Integration of genetic data and genomic annotation in the analysis of genome-wide association studies

Esther Ng

Lady Margaret Hall

DPhil Genomic Medicine and Statistics

THESIS ABSTRACT

This thesis explored the utility of genome wide association studies (GWAS) and meta-analysis to identify variants associated with serum pollutants levels and metabolic phenotypes. The secondary aims of these thesis were to annotate these variants with regulatory information from databases and to investigate the role of copy number variation in influencing gene expression and phenotype.

In the second and third chapters, we identified single nucleotide polymorphisms (SNPs) associated with pollutant and metabolic phenotypes. An example of a novel association was the relationship between SNPs in the *ABCG2* gene and octachlorodibenzo-p-dioxin (OCDD). In the third chapter, we identified associations between metabolic phenotypes and SNPs. An example of an identified association was that between serum apolipoprotein B levels and rs7412, which is consistent with other GWAS. To fine map these loci and assign functional annotation, I created Bayesian credible sets and checked for overlap between SNPs in these credible sets and regulatory marks in the Encyclopaedia of DNA Elements (ENCODE) database. Annotating these credible set SNPs with functional information revealed various histone modifications and transcription factors that overlapped.

This study was also successful in identifying copy number variants (CNVs) from the PIVUS cohort. There were moderately strong associations between CNVs and some of the phenotypes studied. These associations did not appear to be mediated by the SNP within the CNVs, as the latter had higher P values of associations. In addition, I identified several clinically relevant CNV-expression quantitative trait loci (CNV-eQTL) associations a separate cohort of healthy individuals. Some of these associations were cell specific and/or

context specific. Many of these CNVs contained SNPs which are lead SNPs in GWAS studies on a variety of different phenotypes.

In conclusion, this thesis was successful in identifying SNPs and CNVs associated with phenotypes, as well as annotating some of these variants with regulatory information. Further work is needed to clarify the mechanisms of regulation.

ACKNOWLEDGEMENTS

I would like to acknowledge the following:

- My supervisors Dr Andrew Morris and Dr Julian Knight for all their help, advice and guidance.
- My collaborator in Sweden – Dr Lars Lind and his team for his generosity in sharing data and allowing me to help with his project
- Dr Anubha Mahajan – for help in data quality control and imputation
- Dr Benjamin Fairfax – for his help in performing the gene expression and genotyping experiments
- My college for its support
- All participants in the study for donating their time and blood
- My friends and family for their encouragement
- God for his help

PUBLICATIONS FROM THIS THESIS

1. Ng ES, Lind M, Lindgren C, Ingelsson E, Mahajan A, Morris A, Lind L. Genome-wide association study of toxic metals and trace elements reveals novel associations. **Human Molecular Genetics**. 2015 Aug 15;24(16):4739-45

2. Naranbhai V, Fairfax BP, Makino S, Humburg P, Wong D, Ng ES, Hill AV, Knight JC. Genomic modulators of gene expression in human neutrophils. **Nature Communications.** 2015 Jul 7;6:7545

3. Ng ES, Salihovic S, Monica Lind P, Mahajan A, Syvänen AC, Axelsson T, Ingelsson E, Lindgren CM, van Bavel B, Morris AP, Lind L. Genome-wide association study of plasma levels of polychlorinated biphenyls disclose an association with the CYP2B6 gene in a population-based sample. **Environ Res**. 2015 Mar 31;140:95-101.

4. Ng ES, Salihovic S, Monica Lind P, Mahajan A, Syvänen AC, Axelsson T, Lindgren CM, van Bavel B, Morris AP, Lind L. Genome-wide association study of circulating levels of octachlorodibenzo-p-dioxin (OCDD) reveals a novel association with the ABCG2 gene in a population-based sample. **Human Molecular Genetics**. Submitted.

ABBREVIATIONS

2,2',3,3',4,4',5,5',6,6'-Decachlorobiphenyl (PCB-209)

2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl (PCB-206)

2,2',3,3',4,4',5,5'-Octachlorobiphenyl (PCB-194)

2,2',3,4,4',5,5'-Heptachlorobiphenyl (PCB-70)

2,2',3,4,4',5'-Hexachlorobiphenyl (PCB-138)

2,2',4,4',5,5'-Hexachlorobiphenyl (PCB-153)

2,2',4,4',5-Pentachlorobiphenyl (PCB-74)

2,2',4,4',5-Pentachlorobiphenyl (PCB-99)

2,3,3',4,4',5,5'-Heptachlorobiphenyl (PCB-189)

2,3,3',4,4',5-Hexachlorobiphenyl (PCB-156)

2,3,3',4,4',5'-Hexachlorobiphenyl (PCB-157)

2,3,3',4,4'-Pentachlorobiphenyl (PCB-105)

2,3',4,4',5-Pentachlorobiphenyl (PCB-118)

3,3',4,4',5,5'-Hexachlorobiphenyl (PCB-169)

3,3',4,4',5,5'-Hexachlorobiphenyl (PCB-180)

3,3',4,4',5-Pentachlorobiphenyl (PCB126)

adsorption, distribution, metabolism and elimination (ADME)

alpha-linolenic acid (ALA)

Alzheimer’s disease (AD)

analysis of variance (ANOVA)

apolipoproteins B (ApoB)

autism spectrum disorders (ASD)

B Allele Frequency (BAF)

basic linear algebra subroutine (BLAS)

basic local alignment search (BLAST)

body mass index (BMI)

CCCTC-binding factor (CTCF)

chromatin immunoprecipitation sequencing (ChIP-seq)

CNV-expression quantitative trait loci (CNV-eQTL)

comparative genomic hybridization (CGH)

comparative genomic hybridization (CGH)

copy number variant (CNV)

copy number variants (CNVs)

coronary artery disease (CAD)

D5 desaturase (D5D)

D6 desaturase (D6D)

DNA-methyltransferase 1 (DNMT1)

Encyclopaedia of DNA Elements (ENCODE)

expectation-maximization (EM)

expression quantitative trait loci (eQTLs)

false discovery rate (FDR)

fixed effects (FE)

fluorescence in situ hybridization (FISH)

fork stalling and template switching (FOSTES)

gas chromatograph (GC)

genome wide association studies (GWAS)

genomic control (GC)

Hardy Weinberg equilibrium (HWE)

hidden Markov model (HMM)

high density lipoprotein (HDL)

histone acetyltransferases (HATs)

histone deacetylases (HDACs)

homologous recombination (HR)

human leukocyte (HLA)

intercellular adhesion molecule-1 (ICAM-1)

ischemic heart disease (IHD)

linkage disequilibrium (LD)

lipopolysaccarides (LPS)

log R Ratio (LRR)

long chain polyunsaturated fatty acids (PUFAs)

long interspersed nuclear elements (LINE)

low density lipoprotein (LDL)

lymphoblastoid cell lines (LCLs)

lymphoblastoid cell lines (LCLs)

malignant brain tumour (MBT)

mass spectrometer (MS)

metallothioneins (MTs)

microhomology mediated break induced replication (MMBIR)

minor allele frequency (MAF)

National Human Genome Research Institute (NHGRI)

non-allelic homologous recombination (NAHR)

non-homologous end joining (NHEJ)

objective bayes (OB-HMM)

octachlorodibenzo-p-dioxin (OCDD)

octachlorodibenzo-p-dioxin (OCDD)

oligonucleotide microarray analysis (ROMA)

peripheral blood mononuclear cells (PBMCs)

peroxisome proliferator-activated receptor gamma (PPARG)

plant homeodomain (PHD)

poly(ADP-ribose) polymerase 1 (PARP1)

population frequency of B allele (pfb)

posterior probability (PP)

post-translational modifications (PTMs)

Pro-Trp-Trp-Pro (PWWP)

quality control (QC)

random effects (RE)

selective ion mode (SIM)

single nucleotide polymorphisms (SNPs)

solid-phase extraction (SPE)

solute carrier organic anion transporter family 1B1 (SLCO1B1)

sorting intolerant from tolerant (SIFT)

toll-like receptor 4 (TLR4)

transcription factor binding sites (TFBS)

transcription start site (TSS)

triglycerides (TG)

type 1 diabetes mellitus (T1D)

type 2 Diabetes (T2D)

Uppsala Longitudinal Study of Adult Men (ULSAM cohort)

vascular cell adhesion molecule-1 (VCAM-1)

very low density lipoprotein (VLDL)

CHAPTERS

1. Introduction – Genome wide association studies – methods, usefulness and limitations

1.1 Chapter Overview p16

1.2 Introduction to Genome Wide Association Studies (GWAS) p17

1.3 Genotyping technology p18

1.4 Quality Control p19

1.5 Imputation p21

1.6 Tests of Association p24

1.7 Population Stratification p25

1.8 Multiple Testing p26

1.9 Meta-analysis p27

1.10 Limitations of GWAS p32

1.11 Assigning functional significance to GWAS variants p35

1.12 Thesis Overview p41

2. Genome Wide Association Study (GWAS) of serum pollutant levels

2.1 Chapter Overview p43

2.2 Introduction p43

2.3 Materials and Methods p47

2.4 Results p51

2.5 Discussion p57

2.6 Chapter Summary p65

3. GWAS and meta-analysis of serum metabolic traits in 2 cohorts of Swedish individuals

3.1. Chapter Overview p66

3.2 Introduction p66

3.3 Methods and Materials p76

3.4 Results p80

3.5 Discussion p82

3.6 Chapter Summary p90

4. Fine mapping of significant loci and annotation with functional information

4.1 Chapter Overview p93

4.2 Introduction p93

4.3 Methods and Materials p107

4.4 Results p108

4.5 Discussion p112

4.6 Chapter Summary p119

5. Identification of copy number variation and association with pollutant and metabolic phenotypes

5.1 Chapter Overview p120

5.2 Introduction p120

5.3 Methods p132

5.4 Results p134

5.5 Discussion p136

5.6 Chapter Summary p143

6. Identification of CNVs and association with gene expression in a cohort of UK individuals

6.1 Chapter Overview p144

6.2 Introduction p144

6.3 Methods and Materials p154

6.4 Results p157

6.5 Discussion p161

6.6 Chapter Summary p163

7. Conclusion

7.1 Thesis Summary p165

7.2 Further Work p168

7.3 Conclusion p171

DECLARATION

Anubha Mahajan – Quality control and imputation of PIVUS and ULSAM data

Benjamin Fairfax – laboratory experiments, quality control and normalisation of gene expression data from monocytes and B cells

PIVUS and ULSAM teams – genotyping and phenotyping of cohorts

LIST OF FIGURES

Figure 2.1 Manhattan plot of PCB 99 218

Figure 2.2 Regional Association plot of PCB 99 219

Figure 2.3 Hierarchical Clustering of PCBs 219

Figure 2.4a. Regional Association plot of PCB99. 220

Figure 2.4b. Regional Association plot of PCB118 conditioned on PCB99. 221

Figure 2.4c. Regional Association plot of PCB 105 conditioned on PCB 99 222

Figure 2.4d. Regional Association plot of PCB 74 conditioned on PCB 99 223

Figure 2.5 Manhattan plot of Manganese 224

Figure 2.6 Locuszoom plots of manganese for both loci 225

Figure 2.7 Local association plot for cadmium 227

Figure 2.8 Local association plot for mercury 228

Figure 2.9 Local association plot for OCDD 229

Figure 3.1 Manhattan plot of D6 desaturase 230

Figure 3.2 Local association plot for D6 desaturase 231

Figure 3.3 Manhattan plot of D5 desaturase 232

Figure 3.5 Manhattan plot of ApoB 233

Figure 3.6 Local association plot for ApoB 234

Figure 3.7 Manhattan plot of TG 235

Figure 3.8 Local association plot for TG 236

Figure 3.9 Manhattan plot of HDL 237

Figure 3.10 Local association plot for HDL 238

Figure 4.1 Graph of cumulative posterior probability for variants in the credible set for D5 desaturase 239

Figure 4.2 Graph of cumulative posterior probability for variants in the credible set for D5

desaturase 239

Figure 4.3 Graph of cumulative posterior probability for variants in the credible set for D5 desaturase 240

Figure 4.4 Graph of cumulative posterior probability for variants in the credible set for D5 desaturase 240

Figure 5.1 Mechanisms of CNV formation (Bickhart & Liu, 2014) 241

LIST OF TABLES

Table 2.1 Median Values of pollutants and genomic control lambda 52

Table 2.2 Lead SNPs of genome wide significant loci for pollutants 54

Table 3.1 Lead SNPs significantly associated with phenotypes 80

Table 4.1 Credible Sets for each of the phenotypes 110

Table 6.1 Details about CNV-eQTLs identified 159

CHAPTER 1 GENOME WIDE ASSOCIATION STUDIES – METHODS, USEFULNESS AND LIMITATIONS

1.1 Chapter Overview

1.2 Introduction to Genome Wide Association Studies (GWAS)

1.3 Genotyping technology

1.4 Quality Control

1.5 Imputation

1.6 Tests of Association

1.7 Population Stratification

1.8 Multiple Testing

1.9 Meta-analysis

1.10 Limitations of GWAS

1.11 Assigning functional significance to GWAS variants

1.12 Thesis Overview

1.1 Overview

In the last decade, large numbers of GWAS have been conducted, with varying success. For example, from 2006 to 2013, more than 2000 associations have been identified in more than 1500 GWAS publications (Manolio, 2013). They have been successful in identifying loci for a variety of complex human traits, for example, coronary artery disease (Wellcome Trust Case Control Consortium, 2007) and Crohn's Disease (Yamazaki et al., 2005), helping to generate hypotheses about the pathophysiology of these diseases by identifying genes and pathways.

However, there are various limitations of GWAS. For many complex traits, the sum of genetic effects identified through GWAS comprises less than half of the estimated heritability (Stranger, Stahl, & Raj, 2011). A variety of hypotheses have been generated to account for this. The problem of 'missing heritability' for complex traits may be due to gene-environment interaction that is not accounted for in the analysis. These interactions may not be revealed in GWAS due to lack of statistical power if only a small fraction of individuals experience the adverse exposure. In addition to this, effect sizes of associated variants may be underestimated due to incomplete linkage disequilibrium (LD) (correlation between nearby SNPs) between causal variants and marker SNPs. These issues will be further discussed in Section 1.10.

Besides missing heritability, another important issue in the interpretation of GWAS is that a large proportion of the identified variants are in non-coding regions, for which it is difficult to assign biological function and the identity of the modulated gene. Hence, there has been limited progress in terms of translating GWAS results to clinical practice. One solution for this might be to fine map loci and integrate GWAS data with other genomic resources such as Encyclopaedia of DNA elements (ENCODE) database, which includes histone modifications, transcription factor binding sites and DNase hypersensitivity sites (ENCODE, 2011). The overall aim of this thesis is to test this hypothesis by bringing together some of these resources to see if this can inform GWAS of pollutant and metabolic traits.

1.2 Introduction to GWAS

With advancements in genotyping technology, it has now become possible to test variants on

a genome wide scale. GWAS interrogate SNPs across the genome in a large population of individuals, with the aim of identifying variants associated with the phenotype in question, which can either be a disease state or a quantitative trait. Unlike traditional candidate gene studies, GWAS are able to assay the whole genome in a comprehensive way without needing a prior hypothesis about the causal region.

A typical GWAS design is usually either case-control (for diseases or other binary phenotypes) or cohort (for continuous phenotypes). After an appropriate population is selected, individuals are approached for measurement of phenotypes and collection of samples, most commonly blood, from which DNA is extracted. However, saliva samples have also been shown to be a viable source of DNA for GWAS (Bahlo et al., 2010). After DNA extraction, genotyping is then performed, in which SNPs are genotyped on a chip, so as to identify the allele present in the DNA sample.

The success of GWAS relies on the "common variant common disease hypothesis" (Lander, 1996). This hypothesis states that common SNPs, which by convention occur at greater than 5% minor allele frequency (MAF) in the population, must confer some of the genetic risk for common diseases. In this assumption, different common SNPs have small effects on each disease, so some of these effects can be found by testing large number of SNPs in a sufficiently large population.

1.3 Genotyping Technology

Advancement in genotyping technology has allowed for the simultaneous interrogation of hundreds of thousands of SNPs on the same chip. Since the development of such chips, there

have been increases in SNP density, refinement of SNP selection, development of copy number variation (CNV) tagging approaches, increases in chip throughput, as well as reduction in SNP genotyping costs. At present, there are two main microarray platform manufacturers – Illumina and Affymetrix. One commonly used Affymetrix array is the affymetrix 6.0, which can genotype 1.8 million markers, including more than 900 000 SNPs and more than 940 000 probes for CNVs. It includes a high resolution reference map and a copy number polymorphism algorithm developed by the Broad Institute (http://www.affymetrix.com/). One common group of Illumina arrays is the Omni family, which includes a variety of different chips, ranging from the Human Omni5-Quad with 4.3 million markers, to the human Omni-express, with more than 715000 common markers (http://www.illumina.com).

The number of SNPs tested varies depending on the genotyping chip used, as well as the ethnic group being studied. Most studies have so far been performed in individuals of European descent, but there has been an increase in studies that include other ethnicities, such as Asian and African. Some of these GWAS have revealed genetic associations that are specific to certain ethnic groups (Cai et al., 2011; Carty et al., 2012). However, transethnic meta-analysis techniques have been developed, which leverage the difference in structure of LD between ethnic groups to amplify the signal of association at causal variants (Morris, 2011).

1.4 Quality Control

There are various quality control (QC) procedures which should be performed prior to GWAS analysis. These can largely be classified into sample QC and SNP QC. Sample QC issues

include sex inconsistencies, sample relatedness, genotyping efficiency and batch effects. One of the most straightforward ways of checking for sample handling and labelling errors is to check the reported sex of the individual against that predicted by the genetic data in analysis software such as Plink (Purcell et al., 2007). Individuals that have been incorrectly included can then be removed from the dataset. Another QC step involves the removal of related samples in the study dataset. This can be performed by computing pairwise kinship estimates between each individual in the study. This procedure also calculates the proportion of loci in which two individuals share zero, one or two alleles that are identical by descent. Individuals which share two alleles at every locus are either monozygotic twins or are the same sample. Under the null hypothesis, siblings share zero, one and two alleles with frequencies of 25%, 50% and 25% respectively (Turner et al., 2011). Since association analyses typically assume samples to be "independent", having related samples may result in increased type 1 and type 2 errors. In such circumstances, mixed model regression may be more appropriate, or else one individual from the duplicate or related pairs should be removed from the study sample.

Genotyping call rate can be assessed by calculating the proportion of SNPs which fail QC in each individual. Samples with low call rate should be removed from further analysis. While 98-99% is generally the recommended threshold, the exact threshold can vary depending on the quality of DNA samples, variability in human error and genotyping platform used (Akin, Adal, Akin, & Kurtoglu, 2012). Batch effects arise when quality varies between different genotyping batches. Since thousands of DNA samples are typically genotyped, they are partitioned into batches, typically 96 samples on a plate. If there are systematic differences in the case to control ratio or ethnicity of individuals on each plate, batch effects can result, which confound downstream analysis.

There are several ways of examining the dataset for potential batch effects. One method is to calculate the average MAF and genotyping call rate across SNPs in each plate. Another method is to code case/control status by plate and then running GWAS analysis, testing each plate against other plates. If any plate is shown to have an average p value that deviates significantly, it should be further investigated for genotyping problems. Additionally, ethnic outliers can be identified by comparing the sample with reference individuals from genome databases such as the International Hapmap Project, a large international consortium established with the goal of determining common patterns of sequence variation in the human genome. These outliers could potentially be excluded if the population is assumed to be homogeneous.

Quality control on SNPs generally consists of assessing genotyping call rate, MAF and Hardy Weinberg equilibrium (HWE). A recommended threshold for eliminating SNPs with low genotyping call rate is 98-99%, but this threshold may vary from study to study depending on similar factors as for sample genotyping call rate (Akin et al., 2012). In general, poor quality markers should be removed before assessing sample genotyping call rate. Checking for HWE is also necessary in marker QC. Under HWE assumption, genotype and allele frequencies can be estimated from one generation to the next (Wittke-Thompson, Pluzhnikov, & Cox, 2005). Significant departure from HWE can indicate genotyping errors, population stratification or association with the phenotype. In general, HWE departures towards an excess of heterozygosity may reflect a technical problem with the assay rather than true association with the phenotype. Such SNPs should then be eliminated from analysis.

1.5 Imputation

Imputation is a procedure to infer the genotypes of SNPs that are not directly typed by the chip. It allows for increased coverage, power and fine-mapping resolution. Imputation is based on the presence of LD between typed and untyped markers in the reference haplotypes. In most cases, a reference of haplotypes is used to impute SNPs into a study sample of individuals that have been genotyped at a subset of these markers (B. Howie, Fuchsberger, Stephens, Marchini, & Abecasis, 2012). The goal is to predict untyped markers and hence increase the power of the study, the ability to fine map variants and also conduct meta-analysis with studies that use other chips. Other forms of imputation allow for the prediction of HLA alleles, copy number variants and indels (Franke et al., 2008; X. Zheng et al., 2014).

Imputation requires the use of reference panels of densely genotyped haplotypes in order to indirectly assess the missing SNPs on a less densely genotyped sample. In general, imputation is more accurate when the reference panel and study subjects are from ethnically similar populations (L. Huang et al., 2009). On the other hand, some studies have also shown that incorporating a mixture of different diverse populations can improve accuracy in comparison to using a single ethnic panel (L. Huang et al., 2009; Jostins, Morley, & Barrett, 2011).

Most GWAS studies use reference panels from the International Hapmap Consortium (Hapmap) and the 1000 Genomes project (Abecasis et al., 2012; I. H. Consortium, 2005; International HapMap Consortium, 2005). The Hapmap Project was established with the goal of determining common patterns of sequence variation in the human genome. It includes DNA samples from populations in Africa, Asia and Europe. The data was published publicly in 3 phases, 2005, 2007 and 2009. HapMap haplotypes have been commonly used to carry out imputation in populations that have ancestry that is genetically close to those in the panels. These include CEU (individuals of European ancestry), YRI (Yoruba, African

ancestry) and JPT+CHB (Japanese and Chinese, East Asian ancestry) (Thorisson, Smith, Krishnan, & Stein, 2005). The 1000 Genomes Project is an international research effort launched in 2008, with the goal to sequence the genomes of more than a 1000 individuals from a wide range of countries. In 2010, the pilot phase was finished and in 2012, the complete genomes of 1092 individuals were published for public access. The third phase of this project will include 2500 individuals from more diverse populations.

Various methods can be used for imputation. One commonly used software for this purpose is Impute v1 (Marchini, Howie, Myers, McVean, & Donnelly, 2007), which is based on hidden Markov models (HMMs) for estimating recombination rates and modelling LD. Impute v2 uses a similar method to Impute v1, but there are some differences (B. N. Howie, Donnelly, & Marchini, 2009), for example Impute v2 uses haploid imputation, which is faster than the diploid imputation used by Impute v1. Furthermore, Impute v1 integrates over the unknown phase of the genotypes in the study sample, while Impute v2 integrates over the space of phase reconstructions via a Markov chain monte carlo approach. Mach is another imputation methods that uses a hidden Markov model (HMM) similar to Impute v1 and v2 (N. Li & Stephens, 2003). It successively updates the phase of each sample's genotype conditional on the current haplotype estimates of other samples. FastPHASE is an imputation approach that has been incorporated into the association testing programme – BIMBAM (Scheet & Stephens, 2006). It uses the observation that haplotypes tend to cluster into related groups. It specifies a set of clusters meant to represent haplotypes and then models each individual's genotype as a HMM and uses an expectation-maximization algorithm to fit the model. BEAGLE is an imputation method that uses a different approach; it is based on a graphical model of haplotypes, and works by fitting the model iteratively to the set of estimates then resampling new estimates for each individual depending on the model fit (Browning &

Browning, 2008).

There are various factors that influence the accuracy of imputation. Properties of the reference panel, study population, genotyping platform and allele frequency all affect imputation accuracy. For example, imputation error rate is inversely proportional to MAF, as it is more difficult to establish strong LD between rare SNPs in the study sample and reference population. Populations with complex LD structure are more difficult to impute, for example, imputation accuracy in the YRI population is lower than for the CEU population (L. Huang et al., 2009). Reference panel size is inversely proportional to error rate, and using a combination of CEU, YRI as well as JPT+CHB panels can improve accuracy, especially for SNPs with low minor allele frequency (L. Huang et al., 2009). Furthermore, using a reference panel that has been phased using trio information boosts imputation performance (Marchini & Howie, 2010).

1.6 Statistical tests of association

The most straightforward test of association involves single SNP analysis. There are two different categories – case control and quantitative outcomes. When testing case-control phenotypes, a 2 x 3 matrix can be created to demonstrate the relationship between genotype and case-control status, then a chi square or fisher's exact test can be performed. In the case of complex traits, it is thought that risk from individual SNPs are likely to be additive. One way to improve power to detect additive risks is to count alleles rather than genotypes so that each individual contributes twice to a 2 x 2 table and a Pearson 1 degree of freedom test can be applied (Sasieni, 1997). However, this is generally not performed because it makes the assumption of HWE. A more elegant way would be to do the Cochran-Armitage test, which

tests the hypothesis of zero slope for a line that fits the three genotypic risk estimates best (Armitage, 1955). This test sacrifices power if the genotypic risks are not additive. On the other hand, the Fisher's test spreads the research investment over the full range of risk models, with less power to detect additive risks. One approach may be to use the former in the case of rare variants and use the fisher test when the count of genotypes are sufficiently high to have reasonable power for non-additive models. Logistic regression can also be used for case-control GWAS, in which the logit transformation is applied to the disease risk of each individual, and likelihood ratio test of the general model against the null hypothesis is performed. This allows for the incorporation of covariates in the model, as well as multiple SNPs and interaction terms.

In the case of continuous outcomes, linear regression or analysis of variance (ANOVA) can be used. ANOVA compares the null hypothesis of no association with a general alternative, whilst linear regression assumes a linear relationship between the mean value of the trait and genotype. In both cases, the phenotype should be approximately normally distributed, hence transformation of the trait might be necessary prior to running the association analysis.

1.7 Population Stratification

Population structure can generate spurious associations. This occurs when cases disproportionately represent a particular genetic group, so that SNP with allele proportions that differ with ethnic subgroups in the population also differ in case/control status or are associated with the continuous phenotype of interest. This can result in false positive association signals. There are a few ways which population structure can be dealt with and adjusted for in GWAS analysis. Firstly, genomic control (GC) correction can be used (Devlin

& Roeder, 1999a). In this method, GC λ is calculated as the median of the test statistic divided by its expectation under the chi-square distribution with 1 degree of freedom. If GC λ is considerably larger than 1, this indicates the presence of population stratification and test statistics can be divided by GC λ. Another method for dealing with population stratification is through structured association methods. These attribute the genomes of individuals to hypothetical subpopulations and test for association conditionally on allocation of these subpopulations. A third method is to perform multidimensional scaling (MDS) on a matrix of SNPs and to use the MDS components as covariates in the logistic regression model as they represent the population structure (Q. Li & Yu, 2008). Before this is performed, the matrix of SNPs are usually pruned to remove SNPs that are strongly in LD with other SNPs.

1.8 Multiple Testing

One issue to consider is the appropriate genome-wide significance threshold to use when interpreting p values from association studies. Most recent GWASs use a threshold of 5×10^{-8}, in accordance with a 5% type 1 error rate in the context of 1 million association tests performed (Dudbridge & Gusnanto, 2008). However, it may be suitable to use different thresholds, depending on number of tests being performed, LD patterns, genotyping platform and population being studied. One study demonstrated that associations with a p value between 5×10^{-8} and 1×10^{-7} replicated more than 70% when additional data was collected (Panagiotou, Ioannidis, & Project, 2012). Another possible measure is the false discovery rate (FDR) or Q value. The former is an estimate of the false positives divided by the associations that are discovered. The Q value is the threshold FDR used for calling an association. FDR and Q values are currently used for gene microarray studies, to account for multiple testing. One study suggests that FDR can further be augmented by screening with methods such as

least absolute shrinkage and selection operator (Shi et al., 2011). In the contest of a Bayesian framework, the Bayes factor estimates the ratio of the probability of observed data under the null hypothesis to the corresponding probability under the alternative hypotheses. The Bayes factor and p values often provide approximately similar rankings for common SNPs, with minor allele frequencies in the range of 0.1 to 0.5, but may differ for rarer SNPs (Wakefield, 2009).

1.9 Meta-analysis

Due to the need for larger sample sizes and increased power, meta-analysis is becoming an increasingly important tool in genome wide association studies of complex phenotypes (de Bakker et al., 2008). It provides a feasible solution for combining study level summary statistics rather than individual-level genotype and phenotype data, which may not be possible due to ethical constraints. Meta-analysis has reliably identified loci associated with a broad range of complex traits, including height (M. He et al., 2015), body mass index (BMI) (Wen et al., 2014) and intelligence (Davies et al., 2015).

1.9.1 Usefulness

Evidence demonstrates that meta-analysis has comparable statistical power to pooling individual level data across studies (Lin & Zeng, 2010). Several hundred GWAS meta-analyses exist in published literature; in the period between 2009 and 2012, 139 of these studies had sample sizes in the discovery that exceeded 10 000 participants. As of February 2015, more than 15000 SNPs from 2111 publications have been documented in the GWAS catalog (Hindorff, J, HA, PN, & AK, 2015). Many of these SNPs were discovered in meta-

analysis. Many of the associated loci discovered from GWAS meta-analysis can serve as genetic risk predictors, which have the potential to improve risk prediction accuracy and benefit clinical treatment (Jostins et al., 2012). Dense genome wide markers enable a good approximation of narrow sense heritability and valuable insight has been gained on the genetic etiology of diseases by analysing GWAS meta-analysis lead SNPs and associated pathways.

1.9.2 Heterogeneity

One challenge for meta-analysis is heterogeneity, which can arise from different sources. Firstly, phenotype characterisation can introduce significant heterogeneity between studies. This is especially true for complex traits, when the phenotype is hard to define, for example, cognitive or behavioural traits. These variable definitions may lead to heterogeneity of effect size as well as spurious associations. It would be ideal to restrict studies to those in which the phenotype has been well defined in similar ways, but to do so would trade off power due to smaller sample size. Other causes of heterogeneity include population stratification even within populations that are relatively homogeneous, for example, within the UK population (Babron, de Tayrac, Rutledge, Zeggini, & Génin, 2012). Principal components analysis and multi-dimensional scaling can account for population structure in testing for association with common variants, but further steps may be required for rare variants (Babron et al., 2012). Heterogeneity can also be introduced if different genotyping platforms or quality control measures are used. Imputation software also add an additional layer of heterogeneity since they differ in algorithms implemented.

Various statistical metrics of heterogeneity have been developed to assess intra and inter study heterogeneity. These include Cochran's Q statistic, which follows a chi square distribution. However, this test can be underpowered when there are fewer than 20 studies (Jackson, 2006). This statistic can also be used to estimate the posterior odds of heterogeneity in a Bayesian model. An alternative to the Q statistic is I^2, which describes the percentage of total variation across studies due to heterogeneity rather than chance alone (Higgins & Thompson, 2002). It can be calculated from results obtained from a meta-analysis as 100% x (Q – df)/Q, with df being the degrees of freedom. I^2 lies between 0% and 100%, with larger values showing increasing heterogeneity.

1.9.3 Fixed vs Random Effects Models

There are various models which can be used to conduct GWAS meta-analysis. In fixed effects (FE) meta-analysis, the true effect of each allele being studied is assumed to be the same in each study (de Bakker et al., 2008) and all differences are due to chance. There are two main types of FE meta-analysis – sample size weighted Z scores and inverse variance weighted effect sizes. The Z score model calculates a weighted mean of the Z score, where the weighting is most commonly a function of the sample size, hence it takes into account direction of effects (Begum, Ghosh, Tseng, & Feingold, 2012).

One way to limit heterogeneity is to only include studies that are homogeneous in terms of phenotype definition, measurement and population background. Fixed effects models are useful in this context. Alternatively, if one wishes to maximise power by including more studies, even if they have different effect sizes, random effects (RE) models can be used

instead. In this family of methods, heterogeneity is explicitly modelled, since one does not make the assumption that effect sizes are the same (Begum et al., 2012).

Han and Eskin report that when one applies both FE and RE models to a dataset, RE models produce fewer significant p values that FE at variants that show varying effect sizes between studies (B. Han & Eskin, 2011). Because RE models and FE models give the same p values at markers with no heterogeneity, RE rarely ever gives more significant p values than FE at any marker. Consequently, RE models are sometimes perceived as being less powerful than FE models since all associations identified by RE are usually already identified by FE. The underlying reason for this phenomenon is that RE assumes a conservative null-hypothesis model. In RE, a two-step procedure is undertaken – firstly, one estimates effect size and confidence interval while taking into account study heterogeneity. After that, the effect size is normalised into a z score and translated into a p value, a step which is equivalent to assuming heterogeneity under the null hypothesis (DerSimonian & Laird, 1986). However, this may be overly conservative because there should not be heterogeneity under the null hypothesis of no associations when effect sizes are zero. Han and Eskin developed a RE model that relaxed this assumption (B. Han & Eskin, 2011). It estimates effect size and confidence interval in the same way as the conventional RE method. However, instead of subsequently calculating a Z score, it applies a likelihood-ratio test and assumes no heterogeneity under the null hypothesis, hence separating the effect size estimation from hypothesis testing. (B. Han & Eskin, 2011).

1.9.4 METAL

Various software packages have been developed for fixed effects meta-analysis. METAL is a computationally efficient tool which is open source and free downloadable (Willer, Li, & Abecasis, 2010). It allows analysis of large data sets and supports various different input file formats. METAL converts P value and direction of effect into a signed Z score, with large positive Z scores indicating an allele strongly associated with disease and very negative Z scores indicating an allele strongly associated with low disease risk. Z scores are then combined across samples in a weighted sum, with weights being proportional to the square root of the sample size.

On the other hand, the inverse variance weighting method assigns the inverse of the square of the standard error as a weight for each study (Kavvoura & Ioannidis, 2008; Pereira, Patsopoulos, Salanti, & Ioannidis, 2009). Many studies use this method to combine estimates from genotyped and imputed variants (Soranzo 2009, Willer 2008). However, it is not clear whether this is the optimal weighting scheme. Jiao et al performed simulation studies and found that the optimal weight is proportional to both the inverse variance of estimates as well as the expected value (Jiao, Hsu, Hutter, & Peters, 2011). Based on their results, they postulate that the inverse-variance weighting method is a good approximation for the optimal weights.

1.9.5 GWAMA

GWAMA is a memory efficient software tool that performs meta-analysis of summary statistics generated from separate GWAS of either discrete or quantitative traits (Mägi & Morris, 2010). This tool aligns studies to the same reference strand, and hence solves computational problems related to combining results from GWAS chips that may be aligned

to different strands. Furthermore, it accounts for over-dispersion of test statistics due to population structure and between-study variation by including an option for genomic control.

GWAMA implements the inverse variance weighting method as follows (Mägi & Morris, 2010). Let βij denote the effect of the reference allele at the jth SNP in the ith study. The combined allelic effect across all studies at the jth SNP is then calculated as

$$B_j = \frac{\Sigma_{i=1}^{N} \beta_{ij} w_{ij}}{\Sigma_{i=1}^{N} w_{ij}},$$

where $wij = [\mathrm{Var}(\beta ij)]^{-1}$ is the inverse of the variance of the estimated allelic effect in the *i*th study, obtained from the standard error

1.10 Limitations of GWAS

GWAS have been useful in the research setting for identifying variants which may be associated with disease for further investigation through functional studies. In the clinical setting, GWAS have been useful in identifying variants which may have predictive or prognostic implications (Manolio, 2013). Although GWAS have made these contributions, they have several limitations, some of which have been imposed by nature. While GWAS is able to identify common variants of large effect, many complex traits are associated with lower frequency variants with small effects sizes.

For most complex traits, the sum of genetic effects identified through studies comprises less than half of the estimated heritability (Stranger et al., 2011). For example, in Type 2 Diabetes (T2D), the 66 established susceptibility loci account for only 10-20% of the familial aggregation of the disease. (Morris et al., 2012) The problem of 'missing heritability' for

complex traits may be due to variety of different factors.

Firstly, most markers identified from GWAS confer relatively small risks, even though they fulfilled a very strict p value threshold, which corrected for multiple testing. Furthermore, replication studies show that only a small fraction of significant loci in the discovery GWAS can be replicated (Kraft, Zeggini, & Ioannidis, 2009). There are a few possible reasons for this. The most common reasons would be a false positive error due to sampling error. Alternatively, it is possible that the follow-up study had insufficient power to replicate small effects. Thirdly, the variant identified in the original study could be a poor marker for the phenotype due to differences in LD structure. Fourthly, there could be differences in study design between the cohorts, which might lead to inconsistent effects that fail to replicate.

Another limitation of GWAS is that effect sizes of associated variants may be underestimated due to incomplete LD between causal variants and marker SNPs (J. Yang et al., 2011). The variance accounted for by validated SNPs is usually less than heritability, as defined by the proportion of variance that the phenotype exhibits due to genetic variance. Non-additive genetic effects do not contribute to heritability as defined in this way. There are various possible explanations for the failure of validated SNPs to account for heritability. Firstly, it is possible that the causal variant explains such a low proportion of the variation that the effects do not reach the significance threshold. Secondly, the causal variant may not be in complete LD with the SNPs that are genotyped. This might be particularly relevant for rare variants, and can be expressed by difference between the genomic relationship between each pair of subjects at the causal variant and the relationship between subjects calculated from the SNPs. As the causal variant is not known, it is not possible to estimate the LD directly, but it can be imitated by consideration the LD of the genotyped SNPs. It is possible that the causal variants

and SNPs have different properties, so LD among SNPs is only a rough estimate of LD between causal variants and SNPs, and even then, this may not be a good estimate in the case of rare causal variants.

Low-frequency variants (MAF < 0.05) not captured by current genotyping platforms may contribute to the missing heritability. A substantial proportion of these causal variants could be very rare, requiring characterization through sequencing of clinical samples (B. Li & Leal, 2009). As sequencing technologies improve, whole genome and exome sequencing may become feasible for larger samples, with consequent detection of larger number of rare variants. Furthermore, effect sizes for common variants affecting human complex traits are skewed towards small effect sizes. This problem is further compounded by the fact that GWAS are underpowered to detect small effects.

Yet another contributor to missing heritability might be gene-environment interaction. In humans, several interactions have been documented, for example, the effect of *FTO* alleles on body mass is modified according to exercise (Rampersaud et al., 2008). These interactions may not be revealed in GWAS due to lack of statistical power if only a small fraction of individuals experience the adverse exposure.

Another possibility that may account for missing heritability could be the presence of multiple association signals at a single locus – some of which might not reach genome-wide significance. Morris and colleagues provide evidence that type 2 diabetes susceptibility loci harbour multiple independent association signals, for example, *KCNQ1* and *CDKN2A/B*, in addition to putative independent associations at *DGKB* and *MC4R*, suggesting that multiple independent association signals may be present at T2D susceptibility loci (Morris et al.,

2012). Furthermore, the loci identified by these GWAS of complex traits are typically characterised by common variant association signals that map to large genomic regions (many kb) which often include multiple transcripts and/or non-coding regions. An example would be the *CILP2* locus (Figure 1a) identified by Morris et al (2012) (Morris et al., 2012).

Besides missing heritability, GWASs have other problems. Multiple sources of Type 1 and Type 2 errors exist in GWAS due to its inherent complexity. Type 1 errors occur when the null hypothesis is incorrectly rejected (false positive) while type 2 errors are the failure to reject a false null hypothesis (false negative). Quick advances in genotyping technology have resulted in large number of SNPs being interrogated leading to possible false positive errors due to multiple comparisons. However, GWAS are based on the assumption that common traits are determined by common variants, each conferring a small increase in risk. Since the probability of detecting an individual SNP association is relatively small, a large sample size is usually required for adequate statistical power, resulting in false negatives. Hence, avoiding type 1 errors may increase the rate of type 2 errors. This is especially true for low frequency alleles, small sample size or small effect sizes, for example, with an odds ratio of less than 1.1. However, with a large enough sample, GWAS could detect these small effects. Other potential sources of errors include population stratification, incorrect phenotype characterisation, non-genetic covariates and inaccurate genotyping.

1.11 Assigning Functional Significance to GWAS hits

An important issue in GWAS studies is to assign functional significance to GWAS hits so that they can be interpreted in a biological context and used to inform clinical studies.

While LD is crucial in helping GWAS investigators to identify causal variants that are not on the genotyping chip, it also poses a challenge, since it is not clear among SNPs in LD which

SNP brings about the biological phenotype.

Various approaches have been developed to help solve this problem. Although GWAS achieves greatest success when variants in coding regions are identified, which have a plausible biological mechanism, the majority of SNPs identified in GWAS are in un-transcribed regions, some of which may be in regulatory regions. One study demonstrated that 88% of SNPs in the National Human Genome Research Institute (NHGRI) GWAS catalogue were intronic or intergenic (Schaub, Boyle, Kundaje, Batzoglou, & Snyder, 2012). These SNPs may act to bring about the phenotype via regulatory mechanisms. Methods have been developed for identifying SNPs that overlaps regulatory elements such as transcription binding sites (Boyle et al., 2012), and these have been used with good effect for the refinement of GWAS results. However, the presence of a motif, sometimes predicted computationally, does not guarantee that transcription factor binding takes place in vivo. Since the development of techniques such as a chromatin immunoprecipitation sequencing (ChIP-seq) and DNase-hypersensitivity assays, investigators have been able to experimentally characterise functional regions. The presence of SNPs in these regulatory regions has been shown to produce differences in transcription factor binding between different individuals. A recent study of chromatin marks in various cell lines showed a twofold enrichment for enhancer elements among SNPs found to be GWAS hits in other studies (J. Ernst et al., 2011).

One way to assign functional significance to these SNPs would be to annotate these SNPs with information from the ENCODE database. The ENCODE database is an international collaboration of research groups, with an aim to build a comprehensive list of functional elements in the human gene, including elements that act at the mRNA level, as well as

regulators which act at the protein and cell level. These elements are detected though DNA hypersensitivity assays, DNA methylation assays as well as chromatin immunoprecipitation of proteins that interact with DNA and RNA.

Several integrated tools have been developed to interrogate the ENCODE database, as well as other sources of functional annotation. One such tool is Annotate Variation (ANNOVAR), a pipeline which aids functional annotation of GWAS SNPs. (K. Wang, Li, & Hakonarson, 2010) Besides identifying genes in which SNPs lie, it can also perform region based annotation, looking for overlap with histone modifications, predicted microRNA target sites and transcription factor binding sites. Another such tool is RegulomeDB,(Boyle et al., 2012), which includes datasets from ENCODE, and does computational predictions as well as manual annotations to score SNPs with an aim to identify potential functional variants for further testing. The authors demonstrate this by doing a search for the causative variant for systemic lupus erythematosus around the region of the TNFAIP3 gene and narrowed down the search space from more than 2000 variants to three potential candidates.

One study that examined SNPs in the NHGRI GWAS catalogue found that 34% SNPs overlapped regions of DNAse hypersensitivity, while 12% of SNPs overlapped transcription factor occupied regions, reflecting a significant enrichment compared to expected percentages of 23% and 6% respectively (ENCODE, 2011). This enrichment persisted even after biases were corrected for, such as selection of SNPs for GWAS arrays. Furthermore, after partitioning the genome by density of different classes of elements, GWAS SNPs were shown to be enriched in segmentation regions associated with transcription factors and enhancers with this trend being displayed across different cell types.

Another way to assign functional annotation to GWAS results may be to identify expression quantitative trait loci (eQTLs), which are SNPs associated with gene expression. Nicolae et al demonstrated that trait associated SNPs are more likely to be eQTLs and application of this information could enhance the discovery of SNPs associated with complex phenotypes (Nicolae et al., 2010). They developed a website (www.scandb.org) to demonstrate that these findings are consistent across a broad range of human traits. They also analysed data from the Wellcome Trust study and showed that creating a score which reflected the probability of a SNP being an eQTL was useful to inform the nature of mechanisms that could potentially drive the association (Wellcome Trust Case Control Consortium, 2007).

Yet another way to assign functional significance to SNPs is to identify other types of variants which may be in LD with these SNPs. One such type of variation is copy number variation. A copy number variant (CNV) is defined as a genomic segment 1 kb or larger in size that occurs at variable numbers in the genome. A total of 12-15% of the human genome contains CNVs (Redon et al., 2006) and their effect on the human genome in terms of affected base pairs is far greater than SNPs (Conrad et al., 2010). Furthermore, 56% of the CNVs that have been identified so far occur in known genes (Iafrate et al., 2004). CNVs have been associated with a few disease phenotypes, including autoimmune diseases such as HIV susceptibility(Gonzalez et al., 2005) and psoriasis (Hollox et al., 2008), neurological diseases such as autism (Marshall & Scherer, 2012) and schizophrenia (Kirov et al., 2014), as well as congenital diseases such as heart malformations (Tomita-Mitchell et al., 2012).

CNVs can act to bring about a phenotype by different mechanisms. The most direct would be by altering the copy number of the gene that is susceptible to dosage effect. This has been shown to occur in individuals with neurodevelopmental disorders. Duplications and deletions

of the same gene can result in different outcome. For example, duplication of the *MECP2* gene results in Lubs X-linked mental retardation while deletion of the same gene results in Rett syndrome (Friez et al., 2006; Qiu et al., 2012).

CNVs in noncoding sequences can also result in long-range regulation of target genes by altering the copy number of regulators or by disturbing regulatory pathways. So far, two studies have attempted to map copy number variation to gene expression data on a genome wide scale. Stranger et al (2007) interrogated gene expression in lymphoblastoid cell lines of 210 unrelated HapMap individuals from 4 populations (Stranger et al., 2007). SNP genotypes from phase 1 HapMap, together with CNV data from comparative genomic hybridization (CGH) of each HapMap individual, were analysed in parallel with gene expression, as measured by microarrays. Measuring gene expression is a useful step in interpreting the results of GWAS, especially where the variant correlates with gene expression (expression quantitative trait locus). SNPs and CNVs were found to capture 83.6% and 17.7% of the total genetic variation, with 1.3% being captured by both.

A more recent study by Schlattl et al (2011) studied gene expression and CNVs in 120 healthy individuals, identifying CNVs associated with the expression of 110 genes, 13% of the associations involving complex multialellic CNVs (Schlattl, Anders, Waszak, Huber, & Korbel, 2011). CNVs were categorised according to variant type, size and gene overlap. While many small CNVs were associated with variation in gene expression, there was an enrichment of large duplications and deletions, including large intergenic CNVs. The copy number of genes that intersected with CNVs mostly correlated positively with gene expression, and was correlated more significantly with expression than nearby SNPs, suggesting that some of the eQTL effects in other studies may be related to CNV effects.

Finally, adding epigenomic information, such as overlapping hypomethylation or hypermethylation, may be useful in informing functional annotation of SNPs. For example, some individuals with chronic lymphocytic leukemia have a point mutation upstream of the *DAPK1* promoter, which is also associated with methylation at the gene (Raval et al., 2007). In hereditary non-polyposis colorectal cancer, *MLH1* hypermethylation and hence silencing is associated with a SNP within the untranslated region (Hitchins et al., 2011). The relationship between DNA methylation and SNP variation can now be investigated with high throughput methylation arrays, which conduct a comprehensive interrogation of CpG sites across the genome. CpG sites are regions of DNA in which cytosine occurs next to guanine and can be variably methylated in different individuals. Methylation quantitative trait loci can be used to test association between SNP and CpG methylation in different tissue types. Several studies have demonstrated a *cis*-relationship between methylation and genotype, in which the genetic variant and methylated region are close together, usually within 1 Mb of each other. However, there have been examples of trans-regulation across a range of tissue types, where the variant is far from the methylated region, sometimes on a different chromosome (Gibbs et al., 2010). Although epigenome wide studies show promise, there are various challenges in executing and interpreting them. For example, there is considerable tissue heterogeneity in terms of methylation, hence adequate sample size is required, with replication in independent populations. Furthermore environmental exposures such as smoking can affect epigenetic changes. Despite these limitations, there have been successful examples of integrating of methylation and GWAS data, for example, Bell successfully combined data to identify variant-CpG restricted haplotype-specific methylation within *FTO* obesity susceptibility locus (Bell et al., 2010).

1.12 Thesis overview

This thesis aims to explore some of the uses and limitations of GWAS by performing a GWAS and meta-analysis, and then functionally annotating these variants so as to elucidate possible mechanisms through which they act to influence the phenotype.

In Chapter 2, my research question was to identify genetic variants associated with blood pollutant levels as these pollutants are known to have harmful effects on human health and the identification of these variants may shed light on the metabolism of these compounds and may also help to identify individuals more susceptible to toxicity This chapter allowed me to explore association techniques on an imputed dataset, and to generate results that can be interpreted in the biological context.

In Chapter 3, my research question was to identify genetic variants associated with metabolic traits, including D6 and D5 desaturase, triglycerides, HDL and apoB as these traits have been associated with various markers of cardiovascular disease, which is a major cause of death in developed nations. I performed a meta-analysis on two Swedish cohorts so as to identify variants associated with various metabolic phenotypes. This allowed me to explore the technique of meta-analysis to boost power of GWAS. Although most of these traits have been explored in other larger meta-analyses, D6 and D5 desaturase have not been explored in the GWAS context. These enzymes play a crucial role in the elongation and desaturation of fatty acid precursors and have been associated with carcinogenesis and neurological disorders. Details are provided later in the chapter.

In Chapter 4, I addressed the research question of whether the creation of credible sets is informative in fine mapping variants identified in the previous 2 chapters. In addition, I also annotated variants identified in chapters 2 and 3 with functional elements from the ENCODE and eQTL databases, so as to assign biological functionality, especially for variants located in introns or untranscribed regions.

In Chapter 5, I identified CNVs from the PIVUS cohort and perform an association study between copy number and pollutants as well as metabolic phenotypes. This was performed in order to answer the research question of whether CNVs would add an additional layer of information to the study, by possibly identifying genetic associations which were not identified when only SNP associations were tested.

In Chapter 6, I identified CNVs in a separate cohort of UK individuals and test association between copy number and gene expression in these same individuals, identifying several CNV-eQTLs in clinically relevant genes, some of which were cell specific and/or context specific. The research questions in this chapter were firstly, whether the above mentioned eQTLs exist, and secondly, whether additional information is gleaned by performed CNV analysis as opposed to just SNP-eQTL analysis.

In Chapter 7, I summarised the thesis, discuss limitations and suggest further directions for research.

CHAPTER 2 Genome Wide Association Study (GWAS) of serum pollutant levels

2.1 Chapter Overview

2.2 Introduction

2.3 Materials and Methods

2.4 Results

2.5 Discussion

2.6 Chapter Summary

2.1 Chapter Overview

This chapter describes a GWAS on a cohort of healthy elderly individuals - Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) - investigating genetic determinants of blood levels of several environmental pollutants that have been shown to affect physiological functions in humans and other mammals. In this chapter, I explore GWAS association techniques, including additive models and conditional analysis. I also explore how phenotypes which are partially correlated with each other can have complex relationships in terms of how they are influenced by SNPs within the same region. The GWAS hits uncovered in this chapter are explored in greater detail in the later chapters, with functional annotation from the ENCODE database.

2.2 Introduction

This study focuses on three groups of pollutants –polychlorinated biphenyls (PCBs), dioxins

and trace metals. Preliminary studies have demonstrated that the metabolism of these compounds may be influenced by various genetic polymorphisms (L. Lind et al., 2014).

PCBs are a group non-flammable chlorinated hydrocarbons with variable toxicity and high chemical stability. They have largely been banned due to harmful effects on the environment as well as on the human body, including the nervous, reproductive and immunologic systems (*State of the Science of Endocrine Disrupting Chemicals*, 2012). In the past, they were widely utilised in a broad range of industrial products, including electrical equipment, plasticizers, dyes and insulation.

PCBs are highly bioaccumulative because of their high lipophilicity and long half-life, ranging from 5 to 15 years (Ritter et al., 2011). As a result, these compounds contribute to long term toxic exposure, since exposure and accumulation increases with age. PCBs have been detected in the populations of a large number of countries. (Govarts et al., 2012; Patterson et al., 2009; Salihovic, Lampa, et al., 2012).

The metabolism of PCBs varies depending on structure, including factors such as degrees and positions of chlorination and unsubstituted carbon atoms, as well as non-structural factors such as species being studied and age of organism. (Crinnion, 2011; Megson et al., 2013; Ritter et al., 2011). Absorption occurs through the skin, by inhalation, as well as via the gastrointestinal system (J. F. Brown, 1994). They are primarily accumulated in fat tissue due to high lipophilicity. (Crinnion, 2011). Biotransformation mainly takes place in the liver, through conjugation with glucuronic acid and hydroxylation (Letz, 1983). Several members of the cytochrome (CYP) P450 system have been suggested to be involved in the transformation of PCBs (H. Inui, Itoh, Yamamoto, Ikushiro, & Sakaki, 2014; Hideyuki Inui,

Itoh, Yamamoto, Ikushiro, & Sakaki; Kakutani et al., 2014; Warner, Martin, & Wong, 2009). Animal studies have demonstrated that CYP-dependent monooxygenases are involved in the metabolism of some PCBs; 3-methylchloranthrene-treated rats that received 3,3',4,4',5-pentaCB (PCB126) intraperitoneally metabolized PCB126 to 4-OH-3,3',4',5,5'-pentaPCB, which was detected in their feces (Koga, Beppu, & Yoshimura, 1990). The clearance of PCBs has also been shown to depend on their structure - for example, since two adjacent unsubstituted positions facilitate excretion, rates of clearance are inversely proportional to percent chlorination (Letz, 1983).

There have been no comprehensive efforts to scan the genome for variants which may be linked to the metabolism of these compounds. As such, this is the first GWAS that addresses this issue.

The second type of pollutant that this study focuses on are polychlorinated dioxins (PCDDs), which consist of dibenzo-1,4-dioxin (two benzene rings joined by oxygen bridges) with chlorine atoms attached at any of the 8 places on the molecule (INSERM Collective Expert Reports, 2000). PCDDs are widely present in the environment, and are produced in several industries such as herbicide manufacture and metal smelting, in addition to combustion processes involving medical waste and sewage sludge.

PCDDs are bio-accumulative and their effects on health are mediated by the aryl hydrocarbon receptor action. There are a myriad of health effects, including thyroid problems, immune system disorders, endometriosis, diabetes and neurological pathology (INSERM Collective Expert Reports, 2000). Additionally, exposure to high levels of dioxins cause chloracne, a skin pathology involving the eruption of blackheads and cysts on the face and shoulders (Sorg, 2014).

PCDDs are mainly absorbed through dietary intake, which includes a variety of food sources eg. beef, dairy and milk as well as inhalation of cigarette smoke. Besides dietary intake, industrial exposure is also a source of human absorption, especially among workers in this industry. PCDDs are stored in fatty tissues and their half-life depends on chemical structure, but is on average 7 to 11 years (World Health Organisation, 2014). Various factors affect dioxin concentration, including age, body mass and dietary habits. Excretion is primarily through faeces and sebum (Harrad, Wang, Sandaradura, & Leeds, 2003). There are no known genetic variants which affect levels of dioxins. This would be useful in identifying individuals who may be more susceptible to dioxin toxicity and also possibly uncovering more details about dioxin metabolism in the human body.

The third group of pollutants that I focus on in this chapter are toxic metals, including manganese, zinc, nickel, mercury and cadmium. In our environment, sources of toxic metal are ubiquitous, and range from shipyard and construction business to agricultural and domestic industries. Some of these metals are bio-accumulative, and have a variety of toxic effects in the human body (Godt et al., 2006).

