177(1):577–585, 2007.

[169] Il-Gyo Chong and Chi-Hyuck Jun. Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory Systems*, 78(1):103–112, 2005.

[170] Anil K Jain and Richard C Dubes. *Algorithms for clustering data*. Prentice-Hall, Inc., 1988.

[171] Anil K Jain, M Narasimha Murty, and Patrick J Flynn. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

[172] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 289–300, 1995.

[173] Frank Dudbridge and Bobby PC Koeleman. Rank truncated product of p-values, with application to genomewide association scans. *Genetic epidemiology*, 25(4): 360–366, 2003.

[174] Gregory S Barsh, Gregory P Copenhaver, Greg Gibson, and Scott M Williams. Guidelines for genome-wide association studies. *PLoS genetics*, 8(7):e1002812, 2012.

[175] K Miclaus, R Wolfinger, S Vega, M Chierici, C Furlanello, C Lambert, H Hong, Li Zhang, S Yin, and F Goodsaid. Batch effects in the brlmm genotype calling algorithm influence gwas results for the affymetrix 500k array. *The pharmacogenomics journal*, 10(4):336–346, 2010.

[176] Jesper Lund Pedersen. Optimal prediction of the ultimate maximum of brownian motion. *Stochastics and Stochastic Reports*, 75(4):205–219, 2003.

[177] Javed Khan, Jun S Wei, Markus Ringner, Lao H Saal, Marc Ladanyi, Frank Westermann, Frank Berthold, Manfred Schwab, Cristina R Antonescu, Carsten Peterson, et al. Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks. *Nature medicine*, 7(6):673–679, 2001.

[178] Jan J Gerbrands. On the relationships between svd, klt and pca. *Pattern recognition*, 14(1):375–381, 1981.

[179] Stacey B Gabriel, Stephen F Schaffner, Huy Nguyen, Jamie M Moore, Jessica Roy, Brendan Blumenstiel, John Higgins, Matthew DeFelice, Amy Lochner, Maura Faggart, et al. The structure of haplotype blocks in the human genome. *Science*, 296(5576):2225–2229, 2002.

[180] Devin P Locke, Andrew J Sharp, Steven A McCarroll, Sean D McGrath, Tera L Newman, Ze Cheng, Stuart Schwartz, Donna G Albertson, Daniel Pinkel, David M Altshuler, et al. Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *The American Journal of Human Genetics*, 79(2):275–290, 2006.

[181] Chris Power and Jane Elliott. Cohort profile: 1958 british birth cohort (national child development study). *International journal of epidemiology*, 35(1): 34–41, 2006.

[182] John Wright, Nicola Pickard, Angelique Whitfield, and Nigel Hakin. A population-based study of the prevalence, clinical characteristics and effect of ethnicity in epilepsy. *Seizure*, 9(5):309–313, 2000.

[183] Michael A Rogawski and Michael R Johnson. Intrinsic severity as a determinant of antiepileptic drug refractoriness. *Epilepsy Currents*, 8(5):127–130, 2008.

[184] Doug Speed, Clive Hoggart, Slave Petrovski, Ioanna Tachmazidou, Alison Coffey, Andrea Jorgensen, Hariklia Eleftherohorinou, Maria De Iorio, Marian Todaro, Tisham De, et al. A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy. *Human molecular genetics*, 23(1):247–258, 2014.

[185] Anthony G Marson, Asya M Al-Kharusi, Muna Alwaidh, Richard Appleton, Gus A Baker, David W Chadwick, Celia Cramp, Oliver C Cockerell, Paul N Cooper, Julie Doughty, et al. The sanad study of effectiveness of valproate, lamotrigine, or topiramate for generalised and unclassifiable epilepsy: an unblinded randomised controlled trial. *The Lancet*, 369(9566):1016–1026, 2007.

[186] Patrick Kwan and Martin J Brodie. Early identification of refractory epilepsy. *New England Journal of Medicine*, 342(5):314–319, 2000.

[187] Lachlan JM Coin, Dandan Cao, Jingjing Ren, Xianbo Zuo, Liangdan Sun, Sen Yang, Xuejun Zhang, Yong Cui, Yingrui Li, Xin Jin, et al. An exome sequencing pipeline for identifying and genotyping common cnvs associated with disease with application to psoriasis. *Bioinformatics*, 28(18):i370–i374, 2012.

[188] Sam Roweis. Em algorithms for pca and spca. *Advances in neural information processing systems*, pages 626–632, 1998.

[189] Cristen J Willer, Yun Li, and Gonçalo R Abecasis. Metal: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics*, 26(17):2190–2191, 2010.

[190] Heather C Mefford, Hiltrud Muhle, Philipp Ostertag, Sarah von Spiczak, Karen Buysse, Carl Baker, Andre Franke, Alain Malafosse, Pierre Genton, Pierre Thomas, et al. Genome-wide copy number variation in epilepsy: novel susceptibility loci in idiopathic generalized and focal epilepsies. *PLoS genetics*, 6(5): e1000962, 2010.

