DPHIL THESIS

Christopher Grace

Linacre College

DETECTION AND EXPLOITATION OF EXPRESSION QTL IN DRUG DISCOVERY AND DEVELOPMENT

Supervisors: Dr Andrew Morris, Dr Julie Huxley-Jones

and Dr John Whittaker

The Wellcome Trust Centre for Human Genetics,

Roosevelt Drive, Oxford, UK

ABSTRACT

Expression quantitative trait loci (eQTLs) are genetic markers associated with transcription of Ribonucleic Acid (RNA). eQTLs are detected using association analysis to detect correlations between RNA expression data (microarray or RNA-SEQ) and the genotypes of individuals within a study.

Trans-ethnic meta-analysis can increase power to detect genetic variants for eQTLs and improve fine-mapping resolution because of differential patterns of linkage disequilibrium (LD) between diverse populations. Lymphoblastoid cell lines (LCLs) from samples in the Phase II and III HapMap populations have been used to detect *cis* eQTLs using association analysis followed by meta-analysis. Phase III HapMap samples have also been imputed using the 1000 Genomes March 2012 *"all ancestries"* panel.

The goals of this thesis are to perform meta-analysis on multi-ethnic association summary statistics in order to: Increase the power to detect eQTLs, leverage differences in LD between ancestry groups to fine map eQTL variants and investigate and characterize heterogeneity in allelic effect sizes on expression between diverse populations.

In addition to this, eQTLs identified are used to perform integration with signals from genome-wide association studies (GWAS) of complex human traits. A pipeline has been developed where eSNPs from the eQTL datasets are integrated with disease SNPs (dSNPs) from the NHGRI GWAS catalog using reciprocal conditional analysis to determine whether eSNP and dSNP tag or are the same causal variant. Also, eQTLs which are also "absorption, distribution, metabolism, and excretion" (ADME) genes are studied in more detail, specifically looking for heterogeneity and enrichment in this dataset.

The analysis shows that combining association analysis summary statistics using meta-analysis leads to an increase in power to detect eQTLs. Differences in LD between ancestry groups can be

used to improve fine mapping resolution, as measured by “credible sets” of variants most likely to drive the eQTL signal, when all ancestry groups are combined. Considerable heterogeneity between ancestry groups has been detected, much of which is due to differing LD between tag SNP and causal variants across ancestry groups.

Furthermore, the GWAS integration has led to the identification of several dSNP – eSNP pairs for disease such as Ulcerative Colitis, Inflammatory Bowel Disease, Bechet’s Disease, Sarcoidosis, Crohn’s Disease, Grave’s Disease and Primary Biliary Cirrhosis, and have provided potential novel insights of genes through which these disease association signals are mediated. Several eQTLs for genes within the ADME dataset have also been identified some of which have significant heterogeneity.

ACKNOWLEDGEMENTS

I would like to thank my supervisors Dr Andrew Morris (Wellcome Trust Centre for Human Genetics (WTCHG)), Dr Julie Huxley-Jones (GlaxoSmithKline (GSK)) and Dr John Whittaker (GSK) for their help and support in this project.

I would also like to thank Prof Mark McCarthy (WTCHG) and Prof Emmanouil Dermitzakis (University of Geneva) for advice and providing me access to the HapMap Phase III expression data.

Contents

GLOSSARY

1000 Genomes Project *is a project in which the aim is to sequence at least 1000 human genomes, for use in imputation.*

ADME *an abbreviation for absorption, distribution, metabolism, and excretion, which describes the disposition of a pharmaceutical compound within an organism*

Allele Specific Expression (ASE) *specifies where a genetic variant affects expression of a gene for only one of the alleles, also known as acting in cis*

Alternative Splicing *is the process where a single gene can encode several different proteins, typically by the exclusion or inclusion of exons from the mRNA molecule.*

cDNA Microarray *a collection of microscopic DNA probes attached to a solid surface, can be used to evaluate mRNA expression within a cell, by converting to cDNA and then hybridizing to a microarray.*

***cis* eQTL** *An eQTL which is allele specific, also sometimes used to denote a local eQTL. An example could be a variant within a promoter region, which only effect transcription on the same chromosome.*

CNV *Copy Number Variant is a genetic variant in which a section of DNA has been duplicated or deleted from the genome.*

DNA Methylation *the addition of a methyl-group to DNA bases (typically as CpG sites). Can contribute to epigenetic inheritance.*

ENCODE *the Encyclopaedia of DNA Elements is a project that aims to identify all functional elements in the human genome.*

Enhancer *a short region of DNA which can be bound by transcription factors / activator proteins to increase transcription of mRNA*

eQTL *An expression quantitative trait loci, these are genetic markers (such as SNPs, indels and CNVs) in the genome that are associated with transcription of RNA.*

Epigenetics *heritable changes in gene expression, which do not involve a change in DNA sequence. Examples are methylation and histone modification.*

Genotype *the part of the DNA sequence which determines a specified phenotype of an organism.*

Genome Wide Association Studies (GWAS) *the examination of multiple common variants across the genome to detect correlation with a phenotype of interest, typically using logistic or linear regression.*

GTEX *the Genotype-Tissue Expression project provides access to human gene expression and regulation across multiple tissue types.*

Haplotype *the combination of genetic variants found to inherit together on the same chromosome.*

HapMap *an organization that aims to develop and provide a haplotype map of the human genome. Haplotypes generated by this project can be used to impute unknown genotypes in GWAS.*

Histone *proteins which package eukaryotic DNA into structural units called nucleosomes.*

Imputation *the process of replacing missing data with statistically determined values. In genetics imputation is used to determine missing genetic variants, typically using models which include haplotypes of reference genomes, such as HapMap and 1000 Genome project.*

Indel *A type of genetic variant where bases are **in**serted or **del**eted.*

Linkage Disequilibrium (LD) *this is the non-random association of alleles at different loci.*

Linkage Studies *use recombination rates between markers and phenotype to determine the location of a causal variant*

Lymphoblastoid Cell Line (LCL) LCLs are generated by Epstein-Barr virus transformation of B-lymphocytes. They are immortalized human cell lines and so are useful for the storage of genetic data. LCLs are available for individuals within the HapMap and 1000 Genomes project.

Meta-analysis *Is a statistical methods used to combine experimental results from different samples. The greater sample size can lead to an improvement in power.*

Multiple Testing Problem *the increased numbers of false positives expected to occur when performing a statistical test a large number of times.*

Phasing *the process of determining the haplotypes given the genotypes of an organism.*

Promoter *a region of DNA which initiates transcription of a gene, located near transcription start sites.*

RNA Polymerase *protein which generates RNA from a DNA sequence within transcription.*

RNA-SEQ *a technology that allows profiling of the RNA content within a cell using Next Generation Sequencing (NGS) technologies.*

SNP *a Single Nucleotide Polymorphism is a type of genetic variant which differs at a single nucleotide.*

SNP Microarray *a type of microarray which can be used to detect variants such as SNPs within an organisms DNA.*

Statistical Power *the probability that a statistical test correctly rejects the null hypothesis when the alternative hypothesis is correct.*

Strand Alignment *the alignment of variants such as SNPs to a specific positive or negative strand.*

***trans* eQTL** *An eQTL that acts on both alleles equally, also can be used to specify a distant eQTL.*

Transcription *The generation of mRNA from the DNA sequence encoded within a gene.*

Transcription Factor *This is a protein that is involved in regulating transcription*

Translation *The generation of a protein with a specified amino acid sequence encoded within a mRNA molecule.*

CHAPTER 1 INTRODUCTION

1.1 Gene expression and eQTLs

1.1.1 Gene expression

Gene expression is the process by which the Deoxyribonucleic Acid (DNA) information encoded in a gene is used to generate a gene's product. Information is encoded within a gene as the sequence of nucleotide bases: Adenine (A), Thymine (T) Cytosine (C) and Guanine (G). Often the gene's product is a protein (a macro molecule consisting of chains of amino acids), but genes can also encode various types of Ribonucleic Acid (RNA) with different functionalities (table 1.1 gives an overview of different types of RNA).

Type	**Function**
Messenger RNA (mRNA)	Encodes protein amino acid sequence generated in translation
Ribosomal RNA (rRNA)	Parts of structure of ribosome, which perform translation.
Transfer RNA (tRNA)	Adapter for amino acids used to construct proteins in the ribosome.
Micro RNA (miRNA)	Small non-coding, RNA, function in silencing post transcriptional regulation of gene expression.
Small nuclear RNA (snRNA)	Splicing and other functions

Table 1.1: Types of RNA

In the process of generating proteins, gene expression has two stages: transcription and translation. Transcription is the generation of mRNA molecules from the gene's DNA sequence. This involves a protein known as RNA polymerase, which generate RNA molecule corresponding to the DNA sequence. RNA has the same bases as DNA, but with the base Uracil used instead of Thymine. When transcription has completed post-transcription RNA processing may occur, which includes the removal of introns from mRNA and processing of the 3' and 5' ends of the mRNA molecule.

Translation is the building of the protein product encoded by an mRNA generated in transcription. In translation, the mRNA molecule is used to construct proteins using a four nucleotide base code. Each of the 21 amino acids is specified by one or more of a three nucleotide base codon.

Translation occurs in structures known as ribosomes which use tRNAs act as adapters, which link specific amino acids to codons.

The amount of mRNA in a cell can be measured using cDNA microarrays and RNA-sequencing technologies. This gives an indirect indication of the state of the cell. A cDNA Microarray is a collection of microscopic DNA elements (probes) attached to a surface; they can be used to measure the amount of expressed mRNA in a large number of genes simultaneously. The probes are defined to cover the unique region of mRNA generated in transcription. RNA-sequencing technologies interrogate the RNA sequences in a cell at a particular state and time.

RNA-sequencing (RNA-SEQ) has several advantages over cDNA Microarray technology; RNA-SEQ allows the detection of novel transcripts which could not be detected with cDNA microarrays. This is because cDNA microarrays need to create a probe for each of the transcripts to be quantitated and so transcripts need to be known *a priori*. RNA-sequencing, in comparison, detects the entire transcriptome and requires no prior knowledge. Another advantage is that RNA-SEQ has a greater dynamic range then cDNA microarray technology and so can detect very low and high levels of expression with greater accuracy. Finally, RNA-SEQ is also able to detect gene expression features that microarrays cannot such as Allele Specific Expression (ASE) and splicing variants. However cDNA microarray technology is considerably less expensive than RNA-SEQ and is a mature technology.

1.1.2 Detection of Genetic Variants

Genetic variants are regions of DNA which differ between individuals. Types of genetic variants include: SNPs (single nucleotide polymorphisms) variants upon a single nucleotide, which almost all have two alleles only; indels (insertion or deletion), insertions or deletions of bases in the DNA of different lengths; and CNVs (copy-number variants), consisting of relatively large regions (at least 1000 base pairs) that have been deleted or duplicated.

Genetic variants can be detected either by direct genome sequencing or by the use of SNP microarrays. SNP microarrays use a probe which includes the SNP to be assayed and the immediate regions around the SNP.

1.1.3 eQTLs and GWAS

This thesis focuses on expression quantitative trait loci (eQTLs) which incorporate genetic markers (such as SNPs, indels and CNVs) in the genome that are associated with transcription of RNA. This means that the variant will affect the amount of RNA generated in transcription.

eQTLs can be sub-divided into those in which the marker is local to the expressed gene (referred to as *cis* eQTLs in this thesis) and those in which the marker is distant from the expressed gene (referred to as *trans* eQTLs in this thesis). Local eQTLs are typically defined as those where the genetic variant is 1 Mb upstream and downstream of the expressed genes transcription start-site (in humans), whilst distant *a*re those where the genetic variant is located elsewhere in the genome (including locations on different chromosomes).

eQTLs can be further sub-divided by their genetic effect on transcription. eQTLs can affect transcription in an allele specific way (also known as acting in *cis* or ASE). ASE can be caused by a variant within a promoter or transcription factor binding site. Alternatively the eQTL may effect expression intensities of both alleles equally (also known as acting in *trans*). This indicates that the change in expression is being modulated by a diffusible factor (such as a transcription factor). In addition to eQTLs, variants can have other genetic effects on transcription including splicing QTLs, which affect the relative abundance of different mRNA splice-types. It is important to note that eQTLs defined as *cis* based on the distance criteria, need not have a *cis* effect on transcription.

Potential mechanisms of action for eQTLs include the following: variants may alter DNA sequences within promoters and enhancers which affect capability of transcription factors to bind

to these regions, this will lead to ASE in the allele with the variant change. Access to these regions for transcription factors may also be determined by epigenetic causes such as CpG methylation and chromatin accessibility. In addition, mutations to the transcription factor itself, or expression levels of the transcription factor, may lead to expression changes in both alleles equally.

After initial transcription, eQTLs may be caused by changes in splicing, polyadenylation and mRNA stability. Altered expression of miRNAs can also lead to changes in expression level.

Studies that look at eQTLs proceed by genotyping a set of the genetic variants in the study sample using either a SNP microarray or DNA sequencing, followed by assessing mRNA expression in the same cells or cell-lines using cDNA microarray or RNA-Sequencing. The analysis attempts to find correlations between the genotype at a genetic variant and expression of a probe, which can be assessed using association analysis treating RNA expression as a quantitative trait.

Genome-wide association studies (GWAS) are large scale experiments that attempt to identify genetic variants correlated with complex phenotypes (e.g. height, type II diabetes). These experiments are very similar to eQTL studies, with the difference that one phenotype is typically tested in GWAS, whilst in eQTL studies, thousands of mRNA expression profiles are tested. In GWAS the causal variant is not necessarily typed itself, but a tag SNP in Linkage disequilibrium (LD) with it (i.e. the genotypes at the variants are correlated with each other within individuals).

eQTLs are studied for the several reasons. Firstly, expression data provides a high-throughput quantitative trait (QT) which can be used to study the possible general genetic architectures of other QTs (for instance the range of effect size magnitudes across QTLs and the counts of QTLs). Secondly, causal variants for eQTLs can give insight into the biology of transcription, for instance by identifying regulatory elements such as promoters and enhancers and gene expression networks. Finally, eQTLs can be used to provide possible functional explanations of genetic

variants detected in GWAS, helping to fine-map loci and suggest mechanisms through which their effects are mediated

1.1.4 GWAS Integration Analysis

eQTLs can form an intermediate phenotype between genotype and disease. For example, a disease may be in part caused by a different dosage of a specified protein or RNA across individuals. It therefore could provide insight into biological mechanism through which the genetic variant can have impact on the disease or outcome of interest. The concept is that if a GWAS disease SNP (dSNP) and an eQTL SNP (eSNP) tag the same causal variant then there is a possibility that the eQTL may be causal in the effect of the GWAS phenotype. Therefore, integrating eSNPs and dSNPs can provide candidate genes for experimental follow up and evaluate associated pathways for biological plausibility. One issue with GWAS results is that many dSNPs map to intergenic regions, often in gene deserts. It is difficult to determine the impact these variants have on disease. However if the variant is also an eSNP (or is closely correlated to an eSNP) it can provide a potential mechanistic link through which the variant will affect disease that can be tested through functional studies. The NHGRI frequently updates a catalogue of GWAS results (Hindorff *et al* 2014).

1.2 Previous eQTL Studies

1.2.1 Model Organisms

The initial studies of eQTLs used model organisms such as mice (Schadt *et al* 2003, Bystrykh *et al* 2005, Chesler *et al* 2005), rats (Hubner *et al* 2005) and yeast (Brem *et al* 2002, Yvert *et al* 2003) rather than humans. These studies used inbred organisms rather than natural populations. For model organisms, populations with defined relationships can be generated, which increases the power to detect eQTLs, but decreases the resolution of genetic variants detected. In most studies, two inbred lines are mated together in order to generate diversity for mapping eQTLs, referred to as recombinant inbred strains. The majority of studies on model organism use a linkage design,

rather than association analysis, which use recombination rates between markers and phenotype to determine the location of the causal variant in contrast to association analysis which directly measure correlations between genetic variants and phenotypes.