The accumulation of these metals in the body depends on environmental exposure, as well as mechanisms in the body that control adsorption, distribution, metabolism and elimination (ADME), some of which are genetically influenced. For example, several gene transporter families have been implicated in manganese transport in plants (Socha & Guerinot, 2014). It is not known whether these mechanisms are also operating and valid in humans. One way to examine this is to evaluate if genetic variation in the genes encoding to the critical proteins involved in these ADME steps are related to circulating levels of relevant metals.

To identify genetic variation associated with the pollutants described above, I have undertaken a GWAS in a population of Swedish individuals. Blood samples were collected

and genotyped on the Illumina Omni-Express chip. Pollutant levels were measured from the whole blood and association tests were performed with SNPtest using an additive model. Conditional analysis was then performed to search for secondary signals.

2.3 Materials and Methods

2.3.1 Study Subjects

The Prospective Investigation of Vasculature in Uppsala Seniors study was initiated in 2001 to investigate the predictive power of different measurements of vascular characteristics for future cardiovascular events, and secondary aims included measurements of cardiac and metabolic function, as well as serum biomarkers and levels of environmental pollutants including PCBs. All individuals aged 70 living in the community of Uppsala in Sweden were deemed eligible for the study. The subjects were selected from the community register and invited in randomised order between April 2001 and June 2004. They received an invitation letter for participation within 2 months of their 70th birthday. Of the 2,025 subjects invited, 1,016 (507 male, 509 female) subjects agreed to participate. The participants were asked to answer a questionnaire about their medical history, smoking habits and regular medication. Sample details and methods are described more fully elsewhere (L. Lind, Fors, Hall, Marttala, & Stenborg, 2005).

2.3.2.1 Chemical analysis of PCBs and OCDD

16 PCBs were included in this study. They were 2,2',4,4',5-Pentachlorobiphenyl (PCB-74), 2,2',4,4',5-Pentachlorobiphenyl (PCB-99), 2,3,3',4,4'-Pentachlorobiphenyl (PCB-105), 2,3',4,4',5-Pentachlorobiphenyl (PCB-118), 3,3',4,4',5-Pentachlorobiphenyl (PCB126), 2,2',3,4,4',5'-Hexachlorobiphenyl (PCB-138), 2,2',4,4',5,5'-Hexachlorobiphenyl (PCB-153), 2,3,3',4,4',5-Hexachlorobiphenyl (PCB-156), 2,3,3',4,4',5'-Hexachlorobiphenyl (PCB-157), 3,3',4,4',5,5'-Hexachlorobiphenyl (PCB-169), 2,2',3,4,4',5,5'-Heptachlorobiphenyl (PCB-70), 3,3',4,4',5,5'-Hexachlorobiphenyl (PCB-180), 2,3,3',4,4',5,5'-Heptachlorobiphenyl (PCB-189), 2,2',3,3',4,4',5,5'-Octachlorobiphenyl (PCB-194), 2,2 ',3,3',4,4',5,5',6-Nonachlorobiphenyl (PCB-206), and (2,2',3,3',4,4',5,5',6,6'-Decachlorobiphenyl (PCB-209) (Table 2.1). These were measured in 1,016 stored plasma samples. Briefly, frozen plasma was allowed to thaw at room temperature and formic acid was added to 500 µL plasma samples to denature proteins followed sonication. Internal standards, including ^{13}C-labeled 2,3',4',5-Tetrachlorobiphenyl (^{13}C-PCB-70), 2,2',4,5,5'-Pentachlorobiphenyl (^{13}C-PCB-101), ^{13}C-PCB-118, ^{13}C-PCB-105, ^{13}C-PCB-138, ^{13}C-PCB-153, ^{13}C-PCB-156, ^{13}C-PCB-170, ^{13}C-PCB-180, ^{13}C-PCB-194, ^{13}C-PCB-206, ^{13}C-PCB-209 and octachlorodibenzo-p-dioxin (OCDD), were added to the samples and sonicated prior to the solid-phase extraction (SPE) using Oasis® HLB single use cartridges (6 cm3/150 mg, Waters, Milford, MA, USA). The extracts were further purified with multi-layered acidified silica columns containing 56% KOH silica gel, 40% H_2SO_4 silica gel and activated anhydrous Na_2SO_4. Of the final 25 µL volume, 2 µL was injected on a HRGC/HRMS system and measurements were performed on a Micromass Autospec Ultima (Waters, Mildford, MA, USA) mass spectrometer (MS) operating at ≥ 10 000 resolving power using EI ionization at 35 eV. Measurements were performed in selective ion mode (SIM) and quantification was performed according to the isotope dilution method using the ^{13}C-labeled standards. The MS was coupled to a 6890N gas chromatograph (GC) (Agilent Technologies, Atlanta, GA, USA).

The injector was programmed to 275 °C. The PCBs were separated on a 30m x 0.25 i.d. x 0.25 µm DB-5 capillary column (SGE Analytical Science, Victoria, AUS) with the GC oven programmed from 180 °C (2 min) to 260 °C (3.5 °C/ min) and to 300 °C (6.5 °C/min, 2min).

The analytical method employed to all plasma samples was successfully validated in terms of recovery, precision, and reproducibility. Recovery of labelled internal standards ranged between 49-86%. Reproducibility of procedural quality control samples was below 25% relative standard deviation for PCBs above the limit of detection in the samples. Contamination of procedural blank samples did not exceed any PCB levels above 5% of the levels in the authentic samples. Limits of detection were between 0.8-117.7 pg/mL. Further details on the method and its analytical performance have been described previously (Salihovic, Mattioli, et al., 2012).

2.3.2.2 Chemical analysis of metals

Eleven toxic metals and trace elements (Al, Cd, Co, Cu, Cr, Hg, Mn, Mo, Ni, Pb and Zn) in this study were determined in whole blood. The analysis was performed using inductively coupled plasma-sector field mass spectrometry (ICP-MS), after microwave- assisted digestion with nitric acid (Rodushkin, Odman, Olofsson, & Axelsson, 2000). ICP-MS combines a high temperature source with mass spectrometry. The source converts the atoms into ions, which are then detected by the mass spectrometer. Details on the analysis have been described in Lind et al, 2012 (P. M. Lind, Olsén, & Lind, 2012).

2.3.3 Genotype analysis

DNA samples were genotyped according to the manufacturer's instructions on Illumina Infinium Omni Express bead microarrays. Samples were deposited on plates, scanned with the machine and the images were pre-processed with GenomeStudio. Samples were excluded if the call rate was less than 95%, if they had extreme heterozygosity (>3 SD from the mean), if they were ethnic outliers, or if they were gender discordant. SNP quality control measures included exact p-value for deviation from HWE < 10^{-6} and missing genotype rate > 0.01 (MAF < 5%) or missing genotype rate > 0.05 (MAF ≥ 5%). Multidimensional scaling was performed to obtain principal components to adjust for population structure.

SNPs used for MDS had the same quality control filters applied and were LD pruned to r2<0.2 after removal of known regions of LD and large inversions.

Prior to imputation, variants with MAF < 1% were removed from the GWAS scaffold. Samples were pre-phased with SHAPEIT2. Genotype data were then imputed up to the "all ancestries" reference panel from the 1000 Genomes Project Consortium Phase 1 interim release (June 2010) (Abecasis et al., 2012) using IMPUTEv2 (B. Howie et al., 2012).

2.3.4 Statistical analysis

The concentrations (pg/mL) of pollutants were rank inverse normalised to generate a Gaussian distribution for downstream association analyses, and minimise the impact of outliers. This method of normalisation has been used widely and is shown to be reliable (Beasley, Erickson, & Allison, 2009). Association testing for each transformed level was performed in a linear regression framework under an additive model in the minor allele, after adjusting for triglycerides, cholesterol, gender and 2 principle components from multidimensional scaling to account for population structure as covariates. This method was

used for all pollutants as circulating lipids transport some of the pollutants and the presence of abdominal obesity have been associated with some of the pollutants (D. H. Lee et al., 2012). In addition, there are gender differences between PCBs (Salihovic, Lampa, et al., 2012). Age was not included as a covariate because all individuals were of the same age.

Association testing was performed in SNPTEST, allowing for uncertainty in the imputation in a missing data likelihood. In this study, the association analysis was restricted to SNPs with MAF > 1% and imputation quality information score (info) of greater than 0.4 and HWE p value of 10^{-6}. The genomic control inflation factor for each PCB congener was used to assess evidence of residual population structure. (Devlin & Roeder, 1999b) To account for multiple testing of 28 pollutants, I defined genome-wide significance as $p<1.8x10^{-9}$.

Conditional analyses to search for secondary signals of association were performed by including genotype dosage at the lead SNP as an additional covariate in the linear regression model. In the case of PCBs, because of correlations between multiple PCBs, I also performed reciprocal conditional phenotype analysis, by including serum levels of pollutants as an additional covariate in the regression model for another.

2.4 Results

Following imputation, a total of 8,736,858 high-quality SNPs with MAF > 1% were included in the analysis. The genomic control inflation factor for each pollutant was close to 1 (Table 2.1), so no additional correction for residual population structure was required after adjustment for principal components. Significant associations between genetic variants were detected for PCBs, metals as well as for OCDD, one of the PCDDs.

Table 2.1 Median Values of pollutants and genomic control lambda

Pollutant	**Units**	**Median Value**	**GC λ**
PCB-74	pg/ml	91.4	1
2,2',4,4',5-pentachlorobiphenyl		(63.9-128.1)	
PCB-99	pg/ml	90.9	0.995
2,2',4,4',5-pentachlorobiphenyl		(62.4-131.9)	
PCB-105	pg/ml	32	1.001
2,3,3',4,4'-pentachlorobiphenyl		(21.0-46.8)	
PCB-118	pg/ml	200.6	0.998
2,3',4,4',5-pentachlorobiphenyl		(136.4-281)	
PCB-126	pg/ml	40.4	1.005
3,3',4,4',5-pentachlorobiphenyl		(71.9-385.8)	
PCB-138	pg/ml	819.3	1.001
2,2',3,4,4',5'-Hexachlorobiphenyl		(619.2 -1115.8)	
PCB-153	pg/ml	1427.6	0.996
2,2',4,4',5,5'-Hexachlorobiphenyl		(1114.4-1847.9)	
PCB-156	pg/ml	154.3	1.007
2,3,3',4,4',5-Hexachlorobiphenyl		(118.7-197.6)	
PCB-157	pg/ml	28	1.015
2,3,3',4,4',5'-Hexachlorobiphenyl		(21.4-37.0)	
PCB169	pg/ml	171.4	0.997
3,3',4,4',5,5'-Hexachlorobiphenyl		(219.9-636.4)	
PCB-170	pg/ml	497.5	1.002
2,2',3,4,4',5,5'-Heptachlorobiphenyl		(385.7-633.0)	
PCB-180	pg/ml	1165.4	1.001
3,3',4,4',5,5'-Hexachlorobiphenyl		(917.8-1487.8)	
PCB-189	pg/ml	19.3	0.999

2,3,3',4,4',5,5'-Heptachlorobiphenyl		(14.6-25.8)	
PCB194	pg/ml	119.4	1.003
2,2',3,3',4,4',5,5'-Octachlorobiphenyl		(87.6-158.9)	
PCB-206	pg/ml	26.8	0.987
2,2',3,3',4,4',5,5',6-Nonachlorobiphenyl		(20.8-35.2)	
PCB-209	pg/ml	26.2	0.982
(2,2',3,3',4,4',5,5',6,6'-Decachlorobiphenyl)		(19.6-34.7)	
OCDD	pg/ml	2.6	1.01
octa-chloro-di-benzo-dioxin		(1.6-4.2)	
Al	mmol/L	0.637	1.02
Aluminium		(0.504–0.795)	
Cd	nmol/L	2.41	1
Cadmium		(1.72–3.74)	
Co	nmol/L	1.43	1
Cobalt		(1.09–1.90)	
Cr	nmol/L	11.8	1.01
Chromium		(9.58–16.00)	
Cu	nmol/L	12.8	1
Copper		(11.6–14.1)	
Hg	(nmol/L)	8.92	1
Mercury		(5.86–13.65)	
Mn	nmol/L	138	1.01
Manganese		(115–165)	
Mo	nmol/L	9.74	0.991
Molybdenium		(8.09–11.80)	
Ni	nmol/L	90.6	0.982
Nickel		(68.3–109.0)	
Pb	mmol/L	0.083	1.01

Lead		(0.059–0.110)	
Zn	mmol/L	96	1.02
Zinc		(87.7–104.0)	

For PCBs, levels were successfully determined in 922 participants and the majority of the studied congeners were detected in the 70-100% of the samples. Among the 16 PCB congeners measured, PCB-153 was detected in the highest concentrations among the participants (1430 $pgmL^{-1}$) followed by PCB-180, -138, and -118. Together these four compounds accounted for >70% of the entire PCB body burden in the participants. Compared to other studies, the concentrations of the PCBs in the present study were found similar, or comparable, to other studies of general populations from Sweden and Europe (Salihovic, Lampa, et al., 2012). Median values for the studied PCBs are presented in Table 2.1.

Table 2.2 Lead SNPs of genome wide significant loci for pollutants

Pollutants	Beta	P value	SNP	Variance
PCB99	0.34	3.7×10^{-13}	rs8109848	0.058
Manganese	-0.76	5.1×10^{-11}	rs13107325	0.108
Manganese	-0.45	2.17×10^{-14}	rs1776029	
Cadmium	-1.16	1.35×10^{-10}	rs9350504	0.045
Mercury	-1.80	1.76×10^{-9}	rs148534631	0.027
OCDD	0.59	4.3×10^{-16}	rs2231142	0.073

2.4.1 PCBs

We identified genome-wide significant evidence of association for serum levels of PCB-99, mapping to *CYP2B6*. The SNP with the strongest association signal was rs8109848 (β=0.34, SE=0.05, $p=3.7x10^{-13}$), which maps to the intronic region of *CYP2B6* (Figure 2.1 and 2.2). This SNP also showed weaker association with several other PCBs including PCB-138 ($p=3.0x10^{-8}$) and PCB-153 (p=0.000041). Conditional analyses, including the genotype dosage of rs8109848 as a covariate in the regression model, extinguished the association signal for all pollutants at this locus.

Given the known high correlation between different serum pollutant levels (Lampa et al., 2012) (Figure 2.3), I sought to dissect the effect of *CYP2B6* variation on PCB 99 and other PCB congeners through reciprocal conditional phenotype analyses. After adjusting for PCB-99 as an additional covariate in the regression model, the previously observed associations of rs8109848 with PCB-138 and PCB-153 were extinguished (p=0.17 and p=0.0074 respectively). These data may suggest pleiotropy in which the variant had heterogeneous effects on the different pollutant levels, which are correlated.

We next performed conditional phenotype association testing for all other PCB congeners within the *CYP2B6* region, adjusting for PCB-99 levels as an additional covariate in the regression model. When all PCBs were conditioned on PCB-99, further genome-wide significant association signals were revealed for PCB 74, PCB-105 and PCB-118, all mapping to the same genomic region (Figure 2.4). The lead SNPs were rs8109848 (β = -0.21, SE=0.03, $p=3.8x10^{-12}$) for PCB118, rs4802104 (β = -0.19, SE = 0.03, $p = 1.4x10^{-9}$) for PCB-74 and rs4803413 (β= -0.18, SE=0.03, $p = 2.5x10^{-9}$) for PCB 105. These SNPs had no significant

association in unconditional analysis: rs8109848 (p=0.43) for PCB-118, rs4802104 (p = 0.32) for PCB-74 and rs4803413 (p=0.27) for PCB-105.

The three index SNPs are in strong LD in European ancestry populations (r2>0.76), and thus may reflect the same underlying causal variant. Furthermore, when the PCBs were conditioned on PCB-99 and PCB-118 together, the association signals at this locus were extinguished for PCB-74 (rs4802104, p=0.02) and PCB-105 (rs4803413, p=0.52). Furthermore, no additional signals of association achieved genome-wide significance. These data suggest that the identified association signals observed for PCBs at this locus are mediated through PCB-99 and PCB-118.

2.4.2 Metals

A total of 3 metals demonstrated signals at a genome-wide significance level. The GWAS analyses revealed two regions that were associated with manganese level. These mapped to 4q24 and 1q41 (Figure 2.5 and Figure 2.6). The lead SNP in the 4q24 locus was a coding variant, rs13107325 (p value 5.08×10^{-11}, beta of -0.767, MAF of 0.038), located in *SLC39A8*, a gene that encodes a well-known protein involved in manganese and zinc transport (Kambe, 2011). The amino acid change for this SNP was alanine to threonine. The lead SNP in the 1q41 locus was rs1776029 (p value of 2.17×10^{-14}, beta of -0.456, MAF of 0.182). The SNP lies within the intronic region of *SLC30A10,* another well-established ion transporter protein (Quadri et al., 2012). Conditional analysis revealed no secondary signals.

The locus 6q14.1 was associated with cadmium levels. The lead SNP was rs9350504 (p value of $1.35x10^{-10}$, beta of -1.16, MAF of 0.0175) which is an intronic variant within CD109, a glycosyl phosphatidylinositol-linked glycoprotein (Figure 2.7)

The 3q26.32 locus was associated with mercury levels (p value of $1.76x10^{-9}$, beta of -1.80, MAF of 0.0140). The lead SNP is rs148534631, an intron variant within *LINC00578*, an intergenic non-protein coding RNA (Figure 2.8).

2.4.3 OCDD

SNPs in the region of 4q22 are associated with OCDD levels (Figure 2.9). The top SNP is rs2231142 (p value of 4.3e-16), a missense mutation corresponding to G>T, resulting in Gln to Lys at residue 141 of the ATP-binding cassette sub-family G member 2 (*ABCG2*) gene, within exon 5. This residue is highly conserved across different species (Woodward et al., 2009). There were no secondary signals on conditional analysis.

2.5 Discussion

This study reveals multiple associations between genetic loci and pollutant levels, demonstrating that a GWAS approach is useful for investigating genetic influences in the metabolism of common pollutants. In all three sets of pollutants, imputation helped to refine the signal; in the case of PCBs, conditional analysis helped to uncover complex secondary signals which are crucial to understanding the mechanism for SNP action. The GWAS approach is advantageous as it covers all information on variation in all human genes, hence

and thereby no prior hypothesis on which genes that are of importance is needed. This is novel because there have been no previous GWAS for comparable pollutants.

2.5.1 PCBs

PCBs are polyhalogenated aromatic hydrocarbons theoretically present in 209 different congeners related to their number of chlorine atoms and other chemical characteristics affecting their 3D structure. The present study is the first examination in humans of the relationship between genome-wide genetic variation and circulating levels of various PCBs. Consistent with previous experimental results, I show that *CYP2B6* is involved in the metabolism of at least some PCBs in humans. (Gährs, Roos, Andersson, & Schrenk, 2013; Lehmler et al., 2010). *CYP2B6* has several known alleles (http://www.cypalleles.ki.se/*CYP2B6*.htm), many of which influence its activity and are associated with serum levels of a range of pharmaceutical drugs. The commonest allele is *CYP2B6**6 [Q172H and K262R], tagged by rs3745274 and rs2279343, respectively. *CYP2B6* genotyping has been applied successfully in adjusting therapy for certain diseases, for example, it is useful in the treatment of HIV-infected individuals, to reduce the therapeutic dose of efavirenz, resulting in improvement of central nervous system-related side effects (Gatanaga et al., 2007).

Conditional analyses suggest that association signals with pollutants at the *CYP2B6* locus are mediated through PCB 99 and PCB118, however, the exact mechanism for this is unclear. The common denominator for these pollutants seems to be a low number of chlorine atoms, possibly reflecting a shorter half-life than the more lipophilic PCBs with a larger number of chlorine atoms. The presence of chlorine substitution in PCB 99 and PCB 118 for both the

meta and the para position prevents degradation in biota. This also holds true for hexa PCB isomers 153 and 138. The larger molecules (hepta to deca PCB congeners) and dioxin like hexa isomers (PCB 156, 157, 169) show no correlation with variants at the *CYP2B6* locus.

Another common theme among the PCBs associated with the CYP2B6 locus is that they are all highly bioaccumulative. Furthermore, they are largely non-dioxin like. Dioxin-like PCBs such as PCB 118 and PCB 105 are agonists of aryl hydrocarbon receptors but not non-dioxin-like PCBs like PCB 99, PCB 138 and PCB 153 McFarland and Clark (1989) (McFarland & Clarke, 1989) suggested a classification based on the ability to induce CYPs in a barbiturate-like fashion, and PCB 99 and PCB 153 belong to that class. Wolff et al later further subdivided the PCBs in to antiestrogenenic and phenobarbital-type inducers, and PCB 99 and PCB 153 belong to that latter class (Wolff, Camann, Gammon, & Stellman, 1997). It should however be pointed out that not all PCBs considered to be phenobarbital-type inducers demonstrated an association with genetic variation in the *CYP2B6* gene in the present study. Within this group it was mainly those with a moderate number of chlorine atoms that were associated, but not those with a large number of chlorine atoms, like PCB180 and PCB194, having a longer half-life (Wimmerová et al., 2011).

CYP2B6 has also been reported to metabolise a range of environmental chemicals, including insecticides such as malathion (Buratti, D'Aniello, Volpe, Meneguz, & Testai, 2005) and carbofuran (Usmani, Hodgson, & Rose, 2004), herbicides such as atrazine (Lang, Rettie, & Bocker, 1997) and metolachlor (Coleman, Linderman, Hodgson, & Rose, 2000) as well as industrial contaminants (Cho, Rose, & Hodgson, 2006), hence this study has wider implications for the toxicogenomics, and may help to direct further investigations into individual susceptibility to various environmental pollutants.

2.5.2 Manganese

Besides associations with the cytochrome P450 locus, the investigation has also revealed novel associations between metals and transporter genes, most prominently manganese and *SLC39A8* as well as *SLC30A10*. Manganese is a trace metal that is essential to the body, but can also be neurotoxic at higher chronic exposure levels, with symptoms persisting long after cessation of exposure. Occupational studies have demonstrated that exposure to manganese can cause cognitive and motor defects, in addition to changes in mood and short term memory, altered reaction time and reduced hand eye coordination (Antonini, Santamaria, Jenkins, Albini, & Lucchini, 2006; Bowler et al., 2011). There is little known about genetic factors that affect serum manganese levels.

In our study, rs13107325 was associated with manganese levels. Unlike PCBs, there is a clear biological mechanism through which this SNP influences the phenotype. It gives rise to substitution of alanine (hydrophobic) to alanine (hydrophilic). It is in low LD with nearby SNPs in HapMap sample (Carrera et al., 2012). *SLC39A8* is one of the members of the family 39 of solute carrier transporters. It transports manganese and zinc, which are crucial for the functioning of various organs, but may be neurotoxic at levels beyond a certain concentration. Excessive exposure to manganese may lead to Parkinsonian features, as well as other neuropsychiatric symptoms such as agitation and hallucination. *SLC39A8* is responsible for the uptake of manganese and zinc from blood into various tissues. It is expressed at high levels in various tissues (including placenta), but low levels in the brain. Several studies have uncovered an association between this SNP and other phenotypes, including metabolic traits such as blood pressure, body mass index and lipid levels (Ehret et al., 2011; Speliotes et al., 2010; Teslovich et al., 2010). This SNP has also been associated with neurological

phenotypes including schizophrenia (Carrera et al., 2012) and stress (Bruenig, White, Young, & Voisey, 2014). The lead SNP in the 1q41 locus is rs1776029. The SNP lies within the intronic region of *SLC30A10*. *SLC30A10* is also responsible for manganese transport. It is classified as a member of the *SLC30* solute carrier subfamily of the cation diffusion facilitator (Seve, Chimienti, Devergnas, & Favier, 2004). However, unlike rs13107325, the mechanism through which this SNP influences the phenotype is unclear.

2.5.3 Cadmium and Mercury

Similarly, genetic loci associated with cadmium and mercury do not present an obvious biological mechanism through which they exert their effects. The locus 6q14.1 was associated with cadmium levels and the lead SNP, rs9350504 is an intronic variant within CD109, a glycosyl phosphatidylinositol-linked glycoprotein. Cadmium is a common environmental pollutant, with exposure arising from dietary sources such as shellfish, liver and kidney, as well as inhalation of cadmium-contaminated air from industry sectors and tobacco smoke (Olsson et al., 2002). Adverse health effects have been reported in association with low-level environmental cadmium exposure, including liver and kidney damage (A. Engström et al., 2012). In addition, there is increased risk of cancer, including breast, kidney, pancreas, and urinary bladder (Huff, Lunn, Waalkes, Tomatis, & Infante, 2007).

The genetic basis of cadmium ADME is poorly understood. Our study demonstrated an association between cadmium blood levels and a SNP within the *CD109* gene. *CD109* encodes a glycosyl phosphatidylinositol-linked glycoprotein that localises to the surfaces of endothelial cells, T cells and platelets. It is associated with carcinogenesis. For example, upregulation of *CD109* is associated with oral cancer and lung cancer (Hagiwara et al., 2008;

Sato et al., 2007). High expression of *CD109* antigen is associated with poor prognosis of soft tissue sarcoma (Emori et al., 2013), and the presence of *CD109+* endothelial cells is a prognostic indicator for glioblastomas (Cuppini et al., 2013). It is of interest to note that cadmium exposure also has been discussed in the etiology of several cancers (Feki-Tounsi & Hamza-Chaffai, 2014; Rahim, Jalali, & Tangestani, 2013), however, a direct link between CD109, cadmium and cancer has yet to be established.

The 3q26.32 locus was associated with mercury levels with the lead SNP rs148534631 being an intron variant within *LINC00578*, an intergenic non-protein coding RNA. A few genetic variants have previously been associated with mercury toxicokinetics, mainly in the glutathione-related genes (Custodio, Harari, Gerhardsson, Skerfving, & Broberg, 2005; A. Engström et al., 2012; Harari et al., 2012), as well as multispecific transporter family genes (K. Engström et al., 2013). Our study demonstrates an association between mercury levels and a SNP in *LINC00578*, but the function of this intergenic non-protein coding RNA is not clear.

2.5.4 OCDD

In contrast, the association between OCDD and SNPs in *ABCG2* is easier to interpret in terms of biological function. SNPs in the region of 4q22 are associated with OCDD levels The top SNP is rs2231142 (p value of 4.3e-16), a missense mutation corresponding to G>T, resulting in Gln to Lys at residue 141 of the *ABCG2* gene. *ABCG2* is a cell membrane transporter that extrudes endobiotics and xenobiotics such as uric acid and porphyrin compounds (Woodward et al., 2009). It is a member of the ATP-binding cassette family (Sarkadi, Homolya, Szakács, & Váradi, 2006), and plays an important role at various barriers, such as the maternal-foetus

barrier and blood brain barrier. Each molecule of *ABCG2* contains one NBD and one TMD, hence homo-dimerization is needed for full functionality. The nucleotide binding domain (NBD), which lies within the cytosol, hydrolyses ATP and then the transmembrane domain (TMD) provide the pathway for substrate recognition and transport (Kerr, Jones, & George, 2010). Substrate specificity and molecular stability can be significantly influenced by polymorphisms within the *ABCG2* gene. The Q141K variant is located within the NBD, and results in less effective substrate transport (Imai et al., 2009) and reduced membrane expression level (Mizuarai, Aozasa, & Kotani, 2004).

A GWAS of more than 11000 European individuals, with replication in another 11 000 Europeans and 3800 African Americans identified common alleles in *ABCG2* as being associated with serum uric acid levels and risk of gout (Dehghan et al., 2008). The *ABCG2* variant with the strongest effect on serum urate concentrations was the SNP rs2231142 with each copy T allele associated with mean serum urate concentrations approximately 0.30 mg/dL higher ($P = 3 \times 10^{-60}$). The odds of gout were increased by 74% with each copy of the T allele ($P= 4 \times 10^{-15}$). Since then, it has been replicated in various populations (Tabara et al., 2010; Yamagishi et al., 2010).

Besides rs2231142, other SNPs within *ABCG2* have also been found to be associated with uric acid levels in separate GWAS. rs1481012, an intronic SNP with a p value of 1.49e-15 to OCDD in our study, is associated with uric acid levels in a meta-analysis on more than 140 000 individuals (Köttgen et al., 2013). This SNP is also associated with LDL reduction of cholesterol with statin therapy (Chasman et al., 2012). *ABCG2* is associated with hepatic drug efflux, including efflux of rosuvastatin as well as renal transport (Dehghan et al., 2008; Doyle & Ross, 2003).

While *ABCG2* has not been linked to OCDD, there is incidental evidence suggesting that it might play a role in OCDD metabolism. A Japanese study demonstrated that dioxin poisoning from contaminated food resulted in increased serum uric acid as a prominent feature (Imamura et al., 2009). Furthermore, occupational dioxin exposure has been linked to increased serum uric acid levels (Kitamura et al., 2000). Experimental evidence suggests that dioxins are high affinity ligands for the aryl hydrocarbon receptor, which upregulate expression of *ABCG2* in human cancer cell lines. This plays a role in cancer multidrug resistance, as it *ABCG2* limits intracellular accumulation of chemotherapeutic agents (Tan et al., 2010). As such allelic differences in rs2231142 is also associated with chemotherapy response and clinical outcomes in ovarian cancer, renal cancer and leukemia (Garcia-Donas et al., 2011; M. Kim et al., 2002; Tian et al., 2012).

The genetic variants discussed above have large effects on pollutant levels. This suggests that regulation of metabolism may be near-monogenic, as there are no SNPs in other genes with effects that reach genome wide significance. Although the biological interpretation requires further functional study, this could possibly be because the rate-limiting step in metabolism may be performed by an enzyme whose activity is regulated by one particular gene.

There is the potential to use these variants as instrumental variables to ascertain the effect of pollutant levels on diseases such as cancer through Mendelian Randomization (MR). MR is a method to obtain unbiased estimates of the effects of a possible causal variable without conducting a randomised controlled trial. This helps to control for reverse causation and confounding because a genetic variant can only cause a phenotype and not the other way around. In MR, common genetic polymorphisms with large effects on exposure patterns or

effects that are closely correlated with exposures are used. It is important that the genotype must only affect the disease via its effect on the stated exposure. Because genotypes are assigned randomly through meiosis, the population distribution should not be related to factors which confound epidemiology studies, hence MR can be thought of in similar terms to a randomised controlled trial.

There are various limitations to this study. Firstly, blood serum pollutant levels may reflect pollutant exposure, rather than purely metabolism. Although individuals were recruited from Uppsala, it is possible that some of these individuals had moved to Uppsala recently from another region. Secondly, blood pollutant measurements were subject to a range of reasons for variability, for example, date and time of collection, personnel involved and experimental error in analysis. Longitudinal studies may help to improve reliability. Thirdly, while we are measuring serum pollutant levels in the blood, the main reservoir for PCBs is fatty tissue. Ideally, an adipose biopsy would be ideal, however, this is invasive, and would severely limit sample size, hence serum levels were used as surrogate measures.

2.6 Chapter Summary

In summary, this chapter describes a GWAS study of various environmental pollutants. I identify loci that are associated with PCBs, toxic metals and dioxins. Some of these associated SNPs, such as those in *ABCG2* and *SLC39A8,* are coding variants with obvious biological functions. Other are intronic, and may share complex relationships with variants associated with other phenotypes, such as those in the *CYP2B6* locus. Chapter 4 will discuss how these variants can be assigned with biological functionality by annotating them with regulatory elements.

CHAPTER 3 GWAS and meta-analysis of serum metabolic traits in 2 cohorts of Swedish individuals

3.1. Chapter Overview

3.2 Introduction

3.3 Methods and Materials

3.4 Results

3.5 Discussion

3.6 Chapter Summary

3.1 Chapter Overview

This chapter describes a GWAS meta-analysis on two cohorts of Swedish individuals – the PIVUS cohort (described in Chapter 2) and the Uppsala Longitudinal Study of Adult Men (ULSAM) cohort. I describe how association testing and meta-analysis were performed for a set of metabolic phenotypes: D6 and D5 desaturase (D6D, D5D), high density lipoprotein (HDL), triglycerides (TG), apolipoproteins B (ApoB). I discuss the significant associations found, comparing them with other studies, and identifying potential biological mechanisms through which they can act.

3.2 Introduction

Various metabolic phenotypes have been associated with cardiovascular disease. These include serum lipid biomarkers such as HDL and triglycerides, as well as cardiac parameters such as blood pressure and ejection fraction. These traits are perturbed in conditions like diabetes mellitus, high blood pressure and obesity, increasing susceptibility to cardiovascular diseases such as stroke and myocardial infarction. Understanding the genetic basis of these intermediate metabolic phenotypes may help to provide insight into the pathophysiology of these diseases and may suggest mechanisms that could be perturbed through treatment to reduce disease risk.

3.2.1 D6 and D5 Desaturase

D6D and D5D are enzymes which play a crucial role in the endogenous elongation and desaturation of fatty acid precursors into physiologically active long chain polyunsaturated fatty acids (PUFAs). PUFAs refer to fatty acids with multiple desaturations in the aliphatic tail (Leonard, Pereira, Sprecher, & Huang, 2004). There are 2 main types of such fatty acid: long chain PUFAs (with 18 or more carbons in length) and short chain PUFAs (with up to 16 carbon atoms). PUFAs can also be classified into n-6 or n-6 fatty acids, depending on the position of the first double bond proximate to the methyl end (Das et al., 2000). The initial step in PUFA synthesis is the desaturation of linoleic acid (LA) and alpha-linolenic acid (ALA) by *FADS2* (Serhan, Chiang, & Van Dyke, 2008). PUFAs play an important role in modulating inflammatory processes, through a variety of mechanisms, including modification of cyclooxygenase and lipoxigenase activity. These two compounds are central in the metabolic pathways that generate eicosanoids (Jump, 2002). Since n-3 and n-6 fatty acids have competitive effects on the same pathway, yet produce eicosanoids with different

biological effects, it is possible that these 2 classes of PUFA may be vital in the pathogenesis of inflammatory disease (Merino et al., 2011).

PUFAs have a myriad of crucial functions in the human body. Firstly, they are fundamental components of the structure of cell membranes, particularly within the central nervous system (Leonard et al., 2004). Their incorporation can alter the membrane's permeability and selectivity, since saturated fatty acids make the membrane more rigid. In this way they can alter the activity of a host of membrane-associated enzymes. For example, increase in membrane fluidity has been associated with increase in insulin sensitivity because of increased number and affinity of insulin receptors on the surface of cells (Risérus, 2008). Besides being membrane components, PUFAs are also a major source of fuel for energy metabolism as they are beta-oxidised in mitochondria (Leonard et al., 2004). PUFAs can act as ligands for transcription factors, which are involved in fatty acid oxidation, lipogenesis and steroid hormone production. For example, omega-3 PUFA influences several nuclear receptors (HFN-4alpha, PPAR, farnesol X receptor), producing hypotriglyceridemic effects (Jump, 2004). PUFAs also regulate the synthesis of inflammatory markers such as eicosanoids. In particular, arachidonic acid is one of the most important substrates for the synthesis of inflammatory compounds which play a role in cancer and cardiovascular disease (D. Wang & Dubois, 2010). In addition, it is also the substrate for the synthesis of cannabinoids, which play a role in neurogenerative conditions and metabolic abnormalities such as beta cell failure in type 2 diabetes (Pacher & Kunos, 2013).

Desaturases are enzymes of the microsomal system. They are believed to be a component of a network that also includes cytochrome b5 and NADH-cytochrome b5 reductase enzymes. D5D and D6D are key enzymes which allow the formation of long chain metabolites from

dietary ALA and LA. They are encoded by *FADS1* and *FADS2* respectively, and their levels can be estimated by calculating the fraction of serum fatty acid levels - 20:3ω6/18:2ω6 (D6D) and 20:4ω6/20:3ω6 (D5D) (Swenne & Vessby, 2013).

The conversion of alpha-linolenic acid (ALA) to stearidonic acid, or linoleic acid (LA) to gamma-linolenic acid, is the rate limiting step of the omega-3 and omega-6 PUFA metabolic pathways (H. P. Cho, M. Nakamura, & S. D. Clarke, 1999; H. P. Cho, M. T. Nakamura, & S. D. Clarke, 1999). Blood and tissue levels of PUFA are known to be influenced by heritability, with up to 28% of blood levels being due to genetic variation (Schaeffer et al., 2006). Investigations within European populations demonstrate associations between polymorphisms in *FADS1* and *FADS2* and PUFA levels both in red blood cell membranes as well as in serum phospholipids (Malerba et al., 2008). Minor alleles were usually associated with lower arachidonic acid (AA levels) and higher LA levels, suggesting decreased desaturase activity. Later studies have replicated these associations in various other populations, including African and Asian (Kwak et al., 2011; Mathias et al., 2011; Merino et al., 2011). One GWAS confirmed the *FADS* locus as the strongest predictor of plasma phospholipid (Lemaitre et al., 2011).

Desaturase activity can be measured in a variety of ways. Firstly it can be measured in animal studies or in vitro by assaying the rate of conversion of fatty acids to respective products using radio-labelling (Narce et al., 1994). In humans, *FADS* gene expression can be determined the relevant tissue – liver cells. However, this would require a liver biopsy, and hence is not feasible for large scale studies (Kröger & Schulze, 2012). A surrogate measure – product to precursor ratio can be measured, with information obtained from the analysis of lipid composition from cell membranes or human plasma, both of which can be obtained from

blood collection. Most commonly, the product to precursor ratios are calculated on omega-6 fatty acids, and can estimate the specific D5D or D6D activity, as well as the global desaturase activity (Martinelli, Consoli, & Olivieri, 2009).

D5 desaturase levels in serum and skeletal muscle have been associated with insulin sensitivity, with high levels predicting a decreased risk of metabolic syndrome (Warensjö, Risérus, & Vessby, 2005) and lower total risk of cardiovascular death (Warensjö, Sundström, Vessby, Cederholm, & Risérus, 2008). On the other hand, there is an inverse relationship between D5D and diabetes risk. Low D5D activity and high D6D activity has been associated with insulin resistance, and also to predict the development of diabetes (Hodge et al., 2007; Wirfält et al., 2004). In a study on children, increased D6D activity and decreased D5D activity have been found to predict future development of metabolic abnormalities caused by insulin resistance (Saito et al., 2013). In addition, liver abnormalities have also been associated with D5 and D5 desaturase activities. In particular, non-alcoholic steatohepatitis (NASH) is a disease associated with fat accumulation and consequent liver inflammation (Armutcu, Akyol, Ucar, Erdogan, & Akyol, 2013). It can deteriorate into cirrhosis, liver failure and the need for liver transplantation. As with risk of diabetes, NASH has been associated with increased D6D and decreased D5D (López-Vicario et al., 2014). Animal studies have also supported this connection, with increased expression of D5D and D6D as measured by mRNA and protein expression when mice were fed with high fat diet and developed NASH (Kang, Wang, Wu, & Kang, 2004).

Desaturase activity has also been linked to cardiovascular risk. In a population of 50 year old men, D6D activity showed a direct relationship with cardiovascular mortality, with D5D showing an inverse relationship (Warensjö et al., 2008). In the Verona Heart study, the AA/LA

ratio, a measure of global desaturase activity, was higher in patients with coronary artery disease (CAD) (Martinelli et al., 2008). This relationship was independent of other cardiovascular risk factors. The authors suggested that high desaturase activity may predispose to inflammatory stimuli during atherosclerosis.

D5D and D6D also play a role in carcinogenesis, with increased eicosanoid formation being a prominent characteristic in various cancer cells (Eberhart et al., 1994). These can interact with surrounding stromal cells and are key mediators in tumorigenesis as well as metastasis through various mechanisms. Inhibition of D6D has been shown to inhibit tumour growth in mice models of intestinal cancer (Hansen-Petrik et al., 2002). Another study demonstrated that D6D activity is upregulated in various other cancers, such as lung carcinoma and melanomas (C. He et al., 2012). Suppression of D6D activity with ribonucleic acid interference caused a reduction in tumour growth, with AA levels simultaneously lowered in tumour tissues. The tumour-reducing effect of D6D inhibition could be mediated by its influence on inflammation and angiogenesis. In patients with leukemia, increased D6D activity has been found, with an alteration of the composition of fatty acids in blood cells (Agatha, Häfer, & Zintl, 2001). Furthermore, inhibition of D6D has been found to decrease the growth of various leukemia cell types (Agatha, Voigt, Kauf, & Zintl, 2004). In renal cancer, increased D6D and decreased D5D activity has been observed, with increased AA and ALA (Hoffmann et al., 2005). High D6D activity has been observed in breast cancer (Pender-Cudlip et al., 2013).

D6D and D5D activity has also been associated with neurological disorders. Firstly, the availability of omega-3 and omega-6 PUFA influences brain development in early life, with alterations in PUFA content being documented in various different neurological illnesses.

Postmortem investigations on the prefrontal cortex of schizophrenia patients demonstrate higher D6D activity (Liu, Jandacek, Rider, Tso, & McNamara, 2009). Furthermore, *FADS2* mRNA expression was higher in patients independent of medication (McNamara et al., 2013). In bipolar disorder, *FADS2* expression was also higher in the prefrontal cortex (Liu & McNamara, 2011).

3.2.2 Apolipoprotein B

Another phenotype investigated in our study was apoB. This is an apolipoprotein found in two forms in humans – apo B48, which is manufactured in enterocytes, and apo B100, which is manufactured in the liver (Aguie et al., 1995). The former is required for the secretion of chylomicrons (fat particles) into the blood stream. The latter is needed for the initial lipidation of very low density lipoprotein (VLDL) so that it can deliver triglycerides from the liver to the general circulation. After secretion, the triglycerides in VLDL are hydrolysed, sending fatty acids to muscle and fat tissue (M. S. Brown & Goldstein, 1979). The resultant particle is then hydrolysed further to form low density lipoprotein (LDL), which can be cleared by the binding of ApoB to the lipoprotein reception. Decreased secretion of apoB results in decreased production of chylomicrons and VLDL, which can lead to decreased absorption of fat and fat soluble vitamins. Although ApoB containing lipoproteins are important for lipid absorption, high levels in plasma can cause atherosclerosis because retention of these particles under the endothelium is a requisite for the initiation of atherosclerosis (Lusis, 2000). Twin studies suggest that plasma levels of apoB are influenced by genetic factors, with 50-60% of the variation being genetically determined (Lamon-Fava et al., 1991). Missense mutations in the apoB gene (mostly rare) can result in very high cholesterol and increased risk

of ischaemic heart disease (Vega & Grundy, 1986). These mutations confer and important risk for those who carry it, but have small impact on the population as a whole.

Ischemic heart disease (IHD) is one of the leading causes of death in the developed world. Elevated levels of LDL can be used to screen individuals for disease risk. However, since risk of atherosclerosis is related to total number of circulating LDL particles rather than the concentration of cholesterol within these particles, apoB has been suggested to be a better predictor of risk, since it measures the total number of atherogenetic particles since each of these contain a single molecule of apoB (Talmud, Hawe, Miller, & Humphries, 2002).

Several studies have established that plasma levels of apoB are predictive of ischemic heart disease in patients with history of myocardial infarction, as well as in the general population (Chien et al., 2007; Ingelsson et al., 2007; Talmud et al., 2002). It can also predict the risk of fatal ischemic stroke in the general population. The Copenhagen City Heart Study, a prospective study which examined a wide range of cardiovascular phenotypes, found that apoB was superior to LDL in the prediction of any ischemic cardiovascular event in both genders (Benn, Nordestgaard, Jensen, & Tybjaerg-Hansen, 2007).

3.2.3 High Density Lipoproteins

HDL particles are a heterogeneous group that differ in density, shape and content. When Apolipoprotein A1 is secreted by the liver, it binds to cholesterol and phospholipids, forming immature HDL particles (Mahdy Ali, Wonnerth, Huber, & Wojta, 2012). These trigger cholesterol efflux in macrophages and fibroblasts, and then store the cholesterol in their core

after esterification, forming mature HDL particles (Mahdy Ali et al., 2012). The mature HDL then delivers the cholesterol to the liver to be excreted in bile acids.

Because of its involvement in cholesterol clearance, decreased HDL is a risk factor for the development of atherosclerosis in IHD. The Framingham study was the first large scale investigation to confirm this, and furthermore demonstrated that the association between HDL and IHD was stronger than between LDL and IHD (T. Gordon, Castelli, Hjortland, Kannel, & Dawber, 1977). In a follow up study 11 years later, HDL levels were associated with increased mortality (Wilson, Abbott, & Castelli, 1988). The link between HDL and IHD has since been proven with angiographic evidence of increased severity of atherosclerosis with increased HDL levels (Jenkins, Harper, & Nestel, 1978). A recent large meta-analysis of 68 long term prospective studies confirms the importance of measuring HDL when assessing risk of IHD (Di Angelantonio et al., 2009). Besides its cholesterol-clearing effects, HDL also has several other properties which make it cardioprotective.

Firstly, it has anti-inflammatory effects, as it inhibits the expression of endothelial adhesion molecules, which play a role in atherosclerosis. Atherosclerosis has been shown to be an inflammatory condition, with adhesion and migration of immune cells into the vessel wall together with the activation of inflammatory cytokines (Ross, 1999). HDL ameliorates this by inhibiting the expression of endothelial adhesion moleculars such as intercellular adhesion molecule-1 (ICAM-1) and vascular cell adhesion molecule-1 (VCAM-1) (Cockerill, Rye, Gamble, Vadas, & Barter, 1995). One *in vivo* study on induced carotid stenosis in rabbits demonstrated that reconstituted HDL resulted in the inhibition of VCAM-1 and ICAM-1 production and abolishment of reactive oxygen species production in the occluded artery (Nicholls et al., 2005).

Secondly, it has anti-oxidative properties, which are beneficial because oxidation of LDL in the arterial wall is essential for the development of atherosclerotic plaques. HDL is associated with antioxidant enzymes such as paraoxonase and platelet-activating factor acetylhydrolase. Furthermore, it aids in clearing oxidised HDL from the circulation and plays a role in the detoxification of oxidised HDL (Garner, Waldeck, Witting, Rye, & Stocker, 1998). Thirdly, it has anti-apoptotic effects, which help to ameliorate cell death in response to endothelial injury, which leads to atherosclerosis. HDL prevents this apoptosis by blocking intracellular signalling pathways involved in this process (Suc, Escargueil-Blanc, Troly, Salvayre, & Nègre-Salvayre, 1997). Fourthly, HDL is inhibits platelet activation and aggregation and enhances the effects of anticoagulants such as protein C and protein S (Suc et al., 1997).

Plasma levels of HDL appear to have a strong inherited basis, with heritability estimates of 40-60% (Lusis, 2012). The most common genetic disorder affecting HDL is familial hypoalphalipoproteinemia, which is a disease caused by a dominant mutation in the *ABCA1* gene in some families and unknown genes in others (Brooks-Wilson et al., 1999). Several monogenic disorders of low HDL have been described, but these are rare and only explain a small portion of low HDL cases. Candidate-gene analysis as well as genome wide association studies have both been performed to identify polygenic causes of low HDL. A variety of genes thought to play a role in HDL metabolism have been tested in association studies, with inconsistent findings due to small sample size. Some examples are *LCAT*, *ABCA1* and several lipoprotein A genes (Pollex & Hegele, 2007).

3.2.4 Triglycerides

TG are mainly carried in VLDL and chylomicrons, while cholesterol is mainly carried in LDL or HDL. In epidemiological studies, TG is associated with increased risk for CAD, but it is difficult to establish a causal inference because these studies also demonstrate that LDL and HDL are associated with the same phenotype and correlations are known to be present between levels of cholesterol and TG (Sarwar et al., 2007). However, using an analytical approach that accounts for pleiotropic effects of a SNP on triglycerides, HDL and LDL, Do et al demonstrate that SNPs with the same direction and magnitude of association for TG and LDL tend to associate with CAD risk and loci that have an exclusive effect on TG are also associated with CAD (Do et al., 2013). Furthermore, the strength of a SNP's effect on TG was correlated with the magnitude of effect on CAD, even after accounting for its effect on LDL/HDL. This demonstrates that plasma TG levels reflect processes that are likely to be causal in CAD. This is consistent with other reports which have identified SNPs related to TG levels, all of which have been related to clinical CAD or subclinical atherosclerosis. These SNPs lie in the promoter of *APOA5* (Pennacchio et al., 2001), upstream of *TRIB1* (Kathiresan et al., 2008) gene and in the *APOC3* gene.(Pollin et al., 2008).

There are several hypotheses that explain how plasma triglycerides are linked to atherosclerosis. Firstly, they may reflect postprandial cholesterol metabolism as they are highly correlated with the amount of cholesterol in remnant lipoproteins. Another process reflected by TG levels may be the activity of lipoprotein lipase, an enzyme that hydrolyses triglycerides TG in lipoproteins. Higher enzymatic activity of lipoprotein lipase leads to lower TG levels, with a gain in function nonsense polymorphism in the *LPL* gene reducing plasma levels and lowering risk for CAD (Varbo, Benn, & Nordestgaard, 2014).

3.3 Methods and Materials

This meta-analysis includes 2 cohorts, the PIVUS cohort and the Uppsala Longitudinal Study of Adult Men (ULSAM cohort). Association tests were performed separately in each cohort for a number of metabolic-related traits using SNPtest. Subsequently, GWAMA was used to perform meta-analysis on the summary statistics (Mägi & Morris, 2010).

3.3.1 Subjects

The PIVUS cohort was recruited by making a register of all adults aged 70 living in the community of Uppsala Sweden. Subjects were then invited in a randomized order by receiving a letter within 2 months of their 70th birthday. A total of 1016 out of 2025 subjects participated, giving a participation rate of 50.1%. For more details, please refer to Chapter 2.

The ULSAM cohort is a study of healthy elderly men in the Uppsala region of Sweden (Lithell et al., 2000). It was initiated as a health screen focused at identifying metabolic risk factors for cardiovascular disease. In 1970, all 50 year old men living in Uppsala were invited to participate. Of these, 82% of them participated initially. The cohort was subsequently invited back at ages 60, 70 and 77. This study included the collection of a wide range of phenotypes, including blood pressure, insulin metabolism, weight and height, lipid markers, diet, cognitive function and socio-economic factors. At age 70, participants were given a glucose tolerance test and insulin clamp to measure insulin resistance, together with memory function tests, dietary record and ambulatory blood pressure monitoring. During this follow up period more than 300 cases of stroke and more than 400 cases of heart attacks have been recorded.

The screening examination included a medical questionnaire with interview, as well as blood and urine sampling, together with blood pressure and anthrometric measurements and electrocardiogram. Anthropometric measurements included height, weight, waist circumference, and skinfold thicknesses (subscapular, triceps, abdominal). A comprehensive two-dimensional and Doppler echocardiography was performed with a Acuson XP124 cardiac ultrasound unit (Acuson, California, USA) with a 2.5 MHz transducer was used for the majority of the 2-D-, M-mode and Doppler examinations. The ULSAM cohort was genotyped with the Illumina 2.5 million HumanOmni SNP BeadChip.

3.3.2 Genotype analysis

In both cohorts, samples were excluded if the call rate was less than 95%, if they had extreme heterozygosity (>3 SD from the mean), if they were ethnic outliers, or if they were gender discordant. SNP quality control measures included exact p-value for deviation from HWE < 10^{-6} and missing genotype rate > 0.01 (MAF < 5%) or missing genotype rate > 0.05 (MAF ≥ 5%).

The Omni 2.5 M chip was used for the analysis of ULSAM samples. 17 individuals were excluded based on call rate, gender discordance and heterozygosity. 22 individuals were excluded as they were PC outliers. 1484 SNPs were excluded due to mapping problems, 646143 SNPs were excluded as they were monomorphic, 17135 SNPs were excluded due to HWE and genotyping criteria. There were a total of 1531196 SNPs and 958 samples in the QCed set.

The Omni-Express chip was used for the analysis of PIVUS samples. 24 individuals were excluded based on call rate, gender discordance and heterozygosity. 22 individuals were excluded based on gender discordance, call rate and heterozygosity. 87884 SNPs were excluded due to HWE and genotyping rate criteria. There were a total of 123771 SNPs and 1179 samples in the QCed set.

Multidimensional scaling (MDS) was performed to obtain principal components to adjust for population structure. SNPs used for MDS had the same quality control filters applied and were LD pruned to r2<0.2 after removal of known regions of LD and large inversions. Prior to imputation, variants with MAF < 1% were removed from the GWAS scaffold. Samples were pre-phased with SHAPEIT2. Genotype data were then imputed up to the "all ancestries" reference panel from the 1000 Genomes Project Consortium Phase 1 interim release (June 2010) (Abecasis et al., 2012) using IMPUTEv2 (B. Howie et al., 2012).

3.3.3 Statistical Analysis

The concentrations of phenotypes were rank inverse normalised to generate a Gaussian distribution for downstream association analyses, and minimise the impact of outliers. Association testing for each transformed level was performed in a linear regression framework under an additive model in the minor allele, two principal components from multidimensional scaling to account for population structure as covariates.

Association testing was performed in SNPTEST, allowing for uncertainty in the imputation in a missing data likelihood. In this study, the association analysis was restricted to SNPs with MAF > 1%, imputation quality information score (info) of greater than 0.4, and HWE exact p

$< 10^{-10}$. The genomic control inflation factor for each phenotype was used to assess evidence of residual population structure (Devlin & Roeder, 1999b). To account for multiple testing, I defined genome-wide significance as $p<1x10^{-8}$, which is the Bonferroni-corrected p value for 5 traits. Conditional analyses to search for secondary signals of association were performed by including genotype dosage at the lead SNP as an additional covariate in the linear regression model, then performing meta-analysis again on the resultant statistic using fixed effects inverse-variance weighted meta-analysis in GWAMA (Mägi & Morris, 2010). Heterogeneity was also calculated for each phenotype, with one degree of freedom since two studies were included in the meta-analysis. No genomic control was performed as the inflation factors were in the region of 0.95 to 1.05 (table 3.1).

3.4 Results

Table 3.1 Lead SNPs significantly associated with phenotypes

Pheno	GC λ	Chr	Coordinates	A1	A2	MAF	beta	SE	pvalue	var_pivus	var_ulsam	i2	effect
TG	1.05	11	116648917	C	G	0.127	-0.357	0.0496	6.98E-13	3.70E-02	2.40E-02	0	--
HDL	1.03	16	56994528	T	C	0.336	0.256	0.0351	3.06E-13	3.00E-02	3.70E-02	0.121	++
D5D	0.999	11	61571382	A	G	0.331	-0.687	0.0387	2.45E-70	2.15E-01	2.31E-01	0	--
ApoB	0.985	19	45412079	T	C	0.0841	-0.633	0.0717	1.18E-18	9.10E-02	3.20E-02	0.811	--
D6D	1.01	11	61620703	C	CTCTT	0.364	-0.671	0.0388	3.50E-67	2.17E-01	2.04E-01	0	--

3.4.1 D6 Desaturase

D6 desaturase levels were associated with SNPs in the *FADS2* locus (Figure 3.1 and 3.2). The

lead SNP was an intronic deletion in the *FADS2* gene, rs138194593 (p value of $3.50x10^{-67}$). It has a minor allele frequency of 0.36. The majority of the SNPs passing genome-wide significance for association with D6 desaturase levels were located in the *FADS2* locus. These enzymes play a crucial role in the endogenous elongation and desaturation of fatty acid precursors into physiologically active long chain PUFAs. When conditioned on the lead SNP, there were no secondary signals.

The lead SNP for D5 desaturase is rs179549 (p value $2.45x10^{-70}$, MAF of 0.33) (Figure 3.3 and 3.4). It is an intronic variant in the *FADS1* gene. When conditioned on the lead SNP, there were no secondary signals. One study suggested that variation in the *FADS* gene block may influence desaturase activity in Caucasian and Asian individuals (Merino et al., 2011). There is a weak correlation between the inverse normalised values of D5 and D6 desaturase, with p value of 0.0336 and pearson correlation coefficient of -0.0712.