[191] TR Dening, GE Berrios, and JM Walshe. Wilson's disease and epilepsy. *Brain*, 111(5):1139–1155, 1988.

[192] Aglaia Vignoli, Giulia Federica Scornavacca, Angela Peron, Francesca La Briola, and Maria Paola Canevini. Interstitial 6q microdeletion syndrome and epilepsy: A new patient and review of the literature. *American Journal of Medical Genetics Part A*, 161(8):2009–2015, 2013.

[193] Erin L Heinzen, Rodney A Radtke, Thomas J Urban, Gianpiero L Cavalleri, Chantal Depondt, Anna C Need, Nicole M Walley, Paola Nicoletti, Dongliang Ge, Claudia B Catarino, et al. Rare deletions at 16p13. 11 predispose to a diverse spectrum of sporadic epilepsy syndromes. *The American Journal of Human Genetics*, 86 (5):707–718, 2010.

[194] Stephen M Dombrowski, Shailesh Y Desai, Matteo Marroni, Luca Cucullo, Kris Goodrich, William Bingaman, Marc R Mayberg, Ljiljana Bengez, and Damir Janigro. Overexpression of multiple drug resistance genes in endothelial cells from patients with refractory epilepsy. *Epilepsia*, 42(12):1501–1506, 2001.

[195] Yuko Fukata, Hillel Adesnik, Tsuyoshi Iwanaga, David S Bredt, Roger A Nicoll, and Masaki Fukata. Epilepsy-related ligand/receptor complex lgi1 and adam22 regulate synaptic transmission. *Science*, 313(5794):1792–1795, 2006.

[196] Judith H Miles. Autism spectrum disordersa genetics review. *Genetics in Medicine*, 13(4):278–294, 2011.

[197] Catalina Betancur, Takeshi Sakurai, and Joseph D Buxbaum. The emerging role of synaptic cell-adhesion pathways in the pathogenesis of autism spectrum disorders, 2009.

[198] Rui T Peixoto, Portia A Kunz, Hyungbae Kwon, Angela M Mabb, Bernardo L Sabatini, Benjamin D Philpot, and Michael D Ehlers. Transsynaptic signaling by activity-dependent cleavage of neuroligin-1. *Neuron*, 76(2):396–409, 2012.

[199] Kunimichi Suzuki, Yukari Hayashi, Soichiro Nakahara, Hiroshi Kumazaki, Johannes Prox, Keisuke Horiuchi, Mingshuo Zeng, Shun Tanimura, Yoshitake Nishiyama, Satoko Osawa, et al. Activity-dependent proteolytic cleavage of neuroligin-1. *Neuron*, 76(2):410–422, 2012.

[200] Leonie Welberg. Synaptic plasticity: Neuroligin 1 does the splits. *Nature Reviews Neuroscience*, 13(12):811–811, 2012.

[201] Jing Hua Zhao. gap: Genetic analysis package. *J Stat Softw*, 23(8):1–18, 2007.

[202] Rami I Aqeilan, Valentina Donati, Alexey Palamarchuk, Francesco Trapasso, Mohamed Kaou, Yuri Pekarsky, Marius Sudol, and Carlo M Croce. Ww domain–containing proteins, wwox and yap, compete for interaction with erbb-4 and modulate its transcriptional function. *Cancer research*, 65(15):6764–6772, 2005.

[203] Andrzej K Bednarek, Kendra J Laflin, Rachael L Daniel, Qiaoyin Liao, Kathleen A Hawkins, and C Marcelo Aldaz. Wwox, a novel ww domain-containing protein mapping to human chromosome 16q23. 3–24.1, a region frequently affected in breast cancer. *Cancer research*, 60(8):2140–2145, 2000.

[204] Maria I Nunez, John Ludes-Meyers, and C Marcelo Aldaz. Wwox protein expression in normal human tissues. *Journal of molecular histology*, 37(3-4):115–125, 2006.

[205] H Suzuki, K Katayama, M Takenaka, K Amakasu, K Saito, and K Suzuki. A spontaneous mutation of the wwox gene and audiogenic seizures in rats with lethal dwarfism and epilepsy. *Genes, Brain and Behavior*, 8(7):650–660, 2009.

[206] Martial Mallaret, Matthis Synofzik, Jaeho Lee, Cari A Sagum, Muhammad Mahajnah, Rajech Sharkia, Nathalie Drouot, Mathilde Renaud, Fabrice AC Klein, Mathieu Anheim, et al. The tumour suppressor gene wwox is mutated in autosomal recessive cerebellar ataxia with epilepsy and mental retardation. *Brain*, page awt338, 2013.