The first study to report genetic mapping of global gene expression was performed in yeast (Brem *et al* 2002). This study used genetic linkage of gene expression intensities between a lab and a wild train of yeast. A total of 1528 genes showed differential expression between the parental strains, and expression levels of 570 genes were linked to one or more genetic loci (regions of the genome).

In studies of model organisms, regions of high *trans* eQTL concentration have been observed. These are known as eQTL "hot spots" and have been postulated to harbour master regulators (such as transcription factors).

1.2.2 Human Studies

There have been several studies looking at eQTLs in humans, the majority of which use association approaches and expression intensities measured using cDNA microarrays. Cell types studied include LCLs (Stranger *et al* 2007, Pickrell *et al* 2010, Montgomery *et al* 2010, Stranger *et al* 2012), cortical brain tissue (Myers *et al* 2007), liver (Shadt *et al* 2008), fibroblasts and T-cells (Dimas *et al* 2009), skin and fat (Nica *et al* 2011) and monocytes (Zeller *et al* 2010). The effect sizes of eQTLs are typically larger than those of complex trait genetic association studies, and hence smaller sample sizes are sufficient to detect them; sample sizes in published studies have ranged from 60 (Montgomery *et al* 2010) to 5,311 (Westra *et al* 2013).

Several recent studies have measured expression intensities using RNA-SEQ technology: Montgomery *et al* 2010 detected eQTLs in 60 CEPH HapMap individuals, Pickrell *et al* 2010 analysed 69 Yoruba HapMap individuals; and Lappalainen *et al* 2013 measured RNA-SEQ and

sequencing data from 462 individuals from the 1000 Genomes Project. In addition to identifying eQTLs, these datasets have been used to identify splicing QTLs (sQTLs) and ASE.

Studies have looked at specificity of eQTLs between tissue types: the Muther consortium (Nica *et al* 2011) investigated the extent of sharing of *cis* eQTLs across tissues (LCL, skin and fat), and found that approximately 30% of eQTLs were shared across all three tissues, whilst 29% appeared exclusively tissue specific. Another study (Dimas *et al* 2009) looked at the extent of sharing across three tissues (B-Cells, LCLs and T-Cells) and found 69-80% of *cis* eQTLs were cell specific.

Recently, studies have investigated *trans* eQTLs in humans. A recent study (Grundberg *et al* 2012) by the Muther Consortium identified *trans* eQTLs in adipose, LCLs and skin tissues, and found that majority of these were tissue-dependant and with a small effect size. The Muther Consortium also showed that the type 2 diabetes and LDL cholesterol associated *cis* eQTL for the maternally expressed gene *KLF14* acts as a master *trans* regulator in adipose gene expression (Small *et al* 2011).

Sharing of eQTLs across ethnic populations has also been studied. In a study looking at the Phase II HapMap LCL cell lines, sharing between two or more populations was observed in 37% of *cis* eQTLs (Stranger *et al* 2007). More recently, in a study of Phase III HapMap *cis* eQTLs (Stranger *et al* 2012) using 726 individuals from 8 populations, out of 5691 *cis* eQTLs detected, 3240 (57%) were identified in two or more populations, and 331 in all eight populations (6%).

Several studies have investigated the effect of controlled environmental perturbation on eQTLs. Grundberg *et al* 2011 studied the effect of environmental stimuli (growth factors, cytokines and hormones) on *cis* eQTLs in human primary osteoblasts. They found that only a small proportion of the *cis* eQTLs could be considered treatment specific. In contrast, Fairfax *et al* 2014 studied the effect of innate immune stimuli on eQTLs, stimulating CD14+ monocytes with LPS or IFN-γ. This

study detected a profound effect of stimuli on eQTLs with hundreds of context specific eQTLs detected.

An important area of research is in detecting genetic variation affecting the proteome. Genetic variants that lead to a change in protein levels are known as protein QTLs (pQTLs). A recent study (Battle *et al* 2015) integrated expression and protein QTLs in Yoruba LCLs. The study found most eQTLs discovered were also associated with pQTLs. However a weaker effect size was detected in pQTLs when compared to eQTLs suggesting the impact of eQTLs on pQTLs may be buffered

Recently, the Genotype-Tissue Expression project (GTEx) (GTEx Consortium 2013) has been established as a large scale public resource and tissue bank for human genetic variation and gene expression across tissues. It will enable studies of eQTLs, alternative splicing, and the tissue specificity of gene regulatory mechanisms, and aid in the interpretation of GWAS.

1.3 Resources

1.3.1 HapMap

The International HapMap Project provides high-density haplotyping of human populations, for use in imputation in GWAS. Phase II (International HapMap Consortium *et al* 2007) of the project genotyped 4 populations (CEPH (Utah), Chinese, Japanese and Yoruba (Nigeria)) at a density at over 3.1 million SNPs. Genotyping was performed using the Perlegen amplicon-based platform.

Phase III (International HapMap 3 Consortium *et al* 2010) of the project genotyped 7 further populations (African-American, Chinese (Denver), Gujarati Indian, Luhya (Kenya), Mexican, Maasai (Kenya) and Toscani (Italy)), but at a lower density of around 1.6 million SNPs. Genotyping was performed using the Affymetrix Human SNP array 6.0 and the Illumina Human1M single bead chip.

Individuals within the HapMap also have corresponding LCLs. LCLs are immortalized B-lynmphocytes generated by infection with the Epstein-Barr virus (EBV). This means that HapMap

individuals can also be used to look at gene expression and other cellular phenotypes. Many of the eQTL studies conducted so far have focussed on analysing the LCLs of the Phase II and III HapMap samples: Stranger *et al* 2007 detected 1348 *cis* eQTLs in the Phase II HapMap dataset, Veyrieras *et al* 2008 re-analysed the Phase II HapMap using Bayesian framework and created a high resolution map of regions that affect mRNA expression in *cis*, Stranger *et al* 2012 detected 5691 *cis* eQTLs in the Phase III HapMap. More recent studies have used RNA-Sequencing techniques Pickrell *et al* 2010 analysed 69 YRI samples, Montgomery *et al* 2010 analysed 60 CEPH individuals.

LCLs have the advantage that they are immortalized cell lines and so provide a way to store individuals DNA for future genotyping and sequencing, and are relatively inexpensive to maintain. They also allow multiple cellular phenotypes to be analysed, such as methylation and alternative splicing, at a later time when relevant technologies become available. However LCLs have the disadvantage that technical artefacts from storage conditions and the age of the cell lines can impact population specific findings. Also, the utility of general / cell-specific conclusions from these cells may be limited: For instance studies have shown that many eQTLs are cell type specific and that integration of GWAS results with eQTL results, should proceed using a cell-type relevant to that disease. For instance obesity, could be studied using adipose tissue. As LCLs have an immunological origin, these cell types can be used to integrate immunity related GWAS phenotypes.

1.3.2 1000 Genomes

The 1000 Genomes Project provides low coverage (2-6X) whole genome sequencing and deep coverage (50-100X) exome sequencing, together with dense SNP genotype data for 1092 individuals from 14 populations (1000 Genomes Project Consortium 2012). The data from this project is used for imputation of GWAS and identification of variation within regions of interest. The populations in this dataset are CEPH (CEU), Tuscan (TSI), British (GBR), Finnish (FIN), Iberian

(IBS), Yoruba (YRI), Luhya (LWK), African American (ASW), Han Chinese (CHB), Japanese Tokyo (JPT) Han Southern Chinese (CHS), Mexican American (MEX), Columbian (CLM) and Puerto Rican (PUR). In total, 38 million SNPs and 1.4 million indels have been typed in this project. (1000 Genomes Project Consortium 2012). 294 of the individuals in the 1000 Genomes dataset are also in the Phase III HapMap dataset. A recent study detected *cis* eQTLs from RNA-Seq with genotype data from 462 individuals from the 1000 Genomes resource (Lappalainen *et al* 2013). This study detected 3,773 genes having an eQTL for gene expression levels.

1.3.3 ENCODE

The Encyclopaedia of DNA Elements (ENCODE) project (ENCODE Project Consortium 2012) is a project that aims to discover and define functional elements in the human genome. Data generated by this project, includes gene annotation, expressed RNA, chromatin structure and modification data, transcription factor binding sites and methylation sites. ENCODE data can be used to determine the function of any region of the genome, for example eQTLs may be within transcription factor binding sites, which leads to a potential function for this variant.

1.4 Statistical Analysis

1.4.1 Association Analysis

In the case of association studies focusing on eQTLs, expression levels are treated as a quantitative trait and linear regression is performed on genotypes at each variant. Quantitative trait association analysis can be carried out using specialist statistical genetics tools such as PLINK (Purcell *et al* 2007), and SNPTEST (Marchini *et al* 2007), or more standard statistical software such as R (R Development Core Team 2011). Covariates such as sex and population structure can also be included.

1.4.2 The Multiple Testing Problem

Due to the high dimensionality of the data-sets analysed (~20,000 genes and ~2 million SNPs), false-positives are expected to be frequent, and therefore the significance threshold needs to be

chosen carefully. A very conservative approach is to use Bonferroni correction which is derived by dividing the required "experiment-wise" significance level by the number of tests performed. An alternative to this is to use the False Discovery Rate (FDR), which estimates the proportion of significant results that are false-positives. FDR can be calculated using several methods including permutations, the R package QVALUE (Storey et al. 2003) and the Benjamini-Hochberg method (Benjamini *et al* 1995).

1.4.3 Confounders

eQTL results are also prone to multiple confounders which need to be taken into consideration, these include batch effects (experimental artefacts), population stratification, sex and age. One way to correct for these confounders is to add covariates to represent these effects into the regression model. Another issue that can cause false-positives is SNPs located within the probe on the array, which can alter hybridization and therefore cause an association with the SNP which is not biologically meaningful. Microarray expression data require normalization before interpretation can take place. Normalization involves the removal of systematic errors between arrays, such as batch effects.

1.4.4 Meta-analysis

In order to improve the power of eQTL experiments to detect associated variants, it is useful to combine the results of different samples, populations and tissues together by means of meta-analysis. However due to the expected differences in allele-frequency and linkage-disequilibrium between diverse populations and expression profiles across cell-types achieving this is not trivial. One solution is to use fixed-effect meta-analysis to pull together effect sizes across populations / cell-types. Effects of similar magnitude and direction will increase the strength of the association. Heterogeneity can also be analysed across samples using measures such as Cochran's Q statistic, which assesses the differences between effect-sizes between studies. However in the presence of heterogeneity alternative approaches to meta-analysis are required such as random effect.

1.4.5 Imputation

Genotype imputation is the process of predicting unobserved genotypes at variants present in a higher-density phased reference panel. The process uses a set of known reference haplotypes, such as those from the Phase III HapMap Project or the 1000 Genomes Project, and utilises observed genotype data to generate a mosaic of the reference haplotypes for each individual. These haplotypes are then used to predict the genotypes of untyped variants by substituting the alleles present on the reference haplotypes at that position. Imputation is performed to increase power to detect association, improve fine-mapping, (because reference panel is more likely to contain the causal variant) and enable meta-analysis if studies are genotyped with different microarrays.

1.5 Aims of thesis

The thesis has the following aims:

1. Detect *cis* eQTL using the Phase II and III HapMap expression and genotype data using association and meta-analysis.

2. Characterize heterogeneity in eQTL effect sizes observed between HapMap populations.

3. Evaluate improvement in fine-mapping due to imputation.

4. Evaluate the improved fine-mapping resolution offered by trans-ethnic meta-analysis.

5. Detect causal variants for GWAS disease SNPs (dSNPs) from the NHGRI GWAS catalog using the eQTL resource generated in order to identify potential function variants for GWAS diseases.

6. Identify *cis* eQTLs for "absorption, distribution, metabolism, and excretion" (ADME) genes and analyse heterogeneity detected in order to identify eQTLs which could be informative within the scope of personalized drug dosage.

CHAPTER 2 MATERIALS AND METHODS

2.1 Overview

This chapter describes the materials and methods used to detect and fine-map *cis* eQTLs in 210 samples from three populations from Phase II HapMap (HapMap 2007) and 709 samples from eight populations from Phase III HapMap (HapMap3 2010). For each phase II sample, the genotype data was generated using custom high-density oligonucleotide arrays. For each phase III sample, genotype data has been obtained for up to 1,316,017 variants from the Affymetrix Human SNP array 6.0 and the Illumina Human1M-single bead chip array. In addition; genotype data has also been obtained for 294 of the phase III HapMap samples from low-pass whole-genome sequencing made available through the 1000 Genomes Project Consortium (1000 Genomes Project Consortium *et al* 2010). For phase II HapMap micro-array expression data was generated using Illumina Sentrix Human-6 Expression V1 BeadChip. For phase III HapMap micro-array expression data has also been derived from LCLs for 21,800 transcripts using Illumina HumanWG-6 v2 expression array. Full details of the data set are given in Section 2.2.

The analysis aims to detect and localise *cis* eQTLs using these data, taking advantage of increased sample size and differences in patterns of linkage-disequilibrium between diverse populations to increase power and improve fine-mapping resolution. The steps in the analytical pipeline are as follows:

Step one: *Association analysis of expression with genotype data from Phase II and III HapMap samples*. Association of each transcript with variants within 1Mb of the TSS was tested in a linear regression framework under an additive model in the number of minor alleles. Analyses were performed within each population. Full details are provided in Section 2.5.

Step two: *Meta-analysis of association summary statistics across populations*. The results of the Phase II and III HapMap association analysis were combined using fixed effect meta-analysis.

Heterogeneity in allelic effects between populations was determined using the Cochran's Q statistic for each signal detected in the meta-analysis. Full details of fixed effect meta-analysis, and heterogeneity analysis are provided in section 2.8.

Step three: *Imputation of Phase III HapMap genotypes for samples not included in the 1000 Genomes Project*. To increase power and facilitate fine-mapping, Phase III HapMap samples not sequenced in the 1000 Genomes Project were imputed up to the phase 1 integrated reference panel ("all ancestries", March 2012 release). Full details are provided in Section 2.6.

Step four*: Association analysis of expression with Phase III HapMap sequenced and imputed genotype data*. The same process as in step one: association of each transcript with variants within 1Mb of the TSS was tested in a linear regression, this time under a dosage model from imputation in the number of minor alleles. Analyses were performed within each population. Full details are given in Section 2.7.

Step five: *Meta-analysis of association summary statistics across populations after imputation*. The same process as in step two: fixed-effects meta-analysis of association summary statistics across populations. After meta-analysis, two significance thresholds were used to identify cis eQTLs: traditional genome wide significance ($p \leq 5x10^{-8}$) and a false discovery rate of 5%.

Step six: *Annotation of cis eQTLs peak signals*. The peak cis eQTL SNPs are annotated using the ANNOVAR annotation tool, using gene-based and region-based annotation, to provide insight into regulatory mechanisms through which these effects are mediated. Full details are given in Section 2.12.

Step seven: Integration of *cis* eQTLs with disease SNPs from GWAS of complex traits. Reciprocal conditional analysis has been performed with SNPTEST in order to integrate GWAS disease SNPs (dSNPs) with peak cis eQTL SNP (eSNP). Full details are given in Section 2.14.

Step eight: *Trans-ethnic fine-mapping of cis eQTLs*. We performed Bayesian trans-ethnic meta-analysis (MANTRA) with HapMap Phase III sequenced and imputed data. Full details are given in Section 2.9. To understand how Phase III HapMap populations are contributing to fine mapping resolution, credible sets have been calculated, as described in Section 2.11.

Step nine: An analysis of the *cis* eQTLs detected using the ADME genes of interest.