The SNP associated with ApoB levels was Rs7412 (p value $1.18x10^{-18}$, MAF 0.08) (Figure 3.5 and 3.6), which lies within the human *ApoE* locus. Conditional analysis revealed no secondary signals. Rs7412 has been linked with ApoB in various studies. In one exome chip study on African American and Hispanic families using linkage analysis, rs7412 was associated with ApoB with a logarithm of odds score of 4.91. The association was particular strong in African-Americans, in which the p value was 4 $x10^{19}$ (MAF of 0.13), and it accounted for 16.2% of variance in in this population (Hellwege et al., 2014). In another GWAS, it was associated with LDL in African Americans, with a p value of 6.3 x 10e-11 (MAF of 0.10) (Rasmussen-Torvik et al., 2012). In a GWAS on Japanese individuals, this SNP was associated with LDL as well as risk of CAD (p values 1.3 × 10-41 and 6.1 x 10-13

respectively, MAF of 0.05) (Takeuchi et al., 2012).

The lead SNP associated with TG was rs964184, with p value of 6.98 x 10^{-13} and a MAF of 0.13 (Figure 3.7 and 3.8). It is located at 11q23.3 in the intronic region of the *ZPR* gene. There were no secondary signals. This SNP is also the lead SNP associated with TG in a large GWAS meta-analysis on Europeans (MAF of 0.13, p value of 7 x 10^{-240}) (Teslovich et al., 2010). This SNP is associated with several phenotypes in different GWAS.

The lead SNP associated with HDL in our study was rs17231506, at the chr 16q21 locus (Figure 3.9 and 3.10). It is an upstream variant in the *CETP* gene and has a meta-analysis p value of 3.06 x 10^{-13}, MAF 0.13. There were no secondary signals. This SNP has been associated with HDL in an association study on European adolescents ($p < 0.0012$) (Legry et al., 2011). The lead SNP for Teslovich et al (2010), a large meta-analysis in Europeans was rs3764261, with MAF of 0.32 and p value of 7 x 10^{-380}. In our study, this SNP had a p value of 3.09 x 10^{-13} and MAF of 0.34. LD between rs3764261 and rs17231506 was was 0.998 in ULSAM and 0.995 in PIVUS.

3.5 Discussion

3.5.1 D5 and D6 Desaturase

The lead SNP for D5D (rs174549) is associated with several phenotypes in 3 different GWAS studies – heart rate (den Hoed et al., 2013), glycerophosphocholine levels (Hong et al., 2013), compressive strength and appendicular lean mass (Y. Han et al., 2012). It is not clear how genetics is linked to compressive strength or bone marrow density. Specific fatty acids have

been demonstrated to prevent bone loss by reducing osteoclastogenesis, suggesting roles for fat metabolism in bone homeostasis (Rahman, Bhattacharya, Banu, Kang, & Fernandes, 2009). Furthermore, a gene expression microarray study showed higher expression levels of FADS in high vs low bone mineral density group (Lei, Papasian, & Deng, 2011). Glycerophosphocholine and other choline compounds have been associated with various diseases, including HIV (Curnow, Mellor, Stephens, Lorch, & Booth, 2005), multiple sclerosis (Kidd, 2005) and cancer (Aboagye & Bhujwalla, 1999).

Fatty acid desaturation is an essential process that generates precursors for the formation of polyunsaturated fatty acids, which influence neurological development, as well as anti-inflammatory and pro-inflammatory eicosanoids, which may play a role in conditions such as cardiovascular disease, cognitive development, eczema and atopic diseases (Caspi et al., 2007; Lattka, Illig, Heinrich, & Koletzko, 2009; Martinelli et al., 2008). Furthermore, there may be ethnic differences in how each allele regulates desaturase activity, which may better inform the prevalence of cardiovascular disease in different populations (Chiu, Austin, Manuel, & Tu, 2010). For example, Merino et al reported that in a Caucasian cohort, desaturase activity was lower among carriers of the minor alleles, whilst in Asians, the minor allele was associated with both increases and decreases in desaturase activity (Merino et al., 2011).

So far, research has demonstrated associations between variants in *FADS* and n-6 fatty acids mainly. However, 2 studies have shown associations with n-3 fatty acids. Firstly, Tanaka et al showed that a locus in *ELOVL2* gene is involved in the homeostasis of longer chain n-3 fatty acids (Tanaka et al., 2009). Secondly, Merino et al demonstrated a relationship between FADS variants and desaturase activity using n-3 fatty acid data ie. eicosapentaenoic acid to

alpha-linolenic acid index. They observed that rs174537 in *FADS1* and rs174576 in *FADS2* were associated with this index (Merino et al., 2011).

Plasma lipid levels are determined by both genetic as well as lifestyle components. Most studies that investigate the former have used samples in homogeneous populations, hence minimising the confounding effects of cultural factors. The difference in results, for example, between Caucasian and Asian populations may reflect dietary factors rather than genetic influences, however, it is difficult to differentiate clearly between the two.

3.5.2 Apolipoprotein B

The lead SNP for ApoB, rs7412, lies within the human ApoE locus, which codes for a polymorphic protein that consists of three haplotypes: ε2, ε3 and ε4. These haplotypes are formed from combinations of rs7412 and rs429358. Several studies have linked lipid and weight related phenotypes to variants in the region.

In various studies, apoE ε2 carriers (minor allele) had increased plasma apoE compared to the other two genotypes (ε3 and ε4). Furthermore, plasma apoE concentration was associated with BMI (Mannila et al., 2013). The ARIC study described BMI as being associated to these isoforms, with carriers of the minor allele having higher plasma apoE concentration than those with other genotypes. This study also showed that minor allele carriers tended to have lower BMI (Volcik et al., 2006). This association was contended by other studies showing association of the minor allele with higher BMI (Duman, Oztürk, Yilmazer, & Hatemi, 2004) or the major allele with decreased BMI (Gottlieb et al., 2004; Izaks et al., 2011). Yet another study showed the ε4 genotype to be significantly associated with obesity (Alvim et al., 2010;

Barrett et al., 2009). These seemingly conflicting results could be due to the confounding effects of plasma lipids when the relationship between apoE and obesity is studied. It is notable that a recent study failed to show significant interaction between the apoE genotype and BMI on serum lipid levels. However, there was increased predictive power of the regression model for LDL when gene-BMI interaction was added (Petkeviciene et al., 2012). Associations between blood lipid concentrations and apoE genetic isoforms have been established historically (Utermann, Pruin, & Steinmetz, 1979), as has the association between these lipids and BMI. Tejedor et al showed that rs7412 had a lipid independent effect on BMI (Tejedor, Garcia-Sobreviela, Ledesma, & Arbones-Mainar, 2014). Controlling for lipids in the regression model may lead to over-adjustment, but this would only be true if the only causal link between apoE and BMI was through blood lipid levels. However, several other biological pathways have been suggested, that are independent of lipids, but instead modulate adipocyte function. These effects have been demonstrated within in vitro and in vivo models (Z. H. Huang, Maeda, & Mazzone, 2011; Loos, 2012).

ApoE circulates on LDL, as a ligand for LDL receptors, which participates in the removal of LDL from circulation. The SNP rs7214 characterises the ε2 allele, which is known to be associated with lower LDL levels (Bennet et al, 2007) and type III hyperlipoproteinemia (Rall et al, 1992). The SNP rs7412 is a coding SNP which changes the amino acid at position 158 in apoE from Arg to Cys. Individuals with this mutation are less efficient at making and transferring VLDLs and chylomicrons from the blood plasma to the liver. In our study, rs7214 was associated with ApoB levels (ApoE was not measured).

The liver produces VLDL, a spherical particle with a core that is rich in cholesterol and triclycerides (M. S. Brown & Goldstein, 1979). Exposure to lipoprotein lipase hydrolyzes the

triglyceride to fatty acids, which can be used by muscles. During this period, some VLDL is also removed from circulation by binding to liver receptors. Over the next few hours, the remaining VLDL is metabolised in the circulation to LDL which can circulate in the bloodstream for several days. LDL levels have been linked to various atherosclerotic diseases, including stroke and CAD (T. Gordon, Kannel, Castelli, & Dawber, 1981).

ApoB is the primary apolipoprotein that aids in the transport of both VLDL and LDL. It is organizing protein of the shell, and is crucial for the formation of the globule which carries fat molecules(Packard et al., 2000).(Packard et al., 2000)(Packard et al., 2000)(Packard et al., 2000)(Packard et al., 2000) ApoB in the LDL acts as a ligand for LDL receptors in cells throughout the body. High levels of ApoB, particularly when associated with increased LDL, can result in the formation of atherosclerosis plaques. Besides apoB, some VLDL and LDL have other apolipoproteins which are not crucial for their synthesis, but instead modulate the metabolism of the fat contents. ApoE and ApoC are examples of such apolipoproteins (Sacks, 2015). They are not uniformly distributed among VLDL and LDL, even though most particles have at least one molecule. They cluster on a certain VLDL and LDL and form a subspecies of apolipoprotein rich particles. In most people when normal blood cholesterol level, 35 to 60% of VLDL has apoE and 40-80% has ApoC (Khoo, Campos, Judge, & Sacks, 1999; C. Zheng, Khoo, Ikewaki, & Sacks, 2007). These two apolipoproteins may have opposing effects on the metabolism of VLDL and LDL, and may influence the risk of cardiac disease in opposite ways (C. Zheng, Khoo, Furtado, & Sacks, 2010).

ApoE is a ligand for the LDL receptor and has high affinity for it. Chylomicrons and VLDL use apoE as the main receptor for removal from the circulation. Animal experiments which involved altering the content or functionality of apoE showed corresponding changes in apoB

lipoprotein metabolism (Aalto-Setälä et al., 1992; Maeda et al., 1994). Treatment which mimicked ApoE reduced cholesterol levels in the bloodstream and reduced the incidence of atherosclerosis in mice that were deficient for LDL. One study demonstrated that the metabolism of apoE, apoC and apoB were linked (Sacks, 2015). Using kinetic methods, they found that when apoB was converted from larger to smaller lipoproteins, apoC content decreases, allowing apoE and apoB to have access to hepatic receptors that clear the lipoprotein from circulation. VLDL that have apoE but not apoC are cleared quickly before they can be metabolised to smaller particles. VLDL that do not have apoE or apoC are mostly converted by lipolysis to LDL.

Importantly, the apoE genotype is also related to response to statin therapy. In a meta-analysis of GWAS on statin therapy, nine SNPs in the apoE region reached genome wide significance for response to LDL therapy (Postmus et al., 2014). The minor allele of the lead SNP, which was a proxy for rs7412, was associated with larger LDL lowering response. This may be because the apoE ε2 protein is associated with greater amount of cholesterol synthesis by the liver, and hence may also predispose to greater inhibition of production when treated with statins. Mega et al demonstrated the minor allele was associated with higher baseline LDL as well as greater response to statins, supporting this hypothesis (Mega, Morrow, Brown, Cannon, & Sabatine, 2009). The authors also performed conditional haplotype analysis and suggested that most of the variation can be attributed to the number of ε2 copies rather than ε4 copies. Furthermore, individuals who are carriers of ε2 are have a greater proportion of synthesised cholesterol compared to dietary cholesterol (J. F. Thompson et al., 2009). Additionally, there may be more IDL particles contributing to apparent LDL in individuals with the ε2 allele. Statins are particularly good at removing such particles through receptor upregulation.

3.5.3 HDL

The lead SNP associated with HDL in our study was rs17231506 at the chr 16q21 locus. It is an upstream variant in the *CETP* gene. CETP promotes the movement of cholesterol from HDL to apoB containing particles, in exchange for triacylglycerols. This enables receptor mediated hepatic uptake of cholesterol, hence decreasing serum cholesterol level. The SNP rs17231506 is in the *CETP* gene promoter region. It is more likely that a SNP in LD with it, rs183130 (r^2=0.96), is the causal allele. When promoter SNPs associated with HDL in a European population were scanned for transcription factor binding sites, the only one that produced a change was rs183130 (J. F. Thompson et al., 2003). Ghosh et al (Ghosh, May, & Kopp, 1998) described that a motif which contained this SNP bound nuclear factor κβ. In addition, Thompson et al also demonstrated that this region binds the Sp1 family of proteins that are important in regulating *CETP* (J. F. Thompson, Lloyd, Lira, & Milos, 2004). Meta-analyses have demonstrated that variants in the *CETP* gene are associated with other phenotypes, including CETP plasma level and activity, LDL and HDL particle size, as well as risk of CAD (Boekholdt et al., 2004). In addition, other variants in the region, for example variable number tandem repeats (VNTRs), have also been found to be associated with HDL (Frisdal et al., 2005). Besides univariate effects, Frisdal et al (2005) also report interactions between SNPs – they demonstrate that some promoter SNPs interact non-additively with other SNPs, a finding that is also observed in other lipid related genes (Hamon et al., 2004).

Low HDL is known to be associated with lower risk of cardiovascular disease in middle aged and elderly individuals, but the relationship between cardiovascular risk and HDL levels is more consistent in the former category than in the latter category (Abbott, Wilson, Kannel, &

Castelli, 1988; Castelli et al., 1977; S. Gordon, Durairaj, Lu, & Davidson, 2010). The inverse relationship between HDL and CAD can be seen in various cohorts, including Framingham, San Francisco, Albany and Evans County (Castelli et al., 1977). The mechanisms behind which HDL acts to influence cardiac risk are complex. Firstly, HDL has been postulated to reduce the formation of atherogenic plaques (Lusis, Fogelman, & Fonarow, 2004; Tall, 2010). Secondly, genetic factors such as variants within the *CETP* gene are thought to play a role in increasing or decreasing CETP expression (Boekholdt et al., 2004). However, it is not obvious whether lack of CETP results in a rise of HDL, mediating the generation of atherosclerotic plaques, or whether CETP reverses cholesterol transport and hence directly reduces the formation of these plaques. This mechanism is important to evaluate because it is relevant for the development of pharmaceutical molecules that may modulate the effect of CETP.

HDL is influenced by multiple environmental factors, including exercise, alcohol and estrogens (Lamon-Fava, 2002). Recent studies have demonstrated various genetic variants that control HDL activity, although these effects can be obscured if sample size is too small (Boekholdt et al., 2004). Most common *CETP* variants have modest effects on CETP activity or mass. For example, for rs183130, one study reported a 10% difference in HDL between individuals with (mean HDL of 46.3 mg/dl) and without the allele (mean HDL of 49.7 mg/dl) (J. F. Thompson, Wood, Pickering, Dechairo, & Hyde, 2007). Environmental factors such as alcohol and gender may have a greater effect. For example, in the ACCESS trial among European individuals, females were found to have significantly higher HDL than males (54 mg/dl vs 44 mg/dl) (J. F. Thompson et al., 2007). Furthermore, males who consumed more than 10 drinks per week had higher HDL compared to those who did not drink (50.7 mg/dl vs 41.9 mg/dl). Interestingly, rare non-functional *CETP* variants have a large impact on HDL,

and they also have a protective influence against disease (Curb et al., 2004). One should be cautious of over interpreting the effect of CETP on HDL and cardiac risk factors when environmental effects are known to be present. For example, high HDL induced by extreme exercise is not always protective of cardiac events (Hartung & Squires, 1982). Because SNPs within the *CETP* gene exist in LD, association analysis can also be performed on haplotype blocks rather than individual SNPs (J. F. Thompson et al., 2007)

3.5.4 Triglycerides

The lead SNP associated with TG is rs964184, which is located in the intronic region of *ZPR1*. *ZPR1* is an essential regulator of cell proliferation, and plays a signal transduction role in a variety of physiological functions (Galcheva-Gargova et al., 1996). One important transcription factor that binds to the promoter of *ZPR1* is the peroxisome proliferator-activator receptor gamma (pparg), which is crucial for insulin sensitivity and may also regulate weight (Mangelsdorf et al., 1995). The promoter region of *ZPR1* is also bound by hepatocyte nuclear factor 4 alpha which activates various genes involved in cholesterol and fatty acid metabolism (Sladek, Zhong, Lai, & Darnell, 1990). *ZPR1* is located upstream of the ApoA5-A-C3-A1 gene complex, and polymorphisms near this complex have been associated with serum triglycerides (Kathiresan et al., 2009; Teslovich et al., 2010), hence the association between rs964184 and triglycerides might be due to LD with functional SNPs in *apoA5*, a gene known to influence the metabolism of chylomicrons and VLDL .

The SNP rs964184 has been associated with type 2 diabetes, fasting plasma glucose and haemoglobin A1C in Japanese individuals (Tokoro et al., 2015). Furthermore, the association between rs964184 and type 2 DM appears to be independent of serum triglycerides when a

multivariate logistic regression was performed. However, when this association was examined in other GWAS on Caucasians, an association was not detected, and this could be due to the lower frequency of the minor allele in Caucasians (Kristiansson et al., 2012; van de Woestijne et al., 2014). For example, the frequency of the CG and GG genotypes were 23.7% and 2.4% in the Caucasian study compared to the Japanese study, in which the genotypes were 39.1% and 7.2% respectively.

Various studies have demonstrated that polymorphisms near the *ApoA5-A4-C3-A1* gene complex are associated with plasma TG levels (Kathiresan et al., 2009; Teslovich et al., 2010). In particular, one study showed that rs964184 was associated with plasma TG, in patients with vascular disease and in patients with one minor allele, this association was modified by BMI (van de Woestijne et al., 2014). However, presence of the allele was not related to risk of new vascular events. Besides an association with TG, rs964184 was also associated with HDL cholesterol, non-HDL cholesterol and apoB. In this population, the average TG levels were higher than in the general population, so LDL cholesterol levels were less reliable in estimating LDL atherosclerotic lipoprotein levels since there is a higher concentration of TG rich articles and this consequently leads to more small dense LDL particles. Hence, non-HDL cholesterol or apoB levels reflect the cardiovascular better (van de Woestijne et al., 2014). This study also demonstrated that carriers of the minor allele were more like to have metabolic syndrome. This could be due to the higher prevalence of metabolic syndrome since all subjects had vascular disease. Since the SNP influences TG and HDL levels while metabolic syndrome involves other risk factors associated with obesity and insulin, the genetic contribution of this SNP may obscure the underlying common basis for metabolic syndrome. However, since differences in plasma TG between patients homozygous for the major allele and those with one minor allele depends on BMI, it would still suggest

that obesity might be the trigger for increased expression of TG. These results might also suggest why certain obese patients develop metabolic syndrome while others are spared due favourable genetic risk factors. This is line with the model that common genetic variants may interact with environmental variables such as diet and obesity to effect the phenotype (Hegele & Pollex, 2009). In that study, rs964184 was not associated with new vascular events. This could be related to sample size, since vascular events are relatively infrequent.

Two studies have shown that rs964184 was found to be associated with lipid fenofibrate response. Aslibekyan showed that there were significant associations for changes in HDL, cholesterol and triglycerides (Aslibekyan et al., 2012). This study is consistent with Brautbar et al, which demonstrated that the same SNP was implicated in improvement in lipid profile after combination therapy (Brautbar et al., 2011).

3.6 Chapter Summary

This study examined a set of metabolic phenotypes and found significant associations in the lipid related traits - D5 and D6 desaturase, as well as HDL and triglycerides. These associations were consistent with that found in other studies. In addition, I identified novel associations associated with apoB. The association signals extended over large genomic regions, so it was not possible to identify causal variants. In the next chapter, I identify Bayesian credible sets which contain SNPs that are most likely to be the causal variants. Credible sets are developed in a Bayesian context, and can be thought of as equivalent to confidence intervals in a frequentist framework. I then describe annotation of these variants with the ENCODE database so as to suggest possible functional pathways through which they might act.

CHAPTER 4 Fine mapping of significant loci and annotation with functional information

4.1 Chapter Overview

4.2 Introduction

4.3 Methods and Materials

4.4 Results

4.5 Discussion

4.6 Chapter Summary

4.1 Chapter Overview

In this chapter, I aim to fine-map loci associated with environmental pollutants and metabolic traits that were identified in Chapters 2 and 3. To do this, I construct credible sets of variants that are most likely to drive the reported association signals. Credible sets are developed in a Bayesian context, and can be thought of as equivalent to confidence intervals in a frequentist framework. I then annotate SNPs within these credible sets with functional elements located in the ENCODE annotation database, as well as expression quantitative trait loci (eQTLs). These include histone modifications, transcription factor binding sites (TFBS) and DNAse hypersensitivity sites. I also use the Sorting Intolerant from Tolerant (SIFT) algorithm to predict the effect of amino acid changes. The results of these fine-mapping analyses can then be used to infer potential biological mechanisms underlying GWAS signals, and candidate effector transcripts through which their effects are mediated.

4.2 Introduction

Although GWASs have identified a multitude of SNPs associated with a very broad range of phenotypes, they do not provide direct evidence for the genes or biological mechanisms through which these SNPs exert their effect. One of the major challenges involved in interpreting GWAS is LD. Typical genotyping platforms measure only a few SNPs within a region of LD, and it is not clear whether the assayed SNPs are the causal variants or whether the nearby SNPs in LD are the variants which influence the phenotype (Manolio, 2013). Although sequencing can be used to assay the region in greater detail, sequence information is insufficient to distinguish between SNPs that are in very strong LD with other, since it does not directly provide functional information about variants (Service et al., 2014). In addition, sequencing is extremely expensive and is hence unlikely to be practical for the analysis of the large sample sizes needed to fine map complex trait loci. In general, variants located in coding regions are easier to assign functional annotation to. However, the majority of lead SNPs from GWAS studies are located within introns and intergenic regions, with the consequence that the underlying mechanism is not obvious (Schadt et al., 2008). These SNPs in non-coding regions may be involved in regulatory mechanisms that contribute to the phenotype. One way of elucidating these mechanisms is to annotate these variants with regulatory information, such as transcription factor binding sites and histone modifications. If an overlap exists, it is possible that these regulatory elements may influence the effect of a SNP on the phenotype, and hence help to explain how an intronic or intergenic variant causes a biological effect.

4.2.1 Credible Sets

As a result of extensive LD in many loci identified for complex traits through GWAS, the lead SNP may not be the causal variant, and hence may not fully represent the association signal. As such, I used a Bayesian modelling approach to construct credible sets of variants

that are most likely to be driving an association signal and annotated those instead of the lead SNP alone. Assuming that the causal variant has been tested for association, the posterior probability that it will be contained in the 99% credible set is 0.99. (Maller et al., 2012).

This approach takes the beta and standard error from GWAS or meta-analysis, and calculates a Bayes factor, which measures the relative evidence in favour of association at each SNP (Wakefield, 2007). Further details are provided in the materials and methods section. Calculation of the Bayes' factor requires specification of a prior distribution of effect sizes. I used a $N(0,0.2^2)$ prior, which gives high probability to small effect sizes and small probability to large effect sizes (Shungin et al., 2015). A posterior probability (PP) that each SNP drives the association signal can then be calculated by considering the relative Bayes' factors across the locus. The PP also requires specification of a prior probability that each variant is driving the association signal, which here is assumed to be uniform across the locus.

These PPs can be compared across the region and I can select the smallest set of SNPs that account for a fixed percentage of the posterior probability. In this way, the credible set could be interpreted in a similar way to the confidence interval used a frequentist context (Maller et al., 2012). The credible set approach can be used for fine mapping SNPs in a region since I can be 99% confident that the SNP driving the association signal will lie within the credible set. If the credible set contains a small number of SNPs (say less than 20), this fine-mapping approach has been successful and informative in narrowing down the likely causal set of SNPs (Maller et al., 2012). One possible disadvantage is that in the context of meta-analysis, the information for each variant is not equivalent across the locus. This is because data for some variants may not be reported from some studies if they fail quality control, hence lower quality variants may not be represented in the credible set. This problem may be alleviated by imputation, but quality of imputation can still vary depending on other factors such as allele

frequency, LD and the population being studied. This approach has been successfully used in a study on type 2 diabetes, coronary artery disease and Grave's Disease (Maller et al., 2012). The authors demonstrated that in 3 of the 14 identified regions, much of the posterior probability rested on a single SNP and in 4 other regions, the credible sets contained less than 34 SNPs, hence adding fine mapping information to the study.

With the credible sets constructed from our data, I then proceeded to annotate them with functional information from eQTL databases, as well as information from the ENCODE database, including histone modifications, TFBS and DNAse hypersensitivity.

4.2.2 eQTLs

Nicolae et al report that SNPs reproducibly associated with complex traits are statistically more likely to be eQTLs, even though few are attributed to nonsense or missense variants (Nicolae et al., 2010). The same study demonstrated that some complex disorders have susceptibility loci that are more easily discovered by annotating SNPs with eQTL scores. The inclusion of expression information in GWAS analysis contributes in two main ways. First, by characterising a greater number of contributing loci, it may be possible to better discern crucial biological functions affected by genetic risk factors. Second, SNPs may be better characterised and localised within the target gene. For example, in the same study, the authors demonstrated that an intronic SNP, rs3129934, on chromosome 6, acted as a *cis*-eQTL for a number of human leukocyte (HLA) transcripts, and also showed the strongest associated with type 1 diabetes in the MHC region in one GWAS (W. T. C. C. Consortium, 2007) and a lead SNP for multiple sclerosis in another GWAS (Comabella et al., 2008).

A variety of issues should be considered when examining studies that quantify the enrichment of lead SNPs with eQTLs. Firstly, the range for MAF of SNPs associated with complex human traits is higher than the overall MAF of most GWAS platforms because of increased

power of detection with a higher MAF (Nicolae et al., 2010). One factor to consider is the correlation among SNPs due to LD. It has been shown that failure to account for LD affects the variance of the estimate but not the actual estimate of proportions of SNPs that are eQTLs (Nicolae et al., 2010). I have dealt with this issue by creating credible sets, which contain SNPs that are most likely to be causal in nature. There could also be more subtle biases related to LD, with true associations being detected more easily in regions of high LD because these regions are more like to have good coverage on GWAS platforms.

Studies of eQTLs also face challenges that should be taken into account when enriching GWAS results. Firstly, the eQTLs identified in gene expression studies are likely to be a small subset of true eQTLs that could be identified using more sensitive techniques such as RNA-seq. Secondly, eQTLs fall into two categories – cell/tissue specific and those which are shared across tissues. Consequently, it may not be possible to reliably enrich GWAS studies with eQTLs that were detected in a different cell type (Bullaughey, Chavarria, Coop, & Gilad, 2009; Cowley et al., 2009). While it is important to have information on gene expression in the relevant tissue for the disease, it is not always clear what that tissue is because some diseases may be multi-systemic. Further to this, gene expression within lymphoblastoid cell lines may be prone to biological artefacts such as adenosine triphosphate (ATP) levels and ebstein-barr virus (EBV) levels, in which case they might be less suitable for eQTL studies (Choy et al., 2008). It would be ideal if the eQTL study was performed on the same cohort of individuals and the same tissue in which the GWAS was undertaken, but this is often not feasible due to logistics and cost. As such, there are useful public databases which can be interrogated to examine overlaps between GWAS data and eQTLs (G. Consortium, 2013)

4.2.3 Transcription Factor Binding Sites

Another relevant functional annotation for SNPs is overlap with presence of TFBS. These are regions in which transcription factors are known to bind and exert regulatory influence. Transcription factor binding patterns vary between tissue types. In the same way as for expression, one concern is that cell lines from which functional information is obtained do not necessarily correspond to the tissue type most relevant to the phenotype of interest. One way to resolve this problem might be to use functional data across all tissue types, therefore favouring sensitivity at the cost of specificity. An example of this is the functional characterisation of rs2074238, a SNP associated with long QT syndrome (Harismendy et al., 2011). Chromatin Immunoprecipitation sequencing (ChIP-seq) identified the binding of estrogen receptor alpha at this SNP within an epithelial cell line (Schaub et al., 2012). This makes biological sense because long QT syndrome is more common in women and the QT interval has been shown to vary with the menstrual cycle (Kadish et al., 2004; Locati et al., 1998; Nakagawa et al., 2006). Furthermore, estrogen therapy alters the QT interval in postmenopausal women who take it (Kadish et al., 2004). If only functional data in cardiac tissue had been used, this association would have been missed.

The ChIP-seq data generated in public databases such as ENCODE (ENCODE, 2011) so far are relatively sparse. However, the number of datasets available is growing rapidly, and it may be possible to increase specificity by examining datasets relevant for the phenotype in the study. One challenge is to identify specific tissue types relevant for the phenotype. For example, rs3764650 is a functional SNP identified in Alzheimer's disease (Schaub et al., 2012). It overlaps the transcription factor binding site for *HNF4A*, which is a transcription factor known mainly to play a role in liver pathology.

4.2.4 Histone Modifications

Post-translational modifications (PTMs) of histones have emerged as prominent mechanisms by which gene expression can be controlled epigenetically (Allfrey, Faulkner, & Mirsky, 1964). Modifications of basic histone amino acid tails can affect inter-nucleosomal interactions, hence altering the overall structure of chromatin. Besides structural change, modifications also recruit remodelling enzymes that use energy from the hydrolysis of ATP to reposition nucleosomes, hence altering structure yet further.

There are different categories of histone modification, and these affect gene expression in a myriad of ways. Acetylation of lysine is regulated by the antagonistic action of two enzymes – histone deacetylases (HDACs) and histone acetyltransferases (HATs) (X. J. Yang & Seto, 2007). HATs catalyse the transfer of an acetyl group to the lysine side chain and hence neutralise the positive charge and weakens the interaction between histones and DNA. Other than structural alterations, histone modifications may also regulate gene expression through chromatin associated factors that interact with modified histones through distinct domains. This is especially relevant for lysine methylation, since there are more distinct domain types that recognise lysine modification compared to other types of modifications. Some notable examples are plant homeodomain (*PHD*) fingers and Tudor domains, including malignant brain tumour (*MBT*) and Pro-Trp-Trp-Pro (*PWWP*) domains (Champagne & Kutateladze, 2009; J. Kim et al., 2006; Maurer-Stroh et al., 2003).

Within a certain group of methyl-lysine binders, a variety of domains can recognise the same histone modification and the recruitment of factors can either be direct or indirect. An example of indirect recruitment is when H3K4me3, an activating mark, is recognised by a PHD finger, which then recruits other chromatin modifiers such as HDACs and HATs that bind onto the domain (Champagne & Kutateladze, 2009). H3K4me3 is also bound by CHD1, an ATP dependent remodelling enzyme that repositions nucleosomes, which is an example of

direct recruitment (Sims et al., 2005). Lysine methylated in a different position can result in repressive rather than activating effects, as in the case of H3K9me3, which suppresses gene expression. H3K9me3 recognises *HP1*, which binds via its N-terminal chromodomain and may result in chromatin compaction (Bannister et al., 2001). Besides binding chromatin domains, histone modifications can also exert their effects through disruption of interaction between the histone and chromatin factor. For example, H3K4me3 can prevent the NuRD complex from binding to the N-terminal tail (Zegerman, Canas, Pappin, & Kouzarides, 2002). The NuRD complex is a transcriptional repressor, whereby prevention of binding would result in gene activation. Another example is DNMT3L, which also represses gene expression and is disrupted from binding to the histone via H3K4me3 (Adams-Cioaba & Min, 2009).

Histone modifications can also interact with each other through various cross-talk mechanisms. For example, there may be antagonism between modifications that have opposing effects on gene expression. Another way by which they may interact is through dependency, with one modification being reliant on another for its effect to be exerted. For, example, methylation of H3K4 in yeast by scCOMPASS is dependent on the ubiquitylation of H2BK123 by ScRad6/Bre1 (J. S. Lee et al., 2007). This mechanism is conserved in humans. Additionally, the binding of a protein to a particular modification can also be disrupted by a modification beside it. For example although HP1 binds to H3K9me3, its binding is disrupted during mitosis by phosphorylation of H3S10 (Fischle et al., 2005). Besides disruption, there may also be cooperation between modifications so as to recruit chromatin binding factors more effectively. For example, PHF8 binds to H3K4me3 via its PHD finger, which is acetylated on the same tail of the histone (Vermeulen et al., 2010).

Histone modifications can also interact with DNA methylation. For example, UHRF1 binds to nucleosomes which are marked by H3K9me3, and this binding is enhanced when the CpGs

of the adjacent DNA is methylated (Bartke et al., 2010). The opposite mechanism also exists, for example, KDM2A only binds nucleosomes with H3K9me3 when neighbouring DNA is unmethylated (Bartke et al., 2010).

Histone modifications are more commonly located at specific points in the human genome. Most of the human genome can generally be classified as either relaxed (euchromatin) or compact (heterochromatin). Heterochromatin can be classified into two regions: facultative and constitutive. The former contains genes that are differentially expressed through development, and subsequently become silenced. This region may contain genes that are marked by H3K27me3 colocalising with polycomb repressor complexes. The H3K27 methyltransferase enzyme EZH2 resides within the PRC2 complex (Hansen et al., 2008). H3K27me3 recruits the latter to sites of DNA replication, and this helps to maintain this histone modification through the action of the former. On the other hand, constitutive heterochromatin consists mainly of permanently silenced genes in telomeres and centromeres and is often marked by H3K9me3 and HP1, with redistribution of the modified histones into the replicated chromatin after DNA replication (Trojer & Reinberg, 2007). Euchromatin, the relaxed region of the genome which contains active genes, is enriched with H3K4me1 at active transcriptional enhancers (Hon, Hawkins, & Ren, 2009). On the other hand, the active genes themselves are enriched by H3K4me3 and H3K36me3 (Schneider et al., 2004). The areas of demarcation between these heterochromatin and euchromatin, otherwise known as boundary elements, are associated with factors such as CTCF, which play a role in maintaining this boundary between different chromatin environments (Ong & Corces, 2009). Boundary elements are enriched for certain histone modifications, such as H3K9me1, and do not contain histone acetylation.

Histone modifications feature in the pathogenesis of a variety of diseases. In particular, cancer is a disease in which histone modifications are being actively considered while designing the therapeutic regime. The role that histone modifications play in carcinogenesis can broadly be divided into: a) altering expression of genes; and b) disrupting genomic integrity or chromosomal segregation. An example of the former involves H3K27me3 and and EZH2 in humans. Gene silencing is caused by overexpression of EZH2, which lays down the H3K27me3 mark. This results in the progression of several solid malignancies, including bladder, prostate and breast cancer (Kleer et al., 2003; Varambally et al., 2002; Weikert et al., 2005). However, the opposite trend is true for haematological malignancies, in which EZH2 has been suggested to be a tumour suppressor protein (T. Ernst et al., 2010). One possible explanation for this discrepancy is that the levels of H3K27me3 may need to be regulated judiciously within cells, with deviations in either directions being promotional of cancer development. Histone modifications have also been linked to genomic instability and subsequently cancer. For example, mice that are deficient for H3K9methyltransferase show lower levels of H3K9me2/3 and have an increased rate of malignancies (Oda et al., 2009). Histone modifications are particularly interesting in the context of cancer because they are potentially reversible, so epigenetic drugs may reverse these changes. HDAC inhibitors have been demonstrated to be effective in repressing tumour growth, inducing differentiation of cells and promoting tumour apoptosis (Sharma, Groselj, Hamdy, & Kiltie, 2013). They have been approved for use in T cell lymphoma, with other related compounds being in clinical trials. Other epigenetic drugs such as histone methyltransferase inhibitors are currently in the development pipeline (Best & Carey, 2010). There are various issues with epigenetic drugs in the context of cancer. Firstly, it is not clear whether the drugs exert their effects through modulation of histone acetylation or acetylation of other substrates. Secondly, majority of HDAC inhibitors are not specific to particular enzymes, but rather, target a broad range. It is

not clear whether it would be advantageous to develop inhibitors specific to certain HDACs or whether subfamily specific inhibitors should be developed, which may reduce unwanted side effects.

4.2.5 DNAse hypersensitivity sites

DNase I hypersensitive sites signify regions of open chromatin indicated by tag-enriched sites from Dnase-seq experiments. These profiles may be somewhat similar to ChIP-seq experimental profiles. However, there are some important differences. Firstly, ChIP data consists of binary states – either bound or unbound. However, Dnase I is a generic indicator of chromatin state that allows multiple states of chromatin accessibility to be assayed (Shu, Chen, Bo, & Wang, 2011). Secondly, ChIP-seq can employ either tag shifting or tag extension for peak detection, both of which do not reveal the exact site of protein-DNA binding (Madrigal & Krajewski, 2012). Thirdly, ChIP-seq peaks are usually relatively well defined, while DNAse hypersensitivity sites are less visually obvious because of tag enrichment over broad ranges of the genome (Madrigal & Krajewski, 2012). There are various factors that can alter Dnase patterns. Firstly, fold enrichment provides an indication of how accessible the region is, and secondly, depth of footprint provides information on how protected the transcription-factor bound sequence is (Madrigal & Krajewski, 2012).

DNAse Hypersensitivity sites, histone marks and transcription factor binding sites are available from the Encyclopaedia of DNA elements (ENCODE) database, which I used for annotating the variants.

4.2.6 ENCODE

The ENCODE Project was launched in September 2003 by an international consortium of scientists with the aim of identifying functional elements in the human genome sequence,

including coding as well as non-coding transcripts, marks of accessible chromatin and protein-binding sites (ENCODE, 2011). It initially focused on 1% of the genome from 2003 to 2007, and then subsequently expanded to include results from the entire genome. Alongside this, the GENCODE collaboration was also initiated, with the aim of annotating all evidence-based gene features in the human genome with high accuracy, using a combination of experimental validation and mathematical analysis(Harrow et al., 2006). The ENCODE project consisted of three stages – pilot phase (which evaluated strategies to identify genomic elements), development phase (which devised laboratory and computational methods) and production phase (which carried out the annotation). As a result of these efforts, 80% of the human genome is now associated with at least one biochemical function (ENCODE, 2011). A total of 147 cell types have been assayed with a variety of experimental techniques, and almost 3 million Dnase hypersensitive sites have been identified with chromatin accessibility studies. Furthermore, ChIP-seq experiments have been conducted for more than 100 transcription factors to identify binding sites.

Nicolae et al demonstrated that majority of known GWAS associations overlap a functional region or are in strong LD with SNPs that overlap functional regions (Nicolae et al., 2010) such as transcription factor binding sites and histone modifications. Furthermore, they report that for most associations, the SNP deemed to be most functional based on ENCODE data is usually a SNP in LD with the SNP revealed by GWAS to have the lowest p value of association with the phenotype, not the latter lead SNP itself. The same study showed that there was significant enrichment of regulatory function in disease associated regions, and a combination of multiple sources of evidence led to stronger enrichment scores. They used information from RegulomeDB, which is a databases designed for fast annotation of SNPs, which combines different datasets from ENCODE, including DNAse hypersensitivity, ChIP seq as well as eQTLs (Boyle et al., 2012). The first part of the study used data generated by

the ENCODE consortium to identify functional SNPs associated with a phenotype, either through direct association by GWAS or through LD with the top SNP identified by GWAS. Using this method alone identified 81% of functional SNPs when population of European descent, using the CEU population to obtain for LD information (Nicolae et al., 2010). The fraction of associated SNPS for which functional annotation could be provided was similar to that reported in another study. Their results were also consistent with chromatin state-based methods, in which segmentation, a machine learning method, was used to identify disease associations in predicted enhancers by predicting chromatin at every position in the genome based on histone information and then comparing these with GWAS results. Most existing regions for prioritising SNPs are largely based on whether a region is transcribed or not. For methods which use regulatory regions, most approaches are based on motif information and approaches that use experimental data are mainly limited to individual associations.

4.2.7 ANNOVAR

I used Annotate Variants (ANNOVAR) as a tool for the annotation of variants with the ENCODE database. ANNOVAR is a pipeline which can assign functional annotation of SNPs, most commonly GWAS lead SNPs (Kumar, Henikoff, & Ng, 2009; K. Wang et al., 2010). ENCODE can be used as a database for annotating these SNPs, both in terms of their relationship to genes, as well as their relationship to other functional elements such as transcription binding sites, histone modifications and microRNA target sites. This is particularly useful in whole genome sequencing, since a large proportion of the variants identified are outside the protein coding regions. ANNOVAR can evaluate variants which are not reported in public databases like dbSNP, making it very relevant for exome-chip sequencing. It is a command-line driven software tool based on Perl, and is freely available at http://www.openbioinformatics.org/annovar/. To perform the annotation, the user needs to

download annotation data sets, some of which include RefSeq genes, Ensembl genes and RefSeq genes. Different genome builds can be specified, including both hg18 and hg19, just as different species genomes can be selected, including mouse, human, rat and cow. For region based annotation with histone modifications and regions of DNAse hypersensitivity, ANNOVAR needs to download annotation databases from UCSC genome browser. ANNOVAR can also download SIFT scores for non-synonymous mutations, to help assign functional predictions to SNPs (Kumar et al., 2009).

To place a variant with respect to genes, ANNOVAR scans the annotation database and identifies introns, exons, splice site variants and upstream/downstream variants, which are defined as variants less than 1kb away from the gene. For exonic variants, ANNOVAR also reports amino acid changes after scanning mRNA sequences. For intergenic variants, ANNOVAR reports the closest two genes and distances away from them.

4.2.8 Sorting Intolerant from Tolerant (SIFT)

For non-synonymous SNPs, I used the SIFT algorithm to predict the effect of the amino change (Ng & Henikoff, 2002, 2003). Currently, amino acid substitutions account for half of the known gene lesions that cause inheritable diseases (Krawczak et al., 2000). The SIFT algorithm uses sequence homology to predict whether a substitution will affect protein function. It has been applied to human databases of SNPs and was successful in distinguishing mutations involved in disease from polymorphisms that do not have a deleterious effect (Ng & Henikoff, 2002). When applied to a database of missense mutations involved in disease mechanisms, SIFT successfully predicted 69% of the variants as damaging (Ng & Henikoff, 2002). When applied to non-synonymous SNPs in dbSNP, 25% of the variants were predicted to be deleterious. SIFT uses sequence to predict function, unlike other tools which use structure. Since it is not limited by structural information, a

larger number of substitutions can be predicted. Of the non-synonymous SNPs identified in previous studies, 74% were similar to homologs in protein sequence databases, allowing for accurate SIFT prediction (Ng & Henikoff, 2002).

The SIFT algorithm, as implemented by the webserver (http://provean.jcvi.org), was also used to predict the effect of non-synonymous variants within the credible sets (Ng & Henikoff, 2003). The algorithm assumes that important amino acids will be conserved in the protein family, so that changes at conserved positions tend to be predicted as deleterious. If a position in an alignment contains the hydrophobic amino acids isoleucine, valine and leucine, SIFT assumes that this position should only contain amino acids with hydrophobic characteristics. Hence substitutions with other hydrophobic amino acids are usually predicted as tolerated while substitutions with polar or charged amino acids will be predicted as deleterious. SIFT also considers the position at which the change occurs and the type of change. It calculates the probability that an amino acid at a position is tolerated conditional on the most frequent amino acid being tolerated. If this probability falls below a cut-off, the substitution is predicted to be deleterious.

4.3 Methods and Materials

To construct credible sets using the approach set out above, beta and standard errors are required from GWAS (in the case of pollutants) and meta-analysis (in the case of metabolic traits). With the beta and standard errors for each SNP, I proceeded to fine map each locus by calculating a 'credible' set of variants based on Bayes factors and posterior probabilities (Maller et al., 2012). A locus was defined as a specific number of 500 base pairs upstream

and downstream from a lead variant. An approximate Bayes' factor was calculated for each variant in the locus based on the beta and standard error from analysis (Wakefield, 2007).

$$\mathrm{BF}_j = \frac{\sqrt{1-R_j}}{\exp\left(-\frac{R_j\beta_j^2}{2\sigma_j^2}\right)}$$

in which β_j is the allelic effect of the jth SNP, with corresponding standard error of σ_j and $R_j=0.04/(\sigma_j^2+0.01)$. This incorporates a $N(0,0.2^2)$ prior for b_j which gives high probability to small effect sizes and small probability to large effect sizes (Shungin et al., 2015).

The posterior probabilities for each SNP driving the association signal at the locus were then calculated by dividing the Bayes factor of each variant by the sum of all Bayes factors across the locus. After that, the variants were ranked in descending order based on their Bayes factor, and a cumulative probability was calculated for each variant in the ranked list. Variants were then added to the credible set moving down the list until the cumulative probability added up to 99%.

The credible sets were then annotated with the ENCODE database using the ANNOVAR software (ENCODE, 2011; K. Wang et al., 2010). Histone modifications, TFBS, and DNAse hypersensitivity sites were first downloaded from the UCSC website (C. K. Wong et al., 2013). In addition, I checked for overlap between SNPs in the credible sets and variants in the NHGRI as well as in the National Centre for Biotechnology Information (NCBI) eQTL browser (http://www.ncbi.nlm.nih.gov/projects/gap/eqtl/index.cgi).

4.4 Results

4.4.1 Credible Sets

I considered 99% credible sets with fewer than 30 variants to be informative for fine-mapping. At this threshold, credible sets for eight association signals were identified (Table 4.1 and Appendix Table 1). For two of the association signals, the lead SNP alone was included in the credible set: rs13107325 for manganese and rs7412 for ApoB. For six of the loci, there were less than 30 variants in the credible set. However, for three of the phenotypes (TG, mercury and cadmium), the credible sets were too large to be informative (4009, 2860, 1884 respectively).

Table 4.1 Credible Sets for each of the phenotypes

Phenotype	Number of Variants	Size of credible set (number of base pairs)	Chr	Start	End
D6_desaturase	27	69629	11	61551927	61621556
D5_desaturase	6	14955	11	61567753	61582758
ApoB	1	1	19	45412079	45412079
HDL	11	7513	16	56987015	56994528
Triglycerides	4009	998245	11	116150361	117148606
PCB99	19	38060	19	41485194	41523254
Manganese chr4q24 locus	1	1	4	103188709	103188709
Manganese chr1q21 locus	11	32971	1	220072698	220105669
OCDD	8	11033	4	89043634	89054667
Mercury	2860	999471	3	176849740	177849211
Cadmium	1884	999577	6	73958002	74957579

4.4.2 Annotation with ENCODE

Annotation with ENCODE showed the credible sets that contain *FADS1* and *FADS2* (associated with the serum levels of D6 and D5 desaturase) are commonly associated with various different histone modifications, including H3K27ac, H3K36me3, H3K9ac,

H3K4me3, H3K4me1 and H3K27me3. The CETP locus associated with HDL was also marked by H3K27me3. Variants within the credible sets for PCB99, OCDD and manganese were marked by H3K27me3. Variants within the credible set for manganese were marked by H3K36me3 (Appendix Table 2a).

The credible sets of several phenotypes overlapped with TFBS binding sites (Appendix Table 2b). Variants associated with D5 and D6 desaturase overlapped binding sites of RNA polymerase II, STAT1 and SREBP1. Variants associated with HDL overlapped the binding site for CTCF. Variants associated with manganese overlapped RNA polymerase II binding site and variants associated with OCDD overlapped with FOXA1 binding site. Variants in the credible sets of OCDD and PCB99, D6 desaturase and HDL were associated with sites of DNAse hypersensitivity (Appendix Table 2c). Annotation with the eQTL database revealed no eQTLs among the credible sets enclosing SNPs that passed genome wide significance.

SIFT identified the non-synonymous results from our GWAS study and meta-analysis, namely rs13107325 (associated with manganese), rs2231142 (associated with OCDD) and predicted both to be tolerated. These are the lead SNPs for each locus and the posterior probabilities are 0.920 and 0.142 respectively. For rs13107325, the amino acid change is alanine (nonpolar) to threonine (polar) within the *SLC39A8* gene. For rs2231142, the amino acid change is glutamine (nonpolar) to lysine (polar and positively charged) within the *ABCG2* gene. Furthermore rs2231142 has been shown to affect activity of *ABCG2* (Keskitalo, Pasanen, Neuvonen, & Niemi, 2009). 94 GWAS lead SNPs were identified, which overlapped the SNPs in the credible sets (Appendix Table 3).

4.5 Discussion

4.5.1 Credible Sets

The calculated PP in constructing credible sets depends on several assumptions, for example, that there is a single causal SNP at a locus that has been genotyped in the study. This may not be the case for all the phenotypes. Maller et al demonstrated that this assumption failed in some of their regions, for example in the *CDKN2A-B* region for type 2 diabetes (Maller et al., 2012). Allelic heterogeneity may cause these credible sets to be overly large and this problem could be overcome by conditional analysis. The authors also found that weaker GWAS signals did not seem to be driven by poor tagging of a variant with larger effect size, but rather smaller true effects. Their experiment was powered to find larger effect variants that were tagged or directly genotyped well, but such variants were not identified in regions with relatively small effect in the original GWAS. They found evidence for the involvement of multiple SNPs in several regions, suggesting that allelic heterogeneity is relatively common at GWAS loci, which is consistent with our findings of multiple association signals on conditional analysis.

The distributions of the posterior probabilities are different for the various phenotypes. For example, in the case of D5D, D6D and HDL, each variant contributes small percentages of the credibility (Figures 4.1, 4.2 and 4.3). In the case of TG, the top variant contributes approximately 0.4 and the subsequent 4008 variants each contribute small percentages of the credibility (Figure 4.4). These findings should be interpreted in relation to the subsequent annotation. Overlap of functional elements such as histone modifications or TFBS sites with

SNPs with large posterior probabilities is more meaningful than overlap with SNPs that have small posterior probabilities.

4.5.2 Histone Modifications

The credible set variants in *FADS1* and *FADS*2 were commonly associated with various different histone modifications, including H3K27ac, H3K36me3, H3K9ac, H3K4me3 and H3K4me1. These confer various activating and repressing marks on the genes in the locus, with activity that can be fine-tuned epigenetically and can involve complex interplay between different modifications. For example, H3K27ac is normally associated with enhancer activity, but it can also distinguish active enhancers from poised enhancers that are enriched with only H3K4me1 (Creyghton et al., 2010). These poised enhancer networks are reset during nuclear reprogramming, and play a crucial role in developmental programmes.

H3K4me3 is a histone mark mainly found at gene promoters. It can be found at both active and inactive promoters, and its level depends on gene activity (Guenther, Levine, Boyer, Jaenisch, & Young, 2007). There is a close association between sites marked by H3K4me3 and CpG islands, which are areas of the genome enriched by GC base content and high CpG content (Akhtar, Scala, Cocozza, Miele, & Monticelli, 2013). The majority of gene promoters in humans are associated with CpG islands. Significant fraction of all CG islands in humans and mice are not associated with transcription start sites but still exhibit the characteristics of promoters, including H3K4me3 marks. The insertion of GC rich DNA without a promoter is sufficient in creating a novel domain of H3K4me3, which may suggest that a key function of CG islands is to recruit H3K4me3 (Mendenhall et al., 2010). While the mechanisms through which histone marks interact with SNPs is not clear, it is possible that genetic variants such as SNPs may influence the degree and type of histone modification (McVicker et al., 2013), which could then activate or repress gene expression.

4.5.3 Transcription factors

Various phenotypes have credible sets which contain variants that overlap TFBS binding sites. A SNP in the credible set of OCDD located in chr4q22 overlaps a FOXA1 binding site. FoxA1 is a member of the FOX family of transcription factors, which has more than 40 members (Hannenhalli & Kaestner, 2009). It was the first of the FOX family to be discovered. The FOXA subfamily, which consists of FoxA1, FoxA2 and FoxA3, have been found to regulate a large number of genes for developmental specification in a large number of tissues (Kaestner, 2010). FOXAs contain an amino acid DNA-binding domain and a winged helix domain that is conserved within the FOXA family (E. Lai, Prezioso, Tao, Chen, & Darnell, 1991). It lacks the basic amino acids needed for chromatin compaction, hence it binds to nucleosomes and induces an open chromatin configuration that allows it to recruit other transcriptional regulators (Sérandour et al., 2011). FOXA1 binding to DNA has been shown to precede the loss of cytosine methylation and the dimethylation of H3K4 during the differentiation of pluripotent cells (Cirillo et al., 1998). It plays an important role in the development of pancreas and kidney, as well as the liver. It is normally expressed in pancreatic islet cells that produce glucagon (Behr et al., 2004; Gao et al., 2008). Loss of FoxA1 reduces glucagon transcription and secretion. Additionally, it modulates kidney function as mice without FOXA1 are dehydrated since they cannot produce anti-diuretic hormone (Behr et al., 2004). FoxA1 and FoxA2 play a combined role in liver cell specification and differentiation. Although dioxin and dioxin-like compounds are known to cause systemic metabolic dysfunctions and liver enzyme derangements (L. Zhang et al., 2015), it is unclear how OCDD is linked to dioxin as the exact role of the liver in dioxin metabolism is not clearly understood.

The credible set of TG located at 11q23.3 overlaps the binding site of CCCTC-binding factor (CTCF). CTCF is a transcription factor responsible for the bridging the gap between gene expression and organisation of the nucleus (Baniahmad, Steiner, Köhne, & Renkawitz, 1990). It can either activate or repress gene expression. It has insulating properties that allow it to interfere with enhancer-promoter communication and buffer transgenes from the effects of heterochromatin spreading (Ong & Corces, 2014). In addition, CTCF also mediates interchromosomal and intrachromosomal interactions between distant sites in the genome (Zhao et al., 2006). CTCF has a highly conserved DNA-binding domain with 11 zinc-fingers, and is normally located in linker regions surrounded by nucleosomes (Ohlsson, Renkawitz, & Lobanenkov, 2001). 30-60% of CTCF-binding sites have a distribution that is cell-type specific (Cuddapah et al., 2009). Since enhancer elements are enriched for CTCF, this suggests that some CTCF-binding sites may be important in regulating transcription of cell-lineage-specific programmes. Half of CTCF-binding sites are found in intergenic regions, with 40% located within intragenic regions and 15% located near promoters (Chen, Tian, Shu, Bo, & Wang, 2012). CpGs are present in the DNA consensus sequence of the CTCF-binding site, supporting the proposition that the methylation of cytosine residues at the 5th carbon of the base may underlie CTCF target selectivity in different cell types (Engel, West, Felsenfeld, & Bartolomei, 2004). Of note, DNA methylation has an extensive role in regulating CTCF occupancy at a number of genes, including brain derived neurotropic factor, *BCL6* and *CDKN2A* (A. Y. Lai et al., 2010; Rodriguez et al., 2010). One study mapped CTCF occupancy and compared that with DNA methylation data in the same cell lines, demonstrating that 41% of cell-type specific CTCF-binding sites were linked to differential DNA methylation (H. Wang et al., 2012). CTCF can affect the methylation status of DNA by forming a complex with poly(ADP-ribose) polymerase 1 (PARP1) and DNA-

methyltransferase 1 (DNMT1). CTCF activates PARP, which inactivates DNMT1 and maintains methyl-free CpGs (Guastafierro et al., 2008; Zampieri et al., 2012).

The CTCF gene has been shown to play a role in adipocyte differentiation as the CTCF chromatin-binding landscape comprises of a large proportion of sites with highly dynamic binding patterns during adipogenesis (Dubois-Chevalier et al., 2014). A subset of the dynamic CTCF binding sites are gained at cell-type specific regulatory regions, which is consistent with a requirement for CTCF in transcriptional induction of adipocyte differentiation, for example, CTCF is required for the transcriptional activation of nuclear receptor peroxisome proliferator-activated receptor gamma (PPARG) as well as the target genes (Cristancho & Lazar, 2011). During adipogenesis, pre-adipocytes are converted into mature adipocytes, and this process requires the activation of PPARG, which is crucial in he maintenance of mature adipocyte functions such as lipid handling and storage. PPARG is regulated by various cooperating factors, including Retinoid X receptor, CREB binding protein and members of the CEBP family (Haakonsson, Stahl Madsen, Nielsen, Sandelin, & Mandrup, 2013). Some of these processes involve nucleosome eviction and changes in histone modifications. On the other hand, the role of CTCF in the complex regulatory network influencing adipocyte differentiation is less well understood, although one study has established that it interacts with TET methylcytosine dioxygenase and promotes adipogenic transcriptional enhancer DNA hydroxymethylation (Fujiki et al., 2013). While this information is interesting, it does not pinpoint the mechanism by which CTCF may influence the genotype-phenotype relationship, hence further investigation is needed.