[207] Mahmoud Reza Mansouri, Lena Marklund, Peter Gustavsson, Edward Davey, Birgit Carlsson, Catharina Larsson, Irene White, Karl-Henrik Gustavson, and Niklas

Dahl. Loss of zdhhc15 expression in a woman with a balanced translocation t (x; 15)(q13. 3; cen) and severe mental retardation. *European journal of human genetics*, 13(8):970–977, 2005.

[208] Helen V Firth, Shola M Richards, A Paul Bevan, Stephen Clayton, Manuel Corpas, Diana Rajan, Steven Van Vooren, Yves Moreau, Roger M Pettett, and Nigel P Carter. Decipher:database of chromosomal imbalance and phenotype in humans using ensembl resources. *The American Journal of Human Genetics*, 84(4):524–533, 2009.

[209] Nora Fernandez-Jimenez, Ainara Castellanos-Rubio, Leticia Plaza-Izurieta, Galder Gutierrez, Iñaki Irastorza, Luis Castaño, Juan Carlos Vitoria, and Jose Ramon Bilbao. Accuracy in copy number calling by qpcr and prt: a matter of dna. *PloS one*, 6(12):e28910, 2011.

[210] Michele Guescini, Davide Sisti, Marco BL Rocchi, Laura Stocchi, and Vilberto Stocchi. A new real-time pcr method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC bioinformatics*, 9(1):326, 2008.

[211] John A Hartigan and Manchek A Wong. Algorithm as 136: A k-means clustering algorithm. *Applied statistics*, pages 100–108, 1979.

[212] Qiu-sheng RONG, Jun-biao YAN, and Guo-qiang GUO. Research and implementation of clustering algorithm based on dbscan [j]. *Computer Applications*, 4, 2004.

[213] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd*, volume 96, pages 226–231, 1996.

[214] Norio Katayama and Shin'ichi Satoh. The sr-tree: An index structure for high-dimensional nearest neighbor queries. In *ACM SIGMOD Record*, volume 26, pages 369–380. ACM, 1997.

[215] Jörg Sander, Martin Ester, Hans-Peter Kriegel, and Xiaowei Xu. Density-based clustering in spatial databases: The algorithm gdbscan and its applications. *Data mining and knowledge discovery*, 2(2):169–194, 1998.

[216] Aleksandar Lazarevic and Zoran Obradovic. Knowledge discovery in multiple spatial databases. *Neural Computing & Applications*, 10(4):339–350, 2002.

[217] Johannes Kettunen, Taru Tukiainen, Antti-Pekka Sarin, Alfredo Ortega-Alonso, Emmi Tikkanen, Leo-Pekka Lyytikäinen, Antti J Kangas, Pasi Soininen, Peter Würtz, Kaisa Silander, et al. Genome-wide association study identifies multiple loci influencing human serum metabolite levels. *Nature genetics*, 44(3):269–276, 2012.

[218] Jerzy Adamski. Genome-wide association studies with metabolomics. *Genome Med*, 4(4):34, 2012.

[219] Simone Wahl, Susanne Krug, Cornelia Then, Anna Kirchhofer, Gabi Kastenmüller, Tina Brand, Thomas Skurk, Melina Claussnitzer, Cornelia Huth, Margit Heier, et al. Comparative analysis of plasma metabolomics response to metabolic challenge tests in healthy subjects and influence of the fto obesity risk allele. *Metabolomics*, 10(3):386–401, 2014.

[220] Rosa Vázquez-Fresno, Rafael Llorach, Mireia Urpi-Sarda, Olha Khymenets, Mònica Bulló, Dolores Corella, Montserrat Fitó, Miguel Angel Martínez-González, Ramon Estruch, and Cristina Andres-Lacueva. An nmr metabolomics approach reveals a combined-biomarkers model in a wine interventional trial with validation in free-living individuals of the predimed study. *Metabolomics*, pages 1–10, 2014.

[221] Daniel J Vis, Johan A Westerhuis, Doris M Jacobs, John PM van Duynhoven, Suzan Wopereis, Ben van Ommen, Margriet MWB Hendriks, and Age K Smilde. Analyzing metabolomics-based challenge tests. *Metabolomics*, pages 1–14, 2014.

[222] Matej Orešič, Satu Simell, Marko Sysi-Aho, Kirsti Näntö-Salonen, Tuulikki Seppänen-Laakso, Vilhelmiina Parikka, Mikko Katajamaa, Anne Hekkala, Ismo Mattila, Päivi Keskinen, et al. Dysregulation of lipid and amino acid metabolism precedes islet autoimmunity in children who later progress to type 1 diabetes. *The Journal of experimental medicine*, 205(13):2975–2984, 2008.

[223] Carolyn M Slupsky, Andriy Cheypesh, Danny V Chao, Hao Fu, Kathryn N Rankin, Thomas J Marrie, and Paige Lacy. Streptococcus pneumoniae and staphylococcus aureus pneumonia induce distinct metabolic responses. *Journal of proteome research*, 8(6):3029–3036, 2009.

[224] Natalie J Serkova and Tessa Little. Metabolomics in infections (hiv). *Metabolomics*, pages 182–191, 2014.