2.2 Genotype and Expression Data

This section describes the genetic and expression datasets used to detect *cis* eQTLs. The genetic data specifies genotype and haplotype data from the Phase II and III HapMap and March 2012 1000 Genomes. The expression data specifies the mRNA microarray data derived from LCLs.

2.2.1 Phase II HapMap Genetic Data

The Phase II HapMap comprises of genotype data for four populations presented in table 2.1. Release 23 of the genotype data (HapMap 2007) for each of three populations was retrieved from the PLINK website in binary format (BED). The genotype data was generated using custom high-density oligonucleotide arrays.

Population	Description	Sample Count
CEU	Utah residents with ancestry from northern and western Europe	60 (90)
CHB (ASN)	Han Chinese in Beijing	45
JPT (ASN)	Japanese in Tokyo	45
YRI	Yoruba in Ibadan, Nigeria	60 (90)

Table 2.1: Phase II HapMap populations used in the analysis. The number in brackets in the Sample Count column are the total number of individuals including non-founders.

2.2.2 Phase III HapMap Genetic Data

The Phase III HapMap comprises of genotype data for eleven populations (HapMap3 2010) Release 3 of the genotype data was downloaded from the National Centre for Biotechnology Information (NCBI) website in the PLINK binary PED format.

Full details of the genotyping and quality control protocols utilised are described in this reference (The International HapMap 3 Consortium 2010). Briefly, genotyping was performed using the Affymetrix Human SNP Array 6.0 and the Illumina Human1M-single beadchip. Samples were discarded if they were discordant across platforms (< 95% concordance) or of low quality (< 95% call rate). SNP filtering was implemented on a population specific basis: call rate < 95% and p-value for deviation from Hardy Weinberg equilibrium < $1.0x10^{-6}$. The consensus genotype set contains 1,440,616 SNPs that are polymorphic in 1,184 individuals from 11 populations.

In addition to this, ten 100-Kb regions were selected for direct PCR-Sanger capillary sequencing analysis. Specific QC filters have been applied to SNPs identified within these regions: The specific QC filters included a) sample quality (outliers were identified with significantly low SNP call rate), b) completeness > 80% for each SNP in each population, and c) deviation from Hardy-Weinberg equilibrium $p > 0.001$. 5,758 bi-allelic SNPs passed the filters.

For the analyses presented in this thesis, a subset of eight of the eleven populations is used due to the availability of mRNA expression data for those individuals. Table 2.2 provides information regarding the eight populations used in the analysis. The ancestry group specifies broad ethnicities with similar average pair wise difference in allele frequencies between populations (More information in section 2.9). Sample count specifies the number of individuals with both genotype and mRNA expression data available.

Population	Description	Ancestry Group	Sample Count
LWK	Luhya in Webuye, Kenya	African	83
MKK	Maasai in Kinyawa, Kenya	African	135
YRI	Yoruba in Ibadan, Nigeria	African	108
CHB	Han Chinese in Beijing, China	East Asian	79
JPT	Japanese in Tokyo, Japan	East Asian	81
CEU	Utah residents with Northern and Western European ancestry from the CEPH collection	European-Asian	107
GIH	Gujarati Indians in Houston, Texas	European- Asian	75
MEX	Mexican ancestry in Los Angeles, California	European- Asian	41
Total	--	--	709

Table 2.2: Phase III HapMap populations used in the analysis.

2.2.3 1000 Genomes Genetic Data

The 1000 Genomes project (1000 Genomes Project Consortium *et al* 2010) released genetic data for ~38 million variants from low-pass whole-genome sequence data from 1092 individuals from 26 populations. 294 of the 709 Phase III HapMap individuals used in this analysis have been sequenced as part of that project.

Full details of the genotyping and quality control protocols utilised are described in (1000 Genomes Project Consortium 2012). Briefly data were sequenced using low-coverage whole genome and exome sequencing, ~38 polymorphic SNPs were detected, 1.4 million short insertions / deletions and > 14,000 larger deletions were also detected.

Table 2.3 provides a summary of the overlap for Phase III HapMap and 1000 Genomes genotyping for the samples used in the analysis: **Total:** Specifies the number of individuals used in the analysis from each population; **1000G + H3:** The number of individuals with genotype data in both Phase III HapMap and 1000 Genomes datasets; **H3 only:** The number of individuals with genotype data in Phase III HapMap only.

Population	Individuals		
	Total	**1000G + H3**	**H3 only**
LWK	83	60	23
MKK	135	0	135
YRI	108	45	63
CHB	79	32	47
JPT	81	50	31
CEU	107	71	36
GIH	75	0	75
MEX / MXL	41	36	5
Total	709	294	415

Table 2.3: Counts of individuals with genetic and expression data within Phase III HapMap, and overlap with 1000 genomes March 2012 "all ancestries" panel.

2.2.4 Phase II HapMap Expression Data

Expression data derived from LCLs are available for the phase II HapMap populations (Stranger *et al* 2007). The expression data from the LCLs is publicly available in the GEO database (**GSE6536**). The data was generated using Illumina Sentrix Human-6 Expression V1 BeadChip (accession **GPL2507**). This chip has 47,296 probes (26098 from RefSeq, 9576 from Gnomon and 11,622 from UniGene). The raw data was normalised on a log-scale using a quantile normalising method across replicates of a single individual followed by median normalisation across all 270 individuals. The data were extracted using the GEOquery package from Bioconductor in R. Only the RefSeq probes were selected to perform the scan, and of these 20,846 had annotation in Ensembl BioMart on NCBI Build 36, which mapped to 17,681 unique genes.

2.2.5 Phase III HapMap Expression Data

Expression data from LCLs for the eight phase III HapMap populations were provided by the Dermitzakis group, University of Geneva. The data provided include a list of 21,800 unique Illumina HumanWG-6 v2 probes, each annotated with an Ensembl gene identifier, chromosome and position in NCBI Build 36. The annotations for these probes were updated to NCBI Build 37 using Ensembl Biomart.

These data were generated in the following steps (fully described in Stranger *et al* 2012).

Step 1 RNA preparation: RNA was prepared using *in vitro* transcription (IVT) reactions, and for each individual, two reactions were carried out.

Step 2 Expression Quantification: The Illumina Sentrix Human-6 Expression BeadChip Version 2 was used to detect expression intensity data in the LCLs. The array has approximately 48,000 unique bead types, one for each of 47,294 transcripts and the remainder being controls. For each of the individual's samples, two IVT reactions were hybridized to one array each, with the result that each cell-line had two replicate hybridizations.

Step 3 Normalization: The expression intensity values were quantified using the Illumina HumanWG-6 v2 expression array in two technical replicates. Replicates of single individuals were normalized on a $\log_2$ scale using quantile normalization. Quantile normalization transforms the distribution of probe intensities for each array in a set of arrays to be the same, and consequently identical in statistical properties. This was followed by median normalization method across all individuals of the eight populations, where all intensity levels were adjusted by subtracting the median intensity level. In order to remove outliers, rank normalization was performed within each of the populations. This generates a quantitative version of the spearman rank correlation, which shares its properties and also allows linear models to be applied with additional covariates. The expression data for GIH, LWK, MEX and MKK were normalized for admixture using a custom version of EIGENSTRAT (Price *et al* 2006). For more information regarding admixture correction see section 2.4.4.

Step 4 Probe Selection: In total, expression data was collected for 47,294 probes out of these 21,800 probes which mapped to 18,226 unique autosomal Ensembl genes were selected for analysis. Genes mapping to the X and Y chromosomes were excluded.

2.3 Quality Control

This section provides information about Quality Control (QC) measures applied to genotype data in the analyses presented in this thesis.

2.3.1 Phase II HapMap genotypes

Quality control (QC) was performed using PLINK (Purcell *et al* 2007) with the following parameters: a minor allele frequency (MAF) ≥ 0.05; SNP missing rate ≤ 0.05; deviation from Hardy Weinberg equilibrium (p-value ≤ $1x10^{-3}$). Non-founders were removed from the CEU and YRI populations.

2.3.2 Phase III HapMap genotypes

QC of the Phase III HapMap genotypes was performed using PLINK (Purcell *et al* 2007). The following parameters were applied: a minor allele frequency (MAF) ≥ 0.05; a maximum per SNP missing rate of 0.05; a deviation from Hardy Weinberg equilibrium p-value ≤ $1x10^{-3}$. Non-founders (i.e. offspring and siblings) were removed from the populations when present.

2.3.3 SNPs within probes

A cause of false-positive eQTLs are SNPs which are located within the genomic region of probes. These SNPs can alter the binding of the cDNA to the probe in the microarray and therefore present a signal between the SNP and the probe which is not biologically meaningful.

Previous studies using the HapMap LCLs detected SNPs within probes. The Stranger *et al* 2007 study using HapMap 2 cell lines identified 99 probes which had SNPs within them that generated false discoveries, and these were filtered from the analysis. An analysis of the HapMap 3 cell lines (Stranger *et al* 2012) did not filter for SNPs within probes, and identified 1000 Genomes SNPs within 1401 of the 21,800 probes used within the analysis.

In this analysis I have chosen not to filter for probes. Intead, eQTLs identified are inspected to search for 1000 Genomes SNPs within the probe region that might reflect false positive signals.

2.4 Population Structure

Population structure can cause false-positives in association analysis when there are even subtle differences in allele frequency and trait values between sub-populations. This section presents how inflation due to population structure was detected and corrected for in association analyses.

2.4.1 QQ Plots

QQ plots were generated to detect whether p-values from the association and fixed effect meta-analysis differ from the null distribution. $-\log_{10}$ p-values were extracted from the results of the analysis and plotted against values expected if the same number of variants were sampled from a null distribution. QQ plots could indicate that there was population structure, if there was deviation from the y=x lines at relatively insignificant p-values.

2.4.2 Genomic Control

In order to assess whether any association signals were inflated due to population structure, the genomic control (GC) inflation factor (λ) was calculated with the p-values from the association analysis and meta-analysis.

GC is determined by taking the median of the square of the normal quantiles for the p-values (which has a chi squared distribution) and dividing by 0.456 (which is the median of the chi-squared distribution under the null). The assumption is that values at the median of the results will be null and so dividing by the null chi-squared will give an idea of any systemic inflation due to population structure. If GC is much bigger than 1 this suggests that there is population structure.

$$\hat{\lambda} = \frac{median\left(Z_1^2, Z_2^2, \ldots Z_L^2\right)}{0.456} \quad (2.1)$$

2.4.3 Principal Component Analysis (PCA)

In order to detect population structure, principal component analysis (PCA) of the genetic data for each of the populations was performed using the SmartPCA program of EIGENSTRAT (Price *et al* 2006). This generates principal components that reflect the genetic relatedness of the population.

2.4.4 Removal of population structure present in expression

In populations where structure was detected, expression data was normalized by the Dermitzakis group using a customized version of EIGENSTRAT (Price *et al* 2006) which generates principal components on the basis of genetic data. Expression values were adjusted for each population using ten primary axes of variation from that population's corresponding intra-population PCA of the set of whole genome SNP genotypes. The residual normalized expression values were used as input for the association analysis.

2.5 Association Analysis: HapMap

This section presents how association analysis was performed with the Phase II and III HapMap genotype and expression data to detect *cis* eQTLs.

2.5.1 Linear Regression: PLINK --linear

In order to detect *cis* eQTLs, PLINK (Purcell *et al* 2007) –*linear* option performs quantitative trait association analysis using a linear regression model. Probe expression intensities were regressed on SNP genotypes, where the SNPs tested were those at least 1Mb upstream and downstream of the probe's gene TSS.

The linear regression model is specified below:

$$Y_{ik} = \alpha_k + \beta_{jk} X_{ij} + \varepsilon_{ijk} \qquad (2.2)$$

Where Y_{ik} is the transcript expressions for individual *i* and transcript *k*, X_{ij} is the genotype for individual *i* and SNP *j*, coded as 0 (homozygous major allele), 1 (heterozygous) and 2 (homozygous

minor allele) for individual *i* and SNP *j*. The parameter α_k is the intercept for transcript *k*, and β_{jk} is co-efficient forSNP *j* (effect size). The parameter ε_{ijk} is a normally distributed error term.

The linear regression model can be generalised to incorporate covariates to adjust for confounders, in particular to adjust for sex using the *–sex* option in PLINK.

2.5.2 Wald Test

PLINK generates an effect size $\hat{\beta}_{jk}$ and its standard error $se(\hat{\beta}_{jk})$. To assess significance a Wald test is performed on the estimated effect sizes. The Wald test assumes that the difference between the calculated effect size and expected effect size (in this case zero) is approximately normally distributed. A test statistic is generated using the following equation:

$$\frac{\widehat{\beta}_{jk}-\beta_0}{se(\widehat{\beta}_{jk})} \quad (2.3)$$

Where β_0 is the value to which the effect sizes is is compared (in this case zero), $\hat{\beta}_{jk}$is the estimated effect size and $se(\hat{\beta}_{jk})$ the standard error of the estimated effect size. Two sided p-values are determined using the standard normal distribution, and provide an assessment of the probability of the observed data (or more extreme) under the null hypothesis of no association between genotypes at SNP j and expression of transcript k.

2.6 Imputation

This section presents how the Phase III HapMap genotype "scaffold" was imputed with the 1000 Genomes "all ancestries" March 2012 reference panel.

2.6.1 Imputation "scaffold" pipeline

To perform imputation, a "scaffold" of variants which are typed in both the reference panel and the panel to be imputed into is required. In order to generate the "scaffold" a pipeline was applied to the Phase III HapMap genotype data with the following stages:

Step 1: Update NCBI Build 36 to 37: The SNPs in the phase III HapMap dataset were updated from NCBI B36 to B37 with the following stages: 1. Update of Chromosome and base pair position. 2. Update of strand alignment using NCBI B37 annotation for the Affymetrix 6.0 and Illumina 1.0 Million SNP arrays.

Step 2: Minor Allele Frequency (MAF) Filtering: Any SNPs with a MAF ≤ 1% were removed.

Step 3: Conversion from PLINK binary PED format to GEN format: The software GTOOL was used to convert from PLINK PED format to IMPUTE / SNPTEST GEN format.

2.6.2 1000 Genomes imputed genotype QC

After completion of imputation, each imputed SNP has:

- Posterior probability of genotype dosage frequency for each variant.
- A calculated MAF for each variant.
- An INFO score for each variant.

To apply QC to the imputed SNPs, MAF was filtered at two levels: MAF 5% and 1%. A threshold of ≥ 0.4 was chosen for the INFO score (Winkler *et al* 2014). This threshold has been widely used in GWAS of complex human traits (for example Morris *et al* 2012)

2.6.3 IMPUTE V2

The software IMPUTE V2 (Howie *et al* 2009) was used to perform imputation. IMPUTE V2 uses a hidden Markov model to calculate the imputed genotype posterior probability:

$$P(G_i|H,\mu,\rho) = \sum_z P(G_i|Z,\mu), P(Z|H,\rho) \qquad (2.4)$$

where G_i is genotype of individual *i*, H is the complete set of reference haplotypes, ρ is the recombination rate and μ is the mutation rate. Z corresponds to the "hidden states" of the model and can be thought of as pairs of haplotypes selected from the reference panel with alleles that

then make up the imputed genotype. The term $P(Z|H,\rho)$ models how pairs of haplotypes change along the sequence using the recombination rate. The term $P(G_i|Z,\mu)$ allows observed genotypes to differ from the alleles at haplotypes through mutation at rate μ. The rates of recombination and mutation depend on the "effective" population size, which is recommended (IMPUTE V2 website) to be set to 20,000 for multi-ethnic studies. In future releases of impute this will be the default setting.

2.6.4 Imputation quality control (INFO score)

IMPUTE generates a posterior probability for each of the imputed genotypes and a quality metric: the INFO score. This score is based on measuring the relative statistical information about the population allele frequency θ_j at SNP *j* and measures the ratio of the observed and complete information of the allele frequencies.

$$I_A = \frac{E_{G.j}[I(\hat{\theta})] - V_G[U(\hat{\theta})]}{E_{G.j}[I(\hat{\theta})]} \quad (2.5)$$

where I_A is the INFO score, $E_{G.j}\left[I\left(\hat{\theta}\right)\right]$ is the complete information and $V_G[U(\hat{\theta})]$ is the missing information at SNP *j*.