STAT1 is a transcription factor which was found to be colocalised to one of the SNPs that was associated with D6 desaturase levels. It mediates cellular responses to various cytokines,

interferons and growth factors. For example, forms a homodimer when activated with IFN gamma and binds to IFN-gamma activated sequence to drive the expression of target genes, inducing cellular antiviral state (Ramana, Gil, Schreiber, & Stark, 2002). In one study which examined the expression and function of bone marrow macrophages, it was found that treatment of IFN-gamma activated macrophages with a rosiglitazone, ligand of PPARG and an anti-diabetic drug, blunted the expression of various inflammatory genes, many of which were target sites fo STAT-1, suggesting that PPARG acted via transrepression of STAT-1 (Ricote, Li, Willson, Kelly, & Glass, 1998).

Sequence specific transcription factors interact with *cis*-regulatory elements encoded in regulatory DNA – resulting in nucleosome displacement, remodelling of chromatin and nuclease hypersensitivity (Felsenfeld, Boyes, Chung, Clark, & Studitsky, 1996). DHSs have been used to mark regulatory DNA and active *cis*-regulatory elements in many organisms (Gaszner & Felsenfeld, 2006). In eurokaryotes, multiple transcription factors bind regulatory DNA to control gene expression. Recent studies have demonstrated that TF binding is highly clustered in humans (Yan et al., 2013). RNA polymerase II and RNA signals were found to be highly enriched in clustered regions (Chen et al., 2015). Furthermore, as the complexity of the region increases, the signals for RNA polymerase II and RNA also increased in tandem (Chen et al., 2015). RNA polymerase II is a crucial enzyme that transcribes enhancers and produces non-coding RNAs that play a crucial role in enhancer activity (T. K. Kim et al., 2010). High complexity TFBS clustered regions may be involved in regulating RNA poly1merase II, and hence may affect gene expression (Chen et al., 2015). These regions may harbour specific features that are associated with enhancer RNAs – non-coding RNA transcripts which are produced from putative enhancer regions (T. K. Kim et al., 2010) and have higher levels of H3K4me2 and H3K4me1 compared to H3K4me3 (Chen et al., 2015).

Histone modifications also cluster in a specific pattern in TFBS. Active chromatin markers such as H3K4me1/2/3, H3K27ac increase with increasing complexity of the TFBS binding site (Chen et al., 2015). A few repressive markers such as H3K9me3 and H3K26me3 decrease as TFBS complexity increases. However, some other repressive marks, such as H3K27me3 and H4K20me1 increases with increasing TFBS complexity (Chen et al., 2015). DNA methylation is also associated with TFBS in specific ways. Firstly, the density of CpGs within TFBS-clustered regions was higher than in the genome wide background. Secondly, the CpGs in the TFBS-clustered regions were less methylated than the rest of the genome background (Chen et al., 2015). Furthermore, analysis of each TFBS clustered region complexity category demonstrated that increasing density of CpGs was associated with increasing TFBS complexity. However, the methylation level was strongly negatively correlated with TFBS complexity, hence it would suggest that TFBS-clustered regions are selectively protected against DNA methylation, with the magnitude of protection has a positive association with TFBS complexity (Chen et al., 2015). Their findings were consistent with other studies on methylation and TFBS (Neph et al., 2012; Stadler et al., 2011). Although the mechanisms through which TFBS interact with SNPs to influence phenotypes is not completely understood, it is possible that genetic variants may disrupt the binding of transcription factors, which would then affect gene expression.

Although functional annotation may provide useful information when interpreting the results of GWAS, it is important to evaluate whether enrichment is statistically significant or whether it represents chance correlation. Functional annotation such as ENCODE data is potentially biased because it is curated by humans, with possible over-representation of pathways or regulatory elements which are more commonly studied. As such, it is important to consider statistical testing of enrichment. One problem is that is difficult to define an

appropriate null distribution. One current method is implemented by Genomic Annotation Shifter (GoShifter) (Trynka et al., 2015). This algorithm shifts sites of features within each locus so as to generate a null distribution of annotations which overlap associated variants. Using this method, the authors demonstrated that loci associated with rheumatoid arthritis and breast cancer harbor potential causal variants near summits of histone peaks and that variants associated with height are enriched in embryonic stem cell DNA hypersensitivity sites.

4.6 Chapter Summary

In this chapter, I demonstrate the utility of a Bayesian approach to fine mapping, defining credible sets which represent the SNPs most likely to be associated with the phenotype being studied. I then annotated these credible sets with regulatory epigenomic information, which added a layer of biological information that could drive further hypotheses on how these variants may act to bring about the phenotype. In the next chapter, I describe copy number CNVs, which are another type of functional variation in the genome besides SNPs. I perform association testing between copy number and phenotype to examine whether some of the missing heritability not explained by SNPs can be explained by copy number.

CHAPTER 5 Identification of copy number variation and association with pollutant and metabolic phenotypes

5.1 Chapter Overview

5.2 Introduction

5.2.1 Copy number variation in disease

5.2.2 CNV formation

5.2.3 CNV identification

5.3 Methods

5.4 Results

5.5 Discussion

5.6 Chapter Summary

5.1 Chapter Overview

In this chapter, I identify copy number variants (CNVs) from the PIVUS cohort using algorithms which utilise hidden Markov models. I then correlate copy number with the phenotypes described in the previous chapters in an attempt to identify variants which may explain some of genetic basis that is not explained by SNPs.

5.2 Introduction

5.2.1 Copy number variation in disease

A copy number variant (CNV) is defined as a genomic segment 1 kb or larger in size that occurs at variable numbers in the genome in a population?. Although SNPs are the most common variants in the human genome, in 2004, two landmark studies have demonstrated that copy number variants are widespread in the human genome (Iafrate et al., 2004; Sebat et al., 2004), with an average of more than 1000 CNVs in each genome, accounting for more than 4 million base pairs of difference (Conrad et al., 2010). It has been estimated that 12-15% of the human genome contains CNVs (Redon et al., 2006; Stankiewicz & Lupski, 2010) and their effect, in terms of affected base pairs, is far greater than SNPs (Conrad et al., 2010). Furthermore, 56% of the CNVs that have been identified so far occur in known genes (Iafrate et al., 2004).

CNVs have been associated with a range of disease phenotypes, including autoimmune diseases such as HIV susceptibility(Gonzalez et al., 2005) and psoriasis (Hollox et al., 2008), as well as congenital diseases such as heart malformations (Tomita-Mitchell et al., 2012). CNV analysis has been particularly successful in identifying variants associated with autism, schizophrenia and bipolar disorder. Heritability estimates for autism vary from 38% to 90%, with twin studies showing 80-90% concordance in monozygotic twins and 5-15% in dizygotic twins (Malhotra & Sebat, 2012). Despite high heritability, linkage studies suggested that patterns of segregation were not consistent with Mendelian patterns of inheritance (Malhotra & Sebat, 2012). Several CNV studies have been performed to look for non-Mendelian causes of autism spectrum disorders (ASD), specifically focusing on *de novo* mutations. Sebat et al investigated the global frequency of *de novo* CNVs in trios and found that a high rate of *de novo* CNVs in idiopathic ASD cases from simplex families were observed compared to cases from multiplex families or unaffected controls, suggesting that a majority of mutations identified were contributing to the risk of autism (Sebat et al., 2004). Two recent studies using very high resolution arrays reported that the burden of rare *de novo*

CNVs was greater in simplex cases compared to unaffected siblings, with an excess of 16 fold of genes impacted by *de novo* CNVs. Case-control studies have also supported this, with cases found to carry a higher global burden of rare genic copy number variants (Pinto et al., 2010). Several regions have been implicated, with notable regions at 16p11.2 (associated with other microdeletion syndromes), 15q11 to 13 (associated with Prader-Willi Angelman Syndrome) and 22q11.2 (associated with DiGeorge Syndrome (Glessner et al., 2009; Sanders et al., 2011). The most frequent CNVs tended to be large, containing multiple genes, with rare or *de novo* CNVs being smaller.

5.2.2 CNV formation

Figure 5.1 demonstrates various different CNV formation mechanism. A NAHR – non-allelic homologous recombination generates CNVs when genomic segments with high sequence similarities combine. B FOSTES – fork stalling and template switching occurs when DNA replication stalls and the lagging strand of DNA associates with a different region of the genome C NHEJ – Non-homologous end joining occurs when double stranded breaks in DNA sequence prompt NHEJ associated proteins to ligate DNA strands together D Mobile element insertion involves an RNA intermediate that is reverse transcribed into cDNA and inserted into the genome.

CNVs can be formed through various mechanisms. They can largely be divided into homologous and non-homologous recombination (Stankiewicz & Lupski, 2002). Homologous recombination (HR) requires extensive stretches of DNA sequence identity compared to non-homologous recombination, which only needs a few complementary base pairs (Liskay, Letsou, & Stachelek, 1987). Furthermore, most HR mechanisms also require a strand exchange protein to catalyse the invasion of the homologous duplex sequence by the 3’

end of the single strand of DNA (Hastings, Lupski, Rosenberg, & Ira, 2009). HR also occurs during DNA repair when an identical sequence is used to replace a damaged segment. If the sequence is repaired using the homologous sequence in the same position in the sister homologue, there will be usually be no change in structure.

However, if the repair utilizes homologous sequences in different positions, it is known as non-allelic homologous recombination (NAHR) (Stankiewicz & Lupski, 2002). NAHR can occur in meiosis as well as mitosis (Lam & Jeffreys, 2006). It can involve interchromosomal, intrachromatidal and interchromatidal rearrangements (Gu, Zhang, & Lupski, 2008). Another mechanism of CNV formation is non-homologous end joining (NHEJ) (Lieber, 2008). This occurs as a result of aberrant repair of DNA double strand breaks, and is based on the information contained near or within the DNA lesion. This makes it more error prone compared to NAHR (Lieber, 2008). The breakpoints of such CNVs are often located within repetitive elements such as long and short interspersed repeat elements. This suggests that NHEJ may be stimulated by certain genomic architectures (Toffolatti et al., 2002). Yet another mechanism of CNV formation is fork stalling and template switching (FOSTES) (J. A. Lee, Carvalho, & Lupski, 2007). This is induced by single strand breaks during DNA replication and can vary greatly in size and complexity. Coupled with NHEJ, it is proposed to be the predominant mechanism for complex rearrangements.

Hastings et al proposed a further generalization of the FOSTES mechanism, which is known as the MMBIR (microhomology mediated break induced replication) model (Hastings et al., 2009). Retrotransposition is another mechanism of CNV formation, which begins with long interspersed nuclear elements (LINE) that can transpose to a new location by target primed reverse transcription (Lupski, 2010). Mills et al examined data from the 1000 Genomes Project and found that majority of the deletions could be attributed to a non-homology based mechanism or MMBIR (Mills et al., 2011). The majority of small insertions could be

attributed to retransposition activity, and most tandem duplications were formed by FOSTES or MMBIR. On the other hand, large deletions and duplications were likely to be generated by NAHR (Mills et al., 2011). New retrotransposon insertions account for majority of the smaller events. Even though the absolute rate of mutation is low, each mutation can affect thousands of kilobases, hence the functional impact may be disproportionately high (Hastings et al., 2009).

5.2.3 CNV identification

The calling of CNVs has traditionally been performed with cytogenetic karyotype analysis using chromosome banding methods. More recently, advances in fluorescence in situ hybridization (FISH) has allowed me to refine the identification of chromosomal defects (O'Connor, 2008). However, to characterise these defects at the molecular level using traditional molecular biology techniques such as PCR, southern blotting and sequencing can be time consuming and labour intensive.

5.2.3.1 Array comparative genomic hybridization (CGH)

Recent developments have allowed for the characterisation of chromosomal abnormalities with high-throughput hybridization-based approaches such as CGH (M. M. Weiss et al., 1999). In this approach, tumour DNA is labelled with a green fluorochrome, which is then mixed with red labelled normal DNA and hybridised to normal metaphase chromosomes. The red and green labelled fragments then compete to hybridise onto their locus of origin on the chromosome, with the green to red fluorescence ratio as measured with a microscope representing loss or gain of genetic material at a specific locus. CGH has several advantages – it can scan the whole genome for imbalances, and does not require cells that are undergoing

division (Speicher et al., 1993). However, the resolution has been limited to alterations of approximately 5 to 10 Mb (Kirchhoff et al., 1998).

To overcome these limitations, CGH has been combined with microarray techniques in an approach called array CGH. Instead of using metaphase chromosomes, this method uses DNA probes deposited onto glass slides as targets for analysis. These probes vary in size, from oligonucleotides (25 to 85 base pairs) to clones from bacterial artificial chromosomes (80000 to 200000 base pairs) (Theisen, 2008). While the resolution of array CGH is higher than CGH because the probes are smaller, the actual level of resolution depends on the probe size and genomic distance probes. Array CGH can simultaneously detect deletions, duplications and amplifications of any locus on an array, hence it is equivalent to thousands of FISH experiments, hence saving labour and costs. Additionally, it has been shown to be successful in the detection of submicroscopic chromosomal abnormalities in individuals with various birth defects and mental retardation (Shaffer & Lupski, 2000). Two prominent studies in 2004 identified CNVs in the genomes of normal individuals using array-CGH. Using array-CGH on commercial arrays with bacterial artificial chromosomes (BACs), Iafrate et al identified 255 variable regions in 55 individuals and Sebat identified 76 variable regions in 20 individuals (Iafrate et al., 2004; Sebat et al., 2004).

Various developments in array technology have taken place over the last decade with trends towards increased number of features and shorter DNA sequences, so as to increase the resolution of CNV detection. While the resolution of the array depends on number, distribution and length of probes, the actual ability to detect CNVs also depend on probe response characteristics and the signal-to-noise ratio (Ylstra, van den Ijssel, Carvalho, Brakenhoff, & Meijer, 2006).

Current array-CGH methods for identifying CNVs typically have a resolution of 30 kb to several megabases. However, in the case of deletions, it may be possible to detect CNVs as small as a few kilobases by identifying errors in Mendelian inheritance. Conrad et al modelled the size distribution of deletions in their study or 30 European and 30 Yoruban trios. They concluded that the frequency of deletions in the genome increased with decreasing size, suggesting that smaller deletions may be even more numerous. Hence, it is important to develop array-based CNV assays that access copy number changes in this range. Higher density oligonucleotide or SNP arrays are more suitable for this task than large insert clone arrays.

5.2.3.2 Identifying CNVs from oligonucleotide Arrays

Oligonucleotide arrays are one step up from array-CGH in terms of resolution from CNV detection. Early arrays were developed using the mechanical spotting of synthesised oligonucleotides, with a maximum of 60 000 oligonucleotides on each glass side, corresponding to one probe every 50 kb (Carvalho, Ouwerkerk, Meijer, & Ylstra, 2004). More recent commercial oligonucleotide arrays are manufactured by directly synthesising the oligonucleotides onto the glass side. For example, Agilent uses ink-jet technology to do this on a spot-by-spot basis, achieving up to 244 000 oligonucleotides per array (Agilent). NimbleGen generates a digital mask by photolithography and can print 385 000 oligonuclotides per array (Nimblegen). One disadvantage is that poor signal-to-noise ratio of hybridisations can result in variation in CGH ratio. Because of this, data from several oligonucleotides needs to be averaged to reduce measurement variation (Carter, 2007).

To improve signal-to-noise ratio, some groups have used a method called representational oligonucleotide microarray analysis (ROMA) to decrease the complexity of genomic DNA

used for hybridisation (Lucito et al., 2003). With this method, the genomic DNA is digested using a restriction enzyme and the fragments are ligated to adapters and then amplified with universal primers. During this step, only the smaller fragments are amplified, thereby decreasing the complexity of the DNA. These fragments are then hybridised to an oligonucleotide array of probes selected to correspond to the reduced set of amplified fragments (Lucito et al., 2003). However, there are several issues raised by the ROMA technique. Reduction in complexity may lead to differential representation of various areas in the genome, which may be wrongly interpreted as copy number changes. Furthermore, different individuals have different restriction patterns, and some individual probe ratios may be related to fragment rather than copy number changes (Carter, 2007).

5.2.3.3 Identifying CNVs from genotyping platforms

With the increase in number of GWAS studies performed, a large number of cohorts have now been profiled on genotyping chips from Affymetrix and Illumina. Although these arrays were initially designed for assaying SNPs, the intensity information can determine genomic copy number. With the Affymetrix SNP arrays, match and mismatch probe pairs approximately 25 bases long are designed to target each SNP allele (Rouchka, Phatak, & Singh, 2008). Hybridization is performed using singer DNA rather than cohybridization of two DNA as in the case of array-CGH. To improve the signal noise ratio, the DNA is first digested with restriction enzymes and ligated with adapters, with the smaller fragments being subsequently amplified. For CNV detection, the signal intensities of the match and mismatch probes are compared with values from the rest of the cohort, so the relative copy number is determined. Noise can be further reduced by taking into account probe GC content (Nannya et al., 2005). For robust detection of CNVs, significant ratio shifts in sequential probes are needed. For the Affymetrix 500k array, SNPs have a median pacing of 2.5 kb. However, these probes are relatively sparse in regions of complex CNVs, with the result that the

resolution of CNV detection is typically 10-40 kb. To resolve this problem, Affymetrix has now included 500000 non-polymorphic probes in the latest 5.0 chip (Affymetrix). Interestingly, CNVs can also be detected from genotyping using apparent errors in Mendelian inheritance. Conrad et al used this approach to identify 586 potential deletions ranging in size from 300 bp to 1.2 kb (Conrad, Andrews, Carter, Hurles, & Pritchard, 2006), while McCarroll et al used SNPs that were not in Hardy-Weinberg equilibrium, or had other apparent genotyping errors, to identify 541 potential deletions (McCarroll et al., 2006). One limitation is that these approaches can only be used to identify deletions, not duplications.

One issue that arises in the use of array technology to identify CNVs is how a CNV is defined. A variety of different methods currently exist, which vary from simple threshold values (Vermeesch et al., 2005) to complex multi-threshold models (Fiegler et al., 2006). Due to different efficiencies of hybridization, measurement variance may result, which can lead to false positive and negative results. Carter et al identified reproducible local small variations in reported ratios that could be seen as waves of increasing or decreasing ratios. These waves affected the setting of calling thresholds since the peaks of the waves could pass a calling threshold (Carter, 2007). They found that these waves were correlated with GC content and could be removed by normalising datasets. However, it would certainly be useful to develop a gold standard for testing the performance of each platform. Carter et al selected two DNA samples from Coriell Cell Depositories: the first was NA15510, the source of the fosmid library used to confirm genome assembly during finishing of the human genome; the second was NA10851, a well characterised trio-offspring DNA from HapMap. They used these two DNAs for the testing of new array design platforms (Carter, 2007).

Several algorithms have been developed to call CNVs from genotyping platforms, as it allows for the profiling of both CNVs and SNPs. The main current genotyping chips are the Affymetrix GeneChip and Illumina Infinium arrays. Several algorithms have been developed

to detect copy number alterations on these platforms. Two common algorithms are PennCNV(Version 5.14.0, June 2011) (K. Wang et al., 2007) and QuantiSNP.

5.2.3.3.1 QuantiSNP

QuantiSNP is an objective Bayes hidden Markov model (OB-HMM) that infers regions of segmental aneuploidy from genotyping data. It allows for the benefits of Bayesian marginal probability calculus while incorporating calibrated hyperparameters in the priors. It employs a re-sampling data-driven strategy to set prior parameters given a user defined false positive rate. All other parameters are set using maximum marginal likelihood matched with prior training data, thereby allowing for formal power analysis. This is important in qualifying the risk and costs involved in further validation studies for the CNVs detected. To test their method, the authors compared the results obtained from quantiSNP with mapping obtained from molecular genetics technologies. They demonstrated that it could produce accurate CNV detection and breakpoint identification. The model allows for the specification of one of the hidden states as a normal state, thereby allowing parameters to be set that ensure frequentist coverage properties for the excursion of the model from the null state while benefitting from Bayesian inference. In this method, data can be combined from platforms of different resolution.

The objective bayes (OB-HMM) approach uses log R ratios, which are a measure of the magnitude of combined fluorescence intensity signals from both probes, and B allele frequencies, which represent the relative ratio of fluorescent signal of one probe compared to the other. The parameters for the model are learnt using an expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977), with the maximum likelihood sequence of hidden states inferred using a Viterbi algorithm (Rabiner, 1989). They apply the EM and Viterbi algorithms to each chromosome separately. The number of mixture components is

conditioned on the hidden states. A Bayes factors is then assigned to each copy number region identified, providing a probability measure of the strength of evidence for presence of a copy number variant in a region versus the null hypothesis of two copies present. The larger the Bayes factor, the greater the evidence for the existence of a copy number alteration. The model uses fully conjugate prior distributions, which allows for efficient calculations of the posterior through analytical integration. In order to estimate hyperparameters in prior distributions, reference datasets of chromosome X multiple copy cell lines were analysed using maximum marginal likelihood inference.

In any Bayesian approach, prior probability models are developed to estimate unknown parameters and these priors are then updated using Bayes rule to give posterior distributions. In a subjective Bayesian approach, prior distributions are elicited using subjective knowledge such as expert knowledge. However, in an objective Bayes approach, the priors are set objectively, which is particularly useful when there is little prior knowledge available.

The selection of appropriate prior distributions is not straightforward. Jeffreys et al have provided a scale for interpretation of the Bayes factor, but that just provides a descriptive statement for interpretation rather than calibration (Jeffreys, 1961). High false negative rates have been reported in various CNV calling experiments, hence the true length distribution of CNVs remains unclear. As such, semi-Markov type methods are less appropriate since they rely on such knowledge. An objective Bayes approach helps to solve this problem by selecting parameter values which can calibrate the model to false positive error rates. This calibration is accomplished by generating pseudo null datasets with no copy number variation and then applying the algorithm on it, so as to generate an empirical measure of false positive rates. One advantage of QuantiSNP is that the authors have adapted the OB-HMM method to perform analysis on multiple samples, whereby the joint use of data allows for improved

detection CNVs. This means that small CNVs can still be detected if they are found in large numbers of individuals.

5.2.3.3.2 PennCNV

The PennCNV algorithm uses a hidden Markov model to identify copy number on an individual basis (K. Wang et al., 2007). Using an Illumina genotyping platform, the beadstudio software generates a log R Ratio (LRR), a measure of normalized total signal intensity, and the B Allele Frequency (BAF), a measure of normalized allelic intensity ratio. The combination of these two parameters can be used to determine copy numbers and to differentiate loss of heterozygosity regions from regions with copy number of two, identifying the breakpoints where the copy number state changes. The PennCNV model adopts a six state definition by default. A hidden Markov model is created, which incorporates distance between adjacent SNPs, frequency of the B allele in the general population, LRR and BAF. The first parameter determines the probability of having a copy number state change between the two SNPs. The population frequency of B allele is compiled from a large dataset of individuals with normal phenotype and mixed ancestries.

PennCNV also allows for the calling of copy number from family data. It has been demonstrated that a large majority of CNVs in offspring have been inherited from their parents, hence the fraction of CNVs inferred in offspring, but not detected in parents, could reflect the false positive and false negative rates. Using this technique, the authors demonstrated that false positive and negative calls ranged from 20% to 27% on the genotyping platforms tested. They also investigated the effect of CNV length on calling accuracy, and found that most of the false positive and negative calls were small sized. By incorporating family information into the calling algorithm, the total sensitivity of CNV detection increased. Furthermore, there was better characterisation of CNVs not detected in

parents, for example, somatic rearrangements in the immunoglobulin region.

5.3 Methods

Genotype data from the PIVUS cohort was used for CNV identification. Normalisation was performed according to Illumina Genomestudio default settings (Illumina, last accessed 13 June 2015). Raw intensities were used as input to the QuantiSNP, with no MAF or genotyping call rate filters applied prior to this step. This is because SNPs that are located within CNVs may have genotyping clusters that are less obvious to SNP calling software and may hence fail traditional SNP QC. Instead, QC was applied after CNVs were called, as described later. Sample QC was similar to that of the previous chapters with same filters applied.

5.3.1 CNV calling

The QuantiSNP software was used for CNV calling (S. Colella et al., 2007). As input, SNP chromosome position, LRR and BAF were entered into the model. A region without evidence of CNV should show a LRR of around zero and three clusters of BAF (1, 0.5, 0) corresponding to BB, AB and AA respectively. Local GC content files were downloaded and used for GC based correction of the LRR. The default length of 2×10^6 was used to calculate transition probabilities. Ten iterations were used for the expectation maximisation algorithm during learning. I focused on autosomal chromosomes due to difficulties in identifying copy number in sex chromosomes. For each copy number identified, a score is assigned to it which is the log Bayes factor and represents the support for existence of a copy number alteration at the specified location. I removed CNVs with log Bayes factor of less than 10 as this was the

recommended confidence threshold for QuantiSNP.

5.3.2 Identifying CNV segments

Since there is inherent uncertainty in identifying the breakpoints of CNVs as this depends on location and density of SNPs on the genotyping chip, Plink was used to coalesce nearby CNVs into unique segments (Purcell et al., 2007). An input file was prepared, which included sample identity, chromosome, start and end positions as identified by QuantiSNP, number of copies, confidence score (log Bayes factor) and number of probes. Plink was used to create a file that summarised CNVs per individual, including sample ID, number of segments each individual has, total kilobase distance spanned by segment and average segment size. Error checks were performed by checking whether segments were overlapping for the same person, for example, whether a deletion and duplication had been specified in the same region for the same individual or whether the same event is listed twice. Finally, the CNVs were grouped into segments. Plink takes all segments in a given region (in this case the whole genome) and forms pools of overlapping segments. These pools of overlapping segments will be listed in order of decreasing size (number of segments). Each of these pools was treated as a unique CNV.

5.3.3 Association of CNV with phenotypes

Rank inverse transformation was performed on the pollutant levels and metabolic traits as described in previous chapters. Transformed pollutant levels were regressed against gender, cholesterol, triglycerides and the first 2 components of MDS (as described in Chapter X). Residuals were then used for correlation analysis (Spearman) with copy number of all CNVs.

In the case of metabolic traits, the transformed traits were used directly for correlation analysis with no additional adjustment. Genome wide significance thresholds were set at 3.04 x 10^{-6} and 5.42 x 10^{-7} for pollutants and metabolic traits, respectively, to account for multiple testing based on Bonferroni correction, dividing by total number of CNVs identified and number of phenotypes tested.

5.4 Results

There were a total of 3292 CNVs identified. The maximum and minimum CNV lengths were 173 Mb and 0.23 Mb respectively. The largest CNV was located at chr 4p16, and is a known variant described in the database of genomic variants (Redon et al., 2006), encompassing genes such as *SLC2A9* and *DRD5*. The smallest CNV was located at chr 9q21, and is also previously reported variant (K. K. Wong et al., 2007). It encompasses the *GDA* gene. The commonest CNV was located on chr 16p13, with 432 individuals having that CNV. It has been reported in many different studies (K. K. Wong et al., 2007)

5.4.1 PCBs

Amongst the PCBs, the association signal with the lowest p value was between PCB105 and S10009 (p value 3.3x10^{-6}), a CNV located in chr2p16.1 spanning 600 kb. This region overlaps a number of genes, including *CLHC1*, *CCDC88A*, *MTIF2*, and *RPS27A*. It also overlaps a gene that has been associated with metabolic syndrome – *SMEK2* (Kristiansson et al., 2012), which is notable because PCBs have been associated with metabolic traits such as obesity (D. H. Lee et al., 2012). There were no SNPs in the region significantly associated with the same phenotype.

SMEK-2 functions as a regulatory component within a protein phosphatase (PP4) complex which controls hepatic glucose production by regulating the cAMP-response element binding protein (CREB)-regulated transcriptional coactivator 2 (CRTC2) phosphorylation levels. It is involved in recovery from the DNA damage checkpoint (Gingras et al., 2005), and may have physiological roles in the liver, such as regulating hepatic gluconeogenesis, since there is increased expression of *SMEK* genes during fasting and induced insulin resistance (Y. S. Yoon et al., 2010). *SMEK-2* has also been shown to be associated with systolic blood pressure in a GWAS on metabolic syndrome (Kristiansson et al., 2012).

5.4.2 Metals

Amongst the metals, the association signal with the lowest p value was between copper and S10138 (p value $8.9x10^{-5}$). There were no SNPs in the region which were significantly associated with the same phenotype. This CNV was also associated with mercury (p value $8.3x10^{-5}$). It is located at the chr18 q11.2 region and overlaps the potassium channel tetramerization domain-containing 1 (*KCTD1*) gene, which has been shown to be associated with sudden cardiac death (Aouizerat et al., 2011). *KCTD1* is a negative regulator of AP-2, a transcription factor implicated in carcinogenesis (Ding et al., 2009). A CNV on chr12q21.2 is associated with several metabolic phenotypes, including ApoB, D5 and D6 desaturase, p values (0.00034, 0.00016 and 0.00022 respectively). This CNV does not overlap any genes.

PCBs are known to be metabolised in vivo to hydroxy- and sulfur-containing metabolites by Cyp enzymes. (Lehmler et al, 2010). The 19q13.2 locus contains the Cyp2A, Cyp2B and Cyp2F subfamilies, which are mixed together in clusters rather than occurring distinctly. However, it is not clear how Cyp2b6 variants affect PCB metabolism. Of note, CNVs are

known to be located in Cyp2b6. With multi-locus probe assay screening and qPCR assays, Martis et al. (2014) identified Cyp2B6 deletion and duplication alleles limited to exon 1-4 of the *CYP2B6* gene (Martis et al., 2013). They subsequently carried out high-resolution aCGH and mapped the size of the *CYP2B6* deletion/duplication to approximately 70 kb and localized the CNV breakpoints to introns 4/exons 5 of *CYP2B6* and the proximal *CYP2B7P1* pseudogene, both within LCRs, which are likely to mediate non-allelic homologous recombination between *CYP2B6* and *CYP2B7P1*, resulting in the partial deletion allele *CYP2B6*29* and duplicated fusion allele *CYP2B6*30*. *CYP2B6*30* is predicted to have 15 amino-acid changes, for example, the exon 7 p.R378X non-sense alteration may truncate it, hence it may lack a heme binding site and be non-functional.

5.4.3 CNV in the Cyp2B6 region

One example of a CNV which may be associated with phenotypes in our study would be that which is located at the *CYP2B6* region. In the previous chapters, I have demonstrated that the *CYP2B6* region is associated with the metabolism of PCBs. This region is known to contain duplications and deletions that have associated with phenotypes in other studies. For example, wide variability in the distribution of *CYP2A6* deletion alleles have been shown across 5 tested populations, with highest frequency observed in Asians (Martis et al., 2013). This deletion has been associated with lower cancer risks among Asians (Rodriguez-Antona, Gomez, Karlgren, Sim, & Ingelman-Sundberg, 2010). I identified CNVs in 8 individuals within the *CYP2B6* region, but there was no significant association between copy number and levels of PCBs. Since this is a rare CNV, it is possible that this lack of association might be due to a lack of power in the study. There were no significant associations with any of the five metabolic traits tested.

5.5 Discussion

5.5.1 Challenges in elucidating the functional significance of CNVs

There are various challenges associated with elucidating the contribution of CNVs to disease phenotypes. Firstly, it is not clear to what extent CNVs are inherited as opposed to forming de novo. Several studies have suggested that CNVs are predominantly inherited in a Mendelian fashion (Stefansson et al., 2008). However, de novo CNVs have been implicated in several diseases, especially in those where affected individuals have fewer children than unaffected individuals, such as severe autism and schizophrenia, suggesting that the pool of causal alleles might need to be replenished through mutation (McCarroll, Kuruvilla, et al., 2008). A larger fraction of affected than unaffected individuals had CNVs, especially for the sporadic form of the disease, suggesting that the de novo CNVs identified were likely to be causal.

Two large schizophrenia GWAS identified de novo deletions in 1q21.1, 15q13.3 and 22q11 (International Schizophrenia Consortium, 2008; Stefansson, 2008). These variants were observed in the offspring and excluded in both parents, suggesting that they arose de novo and that their contribution to schizophrenia was through sporadic mutation. However, in autism, a GWAS in 751 multiplex families led to the identification of a microdeletion/duplication at 16p11.2 that was highly penetrant, but detected in less than 1% of autistic individuals (L. A. Weiss et al., 2008). This mutation was de novo in most families, and present in only one of the affected offspring.

5.5.2 Rare vs Common CNVs

Another question that arises is to what extent rare and common CNVs contribute to various

phenotypes. Approximately 80% of CNV differences between any two individuals arise from common CNVs (allele frequency of greater than 5%) (McCarroll, Kuruvilla, et al., 2008). Recent technology has made it more convenient to analyse common CNVs for GWAS association. Firstly, SNP arrays have been re-designed to eliminate design bias against genomic segments affected by common CNVs, so that data can now be collected from most CNVs that are larger than 5 kb (Cooper, Zerr, Kidd, Eichler, & Nickerson, 2008). Secondly, high resolution maps of CNVs make it possible to specify which probes interrogate common CNVs (McCarroll, Kuruvilla, et al., 2008). Thirdly, more advanced algorithms have been developed, which treat CNV genotyping as an explicit problem, yielding more accurate data than the older algorithms (Cooper et al., 2008; Korn et al., 2008).

5.5.3 Potential pitfalls with CNV analysis

That said, there are still several potential pitfalls with the analysis of CNV data. For example, CNV genotyping arrays are less mature and have been tested less frequently than SNP genotyping arrays. As such, there is a broader distribution of data quality, with data from some CNVs clustering clearly, and other being poorly resolved. The latter case is prone to differential bias, in which origin and handling of samples may be different between those with case and control status, resulting in false associations with the phenotype. As such, it is important to examine the CNV-phenotype association for evidence of inflation, and it is also important to analyse the quality of raw data in CNVs which have a putative association with disease.

Many CNVs segregate with different allele frequencies in different populations, which appears to be similar to the allele-frequency differentiation of SNPs, and hence may represent

drift in allele frequencies in isolated populations (McCarroll, Kuruvilla, et al., 2008). One conundrum being debated widely is the extent to which CNVs are in LD with SNPs. For example, a large empirical analysis found that common CNVs were almost as well tagged as SNPs of the same frequency, suggesting that disease association of common CNVs could be evaluated by combining GWAS SNP with a map of LD relationship between SNPs and CNVs (McCarroll, Kuruvilla, et al., 2008). This method was used to identify the association between Crohn's disease and a common deletion polymorphism upstream of the *IRGM* gene (McCarroll, Huett, et al., 2008). This polymorphism is in strong LD with SNPs at *IRGM* that were found to be associated with Crohn's disease (W. T. C. C. Consortium, 2007), and the deletion was also associated with Crohn's disease in an independent patient cohort (McCarroll, Huett, et al., 2008). Since the SNPs and CNV are in LD, interpreting a disease causal association is not straightforward. At the *IRGM* locus, no single variant has been determined to be causal (McCarroll, Huett, et al., 2008).

Investigating the relationship between rare CNVs and disease would be the next frontier since most cohorts collected for GWAS are sufficiently powered to investigate low frequency variants, especially if they have penetrant effects and could be evaluated in the context of a gene based burden test to increase statistical power. When a rare CNV is detected in the same region in unrelated individuals, it is most commonly present on a long SNP haplotype, possibly suggesting recent shared ancestry at the locus. This may confound results, since the causal variant may be located at another location in the shared haplotype (McCarroll, Kuruvilla, et al., 2008).

There are various potential confounding effects when studying the relationship between rare CNVs and disease. Firstly, detection of CNV from array data depends on the signal to noise

ratio of each hybridisation, with the number of CNVs detected per sample varying four fold from sample to sample (McCarroll, 2008). Because DNA from cases and controls in GWAS can be analysed in different batches, arises from different clinical sites, such experiments are prone to differential bias (Clayton et al., 2005). This is particularly relevant when assessing the genomic burden of CNVs across the genome since such methods aggregate the confounding effects across thousands of loci. Population structure is another source of confounding in association studies. Populations with African ancestry tend to have more rare CNVs than non-African populations (McCarroll, Kuruvilla, et al., 2008). Hence, when performing CNV analysis, it might be necessary to take into account population structure though methods such as multi-dimensional scaling. Yet another potential source of confounding is relatedness. When GWAS cohorts are analysed for cryptic relatedness, some analyses have detected that some samples are related to each other, and hence may share more CNVs than unrelated samples (McCarroll, 2008).

CNVs can affect the collection of SNP data at a CNV locus by causing SNP intensity data to cluster poorly, yielding genotypes that may seem to violate Mendelian inheritance and hardy Weinberg equilibrium (Conrad et al., 2010; Redon et al., 2006). Early commercial SNP arrays were designed by evaluating potential SNP assays on screening arrays then selecting high performing assays to carry through for the commercial product, hence this may exclude CNV regions. In fact, one study demonstrated that common CNVs were not well covered in early SNP arrays, but rare CNVs were covered with approximately the same density as the rest of the genome (McCarroll, Kuruvilla, et al., 2008). Newer generations of SNP arrays have addressed this problem in various ways. Affymetrix has developed hybrid arrays with SNPs as well as copy number probes, which are non-polymorphic probes designed for CNV measurements used to target regions of known CNVs (McCarroll, Kuruvilla, et al., 2008). In

another approach by Illumina, assays for SNPs within predicted CNV regions were added to genotyping arrays regardless of whether they passed traditional quality control. As such most common large and intermediate sized CNVs have now been represented on various platforms (McCarroll, Kuruvilla, et al., 2008). However, current platforms have limited ability to detect CNVs less than 20 kb, CNVs in duplication rich corners and those in novel regions of the genome that are not part of the human reference sequence (McCarroll, Kuruvilla, et al., 2008).

CNVs can also be identified from next generation sequencing data. There are two main groups of detection methods – paired end mapping (PEM) and depth of coverage (DOC) (Duan, Zhang, Deng, & Wang, 2013). The latter approach is employed by most current algorithms, but the former is more suitable for detecting inversion and small CNVs. Methods based on DOC first pile up the aligned reads at the genomic coordinates, then calculates coverage across sliding windows, yielding the read depth signal (Miller, Hampton, Coarfa, & Milosavljevic, 2011). The read depth signal can be processed in two alternative ways. The first approach involves segmenting it by local change-point algorithms then merging the results(Abyzov, Urban, Snyder, & Gerstein, 2011). The second approach involves event wise testing at each window or several consecutive windows (S. Yoon, Xuan, Makarov, Ye, & Sebat, 2009).

Various factors affect how easy it is to detect CNVs from NGS data. Shorter CNVs are more difficult to detect than longer CNVs. The greater the deviation from copy number 2, the easier it is to detect the CNVs. Higher coverage provides better resolution for break point detection (Duan et al., 2013). Duan et al compared six different methods to identify CNVs from next generation data – readDepth, CNV-seq, FREEC, CNVnator, Event Wise Testing (EWT) and

SegSeq (Duan et al., 2013). ReadDepth uses circular binary segmentation (Miller et al., 2011), CNVnator uses mean shift (Abyzov et al., 2011) and FREEC uses a LASSO algorithm (Boeva et al., 2011). EWT performs event wise testing at each window (S. Yoon et al., 2009). CNV-seq and SegSeq are both DOC based methods (Chiang et al., 2009; Xie & Tammi, 2009). They found that readDepth, EWT and CNVnator performed better in terms of break point estimation than the other tested methods, with lower false positive rate. On the other hand, EWT, SegSeq and FREEC provided better sensitivity, with higher true positive rate. CNVnator, readDepth and CNV-seq provided better copy number estimation. They found that EWT and FREEC were computationally less expensive than the other algorithms, hence were suitable if speed and memory were priorities (Duan et al., 2013).

These different algorithms also have various weaknesses. The LASSO-based method implemented by FREEC was shown to be robust against outliers. However, since the L-1 norm used in LASSO yields bias on the estimates of the amplitude, the copy number estimates of FREEC were not accurate (Boeva et al., 2011). CNV-seq failed to detect copy numbers of 1 and 3 as the statistics were not sensitive to this purpose (Xie & Tammi, 2009). ReadDepth uses CBS to segment the read depth and calls CNVs based on a negative-binomial model, which requires several parameters to be adjusted, such as variance mean ratio (Miller et al., 2011). The authors found that copy number 1 and 4 could be detected using different combinations of parameters, but not copy number of 3 (Duan et al., 2013). On the other hand, SegSeq fails at copy number zero because it uses the log ratio of the case and control read count as the statistic of interest, hence when the copy number is zero, the log ratio is negative infinity (Chiang et al., 2009). There are various ways in which the existing algorithms could be improved. For example, one approach is to combine the PEM approach with the complementary DOC signature to improve the breakpoint accuracy. There are 2 algorithms which implement this – CNVer and CopySeq (Medvedev, Fiume, Dzamba, Smith, & Brudno,

2010; Waszak et al., 2010). Another approach is to process multiple simultaneously, as implemented by JointSLM, which can decrease the false positive rate (Magi, Benelli, Yoon, Roviello, & Torricelli, 2011).

Additionally, because the human genome is complex, there are other factors which need to be considered when calling CNVs from NGS data, such as multiple alignment and GC content correction (Duan et al., 2013). Multiple alignment can occur due to the presence of SNPs and indels, as well as possible sequencing errors. There are several ways to deal with multiple alignments. One option is to restrict analysis only to unique alignments (Abyzov et al., 2011). Another option is to assign a random locus to the multiple loci aligned read and count the read depth signal (Chiang et al., 2009).

5.6 Chapter Summary

I have been successful in identifying CNVs in a cohort of elderly individuals and correlating them with metabolic and pollutant phenotypes. Although these analyses have revealed CNVs which are significantly associated with some phenotypes, it does not reveal any obvious biological links for these relationships. In the next chapter, I discuss how CNVs can be analysed in relation to gene expression, which may help to uncover the underlying biology which links CNV to phenotypes.

CHAPTER 6 Identification of CNVs and association with gene expression in a cohort of UK individuals

6.1 Chapter Overview

6.2 Introduction

6.3 Methods and Materials

6.4 Results

6.5 Discussion

6.6 Chapter Summary

6.1 Chapter Overview

In the previous chapter, we identified CNV from within the PIVUS cohort and studied the association between copy number and metabolic as well as pollutant phenotypes. Most of these associations were of moderate strength. One way of elucidating further how CNVs might contribute to disease phenotypes might be through analysing the relationship between CNVs and gene expression. This should ideally be done on a dataset in which CNV, eQTL data as well as phenotypes are available for each sample. Since the PIVUS cohort does not have available gene expression data, I identified CNVs from a cohort of healthy individuals and correlated gene expression from the same individuals with the copy number identified. This would shed light on how CNVs might be associated with gene expression, and possibly act to bring about phenotypes.

6.2 Introduction

It is important to identify how variants affect gene expression, especially in identifying how significant GWAS signals affect phenotype. It is becoming increasingly clear that such variants may affect gene expression in cell specific manners. For example, in umbilical cord-derived cell cultures, up to 80% of regulatory variants are cell type specific (Dimas et al., 2009)and comparison between skin, fat and lymphoblastoid cell lines (LCLs) demonstrate that less than 1/3 of eQTLs are common between these tissues (Nica et al., 2011). The biological basis for this is yet unclear, but could be related to variation at tissue specific distal enhancers (Dimas et al., 2009).

6.2.1 Cell and context specific eQTLs

Fairfax et al demonstrated cell specific eQTLs in monocytes and eQTLs, with effect size and statistical significance of eQTLs varying as a function of distance from the transcription start site (TSS) (Fairfax et al., 2012). They observed several eQTLs with cell type-dependent directional effects, with the same SNP being associated with opposing directional effects in different cell types. The most significant directional eQTL was at the gene *DFNA5*, a target of promoter methylation in gastric cancer. The same study also identified multiple cell-specific trans eQTLs, the most significant of which were to *KLF4* and *LYZ*. The *LYZ* locus forms a monocyte specific *cis*-eQTL to a probe mapping to the 3' UTR of lyzozyme, and is also a trans-eQTL. This trans-eQTL is a novel master regulatory region involving both *cis* and trans eQTLs at the 12q15 locus for rs10784774 and SNPs spanning lyzozyme. It forms a monocyte specific *cis*-eQTL to a probe mapping to the 3' UTR region of *LYZ* and forms a trans-eQTL to 62 genes throughout the genome. Monocytes scavenge and present antigens through an endosomal pathway via the release of hydrolyzing enzymes from secretory granules (Nichols, Bainton, & Farquhar, 1971). Lysozyme is an essential component of monocyte secretory granules, which degrades bacterial cell walls (Helal, Bader, & Melzig, 2012).

Disease susceptibility loci can be associated with cell-specific differential gene expression. Fairfax et al found that 49.4% of traits are associated with one or more significant *cis*-eQTLs based on listed GWAS markers in the catalog of published genome wide association studies (Fairfax et al., 2012). They also found that 730 genes show significant *cis*-eQTL related to GWAS, of which 33.5% are B cell specific and 45.5% are monocyte specific. Furthermore, some of these GWAS SNPs have differing effects between cell types. For example, rs10781500, associated with ulcerative colitis and ankylosing spondylitis, shows contrasting *cis*-eQTLs for genes at 9q34.3 depending on the cell type.

In a later study on the same cohort, the authors identify context specific eQTLs (Fairfax et al., 2014). Inappropriate immune activity and inflammation are involved in the pathogenesis of various common diseases, including inflammatory bowel disease, rheumatoid arthritis, cancer and coronary heart disease. Monocytes constitutively express receptors which are responsive to pathogens and cytokines, including toll-like receptors, and can be stimulated to exhibit broad gene transcription with cytokine production, increased antigen processing and migration. This pathway activation can affect disease susceptibility and may only be revealed in stimulated cells. Macrophages from different mouse strains have been shown to demonstrate trans-regulatory loci on inflammatory stimuli. Fairfax et al stimulated monocytes with lipopolysaccarides (LPS), a component of gram negative bacterial cell walls that triggers *TLR4* signalling and interferon-gamma, a cytokine that acts through the *JAK-STAT* pathway and is important in mycobacterial infection (Fairfax et al., 2014).

They found a diverse set of immune-related genes that showed eQTLs after induction with LPS or IFN-g, including transcription factor genes such as IRF2, JUN and various interleukin genes. Interestingly, for eQTLs, associations were observed across all conditions, but the direction of effect being different in separate conditions, for example, *STEAP4* and *HIP1*. Pathway analysis revealed that many of these eQTLs were enriched for genes involved in the

proteosome, apoptosis, antigen processing. Trans-eQTLs demonstrated a high degree of context specificity, most notably after IFN-g treatment, possibly driven by *cis*-eQTL modulating transcription factors and cytokine release. *Cis*-eQTLs that regulate the abundance of transcription factor abundance may generate trans associations to transcription factor target genes. The authors report a context-specific trans gene hub to the chr 4q35 locus. This SNP demonstrated a strong stimulus specific *cis* association with *IRF2*, as well as trans association with a network of more than 300 genes. To identify the relationship between context specific eQTLs and GWAS, they resolved eQTLs into peaks of association and tested overlap with GWAS loci, finding significant overlap, which suggest that disease risk alleles may manifest activity in a context-specific manner. For example, a *cis*-eQTL was observed to *ATM*, after exposure to IFN-g, and this SNP is in complete LD to the GWAS SNP for metformin response in diabetes (Zhou et al., 2011). In addition, context-specific eQTLs can also help in the interpretation of disease associations in existing GWAS. For example, for the autoimmune disease-associated gene *CARD9*, there was a second independent stimulus-specific eQTL after IFN-g stimulation in addition to an active eQTL in the naïve state (Jostins et al., 2012).

6.2.2 CNVs and gene regulation

So far, CNVs have been shown to bring about a phenotype by different mechanisms. The most direct would be by altering the copy number of the gene that is susceptible to dosage effects. This has been shown to occur in individuals with neurodevelopmental disorders. Duplications and deletions of the same gene can result in different outcome. For example, duplication of the *MECP2* gene results in Lubs X-linked mental retardation while deletion of the same gene results in Rett syndrome (Friez et al., 2006; Qiu et al., 2012)

CNVs in noncoding sequences can also result in long-range regulation of target genes by

altering the copy number of regulators or by disturbing regulatory pathways. While association studies between CNVs and common disease have not demonstrated large effects (Craddock et al., 2010) , an intermediate step may be study the association between CNVs and gene expression. Attempts to map genome-wide copy number variation to gene expression data on a genome wide scale have predominantly involved analysis of lymphoblastoid cell lines (LCLs) (Gamazon, Nicolae, & Cox, 2011; Schlattl et al., 2011; Stranger et al., 2007). Estimates of the contribution CNVs make to the total genetic variation in gene expression range from 13% to 17.7% (Schlattl et al., 2011; Stranger et al., 2007).

Whilst small CNVs associated with variation in gene expression, CNVs regulating gene expression were enriched for those comprising large duplications and deletions, including large intergenic CNVs. (Schlattl et al., 2011). The same study also found that CNVs which intersected genes typically affected expression with a positive correlation between the two, which was consistent with an earlier study (Stranger et al., 2007). In contrast, CNVs in intergenic regions showed a mixture of positive and negative correlations.

However, there were several examples of genes in exonic regions that exhibited a negative correlation between gene expression and copy number. One example was the region of the *SIGLEC5* and *SIGLEC14* gene. The expression of these 2 adjacent genes is affected by a deletion that disrupts both genes, resulting in a gene hybrid. There was a negative correlation between the expression of SIGLEC5 and the copy number genotype of the gene-intersecting deletion (Schlattl et al., 2011). This could possibly be explained by the fusion of *SIGLEC5* with *SIGLEC14* and the subsequent juxtaposition of the fusion gene with the *SIGLEC14* promoter region. Another example was the *ULK1* gene, with a negative correlation between copy number and expression gene. However, in this case, the authors performed read depth

analysis and found that the CNV boundaries were misannotated and the coding regions were not contained inside the CNV (Schlattl et al., 2011).

The authors also identified a dosage compensation effect in several regions. In most cases, one would expect that CNVs overlapping a gene would affect expression levels in a proportional fashion, however, this relationship may be disrupted by feedback-regulation or buffering mechanisms (Deng & Disteche, 2010). For example, less than a third of all genes in chromosome 21 are expressed in the case of trisomy 21 (Aït Yahya-Graison et al., 2007). Schlattl et al assessed this effect by checking whether transcript levels associated with copy number 1 were markedly higher than half of the level of that observed for copy number 2 (Schlattl et al., 2011). They found that 3 loci (*ACOT1, LRP5L* and *ZNF280B*) demonstrated significant dosage compensation. It is possible that different human cell types may have different propensity for dosage compensation as compensatory loops may vary between developmental stages and tissues (Chaignat et al., 2011).

Schlattl et al also found that approximately half of the CNV eQTLs were also SNP eQTLs within the same dataset (Schlattl et al., 2011). In an earlier study, the fraction was 18% (Stranger et al., 2007). The strong association between CNVs overlapping genes and expression could possibly be confounded by CNV genotypes displaying a slightly reduced genotyping error rate compared with SNP genotypes. They focused on proximal effects within a search range of 200 kb rather than distal CNV effects and found that 70% of the CNV-eQTLs were located within 100 kb of the gene, hence CNV effects on expression were significantly enriched with short range signals (Schlattl et al., 2011). Furthermore, there also was an enrichment of duplications amongst CNV-eQTLs.

On problem with using lymphoblastoid cell lines to estimate the contribution of CNVs to gene expression is that they does not follow the influence of CNVs in multiple tissues or developmental stages, since an individual is the source of 1-2 cell lines. This issue can be partially resolved using animal models since CNVs contribute significantly to their phenotypic variation. In some cases, the CNVs were shown to cause a similar phenotype in the animal model and in humans. Aitman et al showed that lower copy number of *FCGR3* gene predisposed both rats and humans to glomerulonephritis (Aitman et al., 2006). Another advantage of animal models is that CNVs can be engineered into their genome, which allows for the study of different CNV alleles, as well as position effects in a background that is genetically homogeneous (Molina et al., 2008). Two recent reports that assessed the effects of CNVs on the transcriptomes of rats and mice reported an over-representation of differentially expressed genes, with a positive correlation between relative expression level and gene dosage, which was driven by a restriction fraction of the genes (between 5 and 18%) (Guryev et al., 2008; Henrichsen et al., 2009). In about 2/3 of the genes, copy number had no effect on relative expression level in any animal tissues, suggesting dosage compensation mechanisms or incomplete inclusion of regulatory elements in the deletion or duplication event. Genetic imprinting may also modulate the expression at a CNV locus in a parental allele-dependent manner. Additionally, multiple copies of a gene might map to different chromatin environments and many hence be regulated in different ways. For 2-15% of the genes examined, relative expression levels were inversely correlated with the copy numbers of nearby CNVs (Henrichsen et al., 2009). There may be 2 possible reasons for this. Firstly, the presence of immediate early genes (IEG) may account for this. These genes are initially expressed at levels proportional to the number of copies. Subsequently, these IEG can directly or indirectly induce the expression of a repressor which reduces or abolishes the expression of the CNV gene. The second possible reason is that extra copies of a gene impairs their access

to a specific transcription factor through steric hindrance (Sexton, Umlauf, Kurukuti, & Fraser, 2007).

6.2.3 MatrixeQTL

eQTL analysis can be computationally expensive. For most eQTL datasets, millions of SNPs have been genotyped, and several tens of thousand gene expression probes have been assayed, resulting in more than a billion tests. One study that applied family based association tests to 142 samples, testing 40227 transcript-SNP pairs reported the total computation time to be under 24 hours on 20 processors in parallel. Yet another study running mixed model association on a dataset of 110 individuals report a computation time of 'a few hours' using a cluster of 50 processors. Listgarten tested their method on a larger dataset, with 40639 transcripts and 48186 SNPs. They report that the estimation took 10 hours across 1100 processors.

However, MatrixeQTL has been developed for fast eQTL analysis of large datasets, and is 2-3 times faster than existing software, even when covariates are added to the model.

Matrix eQTL optimises the analysis in various ways. The conventional method for conducting a linear regression model involves calculating several variables, including the slope coefficient, intercept, residuals and sum of squares, and subsequently a test statistic such as likelihood ratio or Z score. The p value can then be calculated from the test statistic, and can be expensive as it involves the calculation of incomplete beta functions. Matrix eQTL does not calculate p values for each transcript-SNP pair. Instead, a threshold is found for the test statistic selected for a specified significance level, and test statistics which exceed this

threshold are then taken forward and p values are calculated for these. In this way, the number of p values needed to be calculated is fewer than the number of comparisons made. Another way in which Matrix eQTL gains efficiency is by expressing the most computationally intensive part of the algorithm in terms of large matrix operations. The most expensive operation is matrix multiplication, which depends on the performance of the basic linear algebra subroutine (BLAS). The authors compared the performance of Matlab and different versions of R. The standard installation of R uses a generic BLAS that is not optimised for any central processing unit (CPU). Faster versions of BLAS include ATLAS or GOTO, but they only work for 32 bit and 64 bit respectively. These versions are improvements on generic BLAS – a test performed by the authors took 98s to run on generic BLAS but 16.5s on ATLAS. Matlab and Revolution R (a commercial version of R) are also improvements on the standard BLAS. They both employ Intel Kernel Math Library, which is optimised for intel CPUs.