[225] Jennifer L Spratlin, Natalie J Serkova, and S Gail Eckhardt. Clinical applications of metabolomics in oncology: a review. *Clinical Cancer Research*, 15(2):431–440, 2009.

[226] DS Wishart. Metabolomics: the principles and potential applications to transplantation. *American Journal of Transplantation*, 5(12):2814–2820, 2005.

[227] Bennett Daviss. Growing pains for metabolomics. *The Scientist*, 19(8):25–28, 2005.

[228] Kate W Jordan, Johan Nordenstam, Gregory Y Lauwers, David A Rothenberger, Karim Alavi, Michael Garwood, and Leo L Cheng. Metabolomic characterization of human rectal adenocarcinoma with intact tissue magnetic resonance spectroscopy. *Diseases of the colon and rectum*, 52(3):520, 2009.

[229] Elaine Holmes, Ian D Wilson, and Jeremy K Nicholson. Metabolic phenotyping in health and disease. *Cell*, 134(5):714–717, 2008.

[230] Sean Ekins, Yuri Nikolsky, and Tatiana Nikolskaya. Techniques: application of systems biology to absorption, distribution, metabolism, excretion and toxicity. *Trends in pharmacological sciences*, 26(4):202–209, 2005.

[231] Donald G Robertson. Metabonomics in toxicology: a review. *Toxicological Sciences*, 85(2):809–822, 2005.

[232] Wolfram Weckwerth. Metabolomics in systems biology. *Annual review of plant biology*, 54(1):669–689, 2003.

[233] Pirkko Räsänen, Helinä Hakko, Matti Isohanni, Sheilagh Hodgins, Marjo-Riitta Järvelin, and Jari Tiihonen. Maternal smoking during pregnancy and risk of criminal behavior among adult male offspring in the northern finland 1966 birth cohort. *American journal of psychiatry*, 156(6):857–862, 1999.

[234] JG Eriksson. Early growth, and coronary heart disease and type 2 diabetes: experiences from the helsinki birth cohort studies. *International Journal of Obesity*, 30:S18–S22, 2006.

[235] Aki Vehtari, Ville-Petteri Mäkinen, Pasi Soininen, Petri Ingman, Sanna M Mäkelä, Markku J Savolainen, Minna L Hannuksela, Kimmo Kaski, and Mika Ala-Korpela. A novel bayesian approach to quantify clinical variables and to determine their

spectroscopic counterparts in 1h nmr metabonomic data. *BMC bioinformatics*, 8 (Suppl 2):S8, 2007.

[236] Chiara Sabatti, Susan K Service, Anna-Liisa Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G Jones, Noah A Zaitlen, Teppo Varilo, Marika Kaakinen, et al. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genetics*, 41(1):35–46, 2008.

[237] Beatriz Valcárcel, Peter Würtz, Nafisa-Katrin Seich al Basatena, Taru Tukiainen, Antti J Kangas, Pasi Soininen, Marjo-Riitta Järvelin, Mika Ala-Korpela, Timothy M Ebbels, and Maria de Iorio. A differential network approach to exploring differences between biological states: an application to prediabetes. *PloS one*, 6 (9):e24702, 2011.

[238] Paul F OReilly, Clive J Hoggart, Yotsawat Pomyen, Federico CF Calboli, Paul Elliott, Marjo-Riitta Jarvelin, and Lachlan JM Coin. Multiphen: joint model of multiple phenotypes can increase discovery in gwas. *PLoS One*, 7(5):e34861, 2012.

[239] Patrick Breheny, Prabhakar Chalise, Anthony Batzler, Liewei Wang, and Brooke L Fridley. Genetic association studies of copy-number variation: should assignment of copy number states precede testing? *PloS one*, 7(4):e34262, 2012.

[240] Beverley Balkau, Céline Lange, Leopold Fezeu, Jean Tichet, Blandine de Lauzon-Guillain, Sebastien Czernichow, Frederic Fumeron, Philippe Froguel, Martine Vaxillaire, Stephane Cauchi, et al. Predicting diabetes: Clinical, biological, and genetic approaches data from the epidemiological study on the insulin resistance syndrome (desir). *Diabetes Care*, 31(10):2056–2061, 2008.

[241] Dale R Nyholt. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *The American Journal of Human Genetics*, 74(4):765–769, 2004.

[242] J Li and L Ji. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity*, 95(3):221–227, 2005.

[243] Joseph L Schafer and John W Graham. Missing data: our view of the state of the art. *Psychological methods*, 7(2):147, 2002.

[244] Stef Van Buuren, Jaap PL Brand, CGM Groothuis-Oudshoorn, and Donald B Rubin. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*, 76(12):1049–1064, 2006.