I_A is bounded above at 1 and will equal zero when the sample mean variance of the imputed genotypes equals the variance you would expect if alleles were sampled with frequency $\hat{\theta}$.

The INFO score should be interpreted as follows: an imputed variant with an INFO score of 0.5 has 50% of the power to detect association compared to if it had been directly genotyped. Variants with a lot of uncertainty in imputed genotypes will have a low INFO score.

2.6.5 Merging of datasets

On completion of imputation two datasets were present: the posterior distribution of genotype calls for imputed variants for the Phase III HapMap samples without 1000 Genomes data, and the

directly typed 1000 Genomes genotype data for the remainder. These two datasets were merged using the software GTOOL. The pipeline has the following steps:

1. Retrieve the 1000 Genomes genotype data (from FTP site) and convert to PLINK BED format.

2. Convert the PLINK BED files to GTOOL GEN format using GTOOL –P option.

3. Merge the GEN files with the output *.impute files from imputation using GTOOL –M option.

2.7 Association Analysis: Imputed Data

This section presents how quantitative trait association analysis was performed on imputed genotypes to detect *cis* eQTLs. The major difference with the analysis in section 2.5 is the formatting of the genotype data used as input for the analysis. The analysis on non-imputed genotypes expects genotype data in the PLINK PED format, in the case of imputed genotypes the genotype data is in the IMPUTE and SNPTEST Genotype (GEN) file format. In addition the GEN file generated by IMPUTE V2 contains posterior probabilities for each genotype.

2.7.1 PLINK --dosage

In order to use GEN files with PLINK, the *–dosage* option is used, which permits flexibility to take a "dosage" file typically from imputation packages as input for further PLINK options. Once the –dosage option has been used correctly to read the GEN dosage files, PLINK is used as in section 2.5 to perform linear regression. (See equation 1 in section 2.5), where now X_{ij} denotes the dosage of the minor allele for individual i at SNP j. Significance is again assessed using the Wald test (See equation 2.2 in section 2.5).

2.9 Trans-ethnic meta-analysis (MANTRA)

This section provides an overview of the Bayesian trans-ethnic meta-analysis software MANTRA (Meta-ANalysis of TRansethnic Association studies) (Morris 2011). MANTRA partitions the

populations into ancestry groups, for example using average allele frequency differences between populations. Effect sizes within ancestry groups are assumed to have the same allelic effect size, whilst those in between ancestry groups are allowed to differ.

2.9.1 Bayes Factor (BF)

Evidence for the null model M_0 of all effect sizes being zero ($\boldsymbol{\beta} = 0$) against the alternative model M_1corresponding to ($\boldsymbol{\beta} \neq 0$) can be assessed by calculating a Bayes' Factor (Δ) which is a ratio of marginal likelihoods for the two alternative models.

$$\Delta = \frac{f(\boldsymbol{b},\boldsymbol{s}|M_1)}{f(\boldsymbol{b},\boldsymbol{s}|M_0)} \quad (2.9)$$

$f(\boldsymbol{b},\boldsymbol{s}|M)$ denotes the marginal likelihood of the observed allele effects under model *M*. This is calculated by integrating over the unknown model parameters θ which include the population specific allele effects **β** and additional hyperparameters for the prior distribution.

$$f(\boldsymbol{b},\boldsymbol{s}|M) \propto \int_{\theta} f(\boldsymbol{b},\boldsymbol{s}|\boldsymbol{\theta}) f(\boldsymbol{\theta}|M) \delta\boldsymbol{\theta} \quad (2.10)$$

2.8 Fixed Effect meta-analysis (GWAMA)

This section specifies the software and model used to perform fixed effect meta-analysis of association summary statistics across populations from Phase III HapMap. Fixed effect meta-analysis combines association summary statistics across different populations, under an assumption of homogenous allelic effect sizes.

2.8.1 GWAMA

Fixed-effect meta-analysis was performed on association summary statistics for each population using the software GWAMA (Magi *et al* 2010). GWAMA estimates combined effect-size across populations. In the fixed-effect meta-analysis a weighted effect size (B_{jk}) at SNP *j* for all samples is calculated by summing the individual effect-sizes (β_{ij}) and scaling by the inverse variance of the effect size (w_{ij}) at SNP j and study *i*.

$$B_j = \frac{\sum_{i=1}^{N} \beta_{ij} w_{ij}}{\sum_{i=1}^{N} w_{ij}} \quad (2.6)$$

$$w_{ij} = \left[Var(\beta_{ij})\right]^{-1}$$

To assess significance the square of the estimated effect size (B_j) divided by its variance (V_j) gives an approximate X^2 distribution with one degree of freedom.

$$V_j = \left[\sum_{i=1}^{N} w_{ij}\right]^{-1} \quad (2.7)$$

$$X_j^2 = \frac{B_j^2}{V_j}$$

2.8.2 Heterogeneity

Heterogeneity between the populations was assessed using Cochran's Q statistic (Q_j) at variant *j*. Cochran's Q is calculated by summing the differences between the effect size of the meta-analysis (B_j) at variant j with each samples effect size (β_{ij}) at sample *i* and variant *j*, weighted by the inverse variance.

$$Q_j = \sum_{i=1}^{N} w_{ij} (B_j - \beta_{ij})^2 \quad (2.8)$$

Cochran's Q statistic has an approximate X^2 distribution with $N_j - 1$ degrees of freedom. (N_j is the number of association analysis results merged in the meta-analysis) Significant evidence of heterogeneity is an indication that the assumption of a homogenous effect size across populations is not valid.

2.9.2 Likelihood function

The likelihood ($f(\boldsymbol{b}, \boldsymbol{s}|\boldsymbol{\theta})$) is obtained from a Bayesian partition model and is of the form:

$$f(b_i, s_i|\beta_i) = f(b_i, s_i|K, \boldsymbol{C}, \boldsymbol{\varphi}) \propto \frac{1}{s_i} exp\left[-\frac{(b_i - \sum_{k=1}^{K} T_{ik}\varphi_k)^2}{2s_i^2}\right] \quad (2.11)$$

Each population are assigned to one of $\boldsymbol{K}$ cluster centres $\boldsymbol{C}$ each having a cluster specific allele effect ($\boldsymbol{\varphi}$). $\boldsymbol{T}$ is a tessellation, where T_{ik} = 1 if population P_i is assigned to the cluster with centre $C_j$and 0 otherwise.

When K = 1, there is a single cluster and method is a Bayesian version of fixed effect meta-analysis, similarly when K = the total number of association analysis results (N), all populations effect sizes are modelled to be different and method is a Bayesian random effect meta-analysis.

2.9.3 Prior density function

The Bayes Factor Δ depends on the prior density function ($f(\boldsymbol{\theta}|M)$) of parameter under model M.

Under M_0 $f(\boldsymbol{\theta}|M_1)$ = 1 if β = 0 and 0 otherwise.

Under M_1 population specific allelic effects are determined by the Bayesian partition model.

Under this model the prior denssity of the number of clusters of populations is given by:

$$f(K) = \begin{cases} \frac{1}{2} & if\ K = 1 \\ \frac{2^{N-1}}{2^k(2^{N-1}-1)} & otherwise \end{cases} \qquad (2.12)$$

This prior model gives greater probability to a partition with fewer clusters of populations.

Each of the clusters allelic effects have a prior $N(\mu, \sigma)$ where μ has a prior uniform distribution and σ has a prior exponential distribution with expectation 1.

Combining the components of the prior density function, it follows that:

$$f(\boldsymbol{\theta}|M_1) \propto f(K)(N-K)!\frac{\exp[-\sigma]}{\sigma}\prod_{k=1}^{k}\exp[-\frac{(\varphi_k-\mu)^2}{2\sigma^2}] \qquad (2.13)$$

2.9.4 MCMC Algorithm

It is not possible to calculate the marginal likelihoods directly. However the joint posterior density of **θ** under model M is given by:

$$f(\boldsymbol{\theta}|\boldsymbol{b},\boldsymbol{s},M) \propto f(\boldsymbol{b},\boldsymbol{s}|\boldsymbol{\theta})f(\boldsymbol{\theta}|M) \qquad (2.14)$$

This density appears in the integrand of equation (1) and can be approximated through the use of the Metropolis-Hastings MCMC algorithm.

$$f(\boldsymbol{b},\boldsymbol{s}|M) \approx \left[\frac{1}{R}\sum_{r=1}^{R} f(\boldsymbol{b},\boldsymbol{s}\,|\boldsymbol{\theta}^{(r)})^{-1}\right]^{-1} \qquad (2.15)$$

$$f(\boldsymbol{b}, \boldsymbol{s} \,|\boldsymbol{\theta}^{(r)}) = \prod_{i=1}^{N} f\left(b_i, s_i \middle| K^{(r)}, \boldsymbol{C}^{(r)}, \boldsymbol{\varphi}^{(r)}\right)$$

Where $f\left(b_i, s_i \middle| K^{(r)}, \boldsymbol{C}^{(r)}, \boldsymbol{\varphi}^{(r)}\right)$ is given by equation (2.1).

An estimate of the Bayes Factor can be obtained from two independent runs of the MCMC algorithm once for each model M_0 and M_1

2.10 Multiple Testing Correction

Due to the large number of statistical tests performed, a significance threshold is required that takes into account the large number of false-positives that are expected to be detected. In this study to significance thresholds have been selected to take this into account.

1. Genome Wide Significance: ($p \leq 5x10^{-8}$)

2. False Discovery Rate (FDR) of 5% (Benjamini-Hochberg)

2.10.1 Genome wide significance

Genome wide significance is a standard significance threshold used for replication in genome wide association studies with the value $p \leq 5x10^{-8}$ (Clarke *et al* 2011).

2.10.2 FDR: Benjamini-Hochberg

FDR controls the expected proportion of truly null hypothesis which are falsely rejected. There are several ways to calculate corresponding p-values. In this case the Benjamini-Hochberg method has been selected.

The Benjamin-Hochberg procedure calculates the p-value corresponding to FDR thresholds using the equation:

$$FDR_i \leq \frac{n \, x \, p_i}{n_{Ri}} \qquad (2.16)$$

Where n_{Ri} is the observed number of test statistics declared as significant, n is the total number of tests and p_i is the required significance threshold. FDR_i is the FDR level corresponding to p_i.

2.10.3 Permutation-based approaches

An alternative approach to correct for multiple testing is to use permutation based approaches (Churchill *et al* 1994). Permutation-based approaches empirically identify the null distribution for an experiment, by randomly shuffling the phenotype (in this case gene expression intensity) across individuals in the study and then generating test statistics between the phenotype and genetic markers. This approach was used in both the HapMap based studies Stranger *et al* 2007 and Stranger *et al* 2012.

2.11 Credible Set Analysis

Trans-ethnic meta-analysis leverages differences in LD patterns between populations to fine map causal variants. In order to determine improvements in fine mapping achieved by combining populations using meta-analysis, the Bayesian Trans-ethnic meta-analysis (MANTRA) results have been used to calculate 99% credible sets for *cis* eQTLs.

MANTRA partitions the populations into ancestry groups, and 99% credible sets are calculated over all ancestry groups together, and each ancestry group on its own. This will show the relative contribution of each ancestry groups to fine mapping.

To determine the 99% credible sets, the following analysis was performed:

1. The *i* SNPs at the eQTL locus were ranked by their Bayes Factor from MANTRA (BF).

2. The total BF was summed for N SNPs at the locus.

$$T = \sum_{i=1}^{N} BF_i \qquad (2.17)$$

3. Posterior probability for each SNP was calculated by dividing its BF by the Total BF.

$$PP_i = \frac{BF_i}{T} \quad (2.18)$$

4. The cumulative probability of the ranked SNPs was calculated until ≥99% posterior probability was reached.

$$CI_{99} = \sum_{i=1}^{N} PP_i \geq 99\% \quad (2.19)$$

The result of this analysis is a list of SNPs accounting for 99% posterior probability at each eQTL locus.

2.12 Annotation

The association and meta-analysis identifies peak eQTL SNPs (eSNPs) which are strongly associated with gene expression. In order to determine what the functional consequences are that lead from polymorphisms in the eSNP, the plausible annotated function of the SNP can be investigated. For example if the eSNP is located within an intron, this could indicate that differences in expression may be related to transcript stability, another example could be if the SNP is located within a known promoter or enhancer.

The software ANNOVAR (Wang *et al* 2010) performs the integration of candidate genetic variants with annotation in multiple databases. There are three types of annotation that this tool provides: Gene based which look at the function of the SNPs within near-by genes, Region based that looks at overlap between the SNP and known functional regions and filter based which determines whether variants are within existing variant databases (for example is the SNP in the 1000 Genomes dataset).

For the purpose of this analysis I have looked at gene-based and region-based annotation for the eSNP variants.

2.12.1 Gene-based annotation

This determined whether the eSNP was located in respect to nearby genes. For example, is the gene located in an exon or an intron? If the eSNP is located within intergenic regions then the distance to the two flanking genes is reported.

2.12.2 Region-based annotation

This analysis looks at the overlap between the eSNP variant and functional regions from a selected database. The functional regions can be *cis* acting elements, such as transcription factor binding sites. For this analysis, eight ENCODE region-based annotation tests have been performed. In some of the ENCODE assays, the analysis has been performed on several different cell lines, and the cell line used to generate the annotation needs to be specified. For the case of this analysis the cell line GM12878 has been selected for all of these assays. The GM12878 cell line is a LCL from within the CEU population, and so is relevant to this study.

1. **wgEncodeRegDnaseClustered**: Variants located in ENCODE DNAse I Hypersensitivity sites.

2. **wgEncodeRegTfbsClustered**: ENCODE Transcription Factor ChIP-Seq data.

3. **wgEncodeBroadHmmGm12878HMM**: chromHMM predictions using the GM12878 cell lines to classify non-coding variants using chromatin state.

4. **wgEncodeHaibMethyl450Gm12878SitesRep1:** Methylation data using the Methyl 450K BeadArray using the GM12878 cell line.

5. **wgEncodeBroadHistoneGm12878H3k27acStdPk:** Histone modification H3K27ac, associated with active promoters.

6. **wgEncodeBroadHistoneGm12878H3k4me1StdPk:** Histone modification H3K4me1, associated with active enhancers.

7. **wgEncodeBroadHistoneGm12878H3k4me3StdPk:** Histone modification: H3K4me3, associated with active promoters.

8. **wgEncodeBroadHistoneGm12878H3k9acStdPk.txt:** Histone modification: H3K9ac, associated with active enhancers.

2.13 GWAS Integration:

Many of the disease SNPs (dSNPs) reported in GWAS are located within intergenic regions. A role that these dSNPs may play is in regulation of gene expression, and that changes in mRNA expression will impact disease risk. eQTL SNPs (eSNPs) can be integrated with dSNPs in order to generate hypothesis of which dSNPs have a potential regulatory function through gene expression. The results of the imputed fixed effect meta-analysis were analysed for overlap with dSNPs from the NHGRI GWAS catalogue (Hindorff *et al*). The analysis uses reciprocal conditioning to determine whether the dSNP and the eSNP are tagging the same variant.

2.13.1 Conditional Analysis

The GWAS dSNP catalog was filtered using a criterion of genome wide significance ($p \leq 5x10^{-8}$) to generate list of dSNPs to use in the analysis. This was followed by a scan to determine which of the eSNP and dSNPs are within 1 MB of each other. This resulted in a list of dSNP and eSNP pairs.