Matrix eQTL can test associations through linear additive models or categorical models. It can also test for the significance of interactions between covariates and genotypes. Furthermore, Matrix eQTL can take continuous values as genotype input,and can include covariates in the model, such principal components to account for population structure, age and gender. Some relatedness can be accounted for through heteroscedastic and correlated errors. The problem of multiple testing is dealt with through the calculation of false discovery rates through the Benjamini Hochberg correction, with different thresholds for *cis* and trans-eQTLs.

The authors compared various different eQTL analysis software (Plink, Merlin, R/qtl, snpMatrix, eMap and FastMap). Plink is written in C/C++ but not optimised for eQTL

analysis. Merlin is written in C/C++ and is designed for analysis in pedigrees. R/qtl, snpMatrix and eMAp are R packages. FastMap is written in Java and is optimised for permutation-based testing. Fastmap and eMap do not incorporate covariates within the model. These methods were applied to a dataset containing 2201 genes and 57333 SNPs across 840 samples. They found that matrix eQTL took 9-12s to analyse the test dataset, which was 50 times faster than FastMap, the second fastest software. In third place was R/qtl, which was more than 100 times slower than Matrix eQTL. Plink, Merlin and eMap were more than a 1000 times slower than Matrix eQTL.

6.2.4 Chapter Aims

There has been limited analysis of the effect of CNVs on primary cell gene expression. It is clear that eQTLs demonstrate a high degree of cellular-specificity (Castaldi et al., 2014; Fairfax et al., 2012; Westra & Franke, 2014). Similarly, it is increasingly recognised that eQTL display context-specificity, (Fairfax et al., 2014).

To this end, we investigated the effect of CNVs on cell specific and context specific regulation of gene expression. CNVs were identified from the genotyping data of Caucasian volunteers using an integrated hidden Markov model PennCNV (K. Wang et al., 2007), which incorporates several sources of information, including allelic intensity ratio, distance between SNPs, SNP allele frequency and total signal intensity. Copy number was then associated with genome-wide gene expression profiles of B cells, monocytes, as well as monocytes stimulated with LPS and IFN-g from the same individuals. LPS is a component of gram-negative cell walls that triggers toll-like receptor 4 (TLR4) signalling (Medzhitov, 2001) and IFNg is a cytokine that acts through the Janus kinase-signal transducer and activator of

transcription (JAK-STAT) pathway (Manry et al., 2011). It has anti-bacterial, anti-parasitic and anti-fungal properties, and mutations in genes involving the IFNg circuit confer a Mendelian predisposition to mycobacterial disease (Filipe-Santos et al., 2006).

6.3 Methods and Materials

This study was approved by the Oxfordshire Research Ethics Committee (COREC reference 06/Q1605/55). A total of 432 volunteers of European ancestry were recruited in the Oxfordshire area after written informed consent. The cohort consisted of 194 males and 238 females with a median age of 30 years (mean, 33 years; interquartile range, 19 to 56 years). Detailed experimental methods are available in Fairfax et al (2012) and Fairfax et al (2014). Briefly, peripheral blood mononuclear cells (PBMCs) were purified by the Ficoll gradient method. Magnetic activating cell sorting methods were used to positively separate CD14 and CD19 positive cells as purified monocyte and B-cell populations respectively. Purity of samples was confirmed with flow cytometry, and was 90 to 95% for B cells and 99% for monocytes. Genomic DNA was extracted from 1-2 ml whole blood following the manufacturer's instruction (Gentra Puregene Blood kit) (Qiagen). Total RNA was extracted using the RNeasy mini kit (Qiagen). Genome wide gene expression profiling was performed with HumanHT-12 v4 BeadChips (Illumina), total of 47231 probes. Genotyping was performed with HumanOmniExpress-12v1.0 BeadChips (Illumina), total of 733202 SNPs.

6.3.1 Stimulation Assays

4×10^5 monocytes were stimulated with ultrapure LPS (20 ng/ml) (catalog # tlrl-pelps, Invivogen) for either 2 or 24 hours. Alternatively, monocytes were stimulated with IFN-γ (catalog # 285-IF, R&D Systems) at a concentration of 20 ng/ml for 24 hours. The

experiments were terminated simultaneously, so that all samples had identical incubation periods. A subset of samples was kept in the incubator without stimulation throughout this period as a control to ascertain incubator effects. All assays were performed within 48 hours of blood sample collection.

6.3.2 Gene Expression Analysis

Total RNA from each sample was quantified using the Illumina HumanHT-12 v4 BeadChip gene expression array. Arrays were read, and data were exported using BeadStudio software (Illumina). Samples were subsequently processed in R using the packages lumi, limma, and ComBat. PCA was performed to identify samples within each treatment cohort with outlier expression for removal from further analyses. Raw data were transformed and normalized using robust spline normalization within lumi. For monocyte samples run across 3 batches ComBat was used to regress out batch effects. A linear model incorporating data from 59 untreated, incubated samples and 421 untreated, non-incubated samples was used to estimate the effect of incubation on gene expression. This was then regressed from all treated samples to minimize the influence of incubator effects on the analysis. In total 414 monocyte samples were analysed in the naïve state, 367 samples after exposure to IFN-γ 322 individuals after 24-hour LPS and 263 samples after 2h LPS. Data across all four conditions were available from 228 individuals, permitting cross-treatment comparison. B-cell expression data was available for 288 samples.

6.3.3 CNV Calling

CNV calls were generated with PennCNV Software (Version 5.14.0, June 2011) (K. Wang et al., 2007). The PennCNV algorithm uses a hidden markov model to identify copy number on an individual basis by considering BAF and LRR calculated using GenomeStudio, with the

standard hg18 “all” PennCNV HMM and population frequency of B allele (pfb) files. We implemented the GC model wave adjustment procedure according to PennCNV instructions (K. Wang et al., 2007).

QuantiSNP was used to confirm the calls (Stefano Colella et al., 2007). This also uses a hidden Markov model as its underlying algorithm. We included CNV calls if they were called by both QuantiSNP as well as PennCNV. PennCNV checks were used to exclude samples that failed quality control. We removed samples that had a LRR standard deviation >0.28, BAF median>0.55 or <0.45, BAF drift >0.002 or waviness factor>0.04 or <-0.04. Two samples that failed quality control after the wave adjustment procedure were removed. This resulted in a final sample size of 278 individuals. Due to the complications of hemizygosity in males and X chromosome inactivation in females, all analyses were restricted to autosomes (Need et al., 2009). In addition, we excluded CNVs for which the difference of the log likelihood of the most likely copy number state and least likely copy number state was less than 10. Plink was used to identify overlapping unique segments of CNVs (Purcell et al., 2007).

6.3.4 Probe Selection

Probes were normalized and corrected for batch and incubator effects as previously described (Fairfax et al., 2014). Basic Local Alignment Search (BLAST) was performed and probes that mapped to more than one genomic location were excluded. A further 6137 probes that overlapped regions with SNPs were excluded, as well as probes that mapped to non-autosomal locations. The remaining probes were included if they were present with detection P value of <0.01 in more than 2.5% of samples from a specific treatment or more than 5% of

all samples. These quality control steps resulted in a total of 15421 probes for CNV association analysis.

6.3.5 CNV association analysis

Association analysis was performed using matrix eQTL (Shabalin, 2012). PCA was performed, and the PCs were used as covariates in the model. Based on previous investigations (Fairfax et al., 2014), 30 PCs were used as they maximise the number of significant associations whilst maintaining type 1 error rates. An additive linear model was used, and CNVs were included in the *cis* analysis if they were located within 1 Mb of the expression probe. FDRs were calculated by matrixEQTL using the Benjamini Hochberg procedure (Benjamini & Hochberg, 1995). Associations which had an FDR of <0.1 were deemed to be statistically significant. (Schlattl et al., 2011). In order to compare our findings with SNP eQTL results so as to evaluate the added value of CNV association analysis, we compared our results with the eQTLs obtained from the same dataset as shown in the supplementary materials of Fairfax et al, 2014 (Fairfax et al., 2014). Only eQTLs within a 500 kb distance either side of the CNV were considered. We also tested overlaps between identified CNV-eQTLs and SNPs within the credible sets described in the previous chapter, but none were identified.

6.4 Results

CNVs were identified from the genotyping data of Caucasian volunteers using PennCNV (K. Wang et al., 2007), CNVs were grouped based on segments that span a particular position using the software Plink (Purcell et al., 2007), which takes all segments in a given region and

forms pools of overlapping segments. For each CNV pool, copy number was then associated with genome-wide gene expression profiles determined using HumanHT-12 v4 BeadChip (Illumina) of B cells, monocytes, as well as monocytes stimulated with LPS and IFN-g from the same individuals.

A total of 675 CNV segments were identified from 414 individuals. The minimum CNV length was 1,170 base pairs, with maximum of 145,600 base pairs. The median CNV length was 47960 and the mean was 142700, with the first and third quartiles being 9530 and 206200 respectively.

A linear additive model was fitted with matrixeQTL with gene expression as the outcome and copy number as the predictor (Shabalin, 2012). Only associations where 4 or more individuals with CNVs were reported, corresponding to a CNV frequency > 1%.

A total of 97 unique CNV eQTL association pairs were identified amongst monocytes and B cells, encompassing 86 genes (Appendix Table 4). There were 26 associations for B cells, with 5 of them being in the negative direction. There were 50 associations for IFN-g, with 18 of them being in the negative direction. There were 40 associations for LPS-24h, with 6 of them being in the negative direction. There were 33 associations for LPS-2h, with 8 of them being in the negative direction. There were 41 associations for the naïve state, with 16 of them being in the negative direction. The number of CNV-eQTLs identified in stimulated states but not naïve states were 23, 19 and 18 for IFN, LPS-24h and LPS-2h respectively (Table 6.1)

Table 6.1 Details about CNV-eQTLs identified

	Number of CNV-eQTLs	Number of CNV-eQTLs in negative direction	Number of CNV-eQTLs present in stimulated but not naïve state	Number of CNV-eQTLs with SNP-eQTLs within 500 kb	Number of CNV-eQTLs with SNP-eQTLs within 500 kb that have lower FDR than CNV-eQTL
Bcell	26	5	NA	18	11
IFN	50	18	23	43	9
LPS24	40	6	19	29	4
LPS2	33	8	18	25	5
Naïve	41	16	NA	34	12

To evaluate the disease associations amongst the SNPs contained in our CNV-eQTLs, we downloaded the NHGRI GWAS catalog for genome wide significant SNPs (p value < 5×10^{-8}) and annotated SNPs which were contained within the CNVs with significant associations. 178 GWAS lead SNPs were identified (Appendix Table 5).

In order to compare our findings with SNP eQTL results so as to evaluate the added value of CNV association analysis, we performed SNP-eQTLs analysis using the same linear model and 30 PCs as covariates. Only eQTLs within a 500 kb distance either side of the CNV were considered.

Among the B cell CNV-eQTL associations, 18 of these had SNP-eQTLs identified within 500 kb of the CNV, of which 11 had FDR which was lower for the CNV than for the minimum FDR identified amongst the nearby SNP-eQTLs. Among the IFN CNV-eQTL associations,

43 of these had SNP-eQTLs identified within 500 kb of the CNV, of which 9 had FDR which was lower for the CNV than for the minimum FDR identified amongst the nearby SNP-eQTLs. Among the LPS-24h CNV-eQTL associations, 29 of these had SNP-eQTLs identified within 500 kb of the CNV, of which 4 had FDR which was lower for the CNV than for the minimum FDR identified amongst the nearby SNP-eQTLs. Among the LPS-2h CNV-eQTL associations, 25 of these had SNP-eQTLs identified within 500 kb of the CNV, of which 5 had FDR which was lower for the CNV than for the minimum FDR identified amongst the nearby SNP-eQTLs. Among the naive CNV-eQTL associations, 34 of these had SNP-eQTLs identified within 500 kb of the CNV, of which 12 had FDR which was lower for the CNV than for the minimum FDR identified amongst the nearby SNP-eQTLs. Although it would be ideal to evaluate whether there is significant LD between SNP and CNV, there have been no satisfactory models which take into consideration the CNV-genotyping, phasing and multiple testing issues.

Several disease relevant CNV-gene associations were observed. In four individuals, there was a CNV identified at the chr15q11 region (chr15: 22750305-23272733) that spans 5.2 Mb. Copy number of this CNV associated with the expression of three genes – *NIPA1*, *NIPA2* and *CYFIP1* in both cell types. These genes have both been implicated in a variety of neurological and psychiatric phenotypes (Bozdagi et al., 2012; Chai et al., 2003). In three individuals, there was a deletion (copy number 1). In one individual, there was an insertion (copy number 3). There was a positive association between copy number and gene expression of *NIPA1* and *NIPA2* in naïve cells as well as in all stimulated contexts (NIPA2 $FDR_{B\text{-}cells}$: 4.42×10^{-20}; FDR_{Mono} 1.04×10^{-39}; FDR_{LPS24} 3.96×10^{-34}, $FDR_{IFN\gamma}$ 4.91×10^{-35}), the association being less significant for *NIPA1*. Individuals with this CNV also demonstrated association of expression of the *CYFIP1* gene, also located at chr15q11, in monocytes but not B cells (*CYFIP1*: $FDR_{B\text{-}cells}$ greater than 0.6; FDR_{Mono} 1.34×10^{-34}; FDR_{LPS24} 9.49×10^{-17}). Another example of a

clinically relevant and cell specific CNV-gene association is the relationship between a CNV located at 5q13.2 and expression of NAIP, which is associated with Legionella susceptibility (Diez et al., 2003). This CNV is located at chr5:70305696-70307464 and spans 1.8 kb. There are 9 subjects with copy number 1 and 2 subjects with copy number 3. In monocytes but not B cells, copy number correlates with 2 probes mapping to the *NAIP* gene. The association was present in monocytes (FDR_{Mono} 1.02 x 10^{-25}; FDR_{LPS24} 5.92 x 10^{-12}; FDR_{IFN} 3.53 x 10^{-11}) but not B cells ($FDR_{B\text{-}cells} > 0.6$).

Context specific eQTLs were identified. For example, a CNV located at chr17:14174994-15412809 was associated with expression of *PMP22* in when stimulated with LPS but not in a naïve state (FDR_{LPS24} 0.00993, $FDR_{Mono} > 0.6$). This locus is associated with keratoconus in a GWAS on European and Asian populations (Lu et al., 2013).

A notably strong association across all conditions was observed between a CNV located at chr22: 25658525-25914592 and expression of the *LRP5L* gene both in B cells and monocytes. In monocytes, FDR was 2.29 x 10^{-26}, 3.05 x 10^{-9} and 5.46 x 10^{-5} respectively for naïve, IFNg and LPS24 respectively. For B cells, the FDR was 1.84 x 10^{-21}.

6.5 Discussion

This chapter demonstrates the importance and added value of CNV-eQTL analysis in adding information to GWAS studies. We identified a large number of CNV-eQTLs, some of which were cell specific and context specific. Many of these CNVs overlap GWAS lead SNPs for a variety of important disease phenotypes. Furthermore, a large proportion of these CNV-eQTLs did not have SNP-eQTLs in the vicinity and in the cases of those that did, these SNP-eQTLs had a weaker association than the CNV-eQTLs in a considerable number of cases.

Our results are consistent with other studies. For example, several CNV-eQTL associations identified within monocytes in our study, including to *LRP5L, NSF* and *MSR1*, were also

identified by Blackburn et al in their study on white blood cells among Mexican Americans (Blackburn et al., 2015). We have also observed several significant associations between CNVs and genes that are of potential clinical relevance. Notably a CNV implicated in schizophrenia and autism susceptibility (Kirov et al., 2009) identified at the chr15q11 region was associated with the expression of 2 genes – *CYFIP1* and *NIPA2*. Chromosome 15 contains 5 breakpoints along the long arm (BP1 to BP5), with a cluster of low copy DNA repeats within the region which can facilitate non-allelic homologous recombination. Four genes are located between BP1 and BP2 (*NIPA1*, *NIPA2*, *CYFIP1* and *TUBGCP5*).This replicated data from LCLs where, dependent upon the gene, 24-99% the variance in mRNA for the 4 genes between BP1 and BP2 was attributed to this CNV. In that study the NIPA2 gene accounted for the largest number of significant correlations between mRNA level and phenotype. Our study is novel in that it demonstrates that there is a significant relationship between *NIPA1/NIPA2* and the CNV in both monocytes and B cells, however, the relationship between *CYFIP1* and the CNV exists only in monocytes but not B cells.

Another strong cell specific association was observed between the CNV located at 5q13.2 and expression of *NAIP*. This finding is consistent with that of Boniotto et al (2011), who reported that African populations have a duplication of the *NAIP* gene at a higher frequency than Europeans and Asians which associates with increased transcription of the gene in macrophages (Boniotto et al., 2012). They demonstrated increased *NAIP* protein sensitizes to *Legionella Pneumophila*, increasing cell death and resistance to infection. However, this is the first time that this relationship has been demonstrated to be cell specific and absent in B cells which unlike myeloid derived macrophages are not targets for Legionella as Legionella multiplies in monocytes (Horwitz & Silverstein, 1983). This is relevant in inferring possible mechanisms through which CNVs in this region affect phenotype, such as resistance to infection (Diez et al., 2003).

There are several limitations to this study. Firstly, there is inherent uncertainty in the identification of CNVs and breakpoints. Although both PennCNV and QuantiSNP have both been demonstrated to be relatively accurate in identifying CNVs and localising breakpoints (X. Zhang et al., 2014), the study would certainly benefit from validation using experimental techniques such as array comparative genomic hybridisation (array-CGH). Secondly, it is difficult to differentiate the effects of SNPs and CNVs on gene expression. In some of the CNV-eQTL, there were SNPs within the CNVs which had a lower p value of association with the same gene than the CNV. However, in others, there were no SNPs which had a lower p value of association. The difference in SNP number versus CNV number complicates comparisons of association testing between these two variants. GWAS have demonstrated that CNVs are associated with various phenotypes, with de novo copy number mutations observed in a larger fraction of individuals with autism and schizophrenia compared to unaffected individuals, particularly in the sporadic form of these diseases (Sebat et al., 2007; Xu et al., 2008). Further analysis is needed in large cohorts, looking at both inherited as well as de novo CNVs and correlating with gene expression across various cell types and conditions. This will in doubt be enhanced by large scale whole genome sequencing with parrallel RNA-sequencing in progress, which will assist in determining more accurate detection of CNVs and breakpoints and corresponding expression effects.

6.6 Chapter Summary

This chapter demonstrates the importance and added value of CNV-eQTL analysis in adding information to GWAS studies. We have identified a large number of clinically relevant CNV-eQTLs, some of which were cell specific and context specific. Many of these CNVs overlap GWAS lead SNPs for a variety of important disease phenotypes. Furthermore, a large proportion of these CNV-eQTLs did not have SNP-eQTLs in the vicinity and in the cases of

those that did, these SNP-eQTLs had a weaker association than the CNV-eQTLs. Although they did not overlap the loci identified the previous chapters, they serve as a useful database for further GWAS studies, as a layer of functional annotation that could inform the biological relevance of GWAS lead SNPs.

CHAPTER 7 Conclusion

7.1 Thesis Summary

7.2 Further Work

7.3 Conclusion

7.1 Thesis Summary

In this thesis, I explored the utility of GWAS and meta-analysis to identify variants associated with pollutants and metabolic phenotypes. This is important from a biological point of view because these pollutants have adverse effects on human health and these metabolic phenotypes act as surrogate measures of disease.

I identified variants within the CYP2B6 gene which are associated with various highly bioaccumulative PCBs. This is the first examination in humans of the relationship between genome wide variation and circulating PCB levels. I uncovered novel associations between variants in *CYP2B6* and circulating PCBs. *CYP2B6* has several known alleles which influence its activity and are associated with serum level of pharmaceutical drugs. As such, these associations may be clinically significant as pollutants could possibly compete with the binding sites of these drugs. Additionally, knowledge of these associations could help us to identify individuals who may be more susceptible to pollutant toxicity due to slower metabolism. Another novel association was the relationship between SNPs in the *ABCG2* gene and OCDD. This finding was particularly interesting clinically because the top SNP

(rs2231142) was a lead SNP in several GWAS on gout, which may suggest that metabolic pathways involved in uric acid clearance may be involved in OCDD metabolism.

This thesis also aims to identify variants associated with metabolic traits which are markers of cardiovascular health. I identified novel associations between D6 and D5 desaturase levels and SNPs in the *FADS2* and *FADS1* loci respectively. These enzymes play a crucial role in the endogenous elongation and desaturation of fatty acid precursors into physiologically active long chain PUFAs. The lead SNP for D5D (rs174549) is associated with several phenotypes in 3 different GWAS studies – heart rate (den Hoed et al., 2013), glycerophosphocholine levels (Hong et al., 2013), compressive strength and appendicular lean mass (Y. Han et al., 2012). In this context, MR could potentially be used to establish the relationship between these phenotypes and cardiovascular risk. The use of MR is discussed later.

This study demonstrates that GWAS was successful in identifying strong association signals in several regions, some of which show a clear biological basis while others are less obvious. An example of the former is the association between manganese and *SLC39A8*, which transports manganese and zinc. This is a non-synonymous mutation, with an amino acid change from alanine to threonine. The same analysis has also revealed another region of association in *SLC30A10*. However, the biological link is not clear. Although *SLC30A10* is an ion transporter, the lead SNP, rs1776029, lies within the intronic region and there are no non-synonymous mutations which are in strong LD with it. Furthermore, there are other identified variants which are in regions that do not appear to have a direct biological link, such as cadmium and mercury.

Although this study did not reveal allelic heterogeneity in the form of secondary signals, I revealed signals when conditioning on a phenotype in the case of the *Cyp2b6* locus. Because of the high correlation between serum pollutant levels, I dissected the effect of *Cyp2B6* on PCB9 using reciprocal conditional phenotype analyses. After adjusting for PCB-99 as an additional covariate in the regression model, the previously observed associations of rs8109848 with PCB-138 and PCB-153 were extinguished, which may reflect pleiotropic effects on serum levels of the pollutants. To further elucidate the biological mechanisms for these association patterns, it would be necessary to carry out functional studies, which might include biochemical assays and toxico-kinetic studies.

To fine map these loci and assign functional annotation, I created Bayesian credible sets and checked for overlap between SNPs in these credible sets and regulatory marks in the ENCODE database. This was informative for some of the phenotypes but not others (credible sets too large or small). Annotating these credible set SNPs with functional information revealed various histone modifications and transcription factors that overlapped, however, it is not easy to understand whether these regulatory marks are part of the biological mechanism for the phenotype or whether they are unrelated.

This study has also been successful in identifying CNVs from the PIVUS cohort. Many of these common CNVs have been reported in other investigations, demonstrating consistency. There were moderately strong associations between CNVs and some of the phenotypes studied. These associations did not appear to be mediated by the SNP within the CNVs, as the latter had higher P values of associations.

In addition, I identified several clinically relevant CNV-eQTL associations in a separate

cohort of healthy individuals. Some of these associations were cell specific and/or context specific. Many of these CNVs contained SNPs which are lead SNPs in GWAS studies on a variety of different phenotypes. The added value of CNV analysis over SNP analysis was evaluated by identifying SNP-eQTLs in the vicinity of CNV-eQTLs and comparing the strength of association. I found that in the case of B cells, copy number had a stronger association with gene expression than SNPs in the vicinity, with the exception of one gene. In the case of monocytes, 9 out of the 49 CNV-eQTLs had stronger association than nearby SNPs, demonstrating that CNV analysis adds considerable value to the study, and may help to explain the genetic basis of certain phenotypes which are not explained by SNPs.

7.2 Further Directions for Research

Many of the identified signs were monogenic, and have the potential to be utilised as instrumental variables in MR to establish the relationship between exposure patterns and disease. MR has been used to demonstrate the relationship between endometrial cancer and estradiol (E2) concentrations. The SNP rs727479 was associated with circulating E2 concentrations in post-menopausal controls. In addition, the observed endometrial cancer odds ratio for this SNP is compatible with that predicted by the observed effect on E2 concentrations, which suggest that endometrial cancer is causally linked to E2 (D. J. Thompson et al., 2016). MR has also been used to establish the relationship between adiponectin and cardiovascular mortality through rs822354 (Ortega Moreno et al., 2016). Because some of these pollutant levels may be associated with adverse health outcomes such as cancer, MR has the potential to be used in identifying causal relationships between the two. Furthermore, some of the serum biomarkers I studied are associated with associated with cardiovascular outcomes, MR could be used to analyse the causality of these associations.

One limitation of our study was that I did not replicate our results in an independent cohort. For most of the metabolic traits, the lead SNPs in our study have also been identified in previous GWAS studies on other populations. However, for the pollutant traits, I lack other independent cohorts with similar phenotypes to replicate our results with. In general, replicating results for GWAS studies can be difficult for various reasons. Firstly, imputing genotypes may not be straightforward for rare variants, which may fail quality control. Secondly, for many complex phenotypes, effect size is small, with the cumulative variance explained by SNPs being moderate, requiring large sample sizes for detection. Obtaining large number of study subjects for independent replication can be difficult. Thirdly, some of the associations may be specific to certain populations, and may not be replicated in populations of different ethnicity.

Another limitation of our study and GWAS in general is difficulty in assigning biological functions to SNPs, especially those in non-coding regions. I tried to elucidate more information on the biological basis for these associations by firstly, fine mapping them using a Bayesian approach to constructing credible sets of SNPs, then annotating these sets of SNPs with data from the ENCODE database. Constructing credible sets was successful for some phenotypes, such as manganese and D6 desaturase, but less successful for others, such as cadmium, in which the association signal was weak hence the credible set contained more than a thousand SNPs. Similarly, annotating these variants with information from the ENCODE database revealed some interesting results, with several regions overlapping with known areas of histone modifications and transcription factor binding sites. However, it is not conclusive as to whether the biological effects are mediated through these epigenetic mechanisms. Better and more comprehensive annotation for regulatory elements would

certainly be helpful, with further experiments that include a wider variety of cell types and transcription factors. Various bioinformatics tools have been developed to interrogate these functional elements. I used ANNOVAR, which assigns functional annotation as well as an overlap score but does not prioritise the data or visualise it. Visualising it may help researchers to gain a more intuitive understanding of how these functional elements interact with SNPs of interest.

GWASrap has recently been developed to visualise the annotation of these results and generate statistics, including variant distribution in different functional categories (Mulin, Pak, & Wang, 2013). The software also includes functional predictions for various features, with predicted changes caused by each genetic variant using an additive model. It first estimates the likelihood of disease association for each type of SNP by mapping it to HapMap3. After that, it computes the variant's genomic mapping score and functional prediction score based on the annotation information. The prioritization core is then computed based on the product of these two scores. The authors tested this method by applying it to disease causing SNPs from the OMIM database, revealing prioritization scores that were significantly different from random sampled background SNP. They went on to apply the method to a GWAS on bipolar disorder, obtaining a stronger signals in several highly associated SNPs after the prioritization step.

In terms of CNV detection, one limitation is the relative uncertainty of CNV breakpoint identification. Besides quantiSNP and PennCNV, various different software have been developed for detecting CNVs using GWAS data. Most of these software generally use hidden markov models or circular binary segmentation as the underlying algorithm and have varying performance standards. Zhang et al compared four of these programmes – Partek,

Birdsuite, PennCNV-Affy and HelixTree (Korn et al., 2008; D. Zhang et al., 2011). They evaluated recovery rate in both common CNVs (CNVs identified in Hapmap samples whole-genome sequencing and array CGH) as well as rare CNVs (CNVs in the Bipolar Genome Study). They assessed accuracy of CNV identification for by positive predictive value, based on validation through quantitative PCR, and found that Birdsuite recovered the highest percentage of common CNVs. However, Birdsuite and Partek performed better for rare CNVs. Ideally, CNVs identified from GWAS arrays should be confirmed using more reliable techniques such as array CGH and whole genome sequencing.

There are various challenges in evaluating the association of CNVs and SNPs with phenotypes or gene expression. Whether CNVs are largely in LD with SNPs is a topic that has been debated at length. A large analysis examining common CNVS found that they were almost as well tagged as SNPs of the same frequency (McCarroll, Kuruvilla, et al., 2008), suggesting that disease associations of many common CNVs could be assessed through SNP data. For example, a common deletion polymorphism upstream of *IRGM* was associated with Crohn's disease as it was in perfect LD with SNPs at *IRGM* (McCarroll, Huett, et al., 2008). However, it is not straightforward to quantify the LD between CNVs and SNPs. Various different measures have been proposed (Wineinger, Pajewski, & Tiwari, 2011), however none of them are able to fully account for CNV genotype, phasing and multiple testing.

7.3 Conclusion

In conclusion, this thesis identified several novel GWAS associations, confirmed various

known associations and uncovered several interesting CNV-eQTL associations. These fulfil the original aims of the thesis, which were to use GWAS and meta-analysis to identify the genetic basis for complex traits. In addition, I also address the second aim of this thesis by creating credible sets and annotating them with regulatory information from the ENCODE, eQTL and GWAS lead SNP databases. Further study is needed to confirm these findings and identify biological mechanisms through which they exert their effects.

Reference

Aalto-Setälä, K., Fisher, E. A., Chen, X., Chajek-Shaul, T., Hayek, T., Zechner, R., . . . Breslow, J. L. (1992). Mechanism of hypertriglyceridemia in human apolipoprotein (apo) CIII transgenic mice. Diminished very low density lipoprotein fractional catabolic rate associated with increased apo CIII and reduced apo E on the particles. *J Clin Invest, 90*(5), 1889-1900. doi:10.1172/JCI116066

Abbott, R. D., Wilson, P. W., Kannel, W. B., & Castelli, W. P. (1988). High density lipoprotein cholesterol, total cholesterol screening, and myocardial infarction. The Framingham Study. *Arteriosclerosis, 8*(3), 207-211.

Abecasis, G. R., Auton, A., Brooks, L. D., DePristo, M. A., Durbin, R. M., Handsaker, R. E., . . . Consortium, G. P. (2012). An integrated map of genetic variation from 1,092 human genomes. *Nature, 491*(7422), 56-65. doi:10.1038/nature11632

Aboagye, E. O., & Bhujwalla, Z. M. (1999). Malignant transformation alters membrane choline phospholipid metabolism of human mammary epithelial cells. *Cancer Res, 59*(1), 80-84.

Abyzov, A., Urban, A. E., Snyder, M., & Gerstein, M. (2011). CNVnator: an approach to discover, genotype, and characterize typical and atypical CNVs from family and population genome sequencing. *Genome Res, 21*(6), 974-984. doi:10.1101/gr.114876.110

Adams-Cioaba, M. A., & Min, J. (2009). Structure and function of histone methylation binding proteins. *Biochem Cell Biol, 87*(1), 93-105. doi:10.1139/O08-129

Affymetrix. http://www.affymetrix.com/estore/browse/staticHtmlContentTemplate.jsp?staticHtmlMediaId=m821208&isHtmlStatic=true&navMode=35810&aId=productsNav.

Agatha, G., Häfer, R., & Zintl, F. (2001). Fatty acid composition of lymphocyte membrane phospholipids in children with acute leukemia. *Cancer Lett, 173*(2), 139-144.

Agatha, G., Voigt, A., Kauf, E., & Zintl, F. (2004). Conjugated linoleic acid modulation of cell membrane in leukemia cells. *Cancer Lett, 209*(1), 87-103. doi:10.1016/j.canlet.2003.11.022

Agilent. http://www.chem.agilent.com/cag/aboutapg/abouthistory.html.

Aguie, G. A., Rader, D. J., Clavey, V., Traber, M. G., Torpier, G., Kayden, H. J., . . . Castro, G. (1995). Lipoproteins containing apolipoprotein B isolated from patients with abetalipoproteinemia and

homozygous hypobetalipoproteinemia: identification and characterization. *Atherosclerosis, 118*(2), 183-191.

Aitman, T. J., Dong, R., Vyse, T. J., Norsworthy, P. J., Johnson, M. D., Smith, J., . . . Cook, H. T. (2006). Copy number polymorphism in Fcgr3 predisposes to glomerulonephritis in rats and humans. *Nature, 439*(7078), 851-855. doi:10.1038/nature04489

Akhtar, M. M., Scala, G., Cocozza, S., Miele, G., & Monticelli, A. (2013). CpG islands under selective pressure are enriched with H3K4me3, H3K27ac and H3K36me3 histone modifications. *BMC Evol Biol, 13*, 145. doi:10.1186/1471-2148-13-145

Akin, L., Adal, E., Akin, M. A., & Kurtoglu, S. (2012). A case of diabetes mellitus associated with Rett syndrome. *J Pediatr Endocrinol Metab, 25*(1-2), 197-198. doi:10.1515/jpem.2011.337

Allfrey, V., Faulkner, R., & Mirsky, A. (1964). Acetylation and methylation of histones and their possible role in the regulation of RNA synthesis. *Proc Natl Acad Sci U S A, 51*, 786-794.

Alvim, R. O., Freitas, S. R., Ferreira, N. E., Santos, P. C., Cunha, R. S., Mill, J. G., . . . Pereira, A. C. (2010). APOE polymorphism is associated with lipid profile, but not with arterial stiffness in the general population. *Lipids Health Dis, 9*, 128. doi:10.1186/1476-511X-9-128

Antonini, J. M., Santamaria, A. B., Jenkins, N. T., Albini, E., & Lucchini, R. (2006). Fate of manganese associated with the inhalation of welding fumes: potential neurological effects. *Neurotoxicology, 27*(3), 304-310. doi:10.1016/j.neuro.2005.09.001

Aouizerat, B. E., Vittinghoff, E., Musone, S. L., Pawlikowska, L., Kwok, P. Y., Olgin, J. E., & Tseng, Z. H. (2011). GWAS for discovery and replication of genetic loci associated with sudden cardiac arrest in patients with coronary artery disease. *BMC Cardiovasc Disord, 11*, 29. doi:10.1186/1471-2261-11-29

Armitage, P. (1955). Tests for linear trends in proportions and frequencies. *Biometrics, 11*, 375–386.

Armutcu, F., Akyol, S., Ucar, F., Erdogan, S., & Akyol, O. (2013). Markers in nonalcoholic steatohepatitis. *Adv Clin Chem, 61*, 67-125.

Aslibekyan, S., Goodarzi, M. O., Frazier-Wood, A. C., Yan, X., Irvin, M. R., Kim, E., . . . Arnett, D. K. (2012). Variants identified in a GWAS meta-analysis for blood lipids are associated with the lipid response to fenofibrate. *PLoS One, 7*(10), e48663. doi:10.1371/journal.pone.0048663

Aït Yahya-Graison, E., Aubert, J., Dauphinot, L., Rivals, I., Prieur, M., Golfier, G., . . . Potier, M. C. (2007). Classification of human chromosome 21 gene-expression variations in Down syndrome: impact on disease phenotypes. *Am J Hum Genet, 81*(3), 475-491. doi:10.1086/520000

Babron, M. C., de Tayrac, M., Rutledge, D. N., Zeggini, E., & Génin, E. (2012). Rare and low frequency variant stratification in the UK population: description and impact on association tests. *PLoS One, 7*(10), e46519. doi:10.1371/journal.pone.0046519

Bahlo, M., Stankovich, J., Danoy, P., Hickey, P. F., Taylor, B. V., Browning, S. R., . . . (ANZgene), A. a. N. Z. M. S. G. C. (2010). Saliva-derived DNA performs well in large-scale, high-density single-nucleotide polymorphism microarray studies. *Cancer Epidemiol Biomarkers Prev, 19*(3), 794-798. doi:10.1158/1055-9965.EPI-09-0812

Baniahmad, A., Steiner, C., Köhne, A. C., & Renkawitz, R. (1990). Modular structure of a chicken lysozyme silencer: involvement of an unusual thyroid hormone receptor binding site. *Cell, 61*(3), 505-514.

Bannister, A. J., Zegerman, P., Partridge, J. F., Miska, E. A., Thomas, J. O., Allshire, R. C., & Kouzarides, T. (2001). Selective recognition of methylated lysine 9 on histone H3 by the HP1 chromo domain. *Nature, 410*(6824), 120-124. doi:10.1038/35065138

Barrett, J. C., Clayton, D. G., Concannon, P., Akolkar, B., Cooper, J. D., Erlich, H. A., . . . Consortium, T. D. G. (2009). Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. *Nat Genet, 41*(6), 703-707. doi:10.1038/ng.381

Bartke, T., Vermeulen, M., Xhemalce, B., Robson, S. C., Mann, M., & Kouzarides, T. (2010). Nucleosome-interacting proteins regulated by DNA and histone methylation. *Cell, 143*(3), 470-484. doi:10.1016/j.cell.2010.10.012

Beasley, T. M., Erickson, S., & Allison, D. B. (2009). Rank-based inverse normal transformations are increasingly used, but are they merited? *Behav Genet, 39*(5), 580-595. doi:10.1007/s10519-009-9281-0

Begum, F., Ghosh, D., Tseng, G. C., & Feingold, E. (2012). Comprehensive literature review and statistical considerations for GWAS meta-analysis. *Nucleic Acids Res, 40*(9), 3777-3784. doi:10.1093/nar/gkr1255

Behr, R., Brestelli, J., Fulmer, J. T., Miyawaki, N., Kleyman, T. R., & Kaestner, K. H. (2004). Mild nephrogenic diabetes insipidus caused by Foxa1 deficiency. *J Biol Chem, 279*(40), 41936-41941. doi:10.1074/jbc.M403354200

Bell, C. G., Finer, S., Lindgren, C. M., Wilson, G. A., Rakyan, V. K., Teschendorff, A. E., . . . Consortium, I. T. D. q. (2010). Integrated genetic and epigenetic analysis identifies haplotype-specific methylation in the FTO type 2 diabetes and obesity susceptibility locus. *PLoS One, 5*(11), e14040. doi:10.1371/journal.pone.0014040

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological), 57*(1), 289-300.

Benn, M., Nordestgaard, B. G., Jensen, G. B., & Tybjaerg-Hansen, A. (2007). Improving prediction of ischemic cardiovascular disease in the general population using apolipoprotein B: the Copenhagen City Heart Study. *Arterioscler Thromb Vasc Biol, 27*(3), 661-670. doi:10.1161/01.ATV.0000255580.73689.8e

Best, J. D., & Carey, N. (2010). Epigenetic opportunities and challenges in cancer. *Drug Discov Today, 15*(1-2), 65-70. doi:10.1016/j.drudis.2009.10.010

Bickhart, D. M., & Liu, G. E. (2014). The challenges and importance of structural variation detection in livestock. *Front Genet, 5*, 37. doi:10.3389/fgene.2014.00037

Blackburn, A., Almeida, M., Dean, A., Curran, J. E., Johnson, M. P., Moses, E. K., . . . Göring, H. H. (2015). Effects of copy number variable regions on local gene expression in white blood cells of Mexican Americans. *Eur J Hum Genet*. doi:10.1038/ejhg.2014.280

Boekholdt, S. M., Kuivenhoven, J. A., Hovingh, G. K., Jukema, J. W., Kastelein, J. J., & van Tol, A. (2004). CETP gene variation: relation to lipid parameters and cardiovascular risk. *Curr Opin Lipidol, 15*(4), 393-398.

Boeva, V., Zinovyev, A., Bleakley, K., Vert, J. P., Janoueix-Lerosey, I., Delattre, O., & Barillot, E. (2011). Control-free calling of copy number alterations in deep-sequencing data using GC-content normalization. *Bioinformatics, 27*(2), 268-269. doi:10.1093/bioinformatics/btq635

Boniotto, M., Tailleux, L., Lomma, M., Gicquel, B., Buchrieser, C., Garcia, S., & Quintana-Murci, L. (2012). Population variation in NAIP functional copy number confers increased cell death upon Legionella pneumophila infection. *Hum Immunol, 73*(2), 196-200. doi:10.1016/j.humimm.2011.10.014

Bowler, R. M., Gocheva, V., Harris, M., Ngo, L., Abdelouahab, N., Wilkinson, J., . . . Roels, H. A. (2011). Prospective study on neurotoxic effects in manganese-exposed bridge construction welders. *Neurotoxicology, 32*(5), 596-605. doi:10.1016/j.neuro.2011.06.004

Boyle, A. P., Hong, E. L., Hariharan, M., Cheng, Y., Schaub, M. A., Kasowski, M., . . . Snyder, M. (2012). Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res, 22*(9), 1790-1797. doi:10.1101/gr.137323.112

Bozdagi, O., Sakurai, T., Dorr, N., Pilorge, M., Takahashi, N., & Buxbaum, J. D. (2012). Haploinsufficiency of Cyfip1 produces fragile X-like phenotypes in mice. *PLoS One, 7*(8), e42422. doi:10.1371/journal.pone.0042422

Brautbar, A., Covarrubias, D., Belmont, J., Lara-Garduno, F., Virani, S. S., Jones, P. H., . . . Ballantyne, C. M. (2011). Variants in the APOA5 gene region and the response to combination therapy with statins and fenofibric acid in a randomized clinical trial of individuals with mixed dyslipidemia. *Atherosclerosis, 219*(2), 737-742. doi:10.1016/j.atherosclerosis.2011.08.015

Brooks-Wilson, A., Marcil, M., Clee, S. M., Zhang, L. H., Roomp, K., van Dam, M., . . . Hayden, M. R. (1999). Mutations in ABC1 in Tangier disease and familial high-density lipoprotein deficiency. *Nat Genet, 22*(4), 336-345. doi:10.1038/11905

Brown, J. F. (1994). Determination of PCB Metabolic, Excretion, and Accumulation Rates for Use as Indicators of Biological Response and Relative Risk. *Environ Sci Technol, 28*(13), 2295-2305. doi:10.1021/es00062a013

Brown, M. S., & Goldstein, J. L. (1979). Receptor-mediated endocytosis: insights from the lipoprotein receptor system. *Proc Natl Acad Sci U S A, 76*(7), 3330-3337.

Browning, B. L., & Browning, S. R. (2008). Haplotypic analysis of Wellcome Trust Case Control Consortium data. *Hum Genet, 123*(3), 273-280. doi:10.1007/s00439-008-0472-1

Bruenig, D., White, M. J., Young, R. M., & Voisey, J. (2014). Subclinical psychotic experiences in healthy young adults: associations with stress and genetic predisposition. *Genet Test Mol Biomarkers, 18*(10), 683-689. doi:10.1089/gtmb.2014.0111

Bullaughey, K., Chavarria, C. I., Coop, G., & Gilad, Y. (2009). Expression quantitative trait loci detected in cell lines are often present in primary tissues. *Hum Mol Genet, 18*(22), 4296-4303. doi:10.1093/hmg/ddp382

Buratti, F. M., D'Aniello, A., Volpe, M. T., Meneguz, A., & Testai, E. (2005). Malathion bioactivation in the human liver: the contribution of different cytochrome p450 isoforms. *Drug Metab Dispos, 33*(3), 295-302. doi:10.1124/dmd.104.001693

Cai, Q., Long, J., Lu, W., Qu, S., Wen, W., Kang, D., . . . Zheng, W. (2011). Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum Mol Genet, 20*(24), 4991-4999. doi:10.1093/hmg/ddr405

Carrera, N., Arrojo, M., Sanjuán, J., Ramos-Ríos, R., Paz, E., Suárez-Rama, J. J., . . . Costas, J. (2012). Association study of nonsynonymous single nucleotide polymorphisms in schizophrenia. *Biol Psychiatry, 71*(2), 169-177. doi:10.1016/j.biopsych.2011.09.032

Carter, N. P. (2007). Methods and strategies for analyzing copy number variation using DNA microarrays. *Nat*

Genet, 39(7 Suppl), S16-21. doi:10.1038/ng2028

Carty, C. L., Johnson, N. A., Hutter, C. M., Reiner, A. P., Peters, U., Tang, H., & Kooperberg, C. (2012). Genome-wide association study of body height in African Americans: the Women's Health Initiative SNP Health Association Resource (SHARe). *Hum Mol Genet, 21*(3), 711-720. doi:10.1093/hmg/ddr489

Carvalho, B., Ouwerkerk, E., Meijer, G. A., & Ylstra, B. (2004). High resolution microarray comparative genomic hybridisation analysis using spotted oligonucleotides. *J Clin Pathol, 57*(6), 644-646.

Caspi, A., Williams, B., Kim-Cohen, J., Craig, I. W., Milne, B. J., Poulton, R., . . . Moffitt, T. E. (2007). Moderation of breastfeeding effects on the IQ by genetic variation in fatty acid metabolism. *Proc Natl Acad Sci U S A, 104*(47), 18860-18865. doi:10.1073/pnas.0704292104

Castaldi, P. J., Cho, M. H., Zhou, X., Qiu, W., Mcgeachie, M., Celli, B., . . . Hersh, C. P. (2014). Genetic control of gene expression at novel and established chronic obstructive pulmonary disease loci. *Hum Mol Genet*. doi:10.1093/hmg/ddu525

Castelli, W. P., Cooper, G. R., Doyle, J. T., Garcia-Palmieri, M., Gordon, T., Hames, C., . . . Vicic, W. J. (1977). Distribution of triglyceride and total, LDL and HDL cholesterol in several populations: a cooperative lipoprotein phenotyping study. *J Chronic Dis, 30*(3), 147-169.

Chai, J. H., Locke, D. P., Greally, J. M., Knoll, J. H., Ohta, T., Dunai, J., . . . Nicholls, R. D. (2003). Identification of four highly conserved genes between breakpoint hotspots BP1 and BP2 of the Prader-Willi/Angelman syndromes deletion region that have undergone evolutionary transposition mediated by flanking duplicons. *Am J Hum Genet, 73*(4), 898-925. doi:10.1086/378816

Chaignat, E., Yahya-Graison, E. A., Henrichsen, C. N., Chrast, J., Schütz, F., Pradervand, S., & Reymond, A. (2011). Copy number variation modifies expression time courses. *Genome Res, 21*(1), 106-113. doi:10.1101/gr.112748.110

Champagne, K. S., & Kutateladze, T. G. (2009). Structural insight into histone recognition by the ING PHD fingers. *Curr Drug Targets, 10*(5), 432-441.

Chasman, D. I., Giulianini, F., MacFadyen, J., Barratt, B. J., Nyberg, F., & Ridker, P. M. (2012). Genetic determinants of statin-induced low-density lipoprotein cholesterol reduction: the Justification for the Use of Statins in Prevention: an Intervention Trial Evaluating Rosuvastatin (JUPITER) trial. *Circ Cardiovasc Genet, 5*(2), 257-264. doi:10.1161/CIRCGENETICS.111.961144

Chen, H., Li, H., Liu, F., Zheng, X., Wang, S., Bo, X., & Shu, W. (2015). An integrative analysis of TFBS-clustered regions reveals new transcriptional regulation models on the accessible chromatin landscape.

Sci Rep, 5, 8465. doi:10.1038/srep08465

Chen, H., Tian, Y., Shu, W., Bo, X., & Wang, S. (2012). Comprehensive identification and annotation of cell type-specific and ubiquitous CTCF-binding sites in the human genome. *PLoS One, 7*(7), e41374. doi:10.1371/journal.pone.0041374

Chiang, D. Y., Getz, G., Jaffe, D. B., O'Kelly, M. J., Zhao, X., Carter, S. L., . . . Lander, E. S. (2009). High-resolution mapping of copy-number alterations with massively parallel sequencing. *Nat Methods, 6*(1), 99-103. doi:10.1038/nmeth.1276

Chien, K. L., Hsu, H. C., Su, T. C., Chen, M. F., Lee, Y. T., & Hu, F. B. (2007). Apolipoprotein B and non-high density lipoprotein cholesterol and the risk of coronary heart disease in Chinese. *J Lipid Res, 48*(11), 2499-2505. doi:10.1194/jlr.M700213-JLR200

Chiu, M., Austin, P. C., Manuel, D. G., & Tu, J. V. (2010). Comparison of cardiovascular risk profiles among ethnic groups using population health surveys between 1996 and 2007. *CMAJ, 182*(8), E301-310. doi:10.1503/cmaj.091676

Cho, H. P., Nakamura, M., & Clarke, S. D. (1999). Cloning, expression, and fatty acid regulation of the human delta-5 desaturase. *J Biol Chem, 274*(52), 37335-37339.

Cho, H. P., Nakamura, M. T., & Clarke, S. D. (1999). Cloning, expression, and nutritional regulation of the mammalian Delta-6 desaturase. *J Biol Chem, 274*(1), 471-477.

Cho, T. M., Rose, R. L., & Hodgson, E. (2006). In vitro metabolism of naphthalene by human liver microsomal cytochrome P450 enzymes. *Drug Metab Dispos, 34*(1), 176-183. doi:10.1124/dmd.105.005785

Choy, E., Yelensky, R., Bonakdar, S., Plenge, R. M., Saxena, R., De Jager, P. L., . . . Altshuler, D. (2008). Genetic analysis of human traits in vitro: drug response and gene expression in lymphoblastoid cell lines. *PLoS Genet, 4*(11), e1000287. doi:10.1371/journal.pgen.1000287

Cirillo, L. A., McPherson, C. E., Bossard, P., Stevens, K., Cherian, S., Shim, E. Y., . . . Zaret, K. S. (1998). Binding of the winged-helix transcription factor HNF3 to a linker histone site on the nucleosome. *EMBO J, 17*(1), 244-254. doi:10.1093/emboj/17.1.244

Clayton, D. G., Walker, N. M., Smyth, D. J., Pask, R., Cooper, J. D., Maier, L. M., . . . Todd, J. A. (2005). Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet, 37*(11), 1243-1246. doi:10.1038/ng1653

Cockerill, G. W., Rye, K. A., Gamble, J. R., Vadas, M. A., & Barter, P. J. (1995). High-density lipoproteins inhibit cytokine-induced expression of endothelial cell adhesion molecules. *Arterioscler Thromb Vasc*

Biol, 15(11), 1987-1994.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., . . . Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data.

Colella, S., Yau, C., Taylor, J. M., Mirza, G., Butler, H., Clouston, P., . . . Ragoussis, J. (2007). QuantiSNP: an Objective Bayes Hidden-Markov Model to detect and accurately map copy number variation using SNP genotyping data. *Nucleic Acids Res, 35*(6), 2013-2025. doi:10.1093/nar/gkm076

Coleman, S., Linderman, R., Hodgson, E., & Rose, R. L. (2000). Comparative metabolism of chloroacetamide herbicides and selected metabolites in human and rat liver microsomes. *Environ Health Perspect, 108*(12), 1151-1157.

Comabella, M., Craig, D. W., Camiña-Tato, M., Morcillo, C., Lopez, C., Navarro, A., . . . Group, B. S. (2008). Identification of a novel risk locus for multiple sclerosis at 13q31.3 by a pooled genome-wide scan of 500,000 single nucleotide polymorphisms. *PLoS One, 3*(10), e3490. doi:10.1371/journal.pone.0003490

Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E., & Pritchard, J. K. (2006). A high-resolution survey of deletion polymorphism in the human genome. *Nat Genet, 38*(1), 75-81. doi:10.1038/ng1697

Conrad, D. F., Pinto, D., Redon, R., Feuk, L., Gokcumen, O., Zhang, Y., . . . Consortium, W. T. C. C. (2010). Origins and functional impact of copy number variation in the human genome. *Nature, 464*(7289), 704-712. doi:10.1038/nature08516

Consortium, G. (2013). The Genotype-Tissue Expression (GTEx) project. *Nat Genet, 45*(6), 580-585. doi:10.1038/ng.2653

Consortium, I. H. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi:10.1038/nature04226

Consortium, W. T. C. C. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661-678. doi:10.1038/nature05911

Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E., & Nickerson, D. A. (2008). Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nat Genet, 40*(10), 1199-1203. doi:10.1038/ng.236

Cowley, M. J., Cotsapas, C. J., Williams, R. B., Chan, E. K., Pulvers, J. N., Liu, M. Y., . . . Little, P. F. (2009). Intra- and inter-individual genetic differences in gene expression. *Mamm Genome, 20*(5), 281-295. doi:10.1007/s00335-009-9181-x

Craddock, N., Hurles, M. E., Cardin, N., Pearson, R. D., Plagnol, V., Robson, S., . . . Consortium, W. T. C. C. (2010). Genome-wide association study of CNVs in 16,000 cases of eight common diseases and 3,000 shared controls. *Nature, 464*(7289), 713-720. doi:10.1038/nature08979

Creyghton, M. P., Cheng, A. W., Welstead, G. G., Kooistra, T., Carey, B. W., Steine, E. J., . . . Jaenisch, R. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A, 107*(50), 21931-21936. doi:10.1073/pnas.1016071107

Crinnion, W. J. (2011). Polychlorinated biphenyls: persistent pollutants with immunological, neurological, and endocrinological consequences. *Altern Med Rev, 16*(1), 5-13.

Cristancho, A. G., & Lazar, M. A. (2011). Forming functional fat: a growing understanding of adipocyte differentiation. *Nat Rev Mol Cell Biol, 12*(11), 722-734. doi:10.1038/nrm3198

Cuddapah, S., Jothi, R., Schones, D. E., Roh, T. Y., Cui, K., & Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. *Genome Res, 19*(1), 24-32. doi:10.1101/gr.082800.108

Cuppini, L., Calleri, A., Bruzzone, M. G., Prodi, E., Anghileri, E., Pellegatta, S., . . . Eoli, M. (2013). Prognostic value of CD109+ circulating endothelial cells in recurrent glioblastomas treated with bevacizumab and irinotecan. *PLoS One, 8*(9), e74345. doi:10.1371/journal.pone.0074345

Curb, J. D., Abbott, R. D., Rodriguez, B. L., Masaki, K., Chen, R., Sharp, D. S., & Tall, A. R. (2004). A prospective study of HDL-C and cholesteryl ester transfer protein gene mutations and the risk of coronary heart disease in the elderly. *J Lipid Res, 45*(5), 948-953. doi:10.1194/jlr.M300520-JLR200

Curnow, P., Mellor, H., Stephens, D. J., Lorch, M., & Booth, P. J. (2005). Translocation of the cell-penetrating Tat peptide across artificial bilayers and into living cells. *Biochem Soc Symp*(72), 199-209.

Custodio, H. M., Harari, R., Gerhardsson, L., Skerfving, S., & Broberg, K. (2005). Genetic influences on the retention of inorganic mercury. *Arch Environ Occup Health, 60*(1), 17-23. doi:10.3200/AEOH.60.1.17-23

Das, T., Thurmond, J. M., Bobik, E., Leonard, A. E., Parker-Barnes, J. M., Huang, Y. S., & Mukerji, P. (2000). Polyunsaturated fatty acid-specific elongation enzymes. *Biochem Soc Trans, 28*(6), 658-660.

Davies, G., Armstrong, N., Bis, J. C., Bressler, J., Chouraki, V., Giddaluru, S., . . . Deary, I. J. (2015). Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53 949). *Mol Psychiatry, 20*(2), 183-192. doi:10.1038/mp.2014.188

de Bakker, P. I., Ferreira, M. A., Jia, X., Neale, B. M., Raychaudhuri, S., & Voight, B. F. (2008). Practical aspects of imputation-driven meta-analysis of genome-wide association studies. *Hum Mol Genet, 17*(R2), R122-128. doi:10.1093/hmg/ddn288

Dehghan, A., Köttgen, A., Yang, Q., Hwang, S. J., Kao, W. L., Rivadeneira, F., . . . Fox, C. S. (2008). Association of three genetic loci with uric acid concentration and risk of gout: a genome-wide association study. *Lancet, 372*(9654), 1953-1961. doi:10.1016/S0140-6736(08)61343-4

Dempster, A., Laird, N., & Rubin, D. (1977). Maximum Likelihood from incomplete data via EM algorithm. *Journal of the Royal Statistical Society Series B-Methodological, 39*(1), 1-38.

den Hoed, M., Eijgelsheim, M., Esko, T., Brundel, B. J., Peal, D. S., Evans, D. M., . . . Consortium, C.-A. (2013). Identification of heart rate-associated loci and their effects on cardiac conduction and rhythm disorders. *Nat Genet, 45*(6), 621-631. doi:10.1038/ng.2610

Deng, X., & Disteche, C. M. (2010). Genomic responses to abnormal gene dosage: the X chromosome improved on a common strategy. *PLoS Biol, 8*(2), e1000318. doi:10.1371/journal.pbio.1000318

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Control Clin Trials, 7*(3), 177-188.