[245] Xue-Jun Zhang, Wei Huang, Sen Yang, Liang-Dan Sun, Feng-Yu Zhang, Qi-Xing Zhu, Fu-Ren Zhang, Chi Zhang, Wen-Hui Du, Xiong-Ming Pu, et al. Psoriasis genome-wide association study identifies susceptibility variants within lce gene cluster at 1q21. *Nature genetics*, 41(2):205–210, 2009.

[246] S Orru, E Giuressi, M Casula, A Loizedda, R Murru, M Mulargia, MV Masala, D Cerimele, M Zucca, N Aste, et al. Psoriasis is associated with a snp haplotype of the corneodesmosin gene (cdsn). *Tissue Antigens*, 60(4):292–298, 2002.

[247] Charlotte Lundsgaard, Ole Hamberg, Ole Østergaard Thomsen, Ole Haagen Nielsen, and Hendrik Vilstrup. Increased hepatic urea synthesis in patients with active inflammatory bowel disease. *Journal of hepatology*, 24(5):587–593, 1996.

[248] Dan L Nicolae, Eric Gamazon, Wei Zhang, Shiwei Duan, M Eileen Dolan, and Nancy J Cox. Trait-associated snps are more likely to be eqtls: annotation to enhance discovery from gwas. *PLoS genetics*, 6(4):e1000888, 2010.

[249] Gokhan S Hotamisligil, Peter Arner, Jose F Caro, Richard L Atkinson, and Bruce M Spiegelman. Increased adipose tissue expression of tumor necrosis factor-alpha in human obesity and insulin resistance. *Journal of Clinical Investigation*, 95(5):2409, 1995.

[250] Philip A Kern, Subramanian Ranganathan, Chunling Li, Linda Wood, and Gouri Ranganathan. Adipose tissue tumor necrosis factor and interleukin-6 expression in human obesity and insulin resistance. *American Journal of Physiology-Endocrinology And Metabolism*, 280(5):E745–E751, 2001.

[251] Daniah Trabzuni, Mina Ryten, Robert Walker, Colin Smith, Sabaena Imran, Adaikalavan Ramasamy, Michael E Weale, and John Hardy. Quality control parameters on a large dataset of regionally dissected human control brains for whole genome expression studies. *Journal of neurochemistry*, 119(2):275–282, 2011.

[252] Laura J van't Veer, Hongyue Dai, Marc J Van De Vijver, Yudong D He, Augustinus AM Hart, Mao Mao, Hans L Peterse, Karin van der Kooy, Matthew J Marton, Anke T Witteveen, et al. Gene expression profiling predicts clinical outcome of breast cancer. *nature*, 415(6871):530–536, 2002.

[253] Adaikalavan Ramasamy, Daniah Trabzuni, Sebastian Guelfi, Vibin Varghese, Colin Smith, Robert Walker, Tisham De, Lachlan Coin, Rohan de Silva, Mark R Cookson, et al. Genetic variability in the regulation of gene expression in ten regions of the human brain. *Nature neuroscience*, 2014.

[254] Andreas Rosenwald, Ash A Alizadeh, George Widhopf, Richard Simon, R Eric Davis, Xin Yu, Liming Yang, Oxana K Pickeral, Laura Z Rassenti, John Powell, et al. Relation of gene expression phenotype to immunoglobulin mutation genotype in b cell chronic lymphocytic leukemia. *The Journal of experimental medicine*, 194 (11):1639–1648, 2001.

[255] Andrea H Brand and Norbert Perrimon. Targeted gene expression as a means of altering cell fates and generating dominant phenotypes. *development*, 118(2): 401–415, 1993.

[256] Joseph L DeRisi, Vishwanath R Iyer, and Patrick O Brown. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 278(5338):680–686, 1997.

[257] Patricia J Wittkopp, Belinda K Haerum, and Andrew G Clark. Evolutionary changes in cis and trans gene regulation. *Nature*, 430(6995):85–88, 2004.

[258] Enrico Petretto, Jonathan Mangion, Nicholas J Dickens, Stuart A Cook, Mande K Kumaran, Han Lu, Judith Fischer, Henrike Maatz, Vladimir Kren, Michal Pravenec, et al. Heritability and tissue specificity of expression quantitative trait loci. *PLoS genetics*, 2(10):e172, 2006.

[259] Stephen B Montgomery and Emmanouil T Dermitzakis. From expression qtls to personalized transcriptomics. *Nature Reviews Genetics*, 12(4):277–282, 2011.

[260] Xia Yang, Eric E Schadt, Susanna Wang, Hui Wang, Arthur P Arnold, Leslie Ingram-Drake, Thomas A Drake, and Aldons J Lusis. Tissue-specific expression and regulation of sexually dimorphic genes in mice. *Genome research*, 16(8):995–1004, 2006.

[261] Derek R Drost, Catherine I Benedict, Arthur Berg, Evandro Novaes, Carolina RDB Novaes, Qibin Yu, Christopher Dervinis, Jessica M Maia, John Yap, Brianna Miles, et al. Diversification in the genetic architecture of gene expression and transcriptional networks in organ differentiation of populus. *Proceedings of the National Academy of Sciences*, 107(18):8492–8497, 2010.