In order to determine whether the dSNP had any effect on the signal at the corresponding eSNP, the list of eSNP and dSNP pairs were used as input into a reciprocal conditional analysis. Conditional analysis was performed using the additive frequentist analysis of SNPTEST using the option –condition_on (Marchini *et al* 2007).

2.13.2 Criteria for follow up

The following criteria were used to select possible dSNP and eSNPs with the same functional variant.

1. The dSNP and eSNP are the same variant.

2. Both the dSNP and eSNP have p-value ≥ 0.05 after reciprocal conditional analysis

3. Both the dSNP and eSNP have p-value ≥ $1x10^{-3}$ after reciprocal conditional analysis, and the trait of the dSNP is immunity related.

LD between the dSNP and eSNP was measured using r^2 with Broad SNAP (Johnson *et al* 2008).

2.14 ADME Analysis

2.14.1 Enrichment analysis

Gene enrichment analyses allow the identification of biological processes and entities associated with a list of candidate genes. There are many tools available to perform enrichment analysis, such as DAVID (Huang *et al* 2009), PANTHER (Mi *et al* 2013) and Ingenuity Pathway Analysis (IPA® , QIAGEN Redwood City, www.qiagen.com/ingenuity).

For this analysis gene pathway enrichment analysis was performed using software developed at GlaxoSmithKline known as Multiverse. The reason this software was chosen was because this analysis was performed on site at GlaxoSmithKline and the tool gives access to several private data sources, some of which are GlaxoSmithKline own internal data and other private data.

For the analysis of the eQTL data, a Fisher's exact test was performed by Multiverse. The origin of the enrichment terms used are as follows:

- MeSH: (Medical Subject Headings) is the NLM controlled vocabulary thesaurus used for indexing articles for PubMed

- GeneGO is a company that provides system biology databases

- Reactome is a free online database of biological pathways.

Fisher's exact test is used to analyse contingency table with two categorical variables to test whether proportions observed are significant.

	Target Genes	Non-Target Genes	Row Total
Bucket	A	B	a + b
Not in Bucket	C	D	c + d
Column total	a + c	b + d	(a + b + c + d) = n

Table 2.4: Example of contingency table analysed using fisher exact.

The probability of obtaining values in the contingency table is given by the hyper geometric distribution.

$$p = \frac{\binom{a+b}{a}\binom{c+d}{c}}{\binom{n}{a+c}} \quad (2.20)$$

Where $\binom{n}{k}$ is the binomial coefficient.

CHAPTER 3 PHASE II HAPMAP ANALYSIS

3.1 Overview

The Phase II HapMap samples comprise of 3 populations; Asian (ASN) that consists of Chinese (CHB) and Japanese (JPT) samples, CEU that comprise European descent samples and YRI that comprises Yoruba (Nigeria) samples. See methods section 2.3.1.1 for more information on populations.

The Phase II HapMap ASN, CEU and YRI populations comprise 90, 60 and 60 unrelated individuals, respectively, which have been genotyped at ~2 million SNPs after QC. In addition to this RNA expression has been measured at these three populations using micro-arrays; (20,846 probes mapping to 17,681 genes). Full details of the Phase II HapMap data and QC procedures in Methods section 2.3.1.

In this chapter, quantitative trait association analysis is used to detect *cis* eQTLs defined as 1 Mb upstream and downstream of the probe's putative gene's transcription start-site (TSS) in the Phase II HapMap expression and genotype datasets. This is carried out on each Phase II population separately. Following this, fixed effect meta-analysis is used to combine the three association analysis results. See methods section 2.5 for full details of the association analysis and section 2.8 for the fixed effect meta-analysis.

Several examples of *cis* eQTLs were selected for further investigation, including those with the strongest association signal, strong evidence of heterogeneity in allelic effects between populations, and to highlight the possibility of using differences in LD structure between populations to perform fine mapping.

Positions of SNPs and genes in this section use NCBI Build 36. References to gene function are not cited as the information has been taken from GeneCards and so is assumed to be public

knowledge. HGNC (HUGO Gene Nomenclature Committee) refers to the HUGO symbol of a particular gene.

3.2 Quality Control

Phase II HapMap genotype data were filtered for exclusion using standard QC measures: MAF ≤ 0.05, SNP missing rate ≥ 0.05 and p-value for deviation from Hardy-Weinberg equilibrium ≤ $1x10^{-3}$. See section 2.3.1 for more information on QC metrics.

Table 3.1 presents the numbers of polymorphic SNPs in each population before and after QC was applied. The table also shows the count of individuals within each population. The CEU and YRI both comprise of 30 trios, so non-founders were removed from these populations as part of the QC. Intersection specifies the number of SNPs that are polymorphic in all three populations.

From table 3.1 it can be seen that the ASN population has the lowest numbers of SNPs which pass the QC analysis. This appears to be driven by the higher number of SNPs with MAF ≤ 0.05 in this population.

Population	SNPs before QC	SNPs after QC	Individuals
ASN (CHB + JPT)	3,998,895	1,871,681	90
CEU	3,967,651	2,076,382	60 (90)
YRI	3,880,150	2,268,301	60 (90)
Intersection	3,733,318	1,271,775	210 (270)

Table 3.1: Summary of genotype quality control on phase II HapMap populations. Individuals: Bracketed counts include the non-founders in CEU and YRI.

3.2 Population Structure

We began by assessing each of the Phase II HapMap populations for evidence of population structure using the SmartPCA principal component analysis component of EIGENSTRAT (Price *et al* 2006). See methods section 2.4.3 for more information on PCA. The analysis was carried out on each population separately and all populations together. All variants that passed QC (See table 3.1) were used as input of SmartPCA of EIGENSTRAT (Price *et al* 2006). SmartPCA takes account of LD between SNPs and therefore no LD pruning was undertaken prior to this analysis.

Figure 3.1 shows the population structure of all three populations together, obtained by plotting the first two eigenvectors obtained from SmartPCA against each other. When looking at the three populations together, each population forms a tight cluster. This indicates that there are no clear population outliers that should be excluded from downstream association analysis.

The population structure within each of the three populations is shown in figure 3.2, obtained by plotting the first two eigenvectors obtained from the population-specific SmartPCA against each other. The CEU and YRI populations both formed one cluster with a few outliers, whilst the ASN population formed two clusters corresponding to the CHB and JPT populations. This indicates that there is some fine-scale structure in the ASN population, the impact of which we sought to investigate further through adjustment for SmartPCA eigenvectors in downstream association analyses.

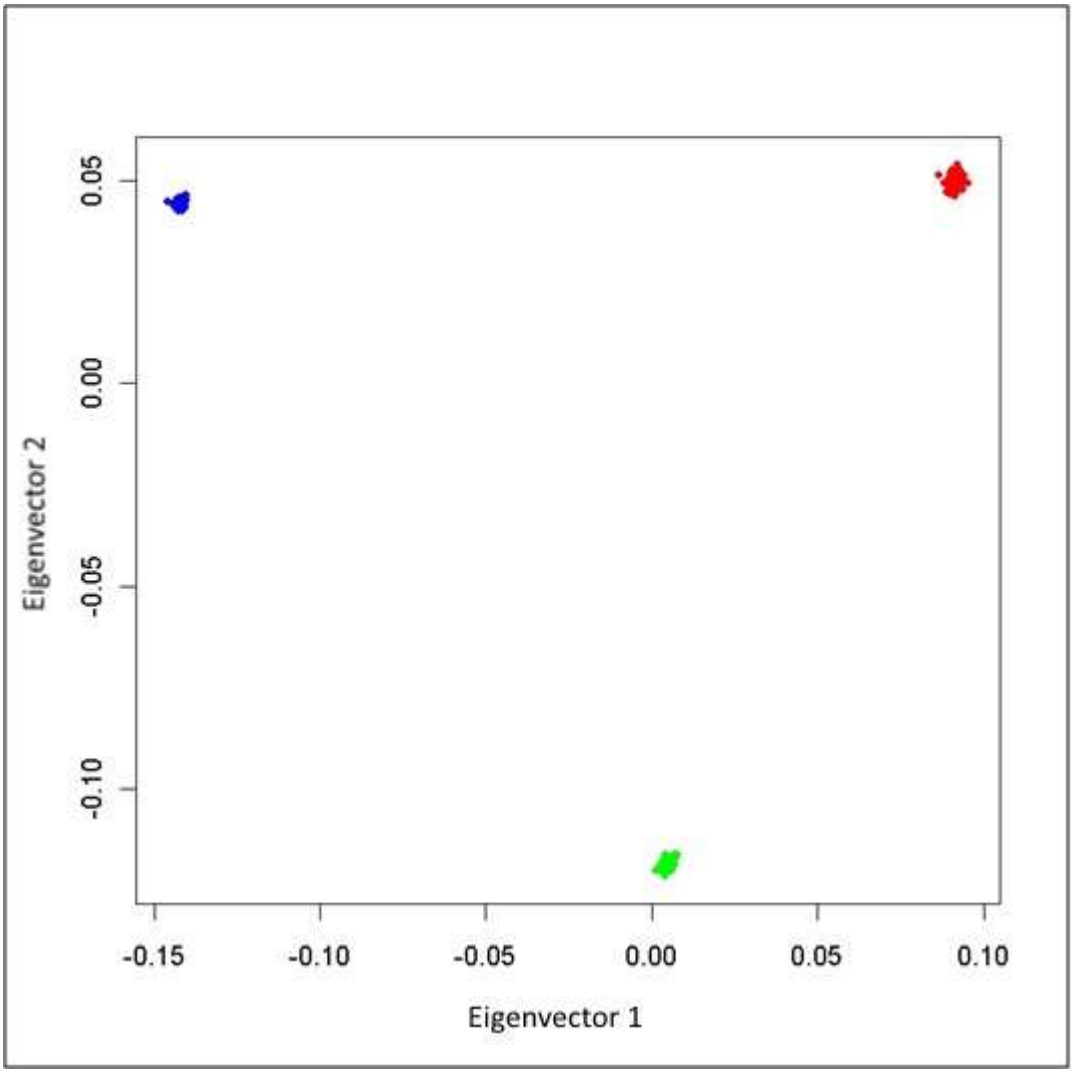


Figure 3.1: Results of SmartPCA principal component analysis of Phase II HapMap genotype data. ASN is red, CEU is green and YRI is blue.

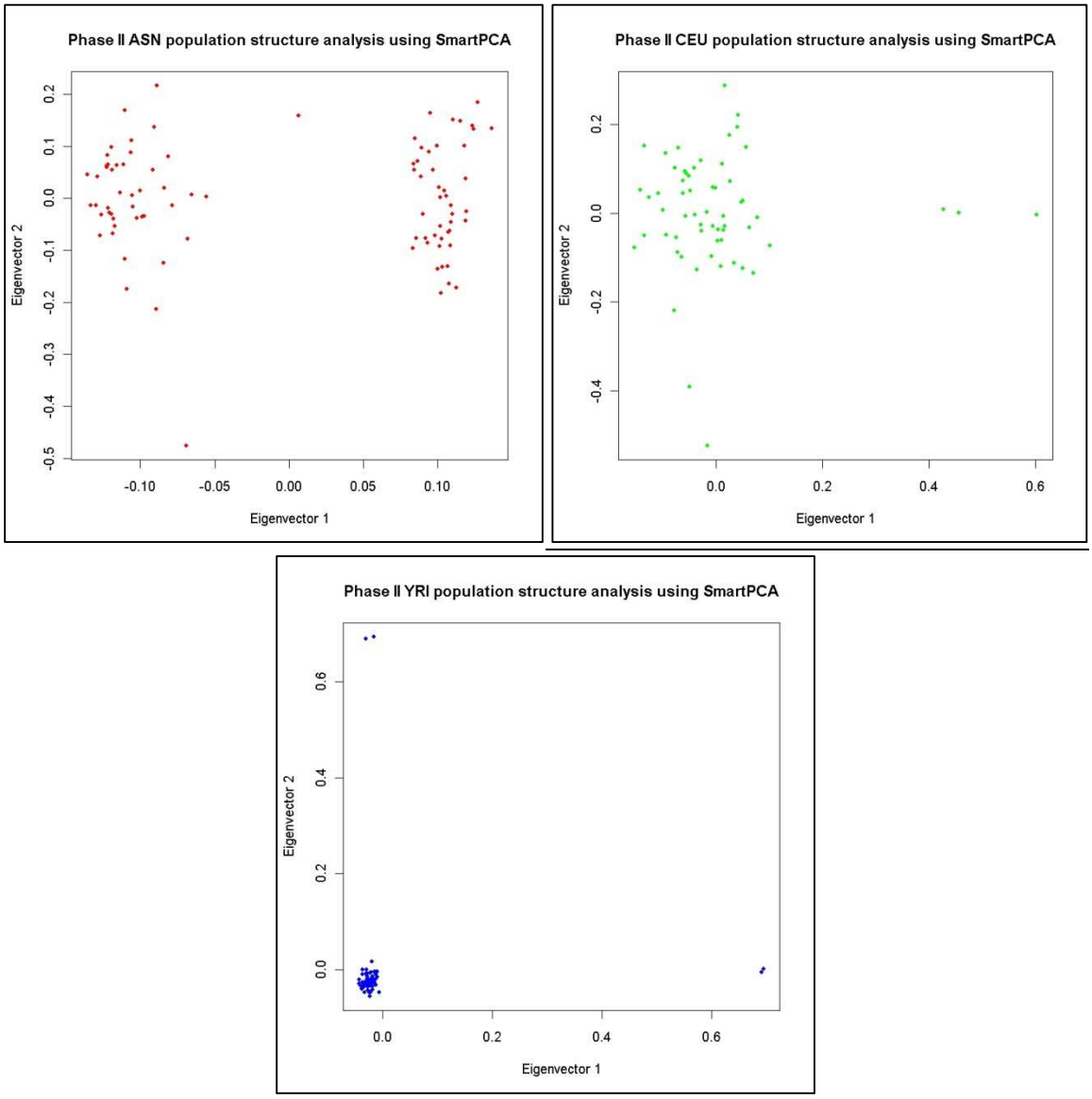


Figure 3.2: Results SmartPCA principal component analysis of Phase II HapMap genotype data. ASN has red dots, CEU has green dots and YRI has blue dots. First two eigenvectors are plotted.

3.3 Multiple Testing

Due to the large number of association tests performed in this analysis, multiple testing needs to be taken into account when assessing significance. For this analysis two thresholds of significance are used to account for multiple testing: GWS (p-value < $5x10^{-8}$) and a FDR of 5%.

Table 3.2 shows the p-values for FDR 5% for results determined using the Benjamini-Hochberg procedure. See section 2.10.2 for full description of the Benjamini-Hochberg procedure. Test count corresponds to the total number of association tests performed within each population, i.e. the number SNP-probe combinations. As can be seen the FDR 5% p-values are less significant than GWS, which suggest that GWS is an overly conservative threshold.

Population	Test count	FDR 5% p-value
ASN	24,272,010	$7.64x10^{-5}$
CEU	26,803,891	$1.87x10^{-5}$
YRI	29,081,921	$1.46x10^{-5}$

Table 3.2: p-values for FDR 5% derived using Benjamini-Hochberg procedure.

3.4 Association Analysis

In order to detect *cis* eQTLs, association analysis was performed with genotype and microarray expression data of samples in the Phase II HapMap. Linear regression was performed using PLINK –assoc option. Three analyses were run: 1) no covariates, 2) sex as covariate, 3) sex and the first two population-specific eigenvectors as covariates. The analysis was performed on each of the three populations separately. This section presents the results of this association analysis. See methods section 2.5 for full information on this analysis.

3.4.1 QQ-Plots

QQ plots were generated using the p-values from the Phase II HapMap association analysis with no correction for sex or population structure. P-values were plotted against the expected null distribution (See figure 3.3). Observed p-values deviate from the null distribution (i.e. the y = x

line) only at approximately $-\log_{10}(p) = 3$ for each population. This would suggest that there is no clear evidence of inflation in association test statistics due to population structure that has not been accounted for in the association analyses. However, to explore this further, we calculated the genomic control inflation factor.