Devlin, B., & Roeder, K. (1999a). Genomic control for association studies.

Devlin, B., & Roeder, K. (1999b). Genomic control for association studies. *Biometrics, 55*(4), 997-1004.

Di Angelantonio, E., Sarwar, N., Perry, P., Kaptoge, S., Ray, K. K., Thompson, A., . . . Collaboration, E. R. F. (2009). Major lipids, apolipoproteins, and risk of vascular disease. *JAMA, 302*(18), 1993-2000. doi:10.1001/jama.2009.1619

Diez, E., Lee, S. H., Gauthier, S., Yaraghi, Z., Tremblay, M., Vidal, S., & Gros, P. (2003). Birc1e is the gene within the Lgn1 locus associated with resistance to Legionella pneumophila. *Nat Genet, 33*(1), 55-60. doi:10.1038/ng1065

Dimas, A. S., Deutsch, S., Stranger, B. E., Montgomery, S. B., Borel, C., Attar-Cohen, H., . . . Antonarakis, S. E. (2009). Common regulatory variation impacts gene expression in a cell type-dependent manner. *Science, 325*(5945), 1246-1250. doi:10.1126/science.1174148

Ding, X., Luo, C., Zhou, J., Zhong, Y., Hu, X., Zhou, F., . . . Zhang, J. (2009). The interaction of KCTD1 with transcription factor AP-2alpha inhibits its transactivation. *J Cell Biochem, 106*(2), 285-295. doi:10.1002/jcb.22002

Do, R., Willer, C. J., Schmidt, E. M., Sengupta, S., Gao, C., Peloso, G. M., . . . Kathiresan, S. (2013). Common variants associated with plasma triglycerides and risk for coronary artery disease. *Nat Genet, 45*(11),

1345-1352. doi:10.1038/ng.2795

Doyle, L., & Ross, D. D. (2003). Multidrug resistance mediated by the breast cancer resistance protein BCRP (ABCG2). *Oncogene, 22*(47), 7340-7358. doi:10.1038/sj.onc.1206938

Duan, J., Zhang, J. G., Deng, H. W., & Wang, Y. P. (2013). Comparative studies of copy number variation detection methods for next-generation sequencing technologies. *PLoS One, 8*(3), e59128. doi:10.1371/journal.pone.0059128

Dubois-Chevalier, J., Oger, F., Dehondt, H., Firmin, F. F., Gheeraert, C., Staels, B., . . . Eeckhoute, J. (2014). A dynamic CTCF chromatin binding landscape promotes DNA hydroxymethylation and transcriptional induction of adipocyte differentiation. *Nucleic Acids Res, 42*(17), 10943-10959. doi:10.1093/nar/gku780

Dudbridge, F., & Gusnanto, A. (2008). Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol, 32*(3), 227-234. doi:10.1002/gepi.20297

Duman, B. S., Oztürk, M., Yilmazer, S., & Hatemi, H. (2004). Apolipoprotein E polymorphism in Turkish subjects with Type 2 diabetes mellitus: allele frequency and relation to serum lipid concentrations. *Diabetes Nutr Metab, 17*(5), 267-274.

Eberhart, C. E., Coffey, R. J., Radhika, A., Giardiello, F. M., Ferrenbach, S., & DuBois, R. N. (1994). Up-regulation of cyclooxygenase 2 gene expression in human colorectal adenomas and adenocarcinomas. *Gastroenterology, 107*(4), 1183-1188.

Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., . . . consortium, C.-H. (2011). Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature, 478*(7367), 103-109. doi:10.1038/nature10405

Emori, M., Tsukahara, T., Murase, M., Kano, M., Murata, K., Takahashi, A., . . . Sato, N. (2013). High expression of CD109 antigen regulates the phenotype of cancer stem-like cells/cancer-initiating cells in the novel epithelioid sarcoma cell line ESX and is related to poor prognosis of soft tissue sarcoma. *PLoS One, 8*(12), e84187. doi:10.1371/journal.pone.0084187

ENCODE. (2011). A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol, 9*(4), e1001046. doi:10.1371/journal.pbio.1001046

Engel, N., West, A. G., Felsenfeld, G., & Bartolomei, M. S. (2004). Antagonism between DNA hypermethylation and enhancer-blocking activity at the H19 DMD is uncovered by CpG mutations. *Nat Genet, 36*(8), 883-888. doi:10.1038/ng1399

Engström, A., Michaëlsson, K., Vahter, M., Julin, B., Wolk, A., & Åkesson, A. (2012). Associations between dietary cadmium exposure and bone mineral density and risk of osteoporosis and fractures among women. *Bone, 50*(6), 1372-1378. doi:10.1016/j.bone.2012.03.018

Engström, K., Ameer, S., Bernaudat, L., Drasch, G., Baeuml, J., Skerfving, S., . . . Broberg, K. (2013). Polymorphisms in genes encoding potential mercury transporters and urine mercury concentrations in populations exposed to mercury vapor from gold mining. *Environ Health Perspect, 121*(1), 85-91. doi:10.1289/ehp.1204951

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., . . . Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature, 473*(7345), 43-49. doi:10.1038/nature09906

Ernst, T., Chase, A. J., Score, J., Hidalgo-Curtis, C. E., Bryant, C., Jones, A. V., . . . Cross, N. C. (2010). Inactivating mutations of the histone methyltransferase gene EZH2 in myeloid disorders. *Nat Genet, 42*(8), 722-726. doi:10.1038/ng.621

Fairfax, B. P., Humburg, P., Makino, S., Naranbhai, V., Wong, D., Lau, E., . . . Knight, J. C. (2014). Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science, 343*(6175), 1246949. doi:10.1126/science.1246949

Fairfax, B. P., Makino, S., Radhakrishnan, J., Plant, K., Leslie, S., Dilthey, A., . . . Knight, J. C. (2012). Genetics of gene expression in primary immune cells identifies cell type-specific master regulators and roles of HLA alleles. *Nat Genet, 44*(5), 502-510. doi:10.1038/ng.2205

Feki-Tounsi, M., & Hamza-Chaffai, A. (2014). Cadmium as a possible cause of bladder cancer: a review of accumulated evidence. *Environ Sci Pollut Res Int, 21*(18), 10561-10573. doi:10.1007/s11356-014-2970-0

Felsenfeld, G., Boyes, J., Chung, J., Clark, D., & Studitsky, V. (1996). Chromatin structure and gene expression. *Proc Natl Acad Sci U S A, 93*(18), 9384-9388.

Fiegler, H., Redon, R., Andrews, D., Scott, C., Andrews, R., Carder, C., . . . Carter, N. P. (2006). Accurate and reliable high-throughput detection of copy number variation in the human genome. *Genome Res, 16*(12), 1566-1574. doi:10.1101/gr.5630906

Filipe-Santos, O., Bustamante, J., Chapgier, A., Vogt, G., de Beaucoudrey, L., Feinberg, J., . . . Casanova, J. L. (2006). Inborn errors of IL-12/23- and IFN-gamma-mediated immunity: molecular, cellular, and clinical features. *Semin Immunol, 18*(6), 347-361. doi:10.1016/j.smim.2006.07.010

Fischle, W., Tseng, B. S., Dormann, H. L., Ueberheide, B. M., Garcia, B. A., Shabanowitz, J., . . . Allis, C. D. (2005). Regulation of HP1-chromatin binding by histone H3 methylation and phosphorylation. *Nature, 438*(7071), 1116-1122. doi:10.1038/nature04219

Franke, L., de Kovel, C. G., Aulchenko, Y. S., Trynka, G., Zhernakova, A., Hunt, K. A., . . . Wijmenga, C. (2008). Detection, imputation, and association analysis of small deletions and null alleles on oligonucleotide arrays. *Am J Hum Genet, 82*(6), 1316-1333. doi:10.1016/j.ajhg.2008.05.008

Friez, M. J., Jones, J. R., Clarkson, K., Lubs, H., Abuelo, D., Bier, J. A., . . . Stevenson, R. E. (2006). Recurrent infections, hypotonia, and mental retardation caused by duplication of MECP2 and adjacent region in Xq28. *Pediatrics, 118*(6), e1687-1695. doi:10.1542/peds.2006-0395

Frisdal, E., Klerkx, A. H., Le Goff, W., Tanck, M. W., Lagarde, J. P., Jukema, J. W., . . . Guerin, M. (2005). Functional interaction between -629C/A, -971G/A and -1337C/T polymorphisms in the CETP gene is a major determinant of promoter activity and plasma CETP concentration in the REGRESS Study. *Hum Mol Genet, 14*(18), 2607-2618. doi:10.1093/hmg/ddi291

Fujiki, K., Shinoda, A., Kano, F., Sato, R., Shirahige, K., & Murata, M. (2013). PPARγ-induced PARylation promotes local DNA demethylation by production of 5-hydroxymethylcytosine. *Nat Commun, 4*, 2262. doi:10.1038/ncomms3262

Galcheva-Gargova, Z., Konstantinov, K. N., Wu, I. H., Klier, F. G., Barrett, T., & Davis, R. J. (1996). Binding of zinc finger protein ZPR1 to the epidermal growth factor receptor. *Science, 272*(5269), 1797-1802.

Gamazon, E. R., Nicolae, D. L., & Cox, N. J. (2011). A study of CNVs as trait-associated polymorphisms and as expression quantitative trait loci. *PLoS Genet, 7*(2), e1001292. doi:10.1371/journal.pgen.1001292

Gao, N., LeLay, J., Vatamaniuk, M. Z., Rieck, S., Friedman, J. R., & Kaestner, K. H. (2008). Dynamic regulation of Pdx1 enhancers by Foxa1 and Foxa2 is essential for pancreas development. *Genes Dev, 22*(24), 3435-3448. doi:10.1101/gad.1752608

Garcia-Donas, J., Esteban, E., Leandro-García, L. J., Castellano, D. E., del Alba, A. G., Climent, M. A., . . . Rodríguez-Antona, C. (2011). Single nucleotide polymorphism associations with response and toxic effects in patients with advanced renal-cell carcinoma treated with first-line sunitinib: a multicentre, observational, prospective study. *Lancet Oncol, 12*(12), 1143-1150. doi:10.1016/S1470-2045(11)70266-2

Garner, B., Waldeck, A. R., Witting, P. K., Rye, K. A., & Stocker, R. (1998). Oxidation of high density lipoproteins. II. Evidence for direct reduction of lipid hydroperoxides by methionine residues of

apolipoproteins AI and AII. *J Biol Chem, 273*(11), 6088-6095.

Gaszner, M., & Felsenfeld, G. (2006). Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat Rev Genet, 7*(9), 703-713. doi:10.1038/nrg1925

Gatanaga, H., Ibe, S., Matsuda, M., Yoshida, S., Asagi, T., Kondo, M., . . . Sugiura, W. (2007). Drug-resistant HIV-1 prevalence in patients newly diagnosed with HIV/AIDS in Japan. *Antiviral Res, 75*(1), 75-82. doi:10.1016/j.antiviral.2006.11.012

Ghosh, S., May, M. J., & Kopp, E. B. (1998). NF-kappa B and Rel proteins: evolutionarily conserved mediators of immune responses. *Annu Rev Immunol, 16*, 225-260. doi:10.1146/annurev.immunol.16.1.225

Gibbs, J. R., van der Brug, M. P., Hernandez, D. G., Traynor, B. J., Nalls, M. A., Lai, S. L., . . . Singleton, A. B. (2010). Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet, 6*(5), e1000952. doi:10.1371/journal.pgen.1000952

Gingras, A. C., Caballero, M., Zarske, M., Sanchez, A., Hazbun, T. R., Fields, S., . . . Aebersold, R. (2005). A novel, evolutionarily conserved protein phosphatase complex involved in cisplatin sensitivity. *Mol Cell Proteomics, 4*(11), 1725-1740. doi:10.1074/mcp.M500231-MCP200

Glessner, J. T., Wang, K., Cai, G., Korvatska, O., Kim, C. E., Wood, S., . . . Hakonarson, H. (2009). Autism genome-wide copy number variation reveals ubiquitin and neuronal genes. *Nature, 459*(7246), 569-573. doi:10.1038/nature07953

Godt, J., Scheidig, F., Grosse-Siestrup, C., Esche, V., Brandenburg, P., Reich, A., & Groneberg, D. A. (2006). The toxicity of cadmium and resulting hazards for human health. *J Occup Med Toxicol, 1*, 22. doi:10.1186/1745-6673-1-22

Gonzalez, E., Kulkarni, H., Bolivar, H., Mangano, A., Sanchez, R., Catano, G., . . . Ahuja, S. K. (2005). The influence of CCL3L1 gene-containing segmental duplications on HIV-1/AIDS susceptibility. *Science, 307*(5714), 1434-1440. doi:10.1126/science.1101160

Gordon, S., Durairaj, A., Lu, J. L., & Davidson, W. S. (2010). High-Density Lipoprotein Proteomics: Identifying New Drug Targets and Biomarkers by Understanding Functionality. *Curr Cardiovasc Risk Rep, 4*(1), 1-8. doi:10.1007/s12170-009-0069-9

Gordon, T., Castelli, W. P., Hjortland, M. C., Kannel, W. B., & Dawber, T. R. (1977). High density lipoprotein as a protective factor against coronary heart disease. The Framingham Study. *Am J Med, 62*(5), 707-714.

Gordon, T., Kannel, W. B., Castelli, W. P., & Dawber, T. R. (1981). Lipoproteins, cardiovascular disease, and

death. The Framingham study. *Arch Intern Med, 141*(9), 1128-1131.

Gottlieb, D. J., DeStefano, A. L., Foley, D. J., Mignot, E., Redline, S., Givelber, R. J., & Young, T. (2004). APOE epsilon4 is associated with obstructive sleep apnea/hypopnea: the Sleep Heart Health Study. *Neurology, 63*(4), 664-668.

Govarts, E., Nieuwenhuijsen, M., Schoeters, G., Ballester, F., Bloemen, K., de Boer, M., . . . ENRIECO. (2012). Birth weight and prenatal exposure to polychlorinated biphenyls (PCBs) and dichlorodiphenyldichloroethylene (DDE): a meta-analysis within 12 European Birth Cohorts. *Environ Health Perspect, 120*(2), 162-170. doi:10.1289/ehp.1103767

Gu, W., Zhang, F., & Lupski, J. R. (2008). Mechanisms for human genomic rearrangements. *Pathogenetics, 1*(1), 4. doi:10.1186/1755-8417-1-4

Guastafierro, T., Cecchinelli, B., Zampieri, M., Reale, A., Riggio, G., Sthandier, O., . . . Caiafa, P. (2008). CCCTC-binding factor activates PARP-1 affecting DNA methylation machinery. *J Biol Chem, 283*(32), 21873-21880. doi:10.1074/jbc.M801170200

Guenther, M. G., Levine, S. S., Boyer, L. A., Jaenisch, R., & Young, R. A. (2007). A chromatin landmark and transcription initiation at most promoters in human cells. *Cell, 130*(1), 77-88. doi:10.1016/j.cell.2007.05.042

Guryev, V., Saar, K., Adamovic, T., Verheul, M., van Heesch, S. A., Cook, S., . . . Cuppen, E. (2008). Distribution and functional impact of DNA copy number variation in the rat. *Nat Genet, 40*(5), 538-545. doi:10.1038/ng.141

Gährs, M., Roos, R., Andersson, P. L., & Schrenk, D. (2013). Role of the nuclear xenobiotic receptors CAR and PXR in induction of cytochromes P450 by non-dioxinlike polychlorinated biphenyls in cultured rat hepatocytes. *Toxicol Appl Pharmacol, 272*(1), 77-85. doi:10.1016/j.taap.2013.05.034

Haakonsson, A. K., Stahl Madsen, M., Nielsen, R., Sandelin, A., & Mandrup, S. (2013). Acute genome-wide effects of rosiglitazone on PPARγ transcriptional networks in adipocytes. *Mol Endocrinol, 27*(9), 1536-1549. doi:10.1210/me.2013-1080

Hagiwara, S., Murakumo, Y., Sato, T., Shigetomi, T., Mitsudo, K., Tohnai, I., . . . Takahashi, M. (2008). Up-regulation of CD109 expression is associated with carcinogenesis of the squamous epithelium of the oral cavity. *Cancer Sci, 99*(10), 1916-1923. doi:10.1111/j.1349-7006.2008.00949.x

Hamon, S. C., Stengard, J. H., Clark, A. G., Salomaa, V., Boerwinkle, E., & Sing, C. F. (2004). Evidence for non-additive influence of single nucleotide polymorphisms within the apolipoprotein E gene. *Ann Hum*

Genet, 68(Pt 6), 521-535. doi:10.1046/j.1529-8817.2003.00112.x

Han, B., & Eskin, E. (2011). Random-effects model aimed at discovering associations in meta-analysis of genome-wide association studies. *Am J Hum Genet, 88*(5), 586-598. doi:10.1016/j.ajhg.2011.04.014

Han, Y., Pei, Y., Liu, Y., Zhang, L., Wu, S., Tian, Q., . . . Deng, H. (2012). Bivariate genome-wide association study suggests fatty acid desaturase genes and cadherin DCHS2 for variation of both compressive strength index and appendicular lean mass in males. *Bone, 51*(6), 1000-1007. doi:10.1016/j.bone.2012.08.127

Hannenhalli, S., & Kaestner, K. H. (2009). The evolution of Fox genes and their role in development and disease. *Nat Rev Genet, 10*(4), 233-240. doi:10.1038/nrg2523

Hansen, K. H., Bracken, A. P., Pasini, D., Dietrich, N., Gehani, S. S., Monrad, A., . . . Helin, K. (2008). A model for transmission of the H3K27me3 epigenetic mark. *Nat Cell Biol, 10*(11), 1291-1300. doi:10.1038/ncb1787

Hansen-Petrik, M. B., McEntee, M. F., Johnson, B. T., Obukowicz, M. G., Masferrer, J., Zweifel, B., . . . Whelan, J. (2002). Selective inhibition of Delta-6 desaturase impedes intestinal tumorigenesis. *Cancer Lett, 175*(2), 157-163.

Harari, R., Harari, F., Gerhardsson, L., Lundh, T., Skerfving, S., Strömberg, U., & Broberg, K. (2012). Exposure and toxic effects of elemental mercury in gold-mining activities in Ecuador. *Toxicol Lett, 213*(1), 75-82. doi:10.1016/j.toxlet.2011.09.006

Harismendy, O., Notani, D., Song, X., Rahim, N. G., Tanasa, B., Heintzman, N., . . . Frazer, K. A. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. *Nature, 470*(7333), 264-268. doi:10.1038/nature09753

Harrad, S., Wang, Y., Sandaradura, S., & Leeds, A. (2003). Human dietary intake and excretion of dioxin-like compounds. *J Environ Monit, 5*(2), 224-228.

Harrow, J., Denoeud, F., Frankish, A., Reymond, A., Chen, C. K., Chrast, J., . . . Guigo, R. (2006). GENCODE: producing a reference annotation for ENCODE. *Genome Biol, 7 Suppl 1*, S4.1-9. doi:10.1186/gb-2006-7-s1-s4

Hartung, G. H., & Squires, W. G. (1982). Physiological measures and marathon running performance in young and middle-aged males. *J Sports Med Phys Fitness, 22*(3), 366-370.

Hastings, P. J., Lupski, J. R., Rosenberg, S. M., & Ira, G. (2009). Mechanisms of change in gene copy number. *Nat Rev Genet, 10*(8), 551-564. doi:10.1038/nrg2593

He, C., Qu, X., Wan, J., Rong, R., Huang, L., Cai, C., . . . Kang, J. X. (2012). Inhibiting delta-6 desaturase activity suppresses tumor growth in mice. *PLoS One, 7*(10), e47567. doi:10.1371/journal.pone.0047567

He, M., Xu, M., Zhang, B., Liang, J., Chen, P., Lee, J. Y., . . . Qi, L. (2015). Meta-analysis of genome-wide association studies of adult height in East Asians identifies 17 novel loci. *Hum Mol Genet, 24*(6), 1791-1800. doi:10.1093/hmg/ddu583

Hegele, R. A., & Pollex, R. L. (2009). Hypertriglyceridemia: phenomics and genomics. *Mol Cell Biochem, 326*(1-2), 35-43. doi:10.1007/s11010-008-0005-1

Helal, R., Bader, G., & Melzig, M. F. (2012). Stimulation of lysozyme release by selected microbial preparations. *Pharmazie, 67*(6), 564-566.

Hellwege, J. N., Palmer, N. D., Raffield, L. M., Ng, M. C., Hawkins, G. A., Long, J., . . . Bowden, D. W. (2014). Genome-wide family-based linkage analysis of exome chip variants and cardiometabolic risk. *Genet Epidemiol, 38*(4), 345-352. doi:10.1002/gepi.21801

Henrichsen, C. N., Vinckenbosch, N., Zöllner, S., Chaignat, E., Pradervand, S., Schütz, F., . . . Reymond, A. (2009). Segmental copy number variation shapes tissue transcriptomes. *Nat Genet, 41*(4), 424-429. doi:10.1038/ng.345

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Stat Med, 21*(11), 1539-1558. doi:10.1002/sim.1186

Hindorff, L., J, M., HA, J., PN, H., & AK, K. (2015). *A Catalog of Published Genome-Wide Association Studies*.

Hitchins, M. P., Rapkins, R. W., Kwok, C. T., Srivastava, S., Wong, J. J., Khachigian, L. M., . . . Ward, R. L. (2011). Dominantly inherited constitutional epigenetic silencing of MLH1 in a cancer-affected family is linked to a single nucleotide variant within the 5'UTR. *Cancer Cell, 20*(2), 200-213. doi:10.1016/j.ccr.2011.07.003

Hodge, A. M., English, D. R., O'Dea, K., Sinclair, A. J., Makrides, M., Gibson, R. A., & Giles, G. G. (2007). Plasma phospholipid and dietary fatty acids as predictors of type 2 diabetes: interpreting the role of linoleic acid. *Am J Clin Nutr, 86*(1), 189-197.

Hoffmann, K., Blaudszun, J., Brunken, C., Höpker, W. W., Tauber, R., & Steinhart, H. (2005). Distribution of polyunsaturated fatty acids including conjugated linoleic acids in total and subcellular fractions from healthy and cancerous parts of human kidneys. *Lipids, 40*(3), 309-315.

Hollox, E. J., Huffmeier, U., Zeeuwen, P. L. J. M., Palla, R., Lascorz, J., Rodijk-Olthuis, D., . . . Schalkwijk, J. (2008). Psoriasis is associated with increased beta-defensin genomic copy number.

Hon, G. C., Hawkins, R. D., & Ren, B. (2009). Predictive chromatin signatures in the mammalian genome. *Hum Mol Genet, 18*(R2), R195-201. doi:10.1093/hmg/ddp409

Hong, M. G., Karlsson, R., Magnusson, P. K., Lewis, M. R., Isaacs, W., Zheng, L. S., . . . Prince, J. A. (2013). A genome-wide assessment of variability in human serum metabolism. *Hum Mutat, 34*(3), 515-524. doi:10.1002/humu.22267

Horwitz, M. A., & Silverstein, S. C. (1983). Intracellular multiplication of Legionnaires' disease bacteria (Legionella pneumophila) in human monocytes is reversibly inhibited by erythromycin and rifampin. *J Clin Invest, 71*(1), 15-26.

Howie, B., Fuchsberger, C., Stephens, M., Marchini, J., & Abecasis, G. R. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet, 44*(8), 955-959. doi:10.1038/ng.2354

Howie, B. N., Donnelly, P., & Marchini, J. (2009). A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet, 5*(6), e1000529. doi:10.1371/journal.pgen.1000529

Huang, L., Li, Y., Singleton, A. B., Hardy, J. A., Abecasis, G., Rosenberg, N. A., & Scheet, P. (2009). Genotype-imputation accuracy across worldwide human populations. *Am J Hum Genet, 84*(2), 235-250. doi:10.1016/j.ajhg.2009.01.013

Huang, Z. H., Maeda, N., & Mazzone, T. (2011). Expression of the human apoE2 isoform in adipocytes: altered cellular processing and impaired adipocyte lipogenesis. *J Lipid Res, 52*(9), 1733-1741. doi:10.1194/jlr.M017160

Huff, J., Lunn, R. M., Waalkes, M. P., Tomatis, L., & Infante, P. F. (2007). Cadmium-induced cancers in animals and in humans. *Int J Occup Environ Health, 13*(2), 202-212. doi:10.1179/oeh.2007.13.2.202

Iafrate, A. J., Feuk, L., Rivera, M. N., Listewnik, M. L., Donahoe, P. K., Qi, Y., . . . Lee, C. (2004). Detection of large-scale variation in the human genome. *Nat Genet, 36*(9), 949-951. doi:10.1038/ng1416

Illumina. (last accessed 13 June 2015). Illumina Genotyping Module User Guide.

Imai, Y., Ohmori, K., Yasuda, S., Wada, M., Suzuki, T., Fukuda, K., & Ueda, Y. (2009). Breast cancer resistance protein/ABCG2 is differentially regulated downstream of extracellular signal-regulated kinase. *Cancer Sci, 100*(6), 1118-1127.

Imamura, T., Matsumoto, S., Akahane, M., Kanagawa, Y., Koike, S., Tajima, B., . . . Furue, M. (2009). Cutaneous symptoms such as acneform eruption and pigmentation are closely associated with blood

levels of 2,3,4,7,8-penta-chlorodibenzofurans in Yusho patients, using data mining analysis. *BMC Res Notes, 2*, 27. doi:10.1186/1756-0500-2-27

Ingelsson, E., Schaefer, E. J., Contois, J. H., McNamara, J. R., Sullivan, L., Keyes, M. J., . . . Vasan, R. S. (2007). Clinical utility of different lipid measures for prediction of coronary heart disease in men and women. *JAMA, 298*(7), 776-785. doi:10.1001/jama.298.7.776

INSERM Collective Expert Reports. (2000). Dioxins in the Environment: What are the health risks? Paris. Retrieved from http://www.ncbi.nlm.nih.gov/books/NBK7128/.

International HapMap Consortium. (2005). A haplotype map of the human genome. *Nature, 437*(7063), 1299-1320. doi:10.1038/nature04226

Inui, H., Itoh, T., Yamamoto, K., Ikushiro, S., & Sakaki, T. (2014). Mammalian cytochrome P450-dependent metabolism of polychlorinated dibenzo-p-dioxins and coplanar polychlorinated biphenyls. *Int J Mol Sci, 15*(8), 14044-14057. doi:10.3390/ijms150814044

Inui, H., Itoh, T., Yamamoto, K., Ikushiro, S.-I., & Sakaki, T. Mammalian cytochrome P450-dependent metabolism of polychlorinated dibenzo-p-dioxins and coplanar polychlorinated biphenyls.

Izaks, G. J., Gansevoort, R. T., van der Knaap, A. M., Navis, G., Dullaart, R. P., & Slaets, J. P. (2011). The association of APOE genotype with cognitive function in persons aged 35 years or older. *PLoS One, 6*(11), e27415. doi:10.1371/journal.pone.0027415

Jackson, D. (2006). The power of the standard test for the presence of heterogeneity in meta-analysis. *Stat Med, 25*(15), 2688-2699. doi:10.1002/sim.2481

Jeffreys, A. (1961). Theory of Probability (3rd Edition ed.). Oxford: Oxford University Press.

Jenkins, P. J., Harper, R. W., & Nestel, P. J. (1978). Severity of coronary atherosclerosis related to lipoprotein concentration. *Br Med J, 2*(6134), 388-391.

Jiao, S., Hsu, L., Hutter, C. M., & Peters, U. (2011). The use of imputed values in the meta-analysis of genome-wide association studies. *Genet Epidemiol, 35*(7), 597-605. doi:10.1002/gepi.20608

Jostins, L., Morley, K. I., & Barrett, J. C. (2011). Imputation of low-frequency variants using the HapMap3 benefits from large, diverse reference sets. *Eur J Hum Genet, 19*(6), 662-666. doi:10.1038/ejhg.2011.10

Jostins, L., Ripke, S., Weersma, R. K., Duerr, R. H., McGovern, D. P., Hui, K. Y., . . . (IIBDGC), I. I. G. C. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature, 491*(7422), 119-124. doi:10.1038/nature11582

Jump, D. B. (2002). Dietary polyunsaturated fatty acids and regulation of gene transcription. *Curr Opin Lipidol,*

13(2), 155-164.

Jump, D. B. (2004). Fatty acid regulation of gene transcription. *Crit Rev Clin Lab Sci, 41*(1), 41-78. doi:10.1080/10408360490278341

Kadish, A. H., Greenland, P., Limacher, M. C., Frishman, W. H., Daugherty, S. A., & Schwartz, J. B. (2004). Estrogen and progestin use and the QT interval in postmenopausal women. *Ann Noninvasive Electrocardiol, 9*(4), 366-374. doi:10.1111/j.1542-474X.2004.94580.x

Kaestner, K. H. (2010). The FoxA factors in organogenesis and differentiation. *Curr Opin Genet Dev, 20*(5), 527-532. doi:10.1016/j.gde.2010.06.005

Kakutani, H., Aozasa, O., Mizuno, A., Akiyama, E., Nakao, T., & Ohta, S. (2014). In vitro and in vivo induction of cytochrome P450 by coplanar polychlorinated/brominated biphenyls (Co-PXBs) providing high TEQ in mother's milk in Japan. *Toxicology, 324*, 68-75. doi:10.1016/j.tox.2014.07.008

Kambe, T. (2011). An overview of a wide range of functions of ZnT and Zip zinc transporters in the secretory pathway. *Biosci Biotechnol Biochem, 75*(6), 1036-1043. doi:10.1271/bbb.110056

Kang, J. X., Wang, J., Wu, L., & Kang, Z. B. (2004). Transgenic mice: fat-1 mice convert n-6 to n-3 fatty acids. *Nature, 427*(6974), 504. doi:10.1038/427504a

Kathiresan, S., Melander, O., Guiducci, C., Surti, A., Burtt, N. P., Rieder, M. J., . . . Orho-Melander, M. (2008). Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nat Genet, 40*(2), 189-197. doi:10.1038/ng.75

Kathiresan, S., Willer, C. J., Peloso, G. M., Demissie, S., Musunuru, K., Schadt, E. E., . . . Cupples, L. A. (2009). Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet, 41*(1), 56-65. doi:10.1038/ng.291

Kavvoura, F. K., & Ioannidis, J. P. (2008). Methods for meta-analysis in genetic association studies: a review of their potential and pitfalls. *Hum Genet, 123*(1), 1-14. doi:10.1007/s00439-007-0445-9

Kerr, I. D., Jones, P. M., & George, A. M. (2010). Multidrug efflux pumps: the structures of prokaryotic ATP-binding cassette transporter efflux pumps and implications for our understanding of eukaryotic P-glycoproteins and homologues. *FEBS J, 277*(3), 550-563. doi:10.1111/j.1742-4658.2009.07486.x

Keskitalo, J. E., Pasanen, M. K., Neuvonen, P. J., & Niemi, M. (2009). Different effects of the ABCG2 c.421C>A SNP on the pharmacokinetics of fluvastatin, pravastatin and simvastatin. *Pharmacogenomics, 10*(10), 1617-1624. doi:10.2217/pgs.09.85

Khoo, C., Campos, H., Judge, H., & Sacks, F. M. (1999). Effects of estrogenic oral contraceptives on the

lipoprotein B particle system defined by apolipoproteins E and C-III content. *J Lipid Res, 40*(2), 202-212.

Kidd, P. M. (2005). Neurodegeneration from mitochondrial insufficiency: nutrients, stem cells, growth factors, and prospects for brain rebuilding using integrative management. *Altern Med Rev, 10*(4), 268-293.

Kim, J., Daniel, J., Espejo, A., Lake, A., Krishna, M., Xia, L., . . . Bedford, M. T. (2006). Tudor, MBT and chromo domains gauge the degree of lysine methylation. *EMBO Rep, 7*(4), 397-403. doi:10.1038/sj.embor.7400625

Kim, M., Turnquist, H., Jackson, J., Sgagias, M., Yan, Y., Gong, M., . . . Cowan, K. (2002). The multidrug resistance transporter ABCG2 (breast cancer resistance protein 1) effluxes Hoechst 33342 and is overexpressed in hematopoietic stem cells. *Clin Cancer Res, 8*(1), 22-28.

Kim, T. K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., . . . Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature, 465*(7295), 182-187. doi:10.1038/nature09033

Kirchhoff, M., Gerdes, T., Rose, H., Maahr, J., Ottesen, A. M., & Lundsteen, C. (1998). Detection of chromosomal gains and losses in comparative genomic hybridization analysis based on standard reference intervals. *Cytometry, 31*(3), 163-173.

Kirov, G., Grozeva, D., Norton, N., Ivanov, D., Mantripragada, K. K., Holmans, P., . . . Consortium, W. T. C. C. (2009). Support for the involvement of large copy number variants in the pathogenesis of schizophrenia. *Hum Mol Genet, 18*(8), 1497-1503. doi:10.1093/hmg/ddp043

Kirov, G., Rees, E., Walters, J. T., Escott-Price, V., Georgieva, L., Richards, A. L., . . . Owen, M. J. (2014). The penetrance of copy number variations for schizophrenia and developmental delay. *Biol Psychiatry, 75*(5), 378-385. doi:10.1016/j.biopsych.2013.07.022

Kitamura, K., Kikuchi, Y., Watanabe, S., Waechter, G., Sakurai, H., & Takada, T. (2000). Health effects of chronic exposure to polychlorinated dibenzo-P-dioxins (PCDD), dibenzofurans (PCDF) and coplanar PCB (Co-PCB) of municipal waste incinerator workers. *J Epidemiol, 10*(4), 262-270.

Kleer, C. G., Cao, Q., Varambally, S., Shen, R., Ota, I., Tomlins, S. A., . . . Chinnaiyan, A. M. (2003). EZH2 is a marker of aggressive breast cancer and promotes neoplastic transformation of breast epithelial cells. *Proc Natl Acad Sci U S A, 100*(20), 11606-11611. doi:10.1073/pnas.1933744100

Koga, N., Beppu, M., & Yoshimura, H. (1990). Metabolism in vivo of 3,4,5,3',4'-pentachlorobiphenyl and toxicological assessment of the metabolite in rats. *J Pharmacobiodyn, 13*(8), 497-506.

Korn, J. M., Kuruvilla, F. G., McCarroll, S. A., Wysoker, A., Nemesh, J., Cawley, S., . . . Altshuler, D. (2008). Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet, 40*(10), 1253-1260. doi:10.1038/ng.237

Kraft, P., Zeggini, E., & Ioannidis, J. P. (2009). Replication in genome-wide association studies. *Stat Sci, 24*(4), 561-573. doi:10.1214/09-STS290

Krawczak, M., Ball, E. V., Fenton, I., Stenson, P. D., Abeysinghe, S., Thomas, N., & Cooper, D. N. (2000). Human gene mutation database-a biomedical information and research resource. *Hum Mutat, 15*(1), 45-51. doi:10.1002/(SICI)1098-1004(200001)15:1<45::AID-HUMU10>3.0.CO;2-T

Kristiansson, K., Perola, M., Tikkanen, E., Kettunen, J., Surakka, I., Havulinna, A. S., . . . Salomaa, V. (2012). Genome-wide screen for metabolic syndrome susceptibility Loci reveals strong lipid gene contribution but no evidence for common genetic basis for clustering of metabolic syndrome traits. *Circ Cardiovasc Genet, 5*(2), 242-249. doi:10.1161/CIRCGENETICS.111.961482

Kröger, J., & Schulze, M. B. (2012). Recent insights into the relation of Δ5 desaturase and Δ6 desaturase activity to the development of type 2 diabetes. *Curr Opin Lipidol, 23*(1), 4-10. doi:10.1097/MOL.0b013e32834d2dc5

Kumar, P., Henikoff, S., & Ng, P. C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc, 4*(7), 1073-1081. doi:10.1038/nprot.2009.86

Kwak, J. H., Paik, J. K., Kim, O. Y., Jang, Y., Lee, S. H., Ordovas, J. M., & Lee, J. H. (2011). FADS gene polymorphisms in Koreans: association with ω6 polyunsaturated fatty acids in serum phospholipids, lipid peroxides, and coronary artery disease. *Atherosclerosis, 214*(1), 94-100. doi:10.1016/j.atherosclerosis.2010.10.004

Köttgen, A., Albrecht, E., Teumer, A., Vitart, V., Krumsiek, J., Hundertmark, C., . . . Consortium, M. (2013). Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet, 45*(2), 145-154. doi:10.1038/ng.2500

Lai, A. Y., Fatemi, M., Dhasarathy, A., Malone, C., Sobol, S. E., Geigerman, C., . . . Wade, P. A. (2010). DNA methylation prevents CTCF-mediated silencing of the oncogene BCL6 in B cell lymphomas. *J Exp Med, 207*(9), 1939-1950. doi:10.1084/jem.20100204

Lai, E., Prezioso, V. R., Tao, W. F., Chen, W. S., & Darnell, J. E. (1991). Hepatocyte nuclear factor 3 alpha belongs to a gene family in mammals that is homologous to the Drosophila homeotic gene fork head. *Genes Dev, 5*(3), 416-427.

Lam, K. W., & Jeffreys, A. J. (2006). Processes of copy-number change in human DNA: the dynamics of {alpha}-globin gene deletion. *Proc Natl Acad Sci U S A, 103*(24), 8921-8927. doi:10.1073/pnas.0602690103

Lamon-Fava, S. (2002). High-density lipoproteins: effects of alcohol, estrogen, and phytoestrogens. *Nutr Rev, 60*(1), 1-7.

Lamon-Fava, S., Jimenez, D., Christian, J. C., Fabsitz, R. R., Reed, T., Carmelli, D., . . . Schaefer, E. J. (1991). The NHLBI Twin Study: heritability of apolipoprotein A-I, B, and low density lipoprotein subclasses and concordance for lipoprotein(a). *Atherosclerosis, 91*(1-2), 97-106.

Lampa, E., Lind, L., Hermansson, A. B., Salihovic, S., van Bavel, B., & Lind, P. M. (2012). An investigation of the co-variation in circulating levels of a large number of environmental contaminants. *J Expo Sci Environ Epidemiol, 22*(5), 476-482. doi:10.1038/jes.2012.41

Lander, E. S. (1996). The new genomics: global views of biology. *Science, 274*(5287), 536-539.

Lang, D. H., Rettie, A. E., & Bocker, R. H. (1997). Identification of enzymes involved in the metabolism of atrazine, terbuthylazine, ametryne, and terbutryne in human liver microsomes. *Chem Res Toxicol, 10*(9), 1037-1044. doi:10.1021/tx970081l

Lattka, E., Illig, T., Heinrich, J., & Koletzko, B. (2009). FADS gene cluster polymorphisms: important modulators of fatty acid levels and their impact on atopic diseases. *J Nutrigenet Nutrigenomics, 2*(3), 119-128. doi:10.1159/000235559

Lee, D. H., Lind, L., Jacobs, D. R., Salihovic, S., van Bavel, B., & Lind, P. M. (2012). Associations of persistent organic pollutants with abdominal obesity in the elderly: The Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study. *Environ Int, 40*, 170-178. doi:10.1016/j.envint.2011.07.010

Lee, J. A., Carvalho, C. M., & Lupski, J. R. (2007). A DNA replication mechanism for generating nonrecurrent rearrangements associated with genomic disorders. *Cell, 131*(7), 1235-1247. doi:10.1016/j.cell.2007.11.037

Lee, J. S., Shukla, A., Schneider, J., Swanson, S. K., Washburn, M. P., Florens, L., . . . Shilatifard, A. (2007). Histone crosstalk between H2B monoubiquitination and H3 methylation mediated by COMPASS. *Cell, 131*(6), 1084-1096. doi:10.1016/j.cell.2007.09.046

Legry, V., Bokor, S., Beghin, L., Galfo, M., Gonzalez-Gross, M., Molnar, D., . . . Group, H. S. (2011). Associations between common genetic polymorphisms in the liver X receptor alpha and its target genes

with the serum HDL-cholesterol concentration in adolescents of the HELENA Study. *Atherosclerosis, 216*(1), 166-169. doi:10.1016/j.atherosclerosis.2011.01.031

Lehmler, H. J., Harrad, S. J., Hühnerfuss, H., Kania-Korwel, I., Lee, C. M., Lu, Z., & Wong, C. S. (2010). Chiral polychlorinated biphenyl transport, metabolism, and distribution: a review. *Environ Sci Technol, 44*(8), 2757-2766. doi:10.1021/es902208u

Lei, S. F., Papasian, C. J., & Deng, H. W. (2011). Polymorphisms in predicted miRNA binding sites and osteoporosis. *J Bone Miner Res, 26*(1), 72-78. doi:10.1002/jbmr.186

Lemaitre, R. N., Tanaka, T., Tang, W., Manichaikul, A., Foy, M., Kabagambe, E. K., . . . Steffen, L. M. (2011). Genetic loci associated with plasma phospholipid n-3 fatty acids: a meta-analysis of genome-wide association studies from the CHARGE Consortium. *PLoS Genet, 7*(7), e1002193. doi:10.1371/journal.pgen.1002193

Leonard, A. E., Pereira, S. L., Sprecher, H., & Huang, Y. S. (2004). Elongation of long-chain fatty acids. *Prog Lipid Res, 43*(1), 36-54.

Letz, G. (1983). The toxicology of PCB's--an overview for clinicians. *West J Med, 138*(4), 534-540.

Li, B., & Leal, S. M. (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet, 5*(5), e1000481. doi:10.1371/journal.pgen.1000481

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics, 165*(4), 2213-2233.

Li, Q., & Yu, K. (2008). Improved correction for population stratification in genome-wide association studies by identifying hidden population structures. *Genet Epidemiol, 32*(3), 215-226. doi:10.1002/gepi.20296

Lieber, M. R. (2008). The mechanism of human nonhomologous DNA end joining. *J Biol Chem, 283*(1), 1-5. doi:10.1074/jbc.R700039200

Lin, D. Y., & Zeng, D. (2010). Meta-analysis of genome-wide association studies: no efficiency gain in using individual participant data. *Genet Epidemiol, 34*(1), 60-66. doi:10.1002/gepi.20435

Lind, L., Fors, N., Hall, J., Marttala, K., & Stenborg, A. (2005). A comparison of three different methods to evaluate endothelium-dependent vasodilation in the elderly: the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS) study. *Arterioscler Thromb Vasc Biol, 25*(11), 2368-2375. doi:10.1161/01.ATV.0000184769.22061.da

Lind, L., Penell, J., Syvänen, A. C., Axelsson, T., Ingelsson, E., Morris, A. P., . . . Lind, P. M. (2014). Genetic variation in the CYP1A1 gene is related to circulating PCB118 levels in a population-based sample.

Environ Res, 133, 135-140. doi:10.1016/j.envres.2014.05.017

Lind, P. M., Olsén, L., & Lind, L. (2012). Circulating levels of metals are related to carotid atherosclerosis in elderly. *Sci Total Environ, 416*, 80-88. doi:10.1016/j.scitotenv.2011.11.064

Liskay, R. M., Letsou, A., & Stachelek, J. L. (1987). Homology requirement for efficient gene conversion between duplicated chromosomal sequences in mammalian cells. *Genetics, 115*(1), 161-167.

Lithell, H., Sundström, J., Arnlöv, J., Björklund, K., Hänni, A., Hedman, A., . . . Reneland, R. (2000). Epidemiological and clinical studies on insulin resistance and diabetes. *Ups J Med Sci, 105*(2), 135-150.

Liu, Y., Jandacek, R., Rider, T., Tso, P., & McNamara, R. K. (2009). Elevated delta-6 desaturase (FADS2) expression in the postmortem prefrontal cortex of schizophrenic patients: relationship with fatty acid composition. *Schizophr Res, 109*(1-3), 113-120. doi:10.1016/j.schres.2008.12.027

Liu, Y., & McNamara, R. K. (2011). Elevated Delta-6 desaturase (FADS2) gene expression in the prefrontal cortex of patients with bipolar disorder. *J Psychiatr Res, 45*(2), 269-272. doi:10.1016/j.jpsychires.2010.06.010

Locati, E. H., Zareba, W., Moss, A. J., Schwartz, P. J., Vincent, G. M., Lehmann, M. H., . . . Hall, W. J. (1998). Age- and sex-related differences in clinical manifestations in patients with congenital long-QT syndrome: findings from the International LQTS Registry. *Circulation, 97*(22), 2237-2244.

Loos, R. J. (2012). Genetic determinants of common obesity and their value in prediction. *Best Pract Res Clin Endocrinol Metab, 26*(2), 211-226. doi:10.1016/j.beem.2011.11.003

Lu, Y., Vitart, V., Burdon, K. P., Khor, C. C., Bykhovskaya, Y., Mirshahi, A., . . . Consortium, N. (2013). Genome-wide association analyses identify multiple loci associated with central corneal thickness and keratoconus. *Nat Genet, 45*(2), 155-163. doi:10.1038/ng.2506

Lucito, R., Healy, J., Alexander, J., Reiner, A., Esposito, D., Chi, M., . . . Wigler, M. (2003). Representational oligonucleotide microarray analysis: a high-resolution method to detect genome copy number variation. *Genome Res, 13*(10), 2291-2305. doi:10.1101/gr.1349003

Lupski, J. R. (2010). Retrotransposition and structural variation in the human genome. *Cell, 141*(7), 1110-1112. doi:10.1016/j.cell.2010.06.014

Lusis, A. J. (2000). Atherosclerosis. *Nature, 407*(6801), 233-241. doi:10.1038/35025203

Lusis, A. J. (2012). Genetics of atherosclerosis. *Trends Genet, 28*(6), 267-275. doi:10.1016/j.tig.2012.03.001

Lusis, A. J., Fogelman, A. M., & Fonarow, G. C. (2004). Genetic basis of atherosclerosis: part I: new genes and

pathways. *Circulation, 110*(13), 1868-1873. doi:10.1161/01.CIR.0000143041.58692.CC

López-Vicario, C., González-Périz, A., Rius, B., Morán-Salvador, E., García-Alonso, V., Lozano, J. J., . . . Titos, E. (2014). Molecular interplay between Δ5/Δ6 desaturases and long-chain fatty acids in the pathogenesis of non-alcoholic steatohepatitis. *Gut, 63*(2), 344-355. doi:10.1136/gutjnl-2012-303179

Madrigal, P., & Krajewski, P. (2012). Current bioinformatic approaches to identify DNase I hypersensitive sites and genomic footprints from DNase-seq data. *Front Genet, 3*, 230. doi:10.3389/fgene.2012.00230

Maeda, N., Li, H., Lee, D., Oliver, P., Quarfordt, S. H., & Osada, J. (1994). Targeted disruption of the apolipoprotein C-III gene in mice results in hypotriglyceridemia and protection from postprandial hypertriglyceridemia. *J Biol Chem, 269*(38), 23610-23616.

Magi, A., Benelli, M., Yoon, S., Roviello, F., & Torricelli, F. (2011). Detecting common copy number variants in high-throughput sequencing data by using JointSLM algorithm. *Nucleic Acids Res, 39*(10), e65. doi:10.1093/nar/gkr068

Mahdy Ali, K., Wonnerth, A., Huber, K., & Wojta, J. (2012). Cardiovascular disease risk reduction by raising HDL cholesterol--current therapies and future opportunities. *Br J Pharmacol, 167*(6), 1177-1194. doi:10.1111/j.1476-5381.2012.02081.x

Malerba, G., Schaeffer, L., Xumerle, L., Klopp, N., Trabetti, E., Biscuola, M., . . . Illig, T. (2008). SNPs of the FADS gene cluster are associated with polyunsaturated fatty acids in a cohort of patients with cardiovascular disease. *Lipids, 43*(4), 289-299. doi:10.1007/s11745-008-3158-5

Malhotra, D., & Sebat, J. (2012). CNVs: harbingers of a rare variant revolution in psychiatric genetics. *Cell, 148*(6), 1223-1241. doi:10.1016/j.cell.2012.02.039

Maller, J. B., McVean, G., Byrnes, J., Vukcevic, D., Palin, K., Su, Z., . . . Consortium, W. T. C. C. (2012). Bayesian refinement of association signals for 14 loci in 3 common diseases. *Nat Genet, 44*(12), 1294-1301. doi:10.1038/ng.2435

Mangelsdorf, D. J., Thummel, C., Beato, M., Herrlich, P., Schütz, G., Umesono, K., . . . Evans, R. M. (1995). The nuclear receptor superfamily: the second decade. *Cell, 83*(6), 835-839.

Mannila, M. N., Mahdessian, H., Franco-Cereceda, A., Eggertsen, G., de Faire, U., Syvänen, A. C., . . . van 't Hooft, F. M. (2013). Identification of a functional apolipoprotein E promoter polymorphism regulating plasma apolipoprotein E concentration. *Arterioscler Thromb Vasc Biol, 33*(5), 1063-1069. doi:10.1161/ATVBAHA.112.300353

Manolio, T. A. (2013). Bringing genome-wide association findings into clinical use. *Nat Rev Genet, 14*(8), 549-

558. doi:10.1038/nrg3523

Manry, J., Laval, G., Patin, E., Fornarino, S., Itan, Y., Fumagalli, M., . . . Quintana-Murci, L. (2011). Evolutionary genetic dissection of human interferons. *J Exp Med, 208*(13), 2747-2759. doi:10.1084/jem.20111680

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nat Rev Genet, 11*(7), 499-511. doi:10.1038/nrg2796

Marchini, J., Howie, B., Myers, S., McVean, G., & Donnelly, P. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet, 39*(7), 906-913. doi:10.1038/ng2088

Marshall, C. R., & Scherer, S. W. (2012). Detection and characterization of copy number variation in autism spectrum disorder. *Methods Mol Biol, 838*, 115-135. doi:10.1007/978-1-61779-507-7_5

Martinelli, N., Consoli, L., & Olivieri, O. (2009). A 'desaturase hypothesis' for atherosclerosis: Janus-faced enzymes in omega-6 and omega-3 polyunsaturated fatty acid metabolism. *J Nutrigenet Nutrigenomics, 2*(3), 129-139. doi:10.1159/000238177

Martinelli, N., Girelli, D., Malerba, G., Guarini, P., Illig, T., Trabetti, E., . . . Olivieri, O. (2008). FADS genotypes and desaturase activity estimated by the ratio of arachidonic acid to linoleic acid are associated with inflammation and coronary artery disease. *Am J Clin Nutr, 88*(4), 941-949.

Martis, S., Mei, H., Vijzelaar, R., Edelmann, L., Desnick, R. J., & Scott, S. A. (2013). Multi-ethnic cytochrome-P450 copy number profiling: novel pharmacogenetic alleles and mechanism of copy number variation formation. *Pharmacogenomics J, 13*(6), 558-566. doi:10.1038/tpj.2012.48

Mathias, R. A., Sergeant, S., Ruczinski, I., Torgerson, D. G., Hugenschmidt, C. E., Kubala, M., . . . Chilton, F. H. (2011). The impact of FADS genetic variants on ω6 polyunsaturated fatty acid metabolism in African Americans. *BMC Genet, 12*, 50. doi:10.1186/1471-2156-12-50

Maurer-Stroh, S., Dickens, N. J., Hughes-Davies, L., Kouzarides, T., Eisenhaber, F., & Ponting, C. P. (2003). The Tudor domain 'Royal Family': Tudor, plant Agenet, Chromo, PWWP and MBT domains. *Trends Biochem Sci, 28*(2), 69-74. doi:10.1016/S0968-0004(03)00004-5

McCarroll, S. A. (2008). Extending genome-wide association studies to copy-number variation. *Hum Mol Genet, 17*(R2), R135-142. doi:10.1093/hmg/ddn282

McCarroll, S. A., Hadnott, T. N., Perry, G. H., Sabeti, P. C., Zody, M. C., Barrett, J. C., . . . Consortium, I. H. (2006). Common deletion polymorphisms in the human genome. *Nat Genet, 38*(1), 86-92.

McCarroll, S. A., Huett, A., Kuballa, P., Chilewski, S. D., Landry, A., Goyette, P., . . . Xavier, R. J. (2008).

Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet, 40*(9), 1107-1112. doi:10.1038/ng.215

McCarroll, S. A., Kuruvilla, F. G., Korn, J. M., Cawley, S., Nemesh, J., Wysoker, A., . . . Altshuler, D. (2008). Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nat Genet, 40*(10), 1166-1174. doi:10.1038/ng.238

McFarland, V. A., & Clarke, J. U. (1989). Environmental occurrence, abundance, and potential toxicity of polychlorinated biphenyl congeners: considerations for a congener-specific analysis. *Environ Health Perspect, 81*, 225-239.

McNamara, R. K., Jandacek, R., Rider, T., Tso, P., Dwivedi, Y., & Pandey, G. N. (2013). Adult medication-free schizophrenic patients exhibit long-chain omega-3 Fatty Acid deficiency: implications for cardiovascular disease risk. *Cardiovasc Psychiatry Neurol, 2013*, 796462. doi:10.1155/2013/796462

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., . . . Pritchard, J. K. (2013). Identification of genetic variants that affect histone modifications in human cells. *Science, 342*(6159), 747-749. doi:10.1126/science.1242429

Medvedev, P., Fiume, M., Dzamba, M., Smith, T., & Brudno, M. (2010). Detecting copy number variation with mated short reads. *Genome Res, 20*(11), 1613-1622. doi:10.1101/gr.106344.110

Medzhitov, R. (2001). Toll-like receptors and innate immunity. *Nat Rev Immunol, 1*(2), 135-145. doi:10.1038/35100529

Mega, J. L., Morrow, D. A., Brown, A., Cannon, C. P., & Sabatine, M. S. (2009). Identification of genetic variants associated with response to statin therapy. *Arterioscler Thromb Vasc Biol, 29*(9), 1310-1315. doi:10.1161/ATVBAHA.109.188474

Megson, D., Kalin, R., Worsfold, P. J., Gauchotte-Lindsay, C., Patterson, D. G., Lohan, M. C., . . . O' Sullivan, G. (2013). Fingerprinting polychlorinated biphenyls in environmental samples using comprehensive two-dimensional gas chromatography with time-of-flight mass spectrometry. *J Chromatogr A, 1318*, 276-283.

Mendenhall, E. M., Koche, R. P., Truong, T., Zhou, V. W., Issac, B., Chi, A. S., . . . Bernstein, B. E. (2010). GC-rich sequence elements recruit PRC2 in mammalian ES cells. *PLoS Genet, 6*(12), e1001244. doi:10.1371/journal.pgen.1001244

Merino, D. M., Johnston, H., Clarke, S., Roke, K., Nielsen, D., Badawi, A., . . . Mutch, D. M. (2011). Polymorphisms in FADS1 and FADS2 alter desaturase activity in young Caucasian and Asian adults.