[262] Jun Ding, Johann E Gudjonsson, Liming Liang, Philip E Stuart, Yun Li, Wei Chen, Michael Weichenthal, Eva Ellinghaus, Andre Franke, William Cookson, et al. Gene expression in skin and lymphoblastoid cells: Refined statistical method reveals extensive overlap in cis-eqtl signals. *The American Journal of Human Genetics*, 87(6):779–789, 2010.

[263] Alexandra C Nica and Emmanouil T Dermitzakis. Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1620):20120362, 2013.

[264] ENCODE Project Consortium et al. The encode (encyclopedia of dna elements) project. *Science*, 306(5696):636–640, 2004.

[265] Joseph E Powell, Anjali K Henders, Allan F McRae, Margaret J Wright, Nicholas G Martin, Emmanouil T Dermitzakis, Grant W Montgomery, and Peter M Visscher. Genetic control of gene expression in whole blood and lymphoblastoid cell lines is largely independent. *Genome research*, 22(3):456–466, 2012.

[266] Luz D Orozco, Shawn J Cokus, Anatole Ghazalpour, Leslie Ingram-Drake, Susanna Wang, Atila Van Nas, Nam Che, Jesus A Araujo, Matteo Pellegrini, and Aldons J Lusis. Copy number variation influences gene expression and metabolic traits in mice. *Human molecular Genetics*, 18(21):4118–4129, 2009.

[267] Jeffrey C Barrett, Sarah Hansoul, Dan L Nicolae, Judy H Cho, Richard H Duerr, John D Rioux, Steven R Brant, Mark S Silverberg, Kent D Taylor, M Michael Barmada, et al. Genome-wide association defines more than 30 distinct susceptibility loci for crohn's disease. *Nature Genetics*, 40(8):955–962, 2008.

[268] Andrew J Walley, Julian E Asher, and Philippe Froguel. The genetic contribution to non-syndromic human obesity. *Nature Reviews Genetics*, 10(7):431–442, 2009.

[269] Eric R Gamazon, Dan L Nicolae, and Nancy J Cox. A study of cnvs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS genetics*, 7(2):e1001292, 2011.

[270] Barbara E Stranger, Stephen B Montgomery, Antigone S Dimas, Leopold Parts, Oliver Stegle, Catherine E Ingle, Magda Sekowska, George Davey Smith, David Evans, Maria Gutierrez-Arcelus, et al. Patterns of cis regulatory variation in diverse human populations. *PLoS genetics*, 8(4):e1002639, 2012.

[271] Tomi Pastinen. Genome-wide allele-specific analysis: insights into regulatory variation. *Nature Reviews Genetics*, 11(8):533–538, 2010.

[272] Tamara Steijger, Josep F Abril, Pär G Engström, Felix Kokocinski, Tim J Hubbard, Roderic Guigó, Jennifer Harrow, Paul Bertone, RGASP Consortium, et al. Assessment of transcript reconstruction methods for rna-seq. *Nature methods*, 2013.

[273] Timothée Flutre, Xiaoquan Wen, Jonathan Pritchard, and Matthew Stephens. A statistical framework for joint eqtl analysis in multiple tissues. *PLoS genetics*, 9 (5):e1003486, 2013.

[274] Shaun Purcell, Benjamin Neale, Kathe Todd-Brown, Lori Thomas, Manuel AR Ferreira, David Bender, Julian Maller, Pamela Sklar, Paul IW De Bakker, Mark J Daly, et al. Plink: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics*, 81(3):559–575, 2007.

[275] Karl W Broman, Hao Wu, Śaunak Sen, and Gary A Churchill. R/qtl: Qtl mapping in experimental crosses. *Bioinformatics*, 19(7):889–890, 2003.

[276] Daniel M Gatti, Andrey A Shabalin, Tieu-Chong Lam, Fred A Wright, Ivan Rusyn, and Andrew B Nobel. Fastmap: fast eqtl mapping in homozygous populations. *Bioinformatics*, 25(4):482–489, 2009.

[277] Volker Strassen. Relative bilinear complexity and matrix multiplication. *Journal für die reine und angewandte Mathematik*, 375:406–443, 1987.

[278] CUDA Nvidia. Programming guide, 2008.

[279] Yonghong Yan, Max Grossman, and Vivek Sarkar. Jcuda: A programmer-friendly interface for accelerating java programs with cuda. In *Euro-Par 2009 Parallel Processing*, pages 887–899. Springer, 2009.

[280] Cross-Disorder Group of the Psychiatric Genomics Consortium et al. Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis. *Lancet*, 381(9875):1371, 2013.

[281] Erin L Heinzen, Dongliang Ge, Kenneth D Cronin, Jessica M Maia, Kevin V Shianna, Willow N Gabriel, Kathleen A Welsh-Bohmer, Christine M Hulette,

Thomas N Denny, and David B Goldstein. Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS biology*, 6(12):e1000001, 2008.