3.4.2 Genomic Control

The genomic control inflation factor was calculated for each population separately to give an indication of possible fine-scale population structure. See methods section 2.4.2 for information on Genomic Control. Genomic control inflation factors for ASN, CEU and YRI populations are shown in table 3.3 for association analyses with adjustment for: 1) no covariates, 2) sex as a covariate, and 3) sex and the first two population-specific eigenvectors as covariates. All three populations have low GC (λ) inflation factor, irrespective of the adjustment of covariates. The highest inflation was observed in the ASN population, which was reduced from 1.08 to 1.04 through inclusion of eigenvectors as covariates in the analysis, and thus accounting for the differences between CHB and JPT within the ASN population. However, for all populations there is minimal evidence of population structure from inspection of QQ plots and calculation of the genomic control inflation factor, so we chose not to adjust association analyses for population-specific eigenvectors. Sex also had no impact on association analyses, with no change in the genomic control inflation factor after including it as a covariate in the regression model. All subsequent results in this chapter are therefore based on unadjusted association analyses.

The ASN population lowest p-value (2.94×10^{-74}) is more significant than CEU (1.79×10^{-35}) and YRI lowest p-values (2.23×10^{-33}). CEU and YRI are approximately the same. This difference is likely to be because of an increase in power caused by larger population size in ASN (90) compared with CEU and YRI (60).

Population	No Covariates	Sex	Population and Sex
ASN	1.081	1.082	1.044
CEU	1.032	1.030	1.030
YRI	1.029	1.029	1.029

Table 3.3: Summary of genomic control values for Phase II HapMap analysis with different covariates added to the linear regression model

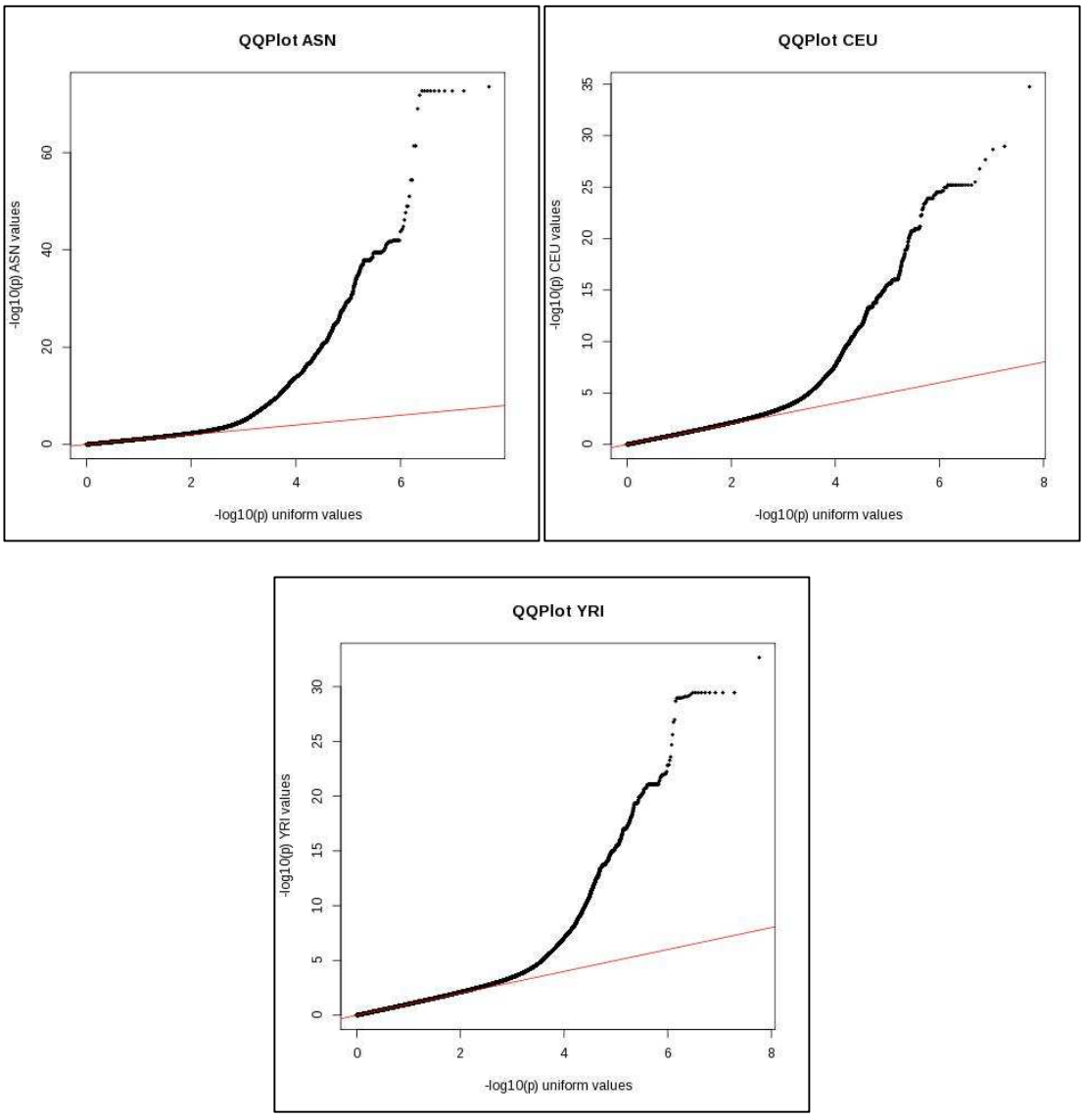


Figure 3.3: QQ plots for Phase II HapMap association analysis.

3.4.3 GWS ($p \leq 5x10^{-8}$)

Table 3.4 presents the counts of *cis* eQTL signals detected with the association analysis with no adjustment for covariates at GWS ($p \leq 5x10^{-8}$). Two different counts are presented:

1. **Probes:** This is the total number of unique probes at which a *cis* eQTL is detected.

2. **Peak SNPs:** Each *cis* eQTL can have multiple SNPs within the peak signals (i.e. SNPs in perfect LD with each other). This count is the total count of peak SNPs detected at the probes with *cis* eQTLs. This effectively gives an overview of LD block size for each of the populations.

The count of eQTL signals detected in ASN is greater than signals detected in CEU and YRI. This is likely to be due to increased power because of the larger sample size of this population.

Population	Probes	Peak SNPs
ASN	483	721
CEU	183	524
YRI	268	488

Table 3.4: Summary of *cis*-eQTLs for Phase II HapMap at GWS ($p \leq 5x10^{-8}$).

In total, there is a non-redundant set of 671 probes showing a significant association in one or more populations. A total of 203 probes with significant association are detected in 2 or more populations, and 60 probes with significant associations are detected in all three populations.

Table 3.5 shows the range of R2 values for eQTLs detected at GWS. ASN has the largest range of R2 values with the smallest lowest R2. CEU and YRI have the same value for the lowest R2.

Population	Lowest	Highest
ASN	0.288	0.9775
CEU	0.4036	0.9379
YRI	0.4036	0.9192

Table 3.5: Summary of R2 range for eQTLs detected at GWS.

Table 3.6 shows the intersection of the results detected in the association analysis at GWS. These results suggest that many *cis* eQTLs are shared, at GWS across populations, but that others may be population-specific.

Population	ASN	CEU	YRI
ASN	483	126	123
CEU	126	183	74
YRI	123	74	268

Table 3.6: Intersection of association analysis results for Phase II HapMap populations at GWS ($p \leq 5x10^{-8}$).

3.4.4 FDR ≤ 5%

Results with FDR 5% are shown in table 3.7. As for the GWS threshold, the highest number of signals detected is within the ASN population, followed by YRI and CEU, and as before is probably due to the increased power in the ASN population because of the larger sample size.

Population	Probes	Peak SNPs
ASN	2,227	2,868
CEU	758	1,464
YRI	1,052	1,592

Table 3.7: Summary of *cis*-eQTLs for Phase II HapMap with FDR 5%, determined using Benjamini-Hochberg.

In total, there is a non-redundant set of 3206 probes showing a significant association in one or more populations. A total of 643 probes with significant associations are detected in 2 or more populations and 188 probes with significant associations are detected across all three populations.

Table 3.8 shows the range of R2 values for eQTLs detected at FDR 5%. The range of R2 has increased as compared with those at GWS. ASN again has the largest range of R2 values.

Population	Lowest	Highest
ASN	0.1637	0.9775
CEU	0.2727	0.9379
YRI	0.2728	0.9192

Table 3.8: Summary of R2 range for eQTLs detected at GWS.

Table 3.9 shows the intersection of the results for probes detected in the association analysis at FDR 5%. As before, these results suggest that many *cis* eQTLs are shared, at FDR ≤ 5%, across populations, but that others may be population-specific.

Population	ASN	CEU	YRI
ASN	2227	379	406
CEU	379	758	234
YRI	406	234	1052

Table 3.9: Intersection of probes detected in the three Phase II HapMap populations at FDR 5%

3.4.5 Stranger 2007 results

The Stranger 2007 study used the same expression and genotype data analysed in this chapter. To compare results table 3.10 presents the results of the Stranger 2007 analysis at a 0.001 permutation threshold, which was reported to be equivalent to a FDR of approximately 5%.

Population	Gene Count
CEU	299
CHB	318
JPT	341
YRI	394

Table 3.10: Counts of eQTLs detected in the Stranger *et al* 2007 study at 0.001 permutation threshold.

In total, there is a non-redundant set of 831 genes showing a significant association in one or more populations. In total, 310 genes with significant associations are detected in 2 or more populations and 62 genes have significant associations across all three populations.

As can be seen, the numbers of eQTLs detected in the Stranger 2007 study are smaller than those detected at FDR 5% in this study. This could be due to differences in assessing significance between these studies. In addition the Stranger results report gene counts rather than probe counts, and the CHB and JPT populations are reported separately in this study.

The R2 of eQTL detected at the 0.001 permutation threshold ranged from 0.27 to almost 1. This range is similar to that observed in the CEU and YRI populations at FDR 5%.

3.5 Fixed effect meta-analysis

In order to combine results across the three Phase II HapMap populations, fixed effect meta-analysis has been performed using the summary statistics generated in the association analysis with no adjustment for covariates. See methods section 2.8 for more information on fixed effect meta-analysis. We implemented fixed effect meta-analysis by combining effect sizes across populations weighted by the inverse of their variance. To do this, the software GWAMA (Magi *et al* 2010) was used.

3.5.1 QQ-Plots

QQ Plots were generated with the p-values from the Phase II HapMap fixed effect meta-analysis. Observed p-values were plotted against the expected null distribution (See figure 3.4). As for the population-specific QQ plots, observed p-values differ from the null distribution only at $-\log_{10}$ p-value of 3, suggesting no clear evidence of structure between populations that has not been accounted for in the association analysis.

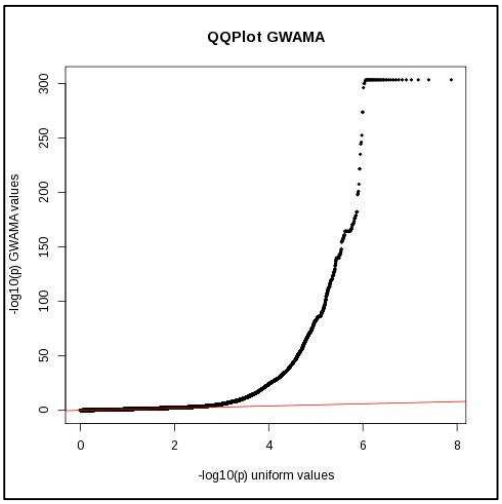


Figure 3.4: QQ plots for Phase II HapMap association analysis and fixed-effect meta-analysis.

3.5.2 Signal counts

Table 3.11 shows the results of the fixed effect meta-analysis for the Phase II HapMap data, at GWS and FDR 5% (p-value < 1.01×10^{-4}), based on SNPs that are reported in at least one

population, number of tests: 37,431,589). In addition to this, the number of populations within which the variant is reported (i.e. is polymorphic and passes QC) is also presented.

Significance Level	Min Population Count	Probes	Peak SNPs
GWS	3	1,024	1,098
	2	1,173	1,298
	1	1,320	1,539
FDR 5%	3	3,719	3,876
	2	4,920	5,219
	1	6,576	7,220

Table 3.11: Results for fixed effect meta-analysis at GWS ($p < 5x10^{-8}$).

The overlap of the results with those presented in the Stranger *et al* 2007 paper was assessed. When the results of the fixed effect meta-analysis at GWS (5E-8) from this study (1,024 probes) are compared with SNPs present in all populations at GWS in the Stranger 2007 study within RefSeq (404 probes), an overlap of 285 probes was detected.

When the probes detected by combining populations in the Stranger *et al* 2007 paper (801 probes), were compared to the meta-analysis at GWS (1,024 probes), 620 probes were detected in both datasets. As the analysis performed within these studies is different, and criteria for filtering the probes was different, some difference between eQTL counts detected should be expected.

3.5.3 Heterogeneity

Heterogeneity of effect sizes between populations was assessed using the Cochran Q-Statistic calculated with GWAMA for SNPs with peak association present in all three populations. See section 2.8.2 for full description of the Cochran's Q statistic. Table 3.12 shows the results with GWS and FDR 5% significance thresholds. For the purpose of this analysis, heterogeneous probes were those in which the peak SNP had a Cochran's Q p-value less than $1x10^{-3}$.

The heterogeneity in allelic effects is presented using the following metrics:

- **Significance Level:** Used to detect *cis* eQTLs in the fixed effect meta-analysis: (GWS or FDR 5%)
- **Min population count:** The minimum number of populations where the SNP is reported. For example, min population count of 3 implies that the SNP is reported in all three populations.
- **Peak Signals:** This is the number of peak signals across all probes, i.e. the signals within perfect LD blocks.
- **Heterogeneous Signals:** Count of the peak signals which have Cochran's Q ($p < 1x10^{-3}$).
- **Expected Signals:** The expected number of signals at Cochran's Q ($p < 1x10^{-3}$), determined by multiplying the number of signals by the Cochran's Q significance level.
- **Binomial Test p-value:** the p-value of the binomial exact test of detecting the number of heterogeneous signals at the Cochran's Q p-value.

Using a binomial test it was established that the number of heterogeneous probes was significantly greater than the number expected by chance.

Significance Level	Min population Count	Peak Signals	Heterogeneous Signals	Expected Signals	Binomial Test p-value
GWS	3	1,098	106	~1	$< 2.2x10^{-16}$
	2	1,298	79	~1.5	$< 2.2x10^{-16}$
	1	1,539	51	~1.5	$< 2.2x10^{-16}$
FDR 5%	3	3,876	166	~4	$< 2.2x10^{-16}$
	2	5,219	133	~5	$< 2.2x10^{-16}$
	1	7,220	79	~7	$< 2.2x10^{-16}$

Table 3.12: Summary of heterogeneity results with GWS ($p < 5x10^{-8}$) and FDR 5%. Heterogeneity is assessed with Cochran's Q ($P < 1x10^{-3}$).

In order to explore the source of heterogeneity in allelic effects between populations further, the heterogeneous signals were characterized using the following criteria.

Zero Effects: Do the effect sizes in any population not significantly differ from zero (i.e. signal is specific to one or more populations)?

Opposite Effects: Are any significant effect sizes in opposite directions in different populations?

Table 3.13 shows the summary for the results from the fixed-effects meta-analysis. The majority of heterogeneous results have effect sizes in one or more population that do not significantly differ from zero (58% for eQTLs at GWS and 73% for eQTLs at FDR 5%). Almost none of the results show effects in opposite directions; only one probe is detected using this criterion (GI_4507400-I, ENSG00000108064, HGNC: *TFAM*). Lack of opposite direction effects is to be expected as SNPs with opposite directions of effect will cancel each other out in fixed effect meta-analysis, and thus association would not have been detected. The remainder of the probes in the analysis correspond to situations where there is a significant association in all three populations, but the magnitude of effect is different.