Mol Genet Metab, 103(2), 171-178. doi:10.1016/j.ymgme.2011.02.012

Miller, C. A., Hampton, O., Coarfa, C., & Milosavljevic, A. (2011). ReadDepth: a parallel R package for detecting copy number alterations from short sequencing reads. *PLoS One, 6*(1), e16327. doi:10.1371/journal.pone.0016327

Mills, R. E., Walter, K., Stewart, C., Handsaker, R. E., Chen, K., Alkan, C., . . . Project, G. (2011). Mapping copy number variation by population-scale genome sequencing. *Nature, 470*(7332), 59-65. doi:10.1038/nature09708

Mizuarai, S., Aozasa, N., & Kotani, H. (2004). Single nucleotide polymorphisms result in impaired membrane localization and reduced atpase activity in multidrug transporter ABCG2. *Int J Cancer, 109*(2), 238-246. doi:10.1002/ijc.11669

Molina, J., Carmona-Mora, P., Chrast, J., Krall, P. M., Canales, C. P., Lupski, J. R., . . . Walz, K. (2008). Abnormal social behaviors and altered gene expression rates in a mouse model for Potocki-Lupski syndrome. *Hum Mol Genet, 17*(16), 2486-2495. doi:10.1093/hmg/ddn148

Morris, A. P. (2011). Transethnic meta-analysis of genomewide association studies. *Genet Epidemiol, 35*(8), 809-822. doi:10.1002/gepi.20630

Morris, A. P., Voight, B. F., Teslovich, T. M., Ferreira, T., Segrè, A. V., Steinthorsdottir, V., . . . Consortium, D. G. R. A. M.-a. D. (2012). Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. *Nat Genet, 44*(9), 981-990. doi:10.1038/ng.2383

Mulin, J., Pak, C., & Wang, J. (2013). GWASrap.

Mägi, R., & Morris, A. P. (2010). GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics, 11*, 288. doi:10.1186/1471-2105-11-288

Nakagawa, M., Ooie, T., Takahashi, N., Taniguchi, Y., Anan, F., Yonemochi, H., & Saikawa, T. (2006). Influence of menstrual cycle on QT interval dynamics. *Pacing Clin Electrophysiol, 29*(6), 607-613. doi:10.1111/j.1540-8159.2006.00407.x

Nannya, Y., Sanada, M., Nakazaki, K., Hosoya, N., Wang, L., Hangaishi, A., . . . Ogawa, S. (2005). A robust algorithm for copy number detection using high-density oligonucleotide single nucleotide polymorphism genotyping arrays. *Cancer Res, 65*(14), 6071-6079. doi:10.1158/0008-5472.CAN-05-0465

Narce, M., Asdrubal, P., Delachambre, M. C., Véricel, E., Lagarde, M., & Poisson, J. P. (1994). Age-related changes in linoleic acid bioconversion by isolated hepatocytes from spontaneously hypertensive and

normotensive rats. *Mol Cell Biochem, 141*(1), 9-13.

Need, A. C., Ge, D., Weale, M. E., Maia, J., Feng, S., Heinzen, E. L., . . . Goldstein, D. B. (2009). A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet, 5*(2), e1000373. doi:10.1371/journal.pgen.1000373

Neph, S., Vierstra, J., Stergachis, A. B., Reynolds, A. P., Haugen, E., Vernot, B., . . . Stamatoyannopoulos, J. A. (2012). An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature, 489*(7414), 83-90. doi:10.1038/nature11212

Ng, P. C., & Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome Res, 12*(3), 436-446. doi:10.1101/gr.212802

Ng, P. C., & Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. *Nucleic Acids Res, 31*(13), 3812-3814.

Nica, A. C., Parts, L., Glass, D., Nisbet, J., Barrett, A., Sekowska, M., . . . Consortium, M. (2011). The architecture of gene regulatory variation across multiple human tissues: the MuTHER study. *PLoS Genet, 7*(2), e1002003. doi:10.1371/journal.pgen.1002003

Nicholls, S. J., Dusting, G. J., Cutri, B., Bao, S., Drummond, G. R., Rye, K. A., & Barter, P. J. (2005). Reconstituted high-density lipoproteins inhibit the acute pro-oxidant and proinflammatory vascular changes induced by a periarterial collar in normocholesterolemic rabbits. *Circulation, 111*(12), 1543-1550. doi:10.1161/01.CIR.0000159351.95399.50

Nichols, B. A., Bainton, D. F., & Farquhar, M. G. (1971). Differentiation of monocytes. Origin, nature, and fate of their azurophil granules. *J Cell Biol, 50*(2), 498-515.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E., & Cox, N. J. (2010). Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet, 6*(4), e1000888. doi:10.1371/journal.pgen.1000888

Nimblegen. http://www.nimblegen.com/news/events/webinar/albert.swf.

O'Connor, C. (2008). Fluorescence *In Situ* Hybridization (FISH). *Nature Education, **1***(1), 171.

Oda, H., Okamoto, I., Murphy, N., Chu, J., Price, S. M., Shen, M. M., . . . Reinberg, D. (2009). Monomethylation of histone H4-lysine 20 is involved in chromosome structure and stability and is essential for mouse development. *Mol Cell Biol, 29*(8), 2278-2295. doi:10.1128/MCB.01768-08

Ohlsson, R., Renkawitz, R., & Lobanenkov, V. (2001). CTCF is a uniquely versatile transcription regulator linked to epigenetics and disease. *Trends Genet, 17*(9), 520-527.

Olsson, I. M., Bensryd, I., Lundh, T., Ottosson, H., Skerfving, S., & Oskarsson, A. (2002). Cadmium in blood and urine--impact of sex, age, dietary intake, iron status, and former smoking--association of renal effects. *Environ Health Perspect, 110*(12), 1185-1190.

Ong, C. T., & Corces, V. G. (2009). Insulators as mediators of intra- and inter-chromosomal interactions: a common evolutionary theme. *J Biol, 8*(8), 73. doi:10.1186/jbiol165

Ong, C. T., & Corces, V. G. (2014). CTCF: an architectural protein bridging genome topology and function. *Nat Rev Genet, 15*(4), 234-246. doi:10.1038/nrg3663

Ortega Moreno, L., Copetti, M., Fontana, A., De Bonis, C., Salvemini, L., Trischitta, V., & Menzaghi, C. (2016). Evidence of a causal relationship between high serum adiponectin levels and increased cardiovascular mortality rate in patients with type 2 diabetes. *Cardiovasc Diabetol, 15*(1), 17. doi:10.1186/s12933-016-0339-z

Pacher, P., & Kunos, G. (2013). Modulating the endocannabinoid system in human health and disease--successes and failures. *FEBS J, 280*(9), 1918-1943. doi:10.1111/febs.12260

Packard, C. J., Demant, T., Stewart, J. P., Bedford, D., Caslake, M. J., Schwertfeger, G., . . . Seidel, D. (2000). Apolipoprotein B metabolism and the distribution of VLDL and LDL subfractions. *J Lipid Res, 41*(2), 305-318.

Panagiotou, O. A., Ioannidis, J. P., & Project, G.-W. S. (2012). What should the genome-wide significance threshold be? Empirical replication of borderline genetic associations. *Int J Epidemiol, 41*(1), 273-286. doi:10.1093/ije/dyr178

Patterson, D. G., Wong, L. Y., Turner, W. E., Caudill, S. P., Dipietro, E. S., McClure, P. C., . . . Needham, L. L. (2009). Levels in the U.S. population of those persistent organic pollutants (2003-2004) included in the Stockholm Convention or in other long range transboundary air pollution agreements. *Environ Sci Technol, 43*(4), 1211-1218.

Pender-Cudlip, M. C., Krag, K. J., Martini, D., Yu, J., Guidi, A., Skinner, S. S., . . . Kang, J. X. (2013). Delta-6-desaturase activity and arachidonic acid synthesis are increased in human breast cancer tissue. *Cancer Sci, 104*(6), 760-764. doi:10.1111/cas.12129

Pennacchio, L. A., Olivier, M., Hubacek, J. A., Cohen, J. C., Cox, D. R., Fruchart, J. C., . . . Rubin, E. M. (2001). An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science, 294*(5540), 169-173. doi:10.1126/science.1064852

Pereira, T. V., Patsopoulos, N. A., Salanti, G., & Ioannidis, J. P. (2009). Discovery properties of genome-wide

association signals from cumulatively combined data sets. *Am J Epidemiol, 170*(10), 1197-1206. doi:10.1093/aje/kwp262

Petkeviciene, J., Smalinskiene, A., Luksiene, D. I., Jureniene, K., Ramazauskiene, V., Klumbiene, J., & Lesauskaite, V. (2012). Associations between apolipoprotein E genotype, diet, body mass index, and serum lipids in Lithuanian adult population. *PLoS One, 7*(7), e41525. doi:10.1371/journal.pone.0041525

Pinto, D., Pagnamenta, A. T., Klei, L., Anney, R., Merico, D., Regan, R., . . . Betancur, C. (2010). Functional impact of global rare copy number variation in autism spectrum disorders. *Nature, 466*(7304), 368-372. doi:10.1038/nature09146

Pollex, R. L., & Hegele, R. A. (2007). Genetic determinants of plasma lipoproteins. *Nat Clin Pract Cardiovasc Med, 4*(11), 600-609. doi:10.1038/ncpcardio1005

Pollin, T. I., Damcott, C. M., Shen, H., Ott, S. H., Shelton, J., Horenstein, R. B., . . . Shuldiner, A. R. (2008). A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science, 322*(5908), 1702-1705. doi:10.1126/science.1161524

Postmus, I., Trompet, S., Deshmukh, H. A., Barnes, M. R., Li, X., Warren, H. R., . . . Consortium, W. T. C. C. (2014). Pharmacogenetic meta-analysis of genome-wide association studies of LDL cholesterol response to statins. *Nat Commun, 5*, 5068. doi:10.1038/ncomms6068

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D., . . . Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet, 81*(3), 559-575. doi:10.1086/519795

Qiu, Z., Sylwestrak, E. L., Lieberman, D. N., Zhang, Y., Liu, X. Y., & Ghosh, A. (2012). The Rett syndrome protein MeCP2 regulates synaptic scaling. *J Neurosci, 32*(3), 989-994. doi:10.1523/JNEUROSCI.0175-11.2012

Quadri, M., Federico, A., Zhao, T., Breedveld, G. J., Battisti, C., Delnooz, C., . . . Bonifati, V. (2012). Mutations in SLC30A10 cause parkinsonism and dystonia with hypermanganesemia, polycythemia, and chronic liver disease. *Am J Hum Genet, 90*(3), 467-477. doi:10.1016/j.ajhg.2012.01.017

Rabiner, L. (1989). A tutorial on hidden markov-models and selected applications in speech recognition *Proceedings of the Ieee, 77*(2), 257-286. doi:10.1109/5.18626

Rahim, F., Jalali, A., & Tangestani, R. (2013). Breast cancer frequency and exposure to cadmium: a meta-analysis and systematic review. *Asian Pac J Cancer Prev, 14*(7), 4283-4287.

Rahman, M. M., Bhattacharya, A., Banu, J., Kang, J. X., & Fernandes, G. (2009). Endogenous n-3 fatty acids protect ovariectomy induced bone loss by attenuating osteoclastogenesis. *J Cell Mol Med, 13*(8B), 1833-1844. doi:10.1111/j.1582-4934.2009.00649.x

Ramana, C. V., Gil, M. P., Schreiber, R. D., & Stark, G. R. (2002). Stat1-dependent and -independent pathways in IFN-gamma-dependent signaling. *Trends Immunol, 23*(2), 96-101.

Rampersaud, E., Mitchell, B. D., Pollin, T. I., Fu, M., Shen, H., O'Connell, J. R., . . . Snitker, S. (2008). Physical activity and the association of common FTO gene variants with body mass index and obesity. *Arch Intern Med, 168*(16), 1791-1797. doi:10.1001/archinte.168.16.1791

Rasmussen-Torvik, L. J., Pacheco, J. A., Wilke, R. A., Thompson, W. K., Ritchie, M. D., Kho, A. N., . . . Chisholm, R. L. (2012). High density GWAS for LDL cholesterol in African Americans using electronic medical records reveals a strong protective variant in APOE. *Clin Transl Sci, 5*(5), 394-399. doi:10.1111/j.1752-8062.2012.00446.x

Raval, A., Tanner, S. M., Byrd, J. C., Angerman, E. B., Perko, J. D., Chen, S. S., . . . Plass, C. (2007). Downregulation of death-associated protein kinase 1 (DAPK1) in chronic lymphocytic leukemia. *Cell, 129*(5), 879-890. doi:10.1016/j.cell.2007.03.043

Redon, R., Ishikawa, S., Fitch, K. R., Feuk, L., Perry, G. H., Andrews, T. D., . . . Hurles, M. E. (2006). Global variation in copy number in the human genome. *Nature, 444*(7118), 444-454. doi:10.1038/nature05329

Ricote, M., Li, A. C., Willson, T. M., Kelly, C. J., & Glass, C. K. (1998). The peroxisome proliferator-activated receptor-gamma is a negative regulator of macrophage activation. *Nature, 391*(6662), 79-82. doi:10.1038/34178

Risérus, U. (2008). Fatty acids and insulin sensitivity. *Curr Opin Clin Nutr Metab Care, 11*(2), 100-105. doi:10.1097/MCO.0b013e3282f52708

Ritter, R., Scheringer, M., MacLeod, M., Moeckel, C., Jones, K. C., & Hungerbühler, K. (2011). Intrinsic human elimination half-lives of polychlorinated biphenyls derived from the temporal evolution of cross-sectional biomonitoring data from the United Kingdom. *Environ Health Perspect, 119*(2), 225-231. doi:10.1289/ehp.1002211

Rodriguez, C., Borgel, J., Court, F., Cathala, G., Forné, T., & Piette, J. (2010). CTCF is a DNA methylation-sensitive positive regulator of the INK/ARF locus. *Biochem Biophys Res Commun, 392*(2), 129-134. doi:10.1016/j.bbrc.2009.12.159

Rodriguez-Antona, C., Gomez, A., Karlgren, M., Sim, S. C., & Ingelman-Sundberg, M. (2010). Molecular

genetics and epigenetics of the cytochrome P450 gene family and its relevance for cancer risk and treatment. *Hum Genet, 127*(1), 1-17. doi:10.1007/s00439-009-0748-0

Rodushkin, I., Odman, F., Olofsson, R., & Axelsson, M. (2000). Determination of 60 elements in whole blood by sector field inductively coupled plasma mass spectrometry. *Journal of Analytical Atomic Spectrometry, 15*(8), 937-944. doi:10.1039/b003561k

Ross, R. (1999). Atherosclerosis--an inflammatory disease. *N Engl J Med, 340*(2), 115-126. doi:10.1056/NEJM199901143400207

Rouchka, E. C., Phatak, A. W., & Singh, A. V. (2008). Effect of single nucleotide polymorphisms on Affymetrix match-mismatch probe pairs. *Bioinformation, 2*(9), 405-411.

Sacks, F. M. (2015). The crucial roles of apolipoproteins E and C-III in apoB lipoprotein metabolism in normolipidemia and hypertriglyceridemia. *Curr Opin Lipidol, 26*(1), 56-63. doi:10.1097/MOL.0000000000000146

Saito, E., Okada, T., Abe, Y., Odaka, M., Kuromori, Y., Iwata, F., . . . Kitamura, Y. (2013). Abdominal adiposity is associated with fatty acid desaturase activity in boys: implications for C-reactive protein and insulin resistance. *Prostaglandins Leukot Essent Fatty Acids, 88*(4), 307-311. doi:10.1016/j.plefa.2013.01.005

Salihovic, S., Lampa, E., Lindström, G., Lind, L., Lind, P. M., & van Bavel, B. (2012). Circulating levels of persistent organic pollutants (POPs) among elderly men and women from Sweden: results from the Prospective Investigation of the Vasculature in Uppsala Seniors (PIVUS). *Environ Int, 44*, 59-67. doi:10.1016/j.envint.2012.01.011

Salihovic, S., Mattioli, L., Lindström, G., Lind, L., Lind, P. M., & van Bavel, B. (2012). A rapid method for screening of the Stockholm Convention POPs in small amounts of human plasma using SPE and HRGC/HRMS. *Chemosphere, 86*(7), 747-753. doi:10.1016/j.chemosphere.2011.11.006

Sanders, S. J., Ercan-Sencicek, A. G., Hus, V., Luo, R., Murtha, M. T., Moreno-De-Luca, D., . . . State, M. W. (2011). Multiple recurrent de novo CNVs, including duplications of the 7q11.23 Williams syndrome region, are strongly associated with autism. *Neuron, 70*(5), 863-885. doi:10.1016/j.neuron.2011.05.002

Sarkadi, B., Homolya, L., Szakács, G., & Váradi, A. (2006). Human multidrug resistance ABCB and ABCG transporters: participation in a chemoimmunity defense system. *Physiol Rev, 86*(4), 1179-1236. doi:10.1152/physrev.00037.2005

Sarwar, N., Danesh, J., Eiriksdottir, G., Sigurdsson, G., Wareham, N., Bingham, S., . . . Gudnason, V. (2007).

Triglycerides and the risk of coronary heart disease: 10,158 incident cases among 262,525 participants in 29 Western prospective studies. *Circulation, 115*(4), 450-458. doi:10.1161/CIRCULATIONAHA.106.637793

Sasieni, P. D. (1997). From genotypes to genes: doubling the sample size. *Biometrics, 53*(4), 1253-1261.

Sato, T., Murakumo, Y., Hagiwara, S., Jijiwa, M., Suzuki, C., Yatabe, Y., & Takahashi, M. (2007). High-level expression of CD109 is frequently detected in lung squamous cell carcinomas. *Pathol Int, 57*(11), 719-724. doi:10.1111/j.1440-1827.2007.02168.x

Schadt, E. E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P. Y., . . . Ulrich, R. (2008). Mapping the genetic architecture of gene expression in human liver. *PLoS Biol, 6*(5), e107. doi:10.1371/journal.pbio.0060107

Schaeffer, L., Gohlke, H., Müller, M., Heid, I. M., Palmer, L. J., Kompauer, I., . . . Heinrich, J. (2006). Common genetic variants of the FADS1 FADS2 gene cluster and their reconstructed haplotypes are associated with the fatty acid composition in phospholipids. *Hum Mol Genet, 15*(11), 1745-1756. doi:10.1093/hmg/ddl117

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Res, 22*(9), 1748-1759. doi:10.1101/gr.136127.111

Scheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet, 78*(4), 629-644. doi:10.1086/502802

Schlattl, A., Anders, S., Waszak, S. M., Huber, W., & Korbel, J. O. (2011). Relating CNVs to transcriptome data at fine resolution: assessment of the effect of variant size, type, and overlap with functional regions. *Genome Res, 21*(12), 2004-2013. doi:10.1101/gr.122614.111

Schneider, R., Bannister, A. J., Myers, F. A., Thorne, A. W., Crane-Robinson, C., & Kouzarides, T. (2004). Histone H3 lysine 4 methylation patterns in higher eukaryotic genes. *Nat Cell Biol, 6*(1), 73-77. doi:10.1038/ncb1076

Sebat, J., Lakshmi, B., Malhotra, D., Troge, J., Lese-Martin, C., Walsh, T., . . . Wigler, M. (2007). Strong association of de novo copy number mutations with autism. *Science, 316*(5823), 445-449. doi:10.1126/science.1138659

Sebat, J., Lakshmi, B., Troge, J., Alexander, J., Young, J., Lundin, P., . . . Wigler, M. (2004). Large-scale copy

number polymorphism in the human genome. *Science, 305*(5683), 525-528. doi:10.1126/science.1098918

Serhan, C. N., Chiang, N., & Van Dyke, T. E. (2008). Resolving inflammation: dual anti-inflammatory and pro-resolution lipid mediators. *Nat Rev Immunol, 8*(5), 349-361. doi:10.1038/nri2294

Service, S. K., Teslovich, T. M., Fuchsberger, C., Ramensky, V., Yajnik, P., Koboldt, D. C., . . . Freimer, N. B. (2014). Re-sequencing expands our understanding of the phenotypic impact of variants at GWAS loci. *PLoS Genet, 10*(1), e1004147. doi:10.1371/journal.pgen.1004147

Seve, M., Chimienti, F., Devergnas, S., & Favier, A. (2004). In silico identification and expression of SLC30 family genes: an expressed sequence tag data mining strategy for the characterization of zinc transporters' tissue expression. *BMC Genomics, 5*(1), 32. doi:10.1186/1471-2164-5-32

Sexton, T., Umlauf, D., Kurukuti, S., & Fraser, P. (2007). The role of transcription factories in large-scale structure and dynamics of interphase chromatin. *Semin Cell Dev Biol, 18*(5), 691-697. doi:10.1016/j.semcdb.2007.08.008

Shabalin, A. A. (2012). Matrix eQTL: ultra fast eQTL analysis via large matrix operations. *Bioinformatics, 28*(10), 1353-1358. doi:10.1093/bioinformatics/bts163

Shaffer, L. G., & Lupski, J. R. (2000). Molecular mechanisms for constitutional chromosomal rearrangements in humans. *Annu Rev Genet, 34*, 297-329. doi:10.1146/annurev.genet.34.1.297

Sharma, N. L., Groselj, B., Hamdy, F. C., & Kiltie, A. E. (2013). The emerging role of histone deacetylase (HDAC) inhibitors in urological cancers. *BJU Int, 111*(4), 537-542. doi:10.1111/j.1464-410X.2012.11647.x

Shi, G., Boerwinkle, E., Morrison, A. C., Gu, C. C., Chakravarti, A., & Rao, D. C. (2011). Mining gold dust under the genome wide significance level: a two-stage approach to analysis of GWAS. *Genet Epidemiol, 35*(2), 111-118. doi:10.1002/gepi.20556

Shu, W., Chen, H., Bo, X., & Wang, S. (2011). Genome-wide analysis of the relationships between DNaseI HS, histone modifications and gene expression reveals distinct modes of chromatin domains. *Nucleic Acids Res, 39*(17), 7428-7443. doi:10.1093/nar/gkr443

Shungin, D., Winkler, T. W., Croteau-Chonka, D. C., Ferreira, T., Locke, A. E., Mägi, R., . . . Consortium, R. (2015). New genetic loci link adipose and insulin biology to body fat distribution. *Nature, 518*(7538), 187-196. doi:10.1038/nature14132

Sims, R. J., Chen, C. F., Santos-Rosa, H., Kouzarides, T., Patel, S. S., & Reinberg, D. (2005). Human but not

yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. *J Biol Chem, 280*(51), 41789-41792. doi:10.1074/jbc.C500395200

Sladek, F. M., Zhong, W. M., Lai, E., & Darnell, J. E. (1990). Liver-enriched transcription factor HNF-4 is a novel member of the steroid hormone receptor superfamily. *Genes Dev, 4*(12B), 2353-2365.

Socha, A. L., & Guerinot, M. L. (2014). Mn-euvering manganese: the role of transporter gene family members in manganese uptake and mobilization in plants. *Front Plant Sci, 5*, 106. doi:10.3389/fpls.2014.00106

Sorg, O. (2014). AhR signalling and dioxin toxicity. *Toxicol Lett, 230*(2), 225-233. doi:10.1016/j.toxlet.2013.10.039

Speicher, M. R., du Manoir, S., Schröck, E., Holtgreve-Grez, H., Schoell, B., Lengauer, C., . . . Ried, T. (1993). Molecular cytogenetic analysis of formalin-fixed, paraffin-embedded solid tumors by comparative genomic hybridization after universal DNA-amplification. *Hum Mol Genet, 2*(11), 1907-1914.

Speliotes, E. K., Willer, C. J., Berndt, S. I., Monda, K. L., Thorleifsson, G., Jackson, A. U., . . . Consortium, P. (2010). Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nat Genet, 42*(11), 937-948. doi:10.1038/ng.686

Stadler, M. B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., . . . Schübeler, D. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. *Nature, 480*(7378), 490-495. doi:10.1038/nature10716

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends Genet, 18*(2), 74-82.

Stankiewicz, P., & Lupski, J. R. (2010). Structural variation in the human genome and its role in disease. *Annu Rev Med, 61*, 437-455. doi:10.1146/annurev-med-100708-204735

State of the Science of Endocrine Disrupting Chemicals. (2012). H. J. Bergman A, Jobling S, Kidd KA, Zoeller RT (Ed.) (pp. 6-7).

Stefansson, H., Rujescu, D., Cichon, S., Pietiläinen, O. P., Ingason, A., Steinberg, S., . . . GROUP. (2008). Large recurrent microdeletions associated with schizophrenia. *Nature, 455*(7210), 232-236. doi:10.1038/nature07229

Stranger, B. E., Forrest, M. S., Dunning, M., Ingle, C. E., Beazley, C., Thorne, N., . . . Dermitzakis, E. T. (2007). Relative impact of nucleotide and copy number variation on gene expression phenotypes. *Science, 315*(5813), 848-853. doi:10.1126/science.1136678

Stranger, B. E., Stahl, E. A., & Raj, T. (2011). Progress and promise of genome-wide association studies for

human complex trait genetics. *Genetics, 187*(2), 367-383. doi:10.1534/genetics.110.120907

Suc, I., Escargueil-Blanc, I., Troly, M., Salvayre, R., & Nègre-Salvayre, A. (1997). HDL and ApoA prevent cell death of endothelial cells induced by oxidized LDL. *Arterioscler Thromb Vasc Biol, 17*(10), 2158-2166.

Swenne, I., & Vessby, B. (2013). Relationship of Δ(6) -desaturase and Δ(5) -desaturase activities with thyroid hormone status in adolescents with eating disorders and weight loss. *Acta Paediatr, 102*(4), 416-418. doi:10.1111/apa.12132

Sérandour, A. A., Avner, S., Percevault, F., Demay, F., Bizot, M., Lucchetti-Miganeh, C., . . . Eeckhoute, J. (2011). Epigenetic switch involved in activation of pioneer factor FOXA1-dependent enhancers. *Genome Res, 21*(4), 555-565. doi:10.1101/gr.111534.110

Tabara, Y., Kohara, K., Kawamoto, R., Hiura, Y., Nishimura, K., Morisaki, T., . . . Miki, T. (2010). Association of four genetic loci with uric acid levels and reduced renal function: the J-SHIPP Suita study. *Am J Nephrol, 32*(3), 279-286. doi:10.1159/000318943

Takeuchi, F., Isono, M., Katsuya, T., Yokota, M., Yamamoto, K., Nabika, T., . . . Kato, N. (2012). Association of genetic variants influencing lipid levels with coronary artery disease in Japanese individuals. *PLoS One, 7*(9), e46385. doi:10.1371/journal.pone.0046385

Tall, A. R. (2010). Functions of cholesterol ester transfer protein and relationship to coronary artery disease risk. *J Clin Lipidol, 4*(5), 389-393. doi:10.1016/j.jacl.2010.08.006

Talmud, P. J., Hawe, E., Miller, G. J., & Humphries, S. E. (2002). Nonfasting apolipoprotein B and triglyceride levels as a useful predictor of coronary heart disease risk in middle-aged UK men. *Arterioscler Thromb Vasc Biol, 22*(11), 1918-1923.

Tan, K. P., Wang, B., Yang, M., Boutros, P. C., Macaulay, J., Xu, H., . . . Ito, S. (2010). Aryl hydrocarbon receptor is a transcriptional activator of the human breast cancer resistance protein (BCRP/ABCG2). *Mol Pharmacol, 78*(2), 175-185. doi:10.1124/mol.110.065078

Tanaka, T., Shen, J., Abecasis, G. R., Kisialiou, A., Ordovas, J. M., Guralnik, J. M., . . . Ferrucci, L. (2009). Genome-wide association study of plasma polyunsaturated fatty acids in the InCHIANTI Study. *PLoS Genet, 5*(1), e1000338. doi:10.1371/journal.pgen.1000338

Tejedor, M. T., Garcia-Sobreviela, M. P., Ledesma, M., & Arbones-Mainar, J. M. (2014). The apolipoprotein E polymorphism rs7412 associates with body fatness independently of plasma lipids in middle aged men. *PLoS One, 9*(9), e108605. doi:10.1371/journal.pone.0108605

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., . . . Kathiresan, S. (2010). Biological, clinical and population relevance of 95 loci for blood lipids. *Nature, 466*(7307), 707-713. doi:10.1038/nature09270

Theisen, A. (2008). Microarray-based Comparative Genomic Hybridization. *Nature Education, 1*(1), 45.

Thompson, D. J., O'Mara, T. A., Glubb, D. M., Painter, J. N., Cheng, T., Folkerd, E., . . . Group, A. (2016). CYP19A1 fine-mapping and Mendelian randomization: estradiol is causal for endometrial cancer. *Endocr Relat Cancer, 23*(2), 77-91. doi:10.1530/ERC-15-0386

Thompson, J. F., Hyde, C. L., Wood, L. S., Paciga, S. A., Hinds, D. A., Cox, D. R., . . . Kastelein, J. J. (2009). Comprehensive whole-genome and candidate gene analysis for response to statin therapy in the Treating to New Targets (TNT) cohort. *Circ Cardiovasc Genet, 2*(2), 173-181. doi:10.1161/CIRCGENETICS.108.818062

Thompson, J. F., Lira, M. E., Durham, L. K., Clark, R. W., Bamberger, M. J., & Milos, P. M. (2003). Polymorphisms in the CETP gene and association with CETP mass and HDL levels. *Atherosclerosis, 167*(2), 195-204.

Thompson, J. F., Lloyd, D. B., Lira, M. E., & Milos, P. M. (2004). Cholesteryl ester transfer protein promoter single-nucleotide polymorphisms in Sp1-binding sites affect transcription and are associated with high-density lipoprotein cholesterol. *Clin Genet, 66*(3), 223-228. doi:10.1111/j.1399-0004.2004.00289.x

Thompson, J. F., Wood, L. S., Pickering, E. H., Dechairo, B., & Hyde, C. L. (2007). High-density genotyping and functional SNP localization in the CETP gene. *J Lipid Res, 48*(2), 434-443. doi:10.1194/jlr.M600372-JLR200

Thorisson, G. A., Smith, A. V., Krishnan, L., & Stein, L. D. (2005). The International HapMap Project Web site. *Genome Res, 15*(11), 1592-1593. doi:10.1101/gr.4413105

Tian, C., Ambrosone, C. B., Darcy, K. M., Krivak, T. C., Armstrong, D. K., Bookman, M. A., . . . DeLoia, J. A. (2012). Common variants in ABCB1, ABCC2 and ABCG2 genes and clinical outcomes among women with advanced stage ovarian cancer treated with platinum and taxane-based chemotherapy: a Gynecologic Oncology Group study. *Gynecol Oncol, 124*(3), 575-581. doi:10.1016/j.ygyno.2011.11.022

Toffolatti, L., Cardazzo, B., Nobile, C., Danieli, G. A., Gualandi, F., Muntoni, F., . . . Patarnello, T. (2002). Investigating the mechanism of chromosomal deletion: characterization of 39 deletion breakpoints in introns 47 and 48 of the human dystrophin gene. *Genomics, 80*(5), 523-530.

Tokoro, F., Matsuoka, R., Abe, S., Arai, M., Noda, T., Watanabe, S., . . . Yamada, Y. (2015). Association of a genetic variant of the ZPR1 zinc finger gene with type 2 diabetes mellitus. *Biomed Rep, 3*(1), 88-92. doi:10.3892/br.2014.379

Tomita-Mitchell, A., Mahnke, D. K., Struble, C. A., Tuffnell, M. E., Stamm, K. D., Hidestrand, M., . . . Mitchell, M. E. (2012). Human gene copy number spectra analysis in congenital heart malformations. *Physiol Genomics, 44*(9), 518-541. doi:10.1152/physiolgenomics.00013.2012

Trojer, P., & Reinberg, D. (2007). Facultative heterochromatin: is there a distinctive molecular signature? *Mol Cell, 28*(1), 1-13. doi:10.1016/j.molcel.2007.09.011

Trynka, G., Westra, H. J., Slowikowski, K., Hu, X., Xu, H., Stranger, B. E., . . . Raychaudhuri, S. (2015). Disentangling the Effects of Colocalizing Genomic Annotations to Functionally Prioritize Non-coding Variants within Complex-Trait Loci. *Am J Hum Genet, 97*(1), 139-152. doi:10.1016/j.ajhg.2015.05.016

Turner, S., Armstrong, L. L., Bradford, Y., Carlson, C. S., Crawford, D. C., Crenshaw, A. T., . . . Ritchie, M. D. (2011). Quality control procedures for genome-wide association studies. *Curr Protoc Hum Genet, Chapter 1*, Unit1.19. doi:10.1002/0471142905.hg0119s68

Usmani, K. A., Hodgson, E., & Rose, R. L. (2004). In vitro metabolism of carbofuran by human, mouse, and rat cytochrome P450 and interactions with chlorpyrifos, testosterone, and estradiol. *Chem Biol Interact, 150*(3), 221-232. doi:10.1016/j.cbi.2004.09.015

Utermann, G., Pruin, N., & Steinmetz, A. (1979). Polymorphism of apolipoprotein E. III. Effect of a single polymorphic gene locus on plasma lipid levels in man. *Clin Genet, 15*(1), 63-72.

van de Woestijne, A. P., van der Graaf, Y., de Bakker, P. I., Asselbergs, F. W., Spiering, W., Visseren, F. L., & Group, S. S. (2014). Rs964184 (APOA5-A4-C3-A1) is related to elevated plasma triglyceride levels, but not to an increased risk for vascular events in patients with clinically manifest vascular disease. *PLoS One, 9*(6), e101082. doi:10.1371/journal.pone.0101082

Varambally, S., Dhanasekaran, S. M., Zhou, M., Barrette, T. R., Kumar-Sinha, C., Sanda, M. G., . . . Chinnaiyan, A. M. (2002). The polycomb group protein EZH2 is involved in progression of prostate cancer. *Nature, 419*(6907), 624-629. doi:10.1038/nature01075

Varbo, A., Benn, M., & Nordestgaard, B. G. (2014). Remnant cholesterol as a cause of ischemic heart disease: evidence, definition, measurement, atherogenicity, high risk patients, and present and future treatment. *Pharmacol Ther, 141*(3), 358-367. doi:10.1016/j.pharmthera.2013.11.008

Vega, G. L., & Grundy, S. M. (1986). In vivo evidence for reduced binding of low density lipoproteins to

receptors as a cause of primary moderate hypercholesterolemia. *J Clin Invest, 78*(5), 1410-1414. doi:10.1172/JCI112729

Vermeesch, J. R., Melotte, C., Froyen, G., Van Vooren, S., Dutta, B., Maas, N., . . . Devriendt, K. (2005). Molecular karyotyping: array CGH quality criteria for constitutional genetic diagnosis. *J Histochem Cytochem, 53*(3), 413-422. doi:10.1369/jhc.4A6436.2005

Vermeulen, M., Eberl, H. C., Matarese, F., Marks, H., Denissov, S., Butter, F., . . . Mann, M. (2010). Quantitative interaction proteomics and genome-wide profiling of epigenetic histone marks and their readers. *Cell, 142*(6), 967-980. doi:10.1016/j.cell.2010.08.020

Volcik, K. A., Barkley, R. A., Hutchinson, R. G., Mosley, T. H., Heiss, G., Sharrett, A. R., . . . Boerwinkle, E. (2006). Apolipoprotein E polymorphisms predict low density lipoprotein cholesterol levels and carotid artery wall thickness but not incident coronary heart disease in 12,491 ARIC study participants. *Am J Epidemiol, 164*(4), 342-348. doi:10.1093/aje/kwj202

Wakefield, J. (2007). A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am J Hum Genet, 81*(2), 208-227. doi:10.1086/519024

Wakefield, J. (2009). Bayes factors for genome-wide association studies: comparison with P-values. *Genet Epidemiol, 33*(1), 79-86. doi:10.1002/gepi.20359

Wang, D., & Dubois, R. N. (2010). Eicosanoids and cancer. *Nat Rev Cancer, 10*(3), 181-193. doi:10.1038/nrc2809

Wang, H., Maurano, M. T., Qu, H., Varley, K. E., Gertz, J., Pauli, F., . . . Stamatoyannopoulos, J. A. (2012). Widespread plasticity in CTCF occupancy linked to DNA methylation. *Genome Res, 22*(9), 1680-1688. doi:10.1101/gr.136101.111

Wang, K., Li, M., Hadley, D., Liu, R., Glessner, J., Grant, S. F., . . . Bucan, M. (2007). PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res, 17*(11), 1665-1674. doi:10.1101/gr.6861907

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res, 38*(16), e164. doi:10.1093/nar/gkq603

Warensjö, E., Risérus, U., & Vessby, B. (2005). Fatty acid composition of serum lipids predicts the development of the metabolic syndrome in men. *Diabetologia, 48*(10), 1999-2005. doi:10.1007/s00125-005-1897-x

Warensjö, E., Sundström, J., Vessby, B., Cederholm, T., & Risérus, U. (2008). Markers of dietary fat quality and fatty acid desaturation as predictors of total and cardiovascular mortality: a population-based

prospective study. *Am J Clin Nutr, 88*(1), 203-209.

Warner, N. A., Martin, J. W., & Wong, C. S. (2009). Chiral polychlorinated biphenyls are biotransformed enantioselectively by mammalian cytochrome P-450 isozymes to form hydroxylated metabolites. *Environ Sci Technol, 43*(1), 114-121.

Waszak, S. M., Hasin, Y., Zichner, T., Olender, T., Keydar, I., Khen, M., . . . Korbel, J. O. (2010). Systematic inference of copy-number genotypes from personal genome sequencing data reveals extensive olfactory receptor gene content diversity. *PLoS Comput Biol, 6*(11), e1000988. doi:10.1371/journal.pcbi.1000988

Weikert, S., Christoph, F., Köllermann, J., Müller, M., Schrader, M., Miller, K., & Krause, H. (2005). Expression levels of the EZH2 polycomb transcriptional repressor correlate with aggressiveness and invasive potential of bladder carcinomas. *Int J Mol Med, 16*(2), 349-353.

Weiss, L. A., Shen, Y., Korn, J. M., Arking, D. E., Miller, D. T., Fossdal, R., . . . Consortium, A. (2008). Association between microdeletion and microduplication at 16p11.2 and autism. *N Engl J Med, 358*(7), 667-675. doi:10.1056/NEJMoa075974

Weiss, M. M., Hermsen, M. A., Meijer, G. A., van Grieken, N. C., Baak, J. P., Kuipers, E. J., & van Diest, P. J. (1999). Comparative genomic hybridisation. *Mol Pathol, 52*(5), 243-251.

Wellcome Trust Case Control Consortium. (2007). Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature, 447*(7145), 661-678. doi:10.1038/nature05911

Wen, W., Zheng, W., Okada, Y., Takeuchi, F., Tabara, Y., Hwang, J. Y., . . . Tanaka, T. (2014). Meta-analysis of genome-wide association studies in East Asian-ancestry populations identifies four new loci for body mass index. *Hum Mol Genet, 23*(20), 5492-5504. doi:10.1093/hmg/ddu248

Westra, H. J., & Franke, L. (2014). From genome to function by studying eQTLs. *Biochim Biophys Acta, 1842*(10), 1896-1902. doi:10.1016/j.bbadis.2014.04.024

Willer, C. J., Li, Y., & Abecasis, G. R. (2010). METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics, 26*(17), 2190-2191. doi:10.1093/bioinformatics/btq340

Wilson, P. W., Abbott, R. D., & Castelli, W. P. (1988). High density lipoprotein cholesterol and mortality. The Framingham Heart Study. *Arteriosclerosis, 8*(6), 737-741.

Wimmerová, S., Lancz, K., Tihányi, J., Sovčíková, E., Kočan, A., Drobná, B., . . . Trnovec, T. (2011). Half-lives of serum PCB congener concentrations in environmentally exposed early adolescents. *Chemosphere, 82*(5), 687-691. doi:10.1016/j.chemosphere.2010.10.099

Wineinger, N. E., Pajewski, N. M., & Tiwari, H. K. (2011). A Method to Assess Linkage Disequilibrium

between CNVs and SNPs Inside Copy Number Variable Regions. *Front Genet, 2*, 17. doi:10.3389/fgene.2011.00017

Wirfält, E., Vessby, B., Mattisson, I., Gullberg, B., Olsson, H., & Berglund, G. (2004). No relations between breast cancer risk and fatty acids of erythrocyte membranes in postmenopausal women of the Malmö Diet Cancer cohort (Sweden). *Eur J Clin Nutr, 58*(5), 761-770. doi:10.1038/sj.ejcn.1601874

Wittke-Thompson, J. K., Pluzhnikov, A., & Cox, N. J. (2005). Rational inferences about departures from Hardy-Weinberg equilibrium. *Am J Hum Genet, 76*(6), 967-986. doi:10.1086/430507

Wolff, M. S., Camann, D., Gammon, M., & Stellman, S. D. Proposed PCB congener groupings for epidemiological studies.

Wolff, M. S., Camann, D., Gammon, M., & Stellman, S. D. (1997). Proposed PCB congener groupings for epidemiological studies. *Environ Health Perspect, 105*(1), 13-14.

Wong, C. K., Vaske, C. J., Ng, S., Sanborn, J. Z., Benz, S. C., Haussler, D., & Stuart, J. M. (2013). The UCSC Interaction Browser: multidimensional data views in pathway context. *Nucleic Acids Res, 41*(Web Server issue), W218-224. doi:10.1093/nar/gkt473

Wong, K. K., deLeeuw, R. J., Dosanjh, N. S., Kimm, L. R., Cheng, Z., Horsman, D. E., . . . Lam, W. L. (2007). A comprehensive analysis of common copy-number variations in the human genome. *Am J Hum Genet, 80*(1), 91-104. doi:10.1086/510560

Woodward, O. M., Köttgen, A., Coresh, J., Boerwinkle, E., Guggino, W. B., & Köttgen, M. (2009). Identification of a urate transporter, ABCG2, with a common functional polymorphism causing gout. *Proc Natl Acad Sci U S A, 106*(25), 10338-10342. doi:10.1073/pnas.0901249106

World Health Organisation. (2014). Dioxins and their effects on human health.

Xie, C., & Tammi, M. T. (2009). CNV-seq, a new method to detect copy number variation using high-throughput sequencing. *BMC Bioinformatics, 10*, 80. doi:10.1186/1471-2105-10-80

Xu, B., Roos, J. L., Levy, S., van Rensburg, E. J., Gogos, J. A., & Karayiorgou, M. (2008). Strong association of de novo copy number mutations with sporadic schizophrenia. *Nat Genet, 40*(7), 880-885. doi:10.1038/ng.162

Yamagishi, K., Tanigawa, T., Kitamura, A., Köttgen, A., Folsom, A. R., Iso, H., & Investigators, C. (2010). The rs2231142 variant of the ABCG2 gene is associated with uric acid levels and gout among Japanese people. *Rheumatology (Oxford), 49*(8), 1461-1465. doi:10.1093/rheumatology/keq096

Yamazaki, K., McGovern, D., Ragoussis, J., Paolucci, M., Butler, H., Jewell, D., . . . Nakamura, Y. (2005).

Single nucleotide polymorphisms in TNFSF15 confer susceptibility to Crohn's disease. *Hum Mol Genet, 14*(22), 3499-3506. doi:10.1093/hmg/ddi379

Yan, J., Enge, M., Whitington, T., Dave, K., Liu, J., Sur, I., . . . Taipale, J. (2013). Transcription factor binding in human cells occurs in dense clusters formed around cohesin anchor sites. *Cell, 154*(4), 801-813. doi:10.1016/j.cell.2013.07.034

Yang, J., Weedon, M. N., Purcell, S., Lettre, G., Estrada, K., Willer, C. J., . . . Consortium, G. (2011). Genomic inflation factors under polygenic inheritance. *Eur J Hum Genet, 19*(7), 807-812. doi:10.1038/ejhg.2011.39

Yang, X. J., & Seto, E. (2007). HATs and HDACs: from structure, function and regulation to novel strategies for therapy and prevention. *Oncogene, 26*(37), 5310-5318. doi:10.1038/sj.onc.1210599

Ylstra, B., van den Ijssel, P., Carvalho, B., Brakenhoff, R. H., & Meijer, G. A. (2006). BAC to the future! or oligonucleotides: a perspective for micro array comparative genomic hybridization (array CGH). *Nucleic Acids Res, 34*(2), 445-450. doi:10.1093/nar/gkj456

Yoon, S., Xuan, Z., Makarov, V., Ye, K., & Sebat, J. (2009). Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res, 19*(9), 1586-1592. doi:10.1101/gr.092981.109

Yoon, Y. S., Lee, M. W., Ryu, D., Kim, J. H., Ma, H., Seo, W. Y., . . . Koo, S. H. (2010). Suppressor of MEK null (SMEK)/protein phosphatase 4 catalytic subunit (PP4C) is a key regulator of hepatic gluconeogenesis. *Proc Natl Acad Sci U S A, 107*(41), 17704-17709. doi:10.1073/pnas.1012665107

Zampieri, M., Guastafierro, T., Calabrese, R., Ciccarone, F., Bacalini, M. G., Reale, A., . . . Caiafa, P. (2012). ADP-ribose polymers localized on Ctcf-Parp1-Dnmt1 complex prevent methylation of Ctcf target sites. *Biochem J, 441*(2), 645-652. doi:10.1042/BJ20111417

Zegerman, P., Canas, B., Pappin, D., & Kouzarides, T. (2002). Histone H3 lysine 4 methylation disrupts binding of nucleosome remodeling and deacetylase (NuRD) repressor complex. *J Biol Chem, 277*(14), 11621-11624. doi:10.1074/jbc.C200045200

Zhang, D., Qian, Y., Akula, N., Alliey-Rodriguez, N., Tang, J., Gershon, E. S., . . . Study, B. G. (2011). Accuracy of CNV Detection from GWAS Data. *PLoS One, 6*(1), e14511. doi:10.1371/journal.pone.0014511

Zhang, L., Hatzakis, E., Nichols, R. G., Hao, R., Correll, J. B., Smith, P. B., . . . Patterson, A. D. (2015). Metabolomics Reveals that Aryl Hydrocarbon Receptor Activation by Environmental Chemicals Induces Systemic Metabolic Dysfunction in Mice. *Environ Sci Technol*. doi:10.1021/acs.est.5b01389

Zhang, X., Du, R., Li, S., Zhang, F., Jin, L., & Wang, H. (2014). Evaluation of copy number variation detection for a SNP array platform. *BMC Bioinformatics, 15*, 50. doi:10.1186/1471-2105-15-50

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., . . . Ohlsson, R. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat Genet, 38*(11), 1341-1347. doi:10.1038/ng1891

Zheng, C., Khoo, C., Furtado, J., & Sacks, F. M. (2010). Apolipoprotein C-III and the metabolic basis for hypertriglyceridemia and the dense low-density lipoprotein phenotype. *Circulation, 121*(15), 1722-1734. doi:10.1161/CIRCULATIONAHA.109.875807

Zheng, C., Khoo, C., Ikewaki, K., & Sacks, F. M. (2007). Rapid turnover of apolipoprotein C-III-containing triglyceride-rich lipoproteins contributing to the formation of LDL subfractions. *J Lipid Res, 48*(5), 1190-1203. doi:10.1194/jlr.P600011-JLR200

Zheng, X., Shen, J., Cox, C., Wakefield, J. C., Ehm, M. G., Nelson, M. R., & Weir, B. S. (2014). HIBAG--HLA genotype imputation with attribute bagging. *Pharmacogenomics J, 14*(2), 192-200. doi:10.1038/tpj.2013.18

Zhou, K., Bellenguez, C., Spencer, C. C., Bennett, A. J., Coleman, R. L., Tavendale, R., . . . investigators, M. (2011). Common variants near ATM are associated with glycemic response to metformin in type 2 diabetes. *Nat Genet, 43*(2), 117-120. doi:10.1038/ng.735

FIGURES

Figure 2.1 Manhattan plot of PCB 99

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

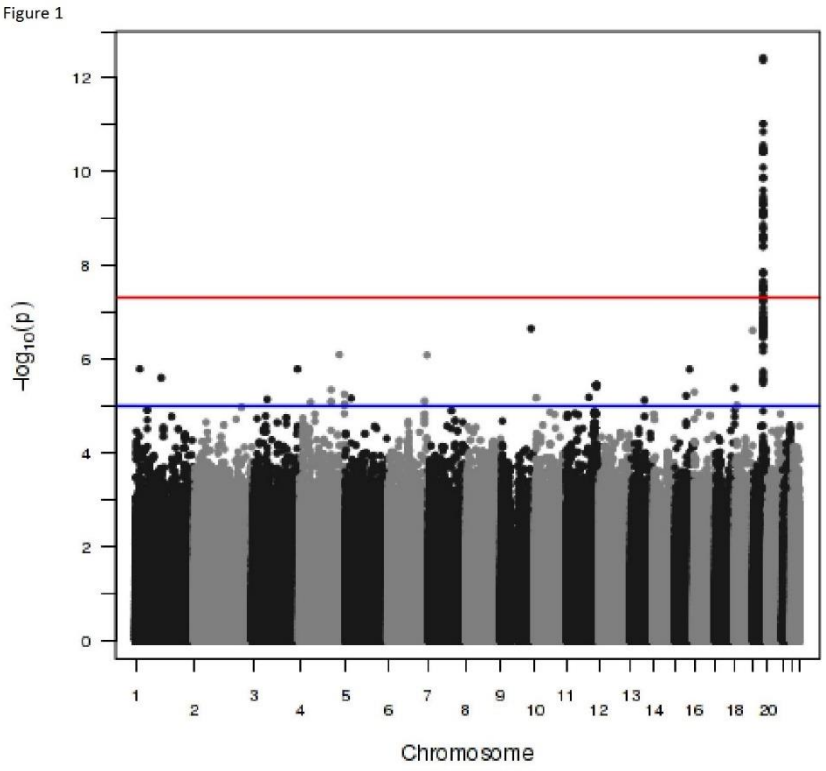


Figure 2.2 Regional Association plot of PCB 99

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r^2 from the 1000 Genomes Project March 2012 release): red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Recombination

rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser.

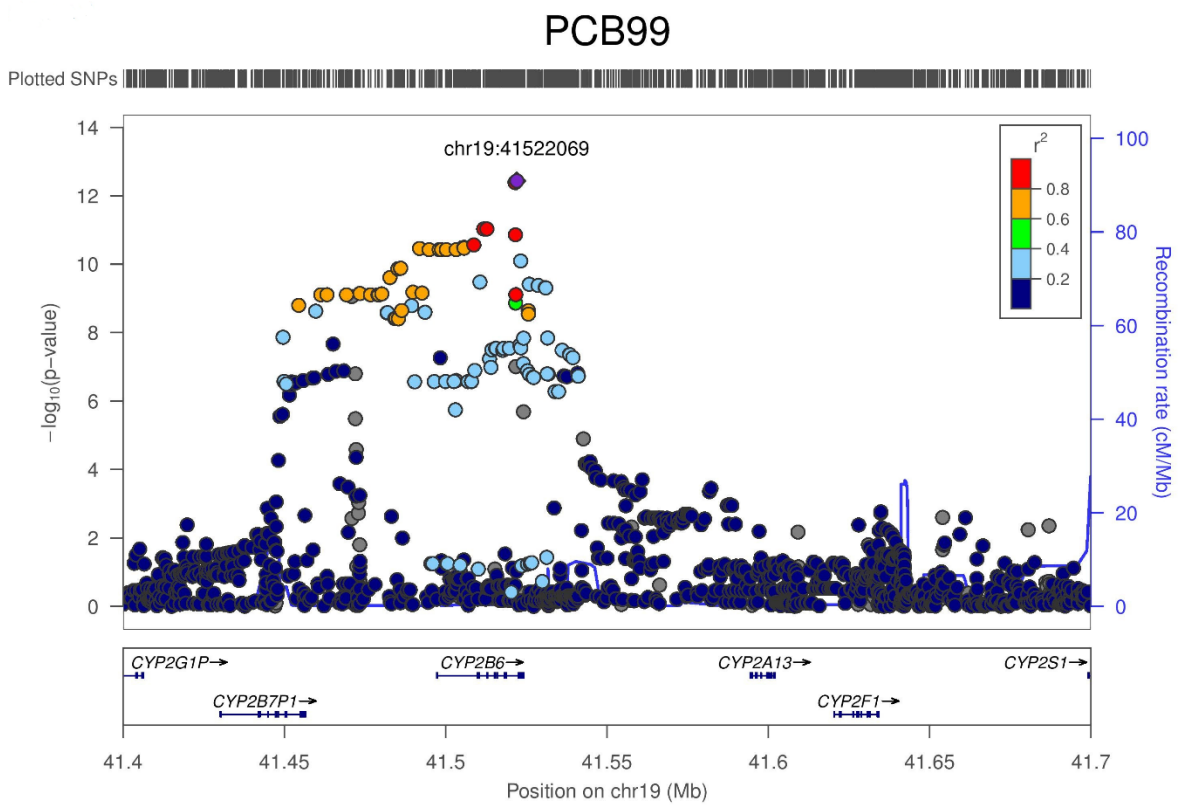


Figure 2.3 Hierarchical Clustering of PCBs

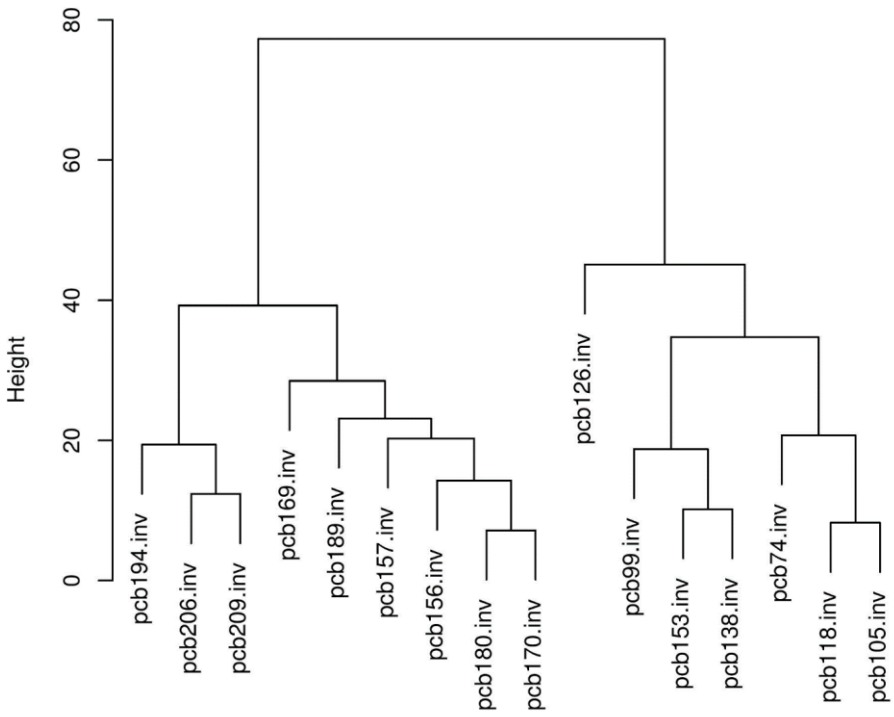


Figure 2.4a. Regional Association plot of PCB118 conditioned on PCB99. Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r^2 from the 1000 Genomes Project March 2012 release): red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser.