[282] Dena G Hernandez, Mike A Nalls, Matthew Moore, Sean Chong, Allissa Dillman, Daniah Trabzuni, J Raphael Gibbs, Mina Ryten, Sampath Arepalli, Michael E Weale, et al. Integration of gwas snps and tissue specific expression profiling reveal discrete eqtls for human traits in blood and brain. *Neurobiology of disease*, 47(1):20–28, 2012.

[283] Tanja Zeller, Philipp Wild, Silke Szymczak, Maxime Rotival, Arne Schillert, Raphaele Castagne, Seraya Maouche, Marine Germain, Karl Lackner, Heidi Rossmann, et al. Genetics and beyond–the transcriptome of human monocytes and disease susceptibility. *PloS one*, 5(5):e10693, 2010.

Appendix A

Lipid Names

Table A.1: Abbreviations of metabonomic lipids 1

Lipid Name	Symbol
Cholesterol esters in medium LDL	M-LDL-CE
Concentration of medium LDL particles	M-LDL-P
Cholesterol in small LDL	S-LDL-C
Concentration of small LDL particles	S-LDL-P
Cholesterol in very large HDL	XL-HDL-C
Free cholesterol in very large HDL	XL-HDL-FC
Phospholipids in very large HDL	XL-HDL-PL
Triglycerides in very large HDL	XL-HDL-TG
Cholesterol esters in very large HDL	XL-HDL-CE
Concentration of very large HDL particles	XL-HDL-P
Cholesterol in large HDL	L-HDL-C
Free cholesterol in large HDL	L-HDL-FC
Phospholipids in large HDL	L-HDL-PL
Cholesterol esters in large HDL	L-HDL-CE
Concentration of large HDL particles	L-HDL-P
Cholesterol in medium HDL	M-HDL-C
Free cholesterol in medium HDL	M-HDL-FC
Phospholipids in medium HDL	M-HDL-PL
Cholesterol esters in medium HDL	M-HDL-CE
Concentration of medium HDL particles	M-HDL-P
Triglycerides in small HDL	S-HDL-TG
Concentration of small HDL particle	S-HDL-P
Apolipoprotein A-I	ApoA.I.
Apolipoprotein B	ApoB.
Apolipoprotein B / Apolipoprotein A-I	ApoB.ApoA.I

Table A.2: Abbreviations of metabonomic lipids 2

Lipid Name	Symbol
Phospholipids in extremely large VLDL	XXL-VLDL-PL
Concentration in extremely large VLDL particles	XXL-VLDL-P
Phospholipids in very large VLDL	XL-VLDL-PL
Triglycerides in very large VLDL	XL-VLDL-TG
Concentration of very large VLDL particles	XL-VLDL-P
Cholesterol in large VLDL	L-VLDL-C
Free cholesterol in large	L-VLDL-FC
Phospholipids in large VLDL	L-VLDL-PL
Triglycerides in large VLDL	L-VLDL-TG
Cholesterol esters in large VLDL	L-VLDL-CE
Concentration of large VLDL particles	L-VLDL-P
Cholesterol in medium VLDL	M-VLDL-C
Free cholesterol in medium VLDL	M-VLDL-FC
Phospholipids in medium VLDL	M-VLDL-PL
Triglycerides in medium VLDL	M-VLDL-TG
Cholesterol esters in medium VLDL	M-VLDL-CE
Concentration of medium VLDL particles	M-VLDL-P
Cholesterol in small VLDL	S-VLDL-C
Free cholesterol in small VLDL	S-VLDL-FC
Phospholipids in small VLDL	S-VLDL-PL
Triglycerides in small VLDL	S-VLDL-TG
Concentration of small VLDL particles	S-VLDL-P
Phospholipids in very small VLDL	XS-VLDL-PL
Triglycerides in very small VLDL	XS-VLDL-TG
Concentration of very small VLDL particles	XS-VLDL-P
Free cholesterol in IDL	IDL-FC
Phospholipids in IDL	IDL-PL
Concentration of IDL particles	IDL-P
Cholesterol in large LDL	L-LDL-C
Free cholesterol in large LDL	L-LDL-FC
Phospholipids in large LDL	L-LDL-PL
Cholesterol esters in large LDL	L-LDL-CE
Concentration of large LDL particles	L-LDL-P
Cholesterol in medium LDL	M-LDL-C
Phospholipids in medium LDL	M-LDL-PL

Appendix B

Linear systems

Consider a $m * n$ (m equations with n variables) system of linear equations

$$Ax = b \tag{B.1}$$

If A is not singular and if $m = n$, then Eq. A.1 is referred to as fully determined system. If $m < n$, then it is said an under-determined system and if $m > n$, then it is said to be a over-determined system.