Significance	Zero Effects	Opposite Effects
GWS	62 / 106 (58%)	1 / 106
FDR 5%	121 / 166 (73%)	1 / 166

Table 3.13: Summary of heterogeneity analysis.

The effect sizes of the heterogeneous signals were also plotted against each other across populations to look for general trends (Figure 3.5). Visually it can be seen that the ASN and CEU populations appear to be in greater agreement with each other than with the YRI population. This is confirmed by the Pearson correlation between the effect sizes for the three populations (ASN:CEU 0.879, ASN:YRI 0.796, CEU:YRI 0.730). For heterogeneous signals, effect sizes for YRI appear to be closer to zero than for ASN and CEU. One interpretation of this is that less of the signals are being detected in the YRI population because of smaller LD blocks in this population, and therefore poorer tagging of the true causal eSNP than in CEU and ASN populations.

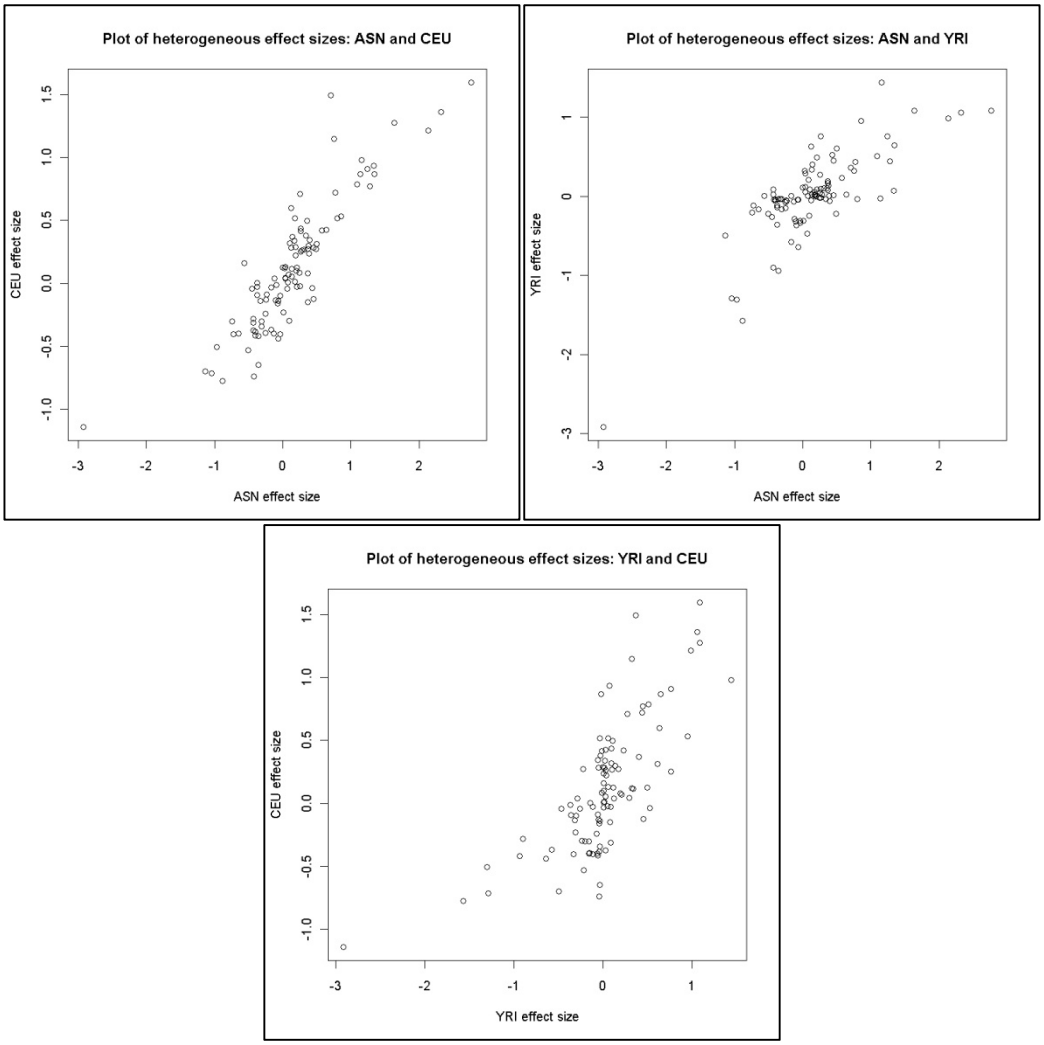


Figure 3.5: Plot of effect sizes of the three Phase II HapMap populations for SNPs with significant associations (GWS) and significant heterogeneity between effect sizes (Cochran's Q p-value ≤ $1x10^{-3}$)

3.6 Examples

In this section I present five examples of *cis* eQTLs detected in the fixed-effects meta-analysis. These examples have been selected for the following reasons:

1. *ORMDL1* (GI_31377759-S) (section 3.6.1) was selected because it has the strongest signal of association from the fixed-effects meta-analysis, and no evidence of heterogeneity in allelic effects between populations. This probe contains a 1000 Genomes SNP rs115591984, which has an allele frequency of 0.018.

2. *PDXDC2P* (GI_40217627-S) (section 3.6.2) and *USMG5* (GI_14249375-S) (section 3.6.3) were selected because they have strong signals of association from the fixed-effects meta-analysis, and the greatest evidence of heterogeneity in allelic effects between populations. The probe GI_40217627-S contains a 1000 Genomes SNP rs187431035 which has an allele frequency of 0.096. GI_14249375-S does not contain any 1000 Genomes SNPs.

3. *SNHG17* (GI_37555934-S) (section 3.6.4) and *XKR9* (GI_29735811-S) (section 3.6.5) were selected to illustrate how LD differences between populations can be used in fine mapping. The probes GI_37555934-S and GI_29735811-S both contain multiple 1000 Genomes SNPs.

3.5.4 Most significant Results

Table 3.14 presents the ten most significant results including those with evidence of heterogeneity (Cochran's Q p-value ≤ 0.05). Out of these, eight probes have evidence of heterogeneity. The probe GI_21553312-S (ENSG00000197195, *CHURC1*) is deprecated in NCBI Build 37. See Appendix table A.1 for the population-specific association analysis results.

Due to the extremely low p-values reported in this table 3.14, and the fact that eight of the ten exhibit significant heterogeneity, the validity of the signals was assessed by identifying whether

any of the probes overlap with a 1000 Genomes SNP. Seven out of the ten probes in the table contain a 1000 Genomes SNP, which suggests that these probes may be false positives. The column 1000 Genomes SNP indicates whether a SNP within the probe has been detected. The table also specifies the R2 value for the association analysis for each of the three populations.

The eight SNPs detected in this table have been annotated using ANNOVAR (Wang *et al* 2010) (see section 2.12). Six of the eight SNPs are located within genes, four of which are intronic. However none of the SNPs are located within exons.

Probe	HGNC	SNP	z-score	p-value	Cochran's Q	Q p-value	1000 Genomes SNP	R2 (ASN, CEU, YRI)	ANNOVAR Annotation
GI_21553312-S	*CHURC1*	rs4899152	38.68	$p < 9.63x10^{-321}$	9.58	$8.31x10^{-3}$	N	0.88,0.86,0.89	Intronic
GI_4504184-S	*GSTT1*	rs407257	36.83	$6.37x10^{-297}$	18.61	$9.10x10^{-5}$	N	0.93,0.78,0.35	ncRNA intronic
GI_4507820-S	*UGT2B17*	rs3100645	30.08	$1.04x10^{-198}$	79.74	$p < 2x10^{-16}$	N	0.90,0.62,0.34	Intergenic
GI_4507822-S	*UGT2B11*	rs3100645	28.84	$9.14x10^{-183}$	64.03	$1.24x10^{-14}$	Y	0.89,0.63,0.38	Intergenic
GI_21536345-A	*TCL6*	rs2296310	28.60	$7.30x10^{-180}$	31.81	$1.24x10^{-7}$	Y	0.84,0.60,0.83	ncRNA intronic
GI_39780596-S	*PNMAL1*	rs8107491	28.39	$3.51x10^{-177}$	34.07	$3.99x10^{-8}$	Y	0.89,0.33,0.32	Upstream
GI_31377831-S	*PKHD1L1*	rs2607624	28.05	$5.37x10^{-173}$	11.04	$4.01x10^{-3}$	Y	0.87,0.41,0.72	Intronic
GI_31377759-S	*ORMDL1*	rs6942	27.98	$2.94x10^{-172}$	1.72	0.42	Y	0.83,0.83,0.44	UTR3
GI_38348363-S	*MXRA7*	rs2240773	27.65	$2.91x10^{-168}$	2.22	0.33	Y	0.86,0.72,0.42	UTR3
GI_4507824-S	*UGT2B7*	rs3100645	27.10	$1.02x10^{-161}$	60.56	$7.06x10^{-14}$	Y	0.88,0.61,0.37	Intergenic

Table 3.14: Strongest signals, including heterogeneity $p \leq 0.05$. 1000 Genomes SNP specify where at least one SNP is located within the probe location. R2 is the coefficient of determination for the three populations' association analysis results.

Table 3.15 presents the ten most significant results excluding those with evidence of heterogeneity (Cochran's Q p-value ≥ 0.05). It can be seen that the probes GI_28872737-I (*SEMA4G*) and GI_28872735-A (*MRPL43*) have the same lead eSNP (rs2863095). See Appendix table A.2 for the population-specific association analysis results. In this case, it was found that eight of the ten probes contained one or more 1000 Genome SNPs which indicates that these eQTLs maybe false-positives.

Probe	HGNC	SNP	z-score	p-value	Cochran's Q	Q p-value	1000 Genomes SNP	R2 (ASN, CEU, YRI)
GI_31377759-S	*ORMDL1*	rs6942	27.98	$2.94x10^{-172}$	1.72	0.42	Y	0.83,0.83,0.44
GI_38348363-S	*MXRA7*	rs2240773	27.65	$2.91x10^{-168}$	2.22	0.33	Y	0.86,0.72,0.42
GI_18426974-S	*HLA-DQA1*	rs9272346	26.98	$2.60x10^{-160}$	0.34	0.84	Y	0.77,0.78,0.79
GI_28872737-I	*SEMA4G*	rs2863095	26.64	$2.82x10^{-156}$	0.05	0.97	N	0.75,0.74,0.82
GI_28872735-A	*MRPL43*	rs2863095	25.21	$4.03x10^{-140}$	1.70	0.43	N	0.74,0.65,0.82
GI_20070185-S	*PTER*	rs1055340	-24.33	$1.01x10^{-130}$	3.22	0.20	Y	0.77,0.58,0.78
GI_15082255-I	*LCMT1*	rs277886	24.11	$2.06x10^{-128}$	4.98	0.08	Y	0.81,0.66,0.61
GI_22265329-I	*NUDT2*	rs10972063	-23.29	$6.79x10^{-120}$	4.39	0.11	Y	0.84,0.46,0.21
GI_42659736-S	*XRRA1*	rs2304683	-20.53	$1.38x10^{-93}$	1.28	0.53	Y	0.72,0.69,0.51
GI_29735811-S	*XKR9*	rs6998786	20.02	$4.23x10^{-89}$	3.00	0.22	Y	0.65,0.67,0.68

Table 3.15: Strongest signals, excluding heterogeneity Cochran's Q $p \geq 0.05$. 1000 Genomes SNP specify where at least one SNP is located within the probe location. R2 is the coefficient of determination for the three populations' association analysis results.

Table 3.16 gives an overview of the ten most heterogeneous signals of the fixed-effect meta-analysis. Within these signals, there is one probe where opposite effect sizes are detected, namely is GI_4507400-I (ENSG00000108064, HGNC: *TFAM*). Note that probe GI_33859747-S (ENSG00000175487, HGNC: *LOC65996*) is deprecated in NCBI Build 37. Also, probes GI_38176290-I and GI_38176291-A are for the same gene (ENSG00000105971, HGNC: *CAV2*). See Appendix table A.3 for population-specific association analysis results. In this case, one or more 1000 Genomes SNPs were detected within five of the probes, indicating that these probes maybe false-positives.

Probe	HGNC	SNP	z-score	p-value	Cochran's Q	Q p-value	1000 Genomes SNP	R2 (ASN, CEU, YRI)
GI_40217627-S	*PDXDC2P*	rs6499292	20.38	$2.87x10^{-92}$	386.64	$p < 2x10^{-16}$	Y	0.86,0.79,0.01
GI_14249375-S	*USMG5*	rs7831	22.59	$6.85x10^{-113}$	170.66	$p < 2x10^{-16}$	N	0.88,0.37,0.01
GI_38176290-I	*CAV2*	rs12670840	8.28	$1.28x10^{-16}$	127.15	$p < 2x10^{-16}$	N	0.68,0.14,0.001
GI_38176291-A	*CAV2*	rs12670840	9.41	$5.26x10^{-21}$	125.87	$p < 2x10^{-16}$	Y	0.70,0.15,0.01
GI_4507400-I (*)	*TFAM*	rs10826176	-6.55	$6.06x10^{-11}$	108.59	$p < 2x10^{-16}$	Y	0.62,0.08,0.001
GI_4507820-S	*UGT2B17*	rs3100645	30.08	$1.04x10^{-198}$	79.74	$p < 2x10^{-16}$	N	0.90,0.62,0.34
GI_33859747-S	*LOC65996*	rs7249714	-18.56	$7.19x10^{-77}$	75.60	$p < 2x10^{-16}$	N	0.68,0.79,0.02
GI_16306561-S	*RPL37A*	rs284565	-9.81	$1.11x10^{-22}$	72.78	$1.11x10^{-16}$	Y	0.60,0.03,0.37
GI_5729867-S	*HERC2*	rs12593929	-6.36	$2.14x10^{-10}$	68.25	$1.55x10^{-15}$	--	0.02,0.002,0.65
GI_4507822-S	*UGT2B11*	rs3100645	28.84	$9.14x10^{-183}$	64.03	$1.24x10^{-14}$	Y	0.89,0.63,0.38

Table 3.16: Summary of the top 10 heterogeneous results at GWS; (*) GI_4507400-I has an effect size with opposite direction. 1000 Genomes SNP specify where at least one SNP is located within the probe location. R2 is the coefficient of determination for the three populations' association analysis results.

3.6.1 GI_31377759-S, (ENSG00000128699, ORMDL1).

The eQTL for the probe GI_31377759-S with eSNP rs6942 has been selected as the first example. It has been selected because it has the strongest signal of association in the fixed effect meta-analysis (p-value=2.94×10^{-172}), whist also not having any evidence of heterogeneity in allelic effects between populations (Cochran's Q p-value= 0.42).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000128699 and the HGNC symbol *ORMDL1*. The full name of the gene is ORMDL Sphingolipid Biosynthesis Regulator 1, and the gene's start site is chr2:190343295. The eSNP rs6942 position is chr2:190344656, and is in the 3' UTR of *ORMDL1.* *ORMDL1* is a negative regulator of sphingolipid synthesis.

Figure 3.6 is a forest plot for the results of association analysis and fixed effect meta-analysis for the probe GI_31377759-S with the eSNP rs6942. YRI has a larger standard error than ASN and CEU whilst ASN and CEU have similar effect sizes and standard errors; however the p-value for ASN is more significant, which is likely due to the larger sample size for that population.