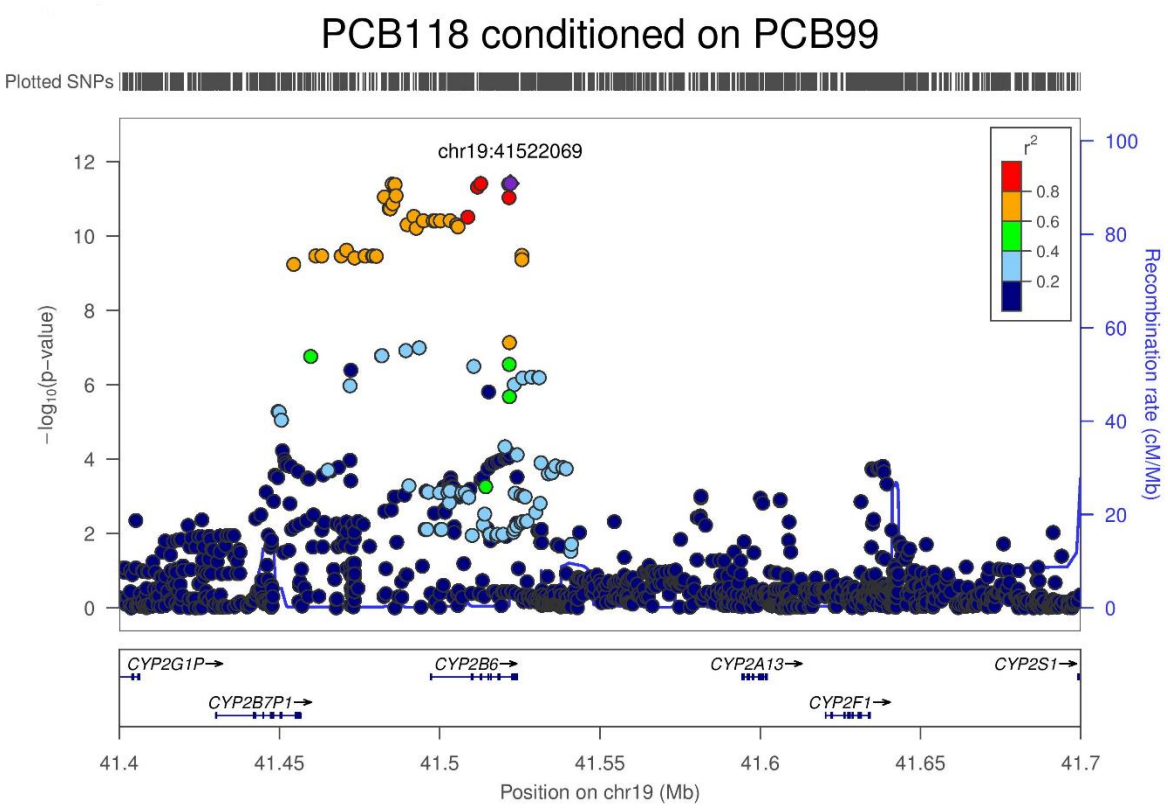


Figure 2.4b. Regional Association plot of PCB 105 conditioned on PCB 99

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r^2 from the 1000 Genomes Project March 2012 release): red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser.

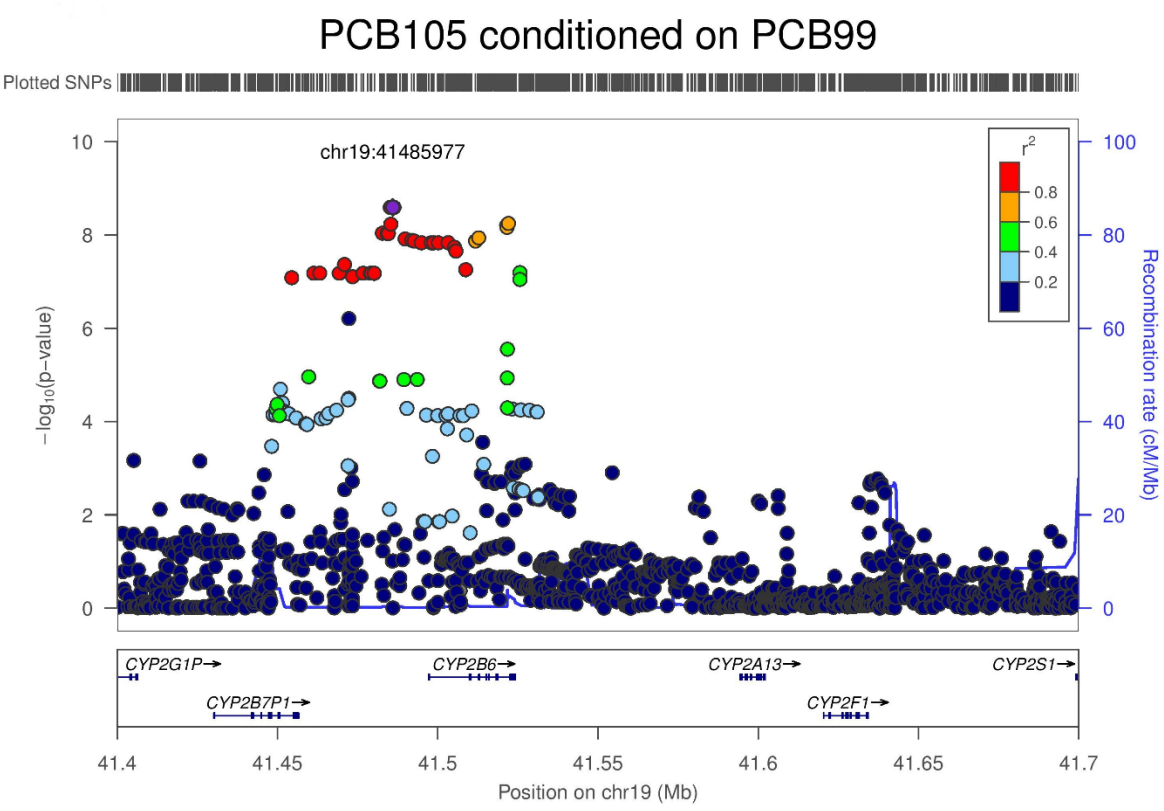


Figure 2.4c. Regional Association plot of PCB 74 conditioned on PCB 99

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r^2 from the 1000 Genomes Project March 2012 release): red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser.

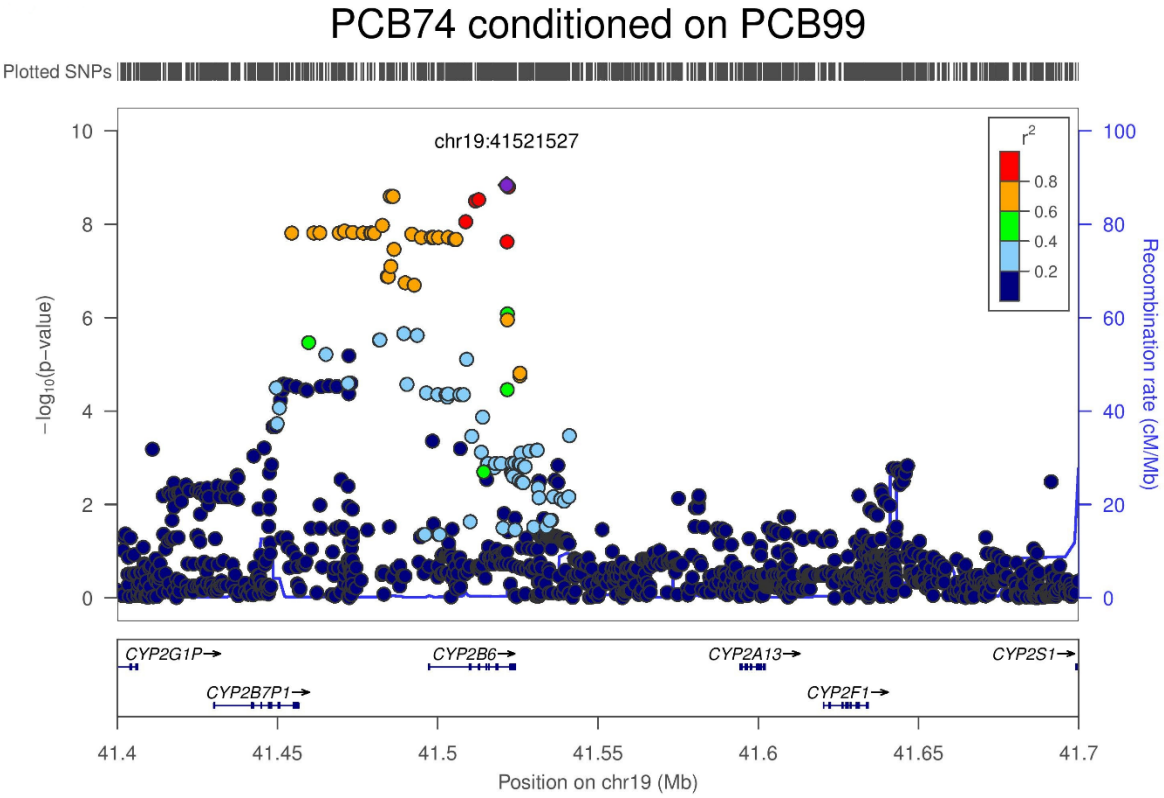


Figure 2.5 Manhattan plot of Manganese

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

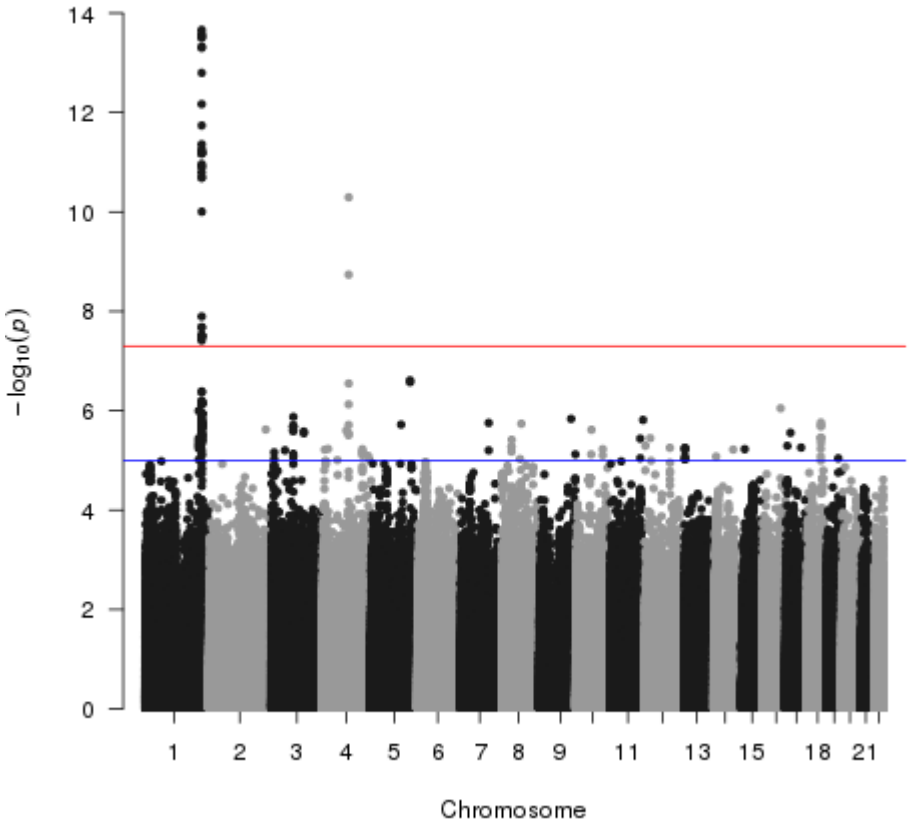
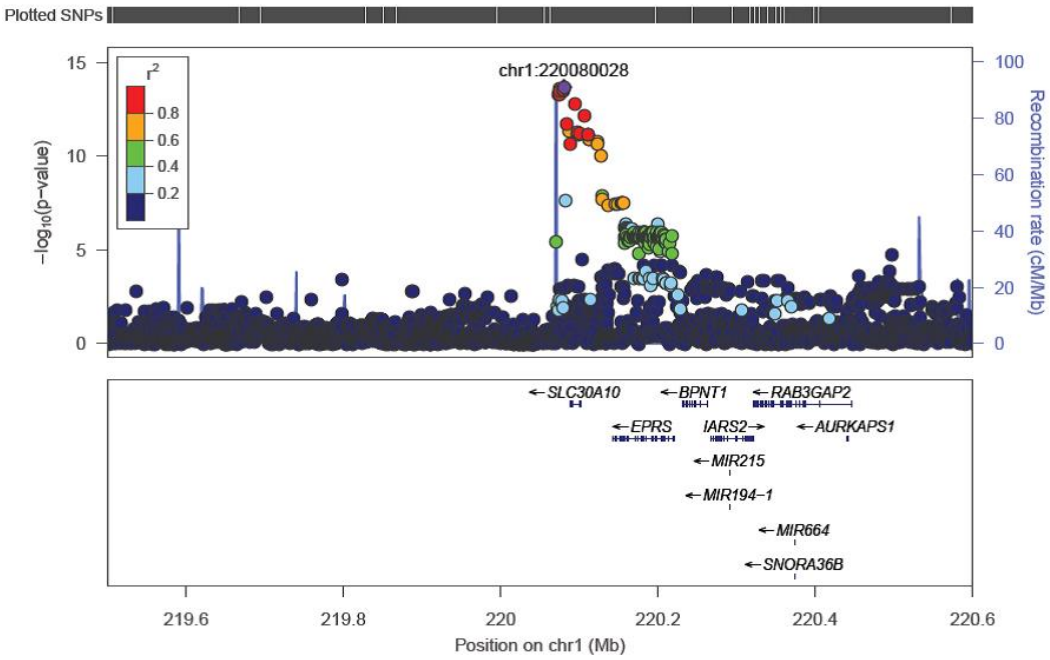


Figure 2.6 Locuszoom plots of manganese for both loci

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r^2 from the 1000 Genomes Project March 2012 release): red $r^2 \geq 0.8$; gold $0.6 \leq r^2 < 0.8$; green $0.4 \leq r^2 < 0.6$; cyan $0.2 \leq r^2 < 0.4$; blue $r^2 < 0.2$; grey r^2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser.



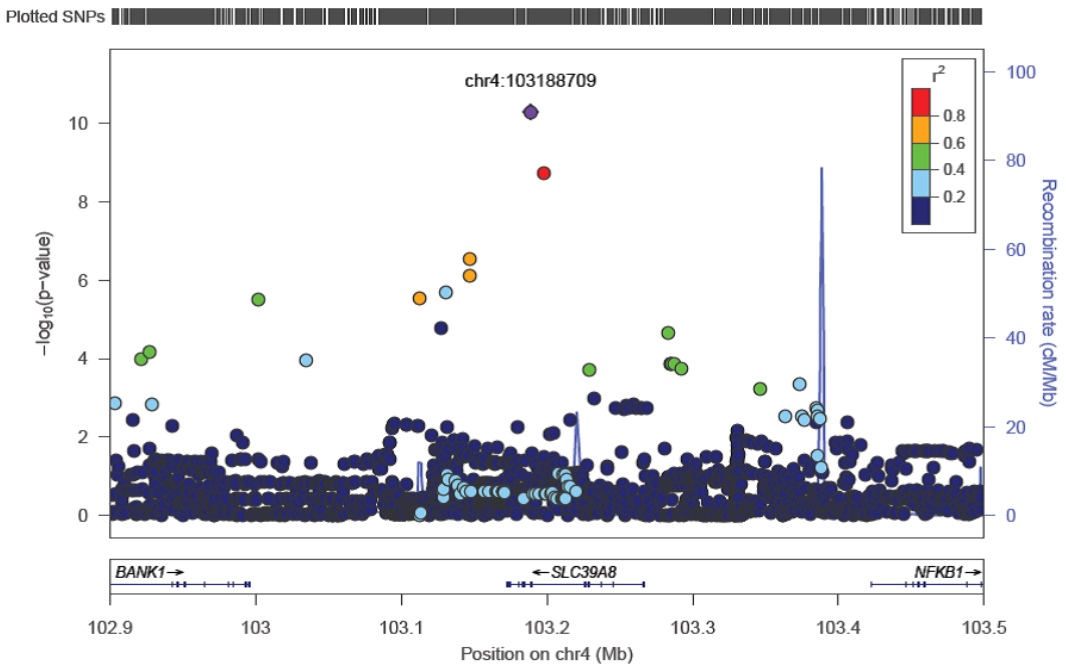

Plotted SNPs
chr4:103188709
r2
0.8
0.6
0.4
0.2
−log10(p−value)
Recombination rate (cM/Mb)
BANK1
SLC39A8
NFKB1
Position on chr4 (Mb)

Figure 2.7 - Local association plot for cadmium

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser.

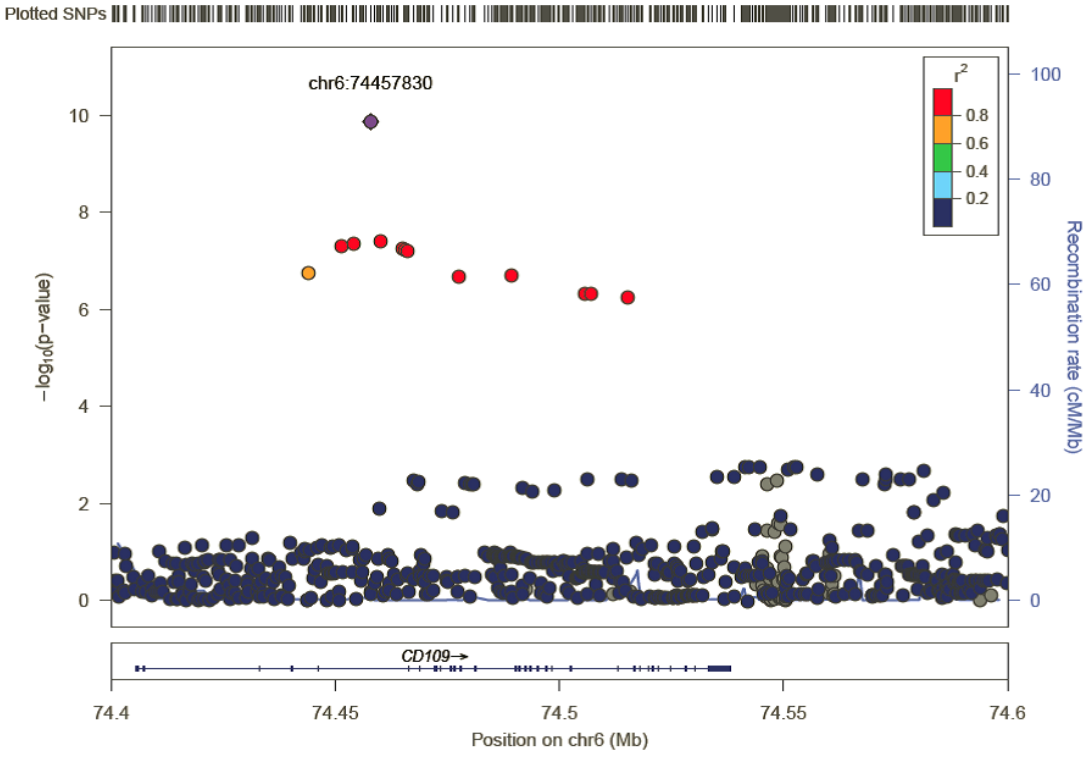


Figure 2.8 – Local association plot for mercury

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

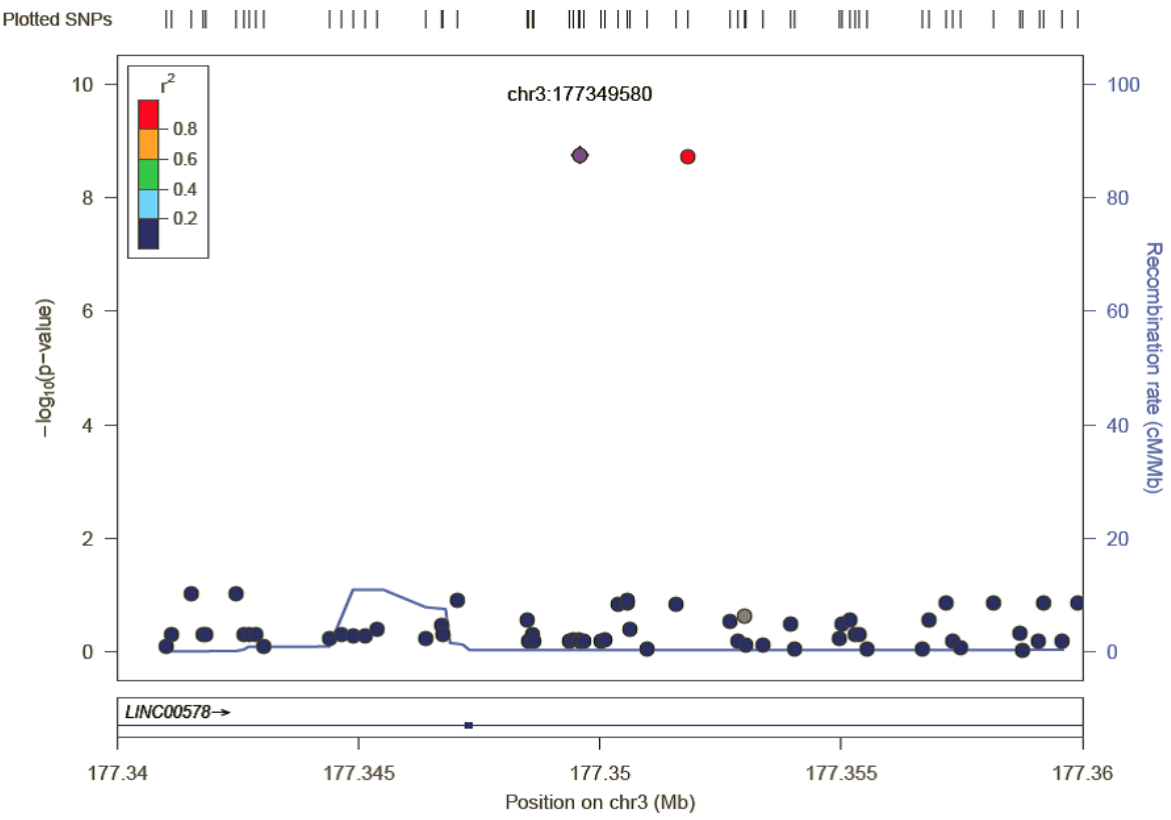


Figure 2.9 – Local association plot for OCDD

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

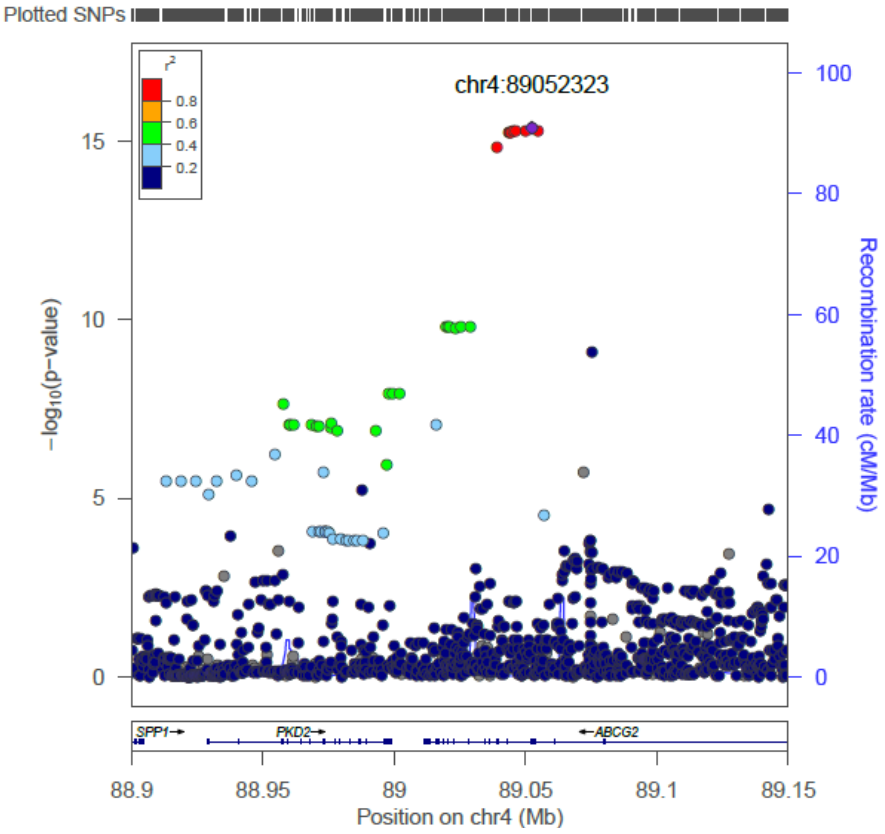


Figure 3.1 Manhattan plot of D6 desaturase

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

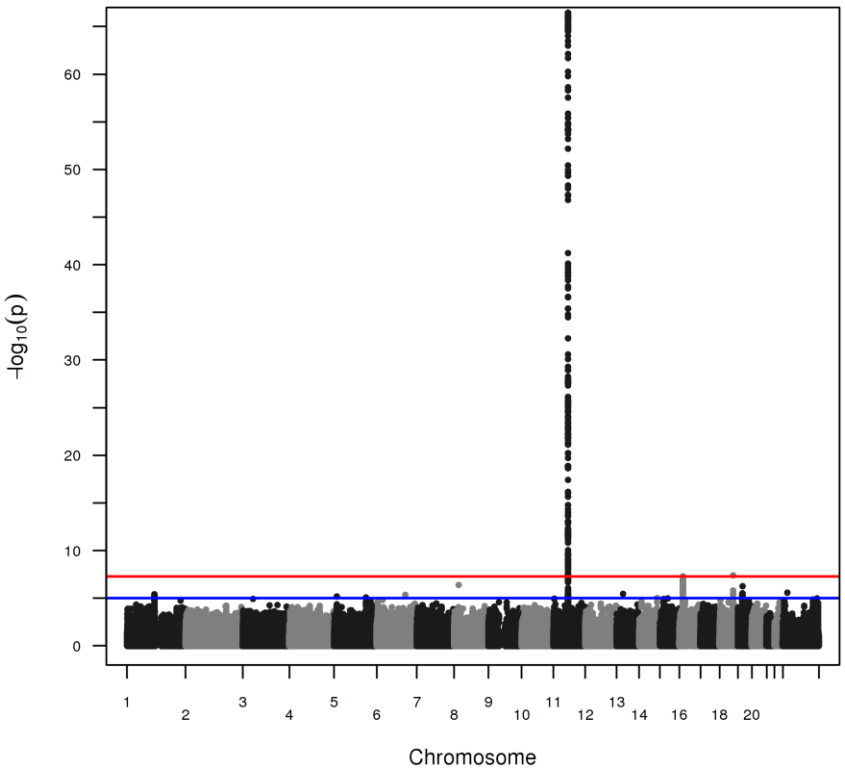


Figure 3.2 – Local association plot for D6 desaturase

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

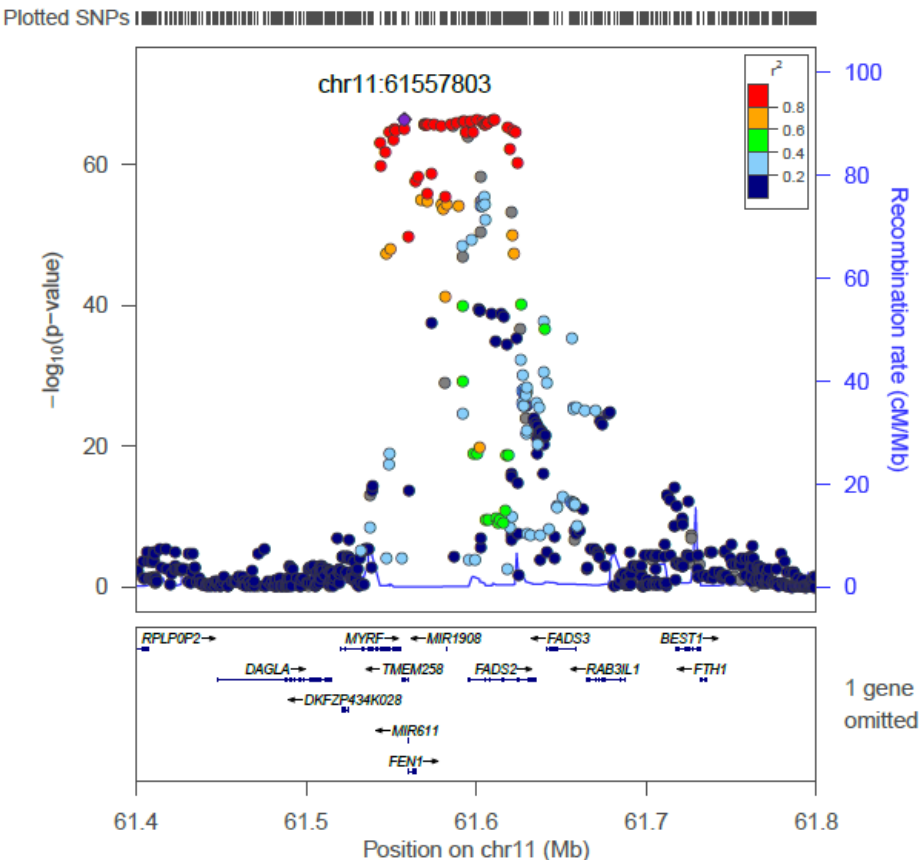


Figure 3.3 Manhattan plot of D5 desaturase

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

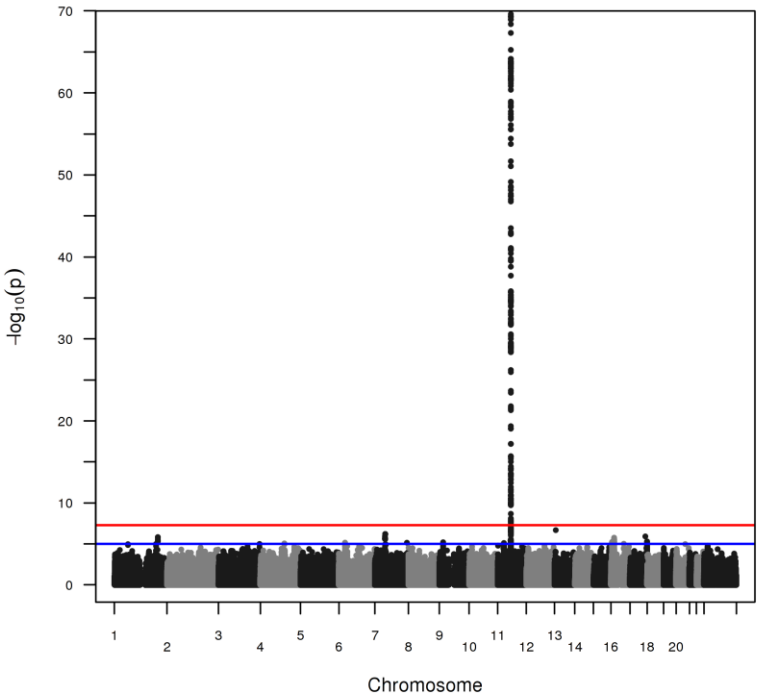


Figure 3.4 – Local association plot for D5 desaturase

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

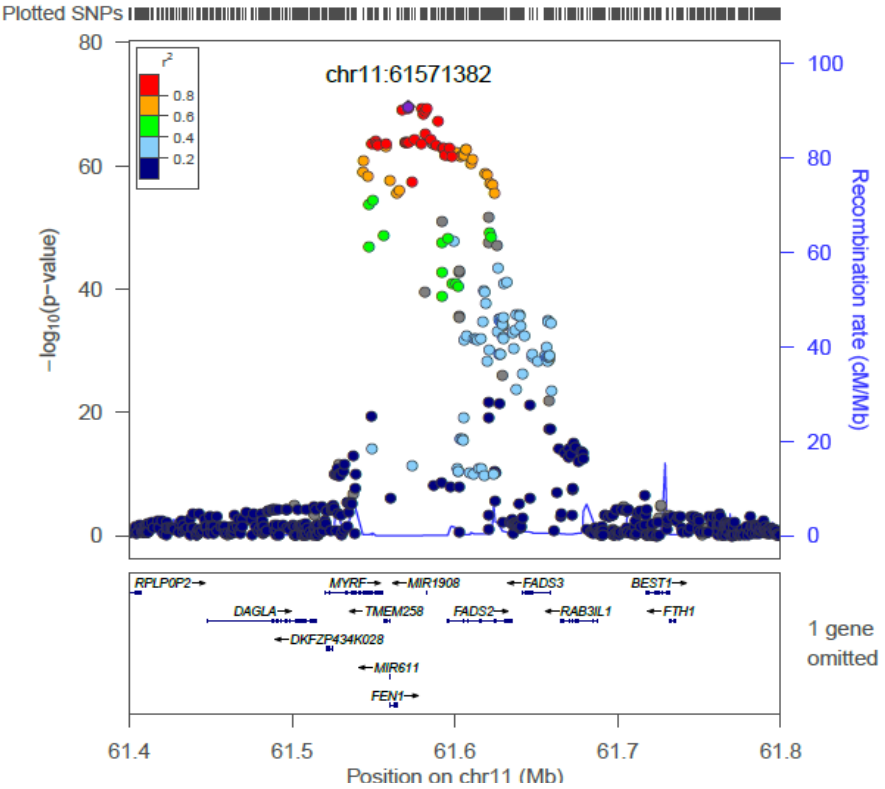


Figure 3.5 Manhattan plot of ApoB

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

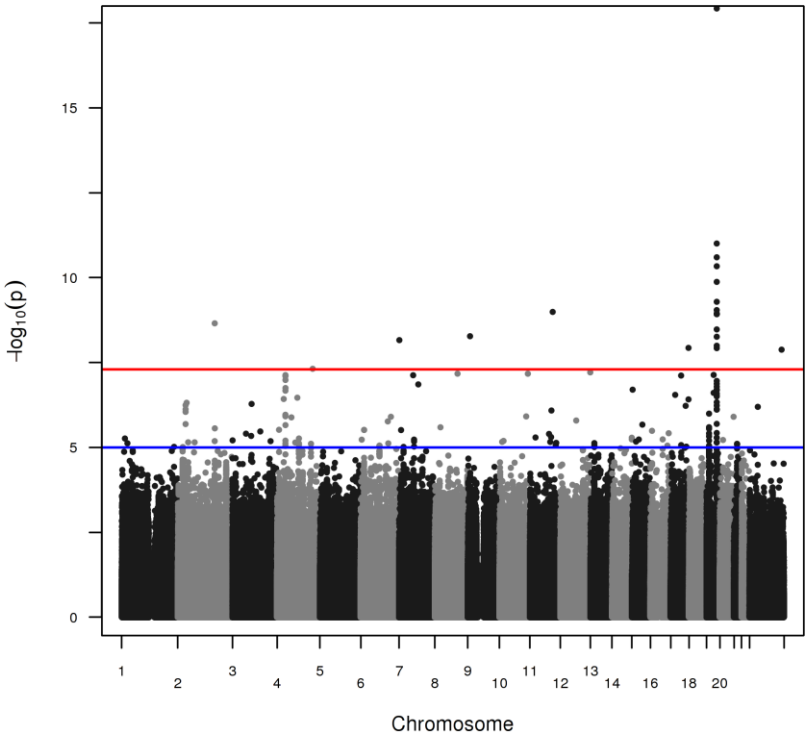


Figure 3.6 – Local association plot for ApoB

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

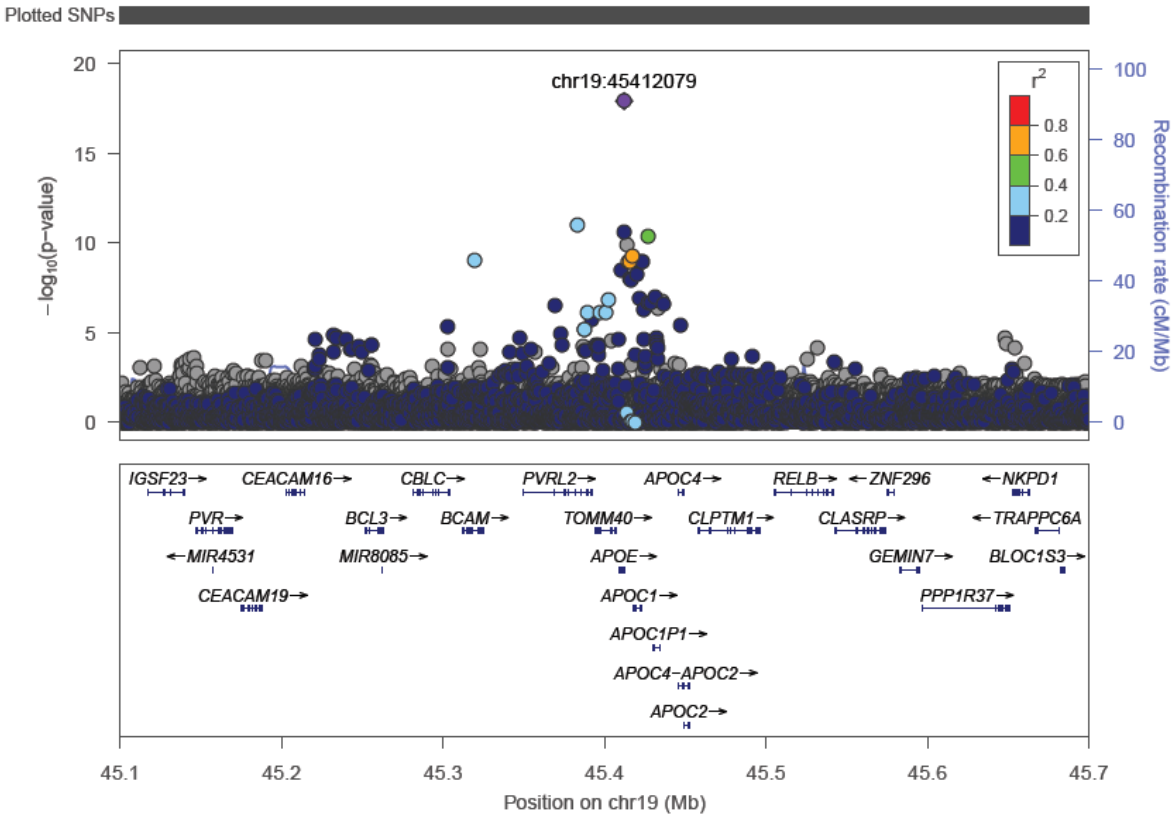


Figure 3.7 Manhattan plot of TG

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

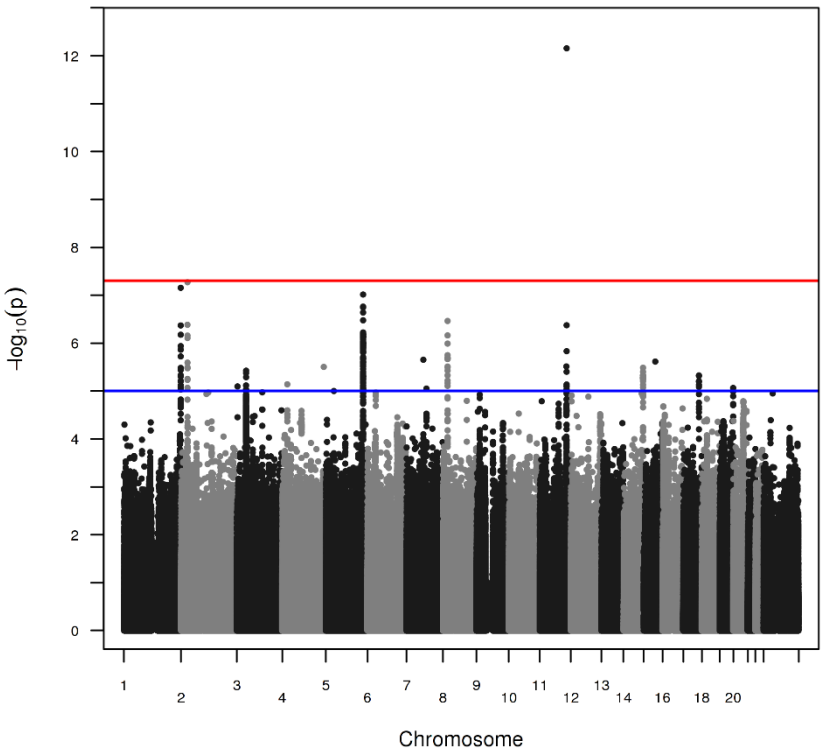


Figure 3.8 – Local association plot for TG

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

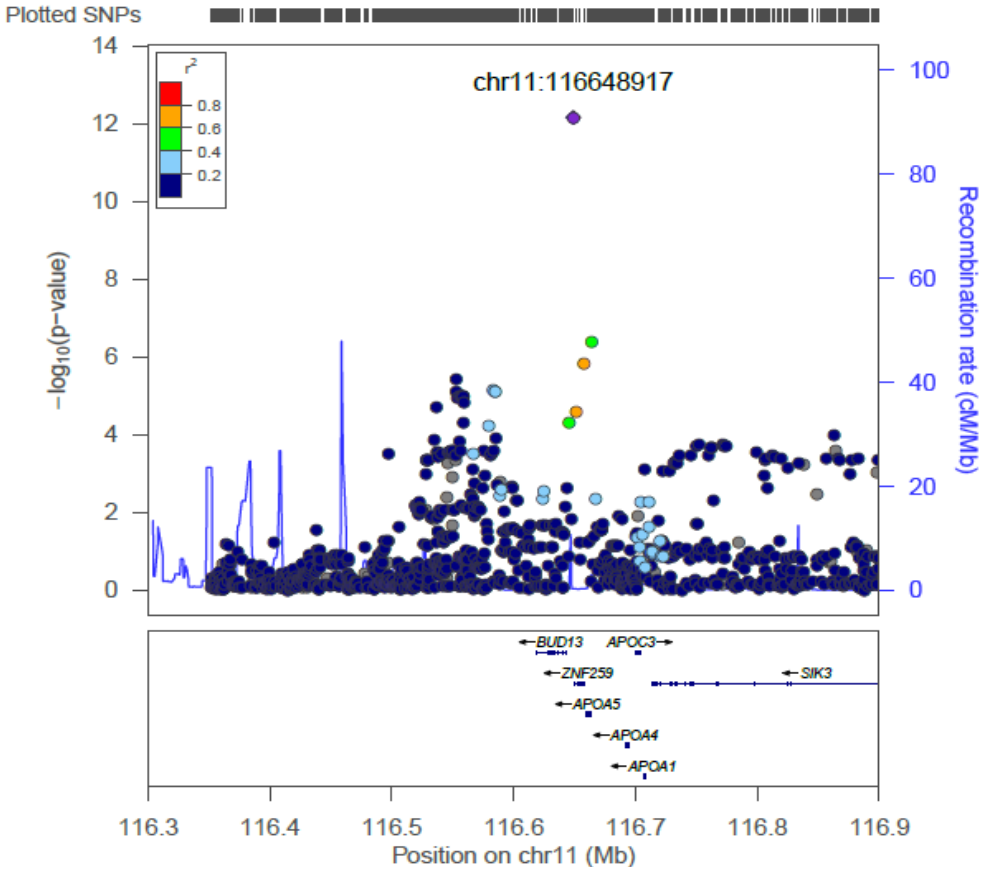


Figure 3.9 Manhattan plot of HDL

Each point corresponds to a SNP passing QC, plotted according to genomic position on the x-axis and the strength of association (-log10 p-value) on the y-axis. The red line indicates the genome wide significance threshold ($5x10^{-8}$), while the blue line indicates a threshold of 10^{-05}

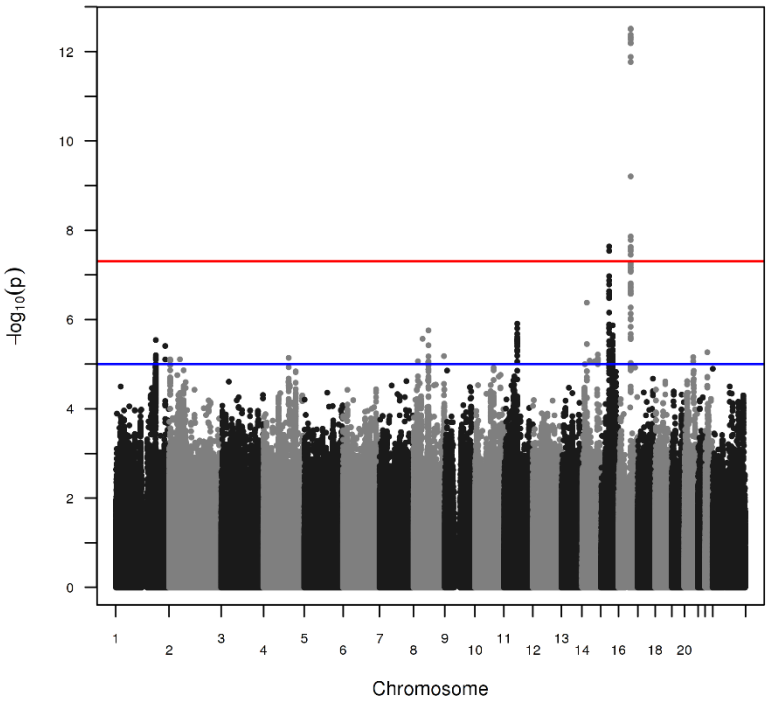


Figure 3.10 – Local association plot for HDL

Each point represents a SNP plotted with their p-value (on a -log10 scale) as a function of genomic position (NCBI Build 37). In each panel, the lead SNP is represented by the purple diamond. The colour coding of all other SNPs (circles) indicates LD with the lead SNP (estimated by CEU r2 from the 1000 Genomes Project March 2012 release): red r2≥0.8; gold 0.6≤r2<0.8; green 0.4≤r2<0.6; cyan 0.2≤r2<0.4; blue r2<0.2; grey r2 unknown. Recombination rates are estimated from the International HapMap Project and gene annotations are taken from the University of California Santa Cruz genome browser

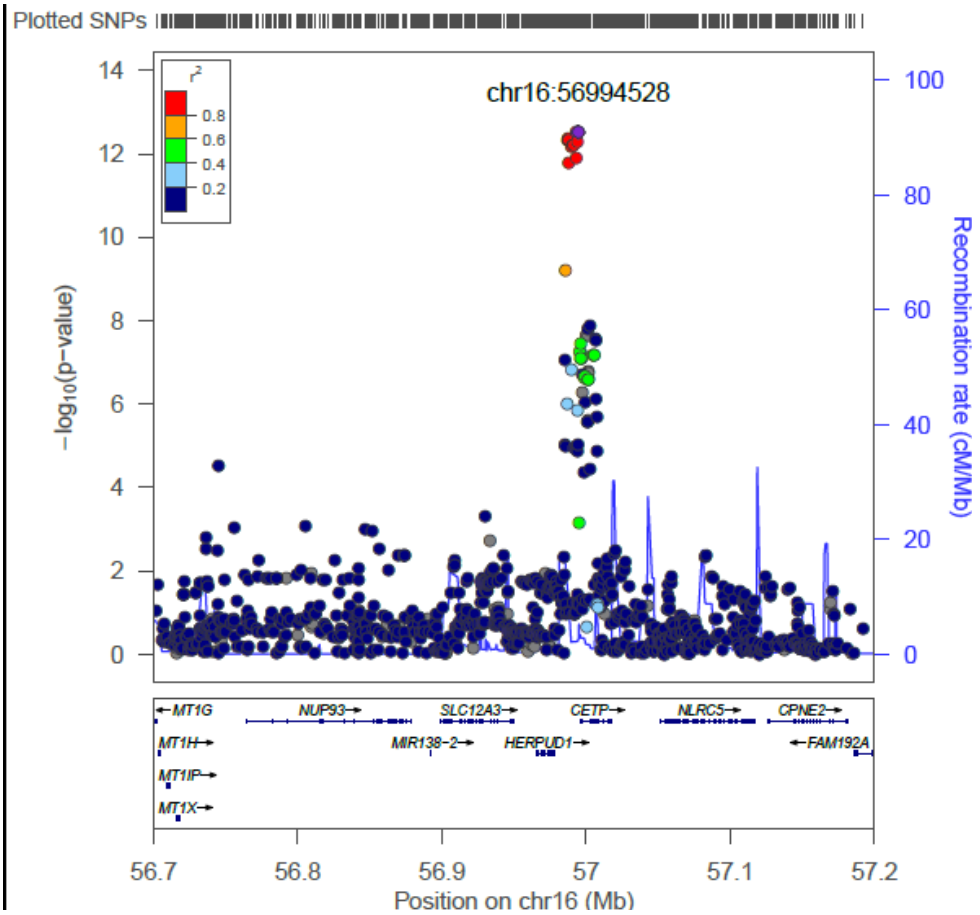


Figure 4.1 Graph of cumulative posterior probability for variants in the credible set for D5 desaturase

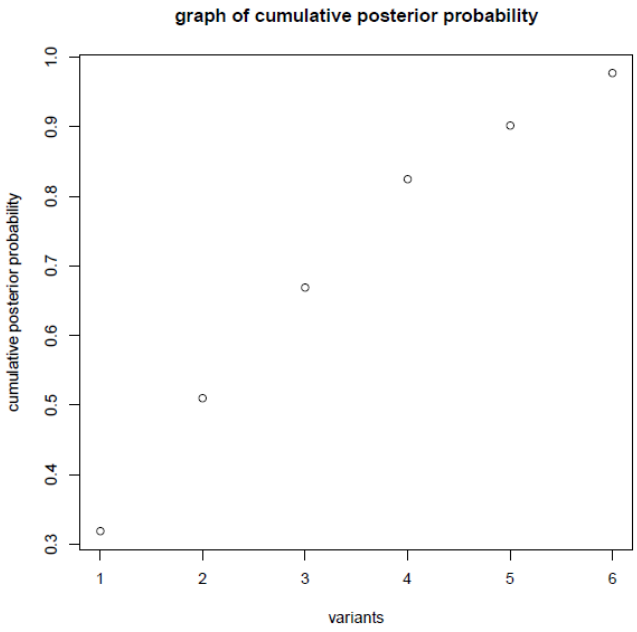


Figure 4.2 Graph of cumulative posterior probability for variants in the credible set for D6 desaturase

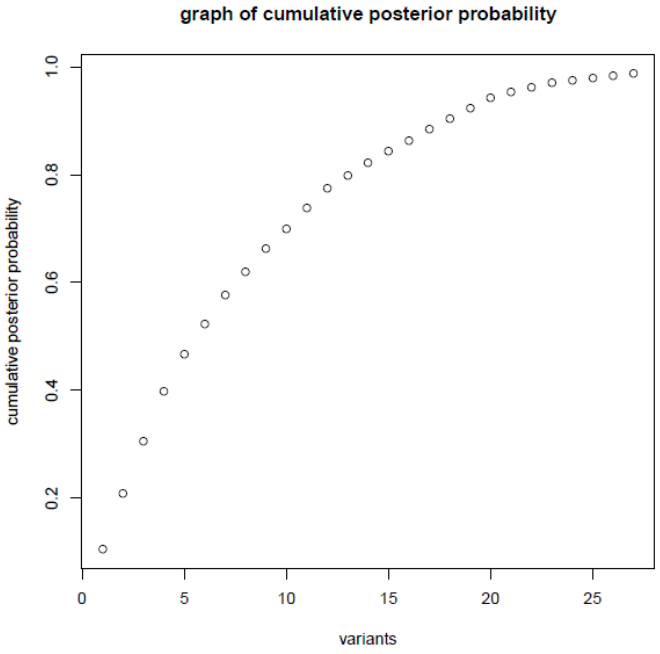


Figure 4.3 Graph of cumulative posterior probability for variants in the credible set for HDL

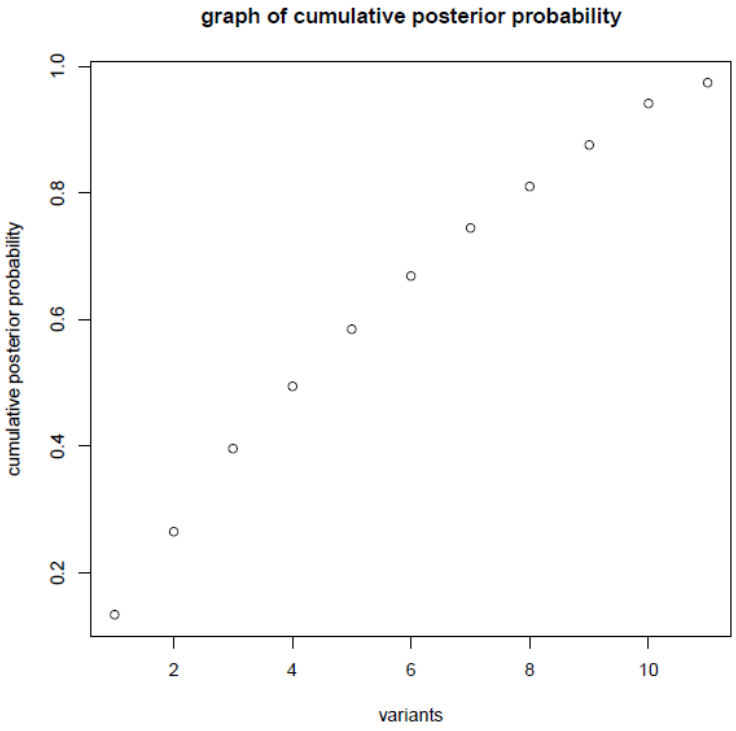


Figure 4.4 Graph of cumulative posterior probability for variants in the credible set for TG

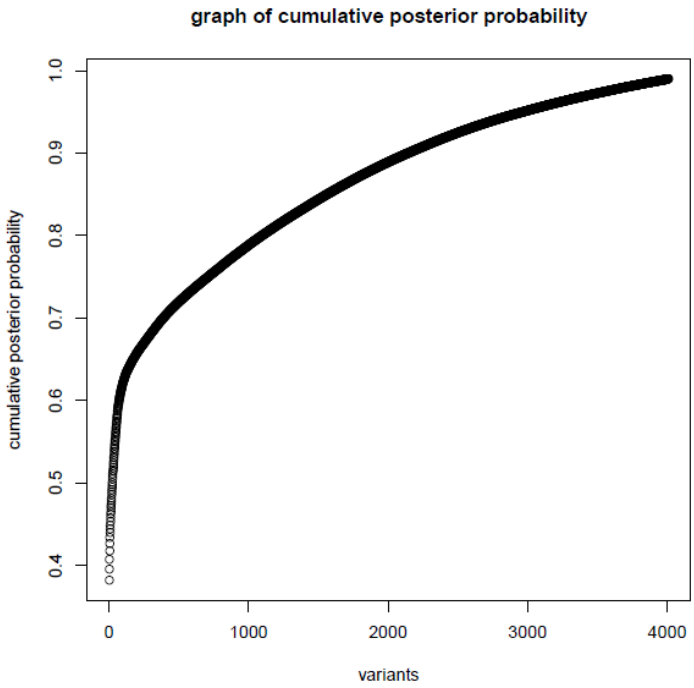
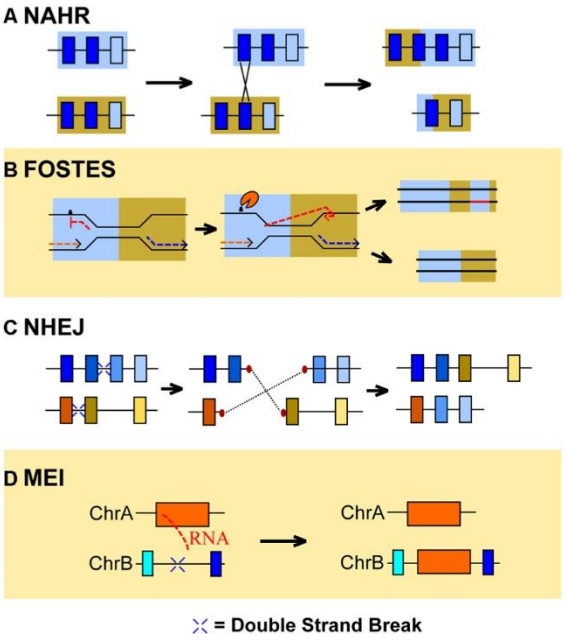


Figure 5.1 Mechanisms of CNV formation (Bickhart & Liu, 2014)



MEI – Mobile Element Insertion

NHEJ – Non-homologous end joining

FOSTES – fork stalling and template switching

NAHR – non-allelic homologous recombination