Solving fully determined systems

When $m * n$ system of linear equations with A not singular and $m = n$, then there exists a unique solution. The solution is given by

$$(A^{-1}A)x = A^{-1}b \tag{B.2}$$

when re-arranged gives

$$x = A^{-1}b \tag{B.3}$$

This can be solved using two simple steps

1. invert matrix A
2. premultiply the vector b by A^{-1}

However, this process is hardly ever used in practice. Some direct methods work by using Gaussian elimination which simply rearranges the matrix A. The main steps are

1. Transform the system of equations into triangular form
2. Solve the triangular set of equations through a series of variable substitutions

One of the main problems is that both sides Equation B.3 is modified and the system is forced into triangular form. This would require repetitions of the entire procedure if we want to solve Equation B.3 with a new b vector every time. The LU decomposition solves this issue by capturing intermediate states of A as the transformation to triangular form progresses. This makes the operations on A independent from of b and this allows Equation B.3 to be solved for many different b vectors.

The main steps in LU factorization produces a lower triangular matrix L and upper triangular matrix U such that

$$LU = A \tag{B.4}$$

Substituting equation B.3 into B.4 gives

$$(LU)x = b \tag{B.5}$$

and by associate the factors we get

$$L(Ux) = b \tag{B.6}$$

There are efficient algorithms for solving triangular systems e.g. forward substitutions for lower triangular systems and backward substitutions for upper triangular systems. Equation B6 suggests an algorithm for solving equation B.6 which we define vector y such that

$$y = Ux \tag{B.7}$$

Substituting Equation B.7 into B.6 yields

$$Ly = b \tag{B.8}$$

Since b is known and L is lower triangular, Equation B.8 can be solved for y by forward substitution. Once we get y, equation B.8 can be solved for x by back substitution.

To summarize:

1. Compute LU decomposition of A
2. Solve equation B.8 for y by forward substitutions
3. Solve equaiton B.7 for x by back substitutions

Solving under-determined systems

A $m * n$ system of linear equations where $m > n$ (or $m = n$ and A is singular) is under-determined. Under-determined systems have q linearly independent families of solutions where

$$q = n - r$$

and

$$\text{r=rank(A)}$$

The value of q is referred to as the nullity of matrix A. The q linearly dependent equations in A are the null space of A. Solving an under-determined set of equations which usually narrows down to solving a fully determined $r * r$ system (known as range of A) and adding this solution to any linear combination of the other q vectors of A. The numerical procedure which solves this problem is known as SVD which constructs a set of orthogonal bases for the null space and range of A.

Solving over-determined systems

A system of $m * n$ linear equations with $m < n$ is overdetermined. There are more equations than there are unknowns. Solving this equation is tantamount to reducing the system to an $m * m$ problem and then solving the reduced set of equations. A common technique for constructing reduced set of equations is the least-squares approach. The least-squares equations are derived by pre-multiplying B.1 by A^T

$$(A^T A)x = A^T b \tag{B.9}$$

Equation B.9 is also known as the normal equations of the least squares solution. The least squares approach minimizes the differences between the left and right side of the Equation B.1.

Computational complexity of linear systems

The decomposition algorithm for solving linear equations is motivated by the computational inefficiency of matrix inversion. In this case, the matrix in normal equation is rarely solved by inverting as a dense matrix requires

$$2n^3 + O(n^2)$$

floating point operations. Whereas LU decomposition requires

$$\frac{2}{3}n^3 + \frac{1}{2}n^2 + \frac{1}{6}n$$

$$\frac{2}{3}n^3 + O(n^2)$$

operations. Computing x from the factorization requires

$$2n^2 + n$$

or

$$2n^2 + O(n)$$

floating point operations, which is equivalent to the product $A^{-1}b$. Thus solving linear system of equations by matrix inversion requires approximately three times the number of operations than LU. In case of sparsity of A, inversion destroys this whereas LU preserves the sparsity. Special cases of LU exist for solving sparse matrices.

Appendix C

Publications

Publications arising from the work done in this thesis:

1. Doug Speed, Clive Hoggart, Slave Petrovski, Ioanna Tachmazidou, Alison Coey, Andrea Jorgensen, Hariklia Eleftherohorinou, Maria De Iorio, Marian Todaro, **Tisham De**, et al. **A genome-wide association study and biological pathway analysis of epilepsy prognosis in a prospective cohort of newly treated epilepsy**. *Human Molecular Genetics, 23(1):247 258, 2014.*

2. Adaikalavan Ramasamy, Daniah Trabzuni, Sebastian Guel, Vibin Varghese, Colin Smith, Robert Walker, **Tisham De**, Lachlan Coin, Rohan de Silva, Mark R Cookson, et al. **Genetic variability in the regulation of gene expression in ten regions of the human brain**. *Nature Neuroscience, 2014.*

3. **Tisham De*** et al. **Metabonomic signatures of CNVs in TSPAN8 exonic regions associated with Type 2 Diabetes**. *Manuscript under preparation*