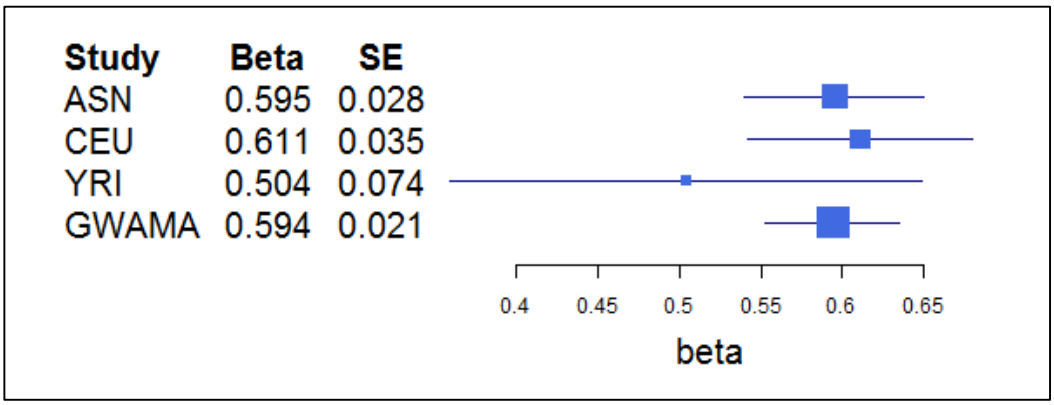


Figure 3.6: Forest plot of association and fixed effect meta-analyses for probe GI_31377759-S with phase II HapMap eSNP rs6942.

Table 3.17 shows p-values and allele frequencies for rs6942 with GI_31377759-S. Allele frequencies range from 0.26 to 0.87. All three populations are significant at GWS ($p \leq 5 \times 10^{-8}$). Both ASN and YRI have the same peak SNP as the fixed effect peak SNP.

Population	Minor Allele (C/G)	Allele Frequencies (Allele C)	Beta (Allele C)	SE	p-value	Peak SNP
ASN	C	0.26	0.595	0.028	$6.53x10^{-36}$	Y
CEU	C	0.26	0.611	0.035	$1.26x10^{-24}$	N
YRI	G	0.87	0.504	0.074	$5.99x10^{-9}$	Y
Fixed Effect Meta-analysis	--	--	0.594	0.032	$2.94x10^{-172}$	Y

Table 3.17: Table of allele frequencies and p-values for phase II HapMap SNP rs6942 with probe GI_31377759-S.

The following pages show signal plots for the results of the association analysis (figure 3.7) and fixed effect meta-analysis (figure 3.8) for the probe GI_31377759-S (*ORMDL1*). The following observations can be made. First, all three populations have a peak signal within an LD block. These LD blocks are broadly similar across the three populations. In YRI all variants in the LD block have the same significance of association, but are not as significant as for the two other populations. Consequently, YRI does not contribute to the peak signal from the fixed-effects meta-analysis as much as the other populations. Second, the peak SNP for the ASN population is also the peak SNP in the fixed effect meta-analysis, presumably because of the larger sample size for this population. The ASN and CEU peak SNPs are at different locations, but both within the same LD block, which indicates that the peak SNP is somewhere within in. However, the exact location cannot be determined from the data available. From the fixed effect meta-analysis signal plots, the peak eSNP is still rs6942, even when including variants that are not reported in all populations, indicating that the peak SNP is the optimum from this dataset. The LD block covers a region of approximately 200 KB (well beyond ORMDL1) and encompasses in addition the genes ASNSD1, ANKAR, OSGEPL1 and PMS1.

Taken together, these data would suggest that there is a shared underlying causal eSNP across populations for this probe, but because different peak SNPs are identified in different populations, that the causal eSNP has not been genotyped in Phase II HapMap.

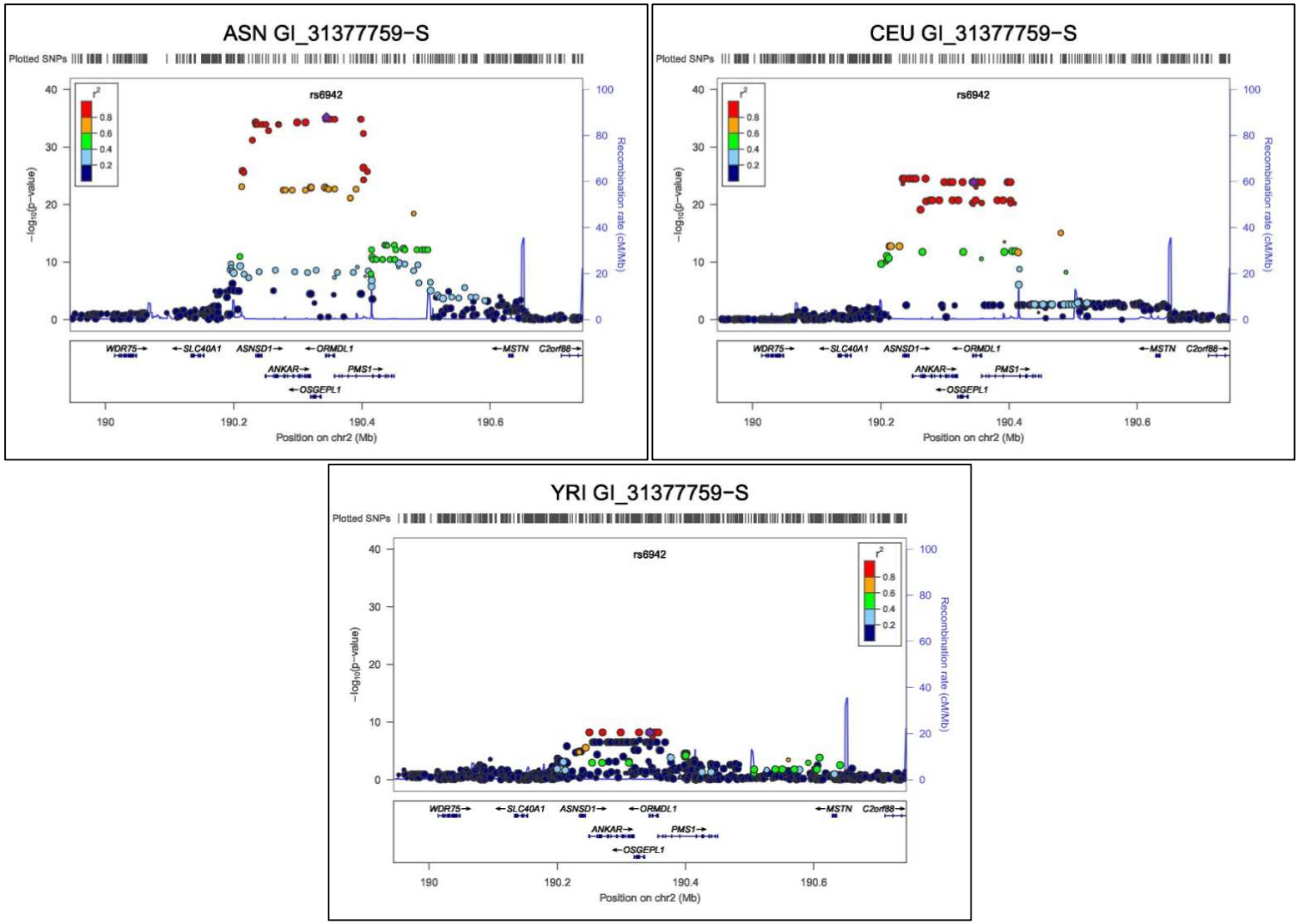


Figure 3.7: Signal plots for association analysis with Phase II HapMap SNPs for probe GI_31377759-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data June 2010).

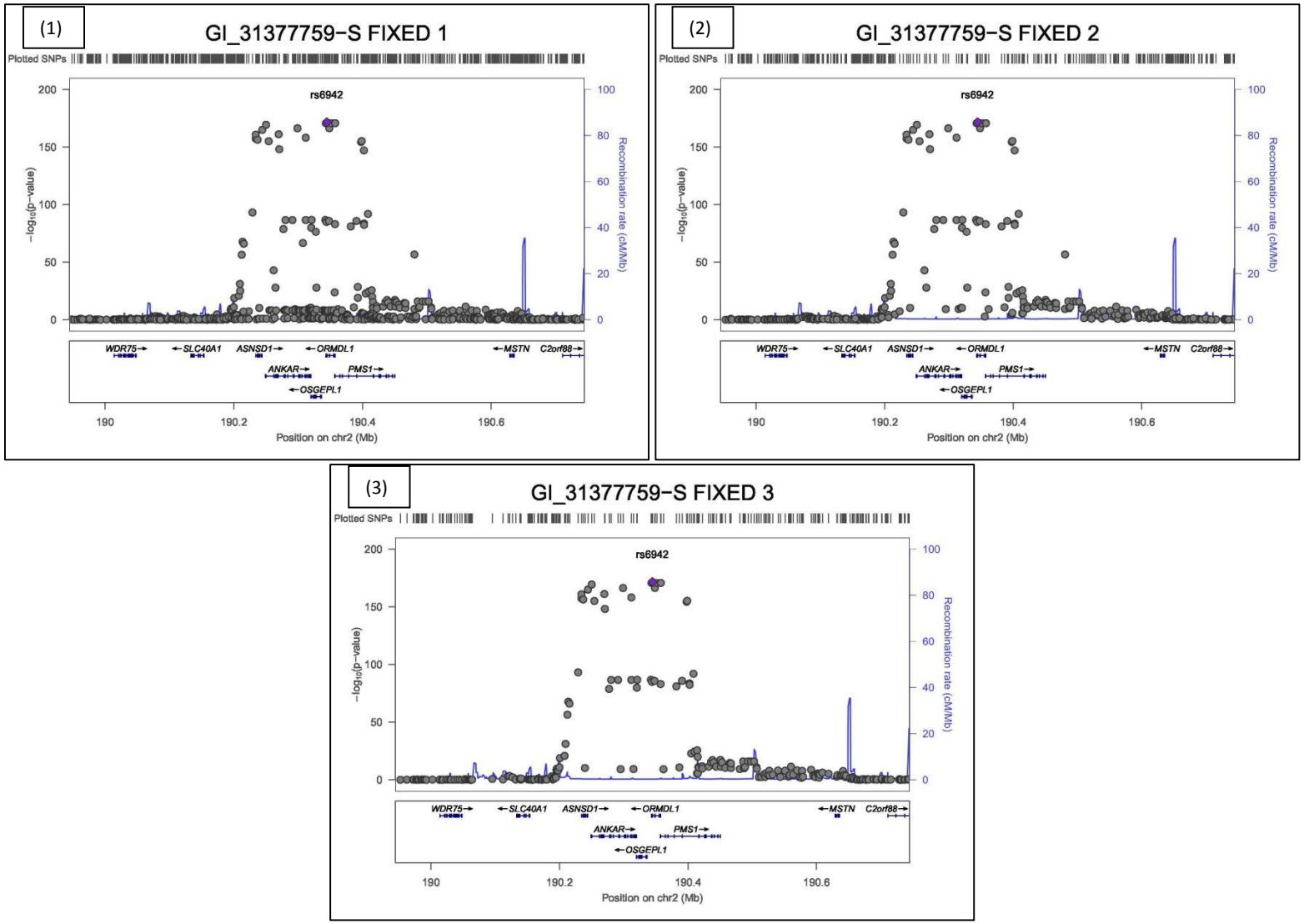


Figure 3.8: Signal plot of fixed effect meta-analysis for the probe GI_31377759-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) two or more SNPs not missing, (3) No missing SNPs: 3 SNPs present.

3.6.2 GI_40217627-S (ENSG00000196696, PDXDC2P).

The eQTL for probe GI_40217627-S has been selected as an example because in the fixed effect meta-analysis it has the most significant heterogeneity for its peak signal out of all the probes with the SNP rs6499292. (p-value=2.87×10^{-92}, Cochran's Q p-value < 2×10^{-16})

Gene name and location

The probe's gene has the Ensembl ID ENSG00000196696 and the HGNC symbol *PDXDC2P*. The full name of the gene is Pyridoxal-Dependent Decarboxylase Domain Containing 2, Pseudogene, and the gene's start position is chr16:68567713. The eSNP rs6499292 position is chr16:68661559, and is intergenic between *PDXDC2P* (4,207 base pairs) and *PDPR* (43,471 base pairs). *PDXDC2P* is a pseudogene. Pseudogenes are genes that have lost their protein coding capability. In this case, it seems that transcription can occur on this pseudo-gene, so the loss of protein coding capability must have occurred downstream of transcription.

Figure 3.9 shows a forest plot for GI_40217627-S with the eSNP rs6499292. As can be seen, the YRI effect size does not differ significantly from zero (p=0.529), whilst ASN and CEU both have significant association signals with effects in the same direction.

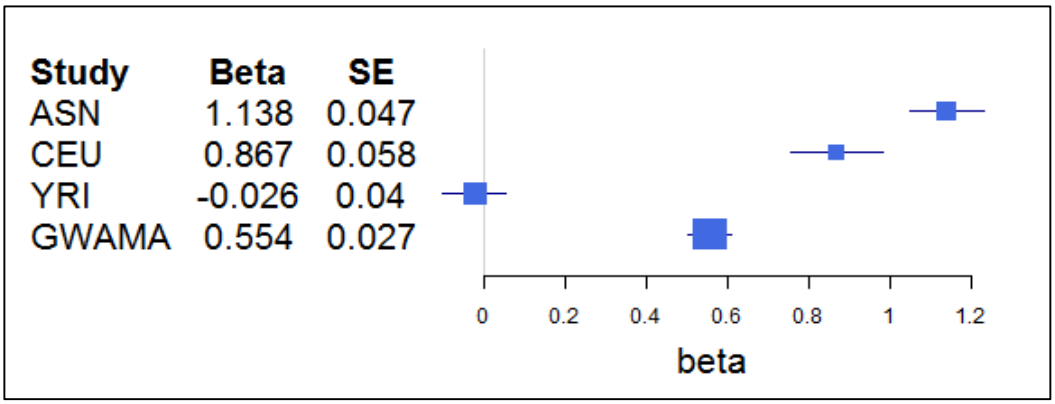


Figure 3.9: Forest plot of association and fixed effect meta-analyses for probe GI_40217627-S with phase II HapMap SNP rs6499292.

Table 3.18 shows the p-value, allele frequencies and peak SNPs for GI_40217627-S at the eSNP rs6499292. Allele frequencies range from 0.18 to 0.66. Both ASN and CEU populations have effect

sizes that achieve GWS. The ASN population peak SNP matches that of the fixed effect meta-analysis peak SNP.

Population	Minor Allele (A/G)	Allele Frequencies (Allele G)	Beta (Allele G)	SE	p-value	Peak SNP
ASN	G	0.18	1.138	0.047	2.74×10^{-40}	Y
CEU	A	0.66	0.867	0.058	3.16×10^{-21}	N
YRI	A	0.61	-0.026	0.04	0.529	N
Fixed Effect Meta-analysis	--	--	0.554	0.027	2.87×10^{-92}	Y

Table 3.18: Table of allele frequencies and p-values for phase II HapMap SNP rs6499292 with probe GI_40217627-S.

The following pages show signal plots for the results of the association analysis (figure 3.10) and fixed effect meta-analysis (figure 3.11) for the probe GI_40217627-S. The following observations can be made. First, from the YRI association signal plot, there are no visible signals detected anywhere in the region. In the ASN signal plot, it can be seen that the eSNP rs6499292 is the peak SNP. In the CEU signal plot rs6499292 is not the peak SNP, but is within an LD block with the peak SNP for this population. Second, from the fixed effect meta-analysis signal plots, the peak eSNP is still rs6499292, even when including variants that are not reported in all populations, indicating that the observed heterogeneity is not due to SNPs failing QC or monomorphic in some populations. Taken together, these results suggest that the heterogeneity in allelic effects between populations is likely due to two factors: (i) there is no association signal in the YRI population, which could indicate that the effect is not present in this population, or that there is not a good tag for the causal eSNP in this population; (ii) there is a difference in the magnitude of the allelic effect between ASN and CEU, possibly due to different LD between the eSNP rs6499292 and the causal variant in these two populations.

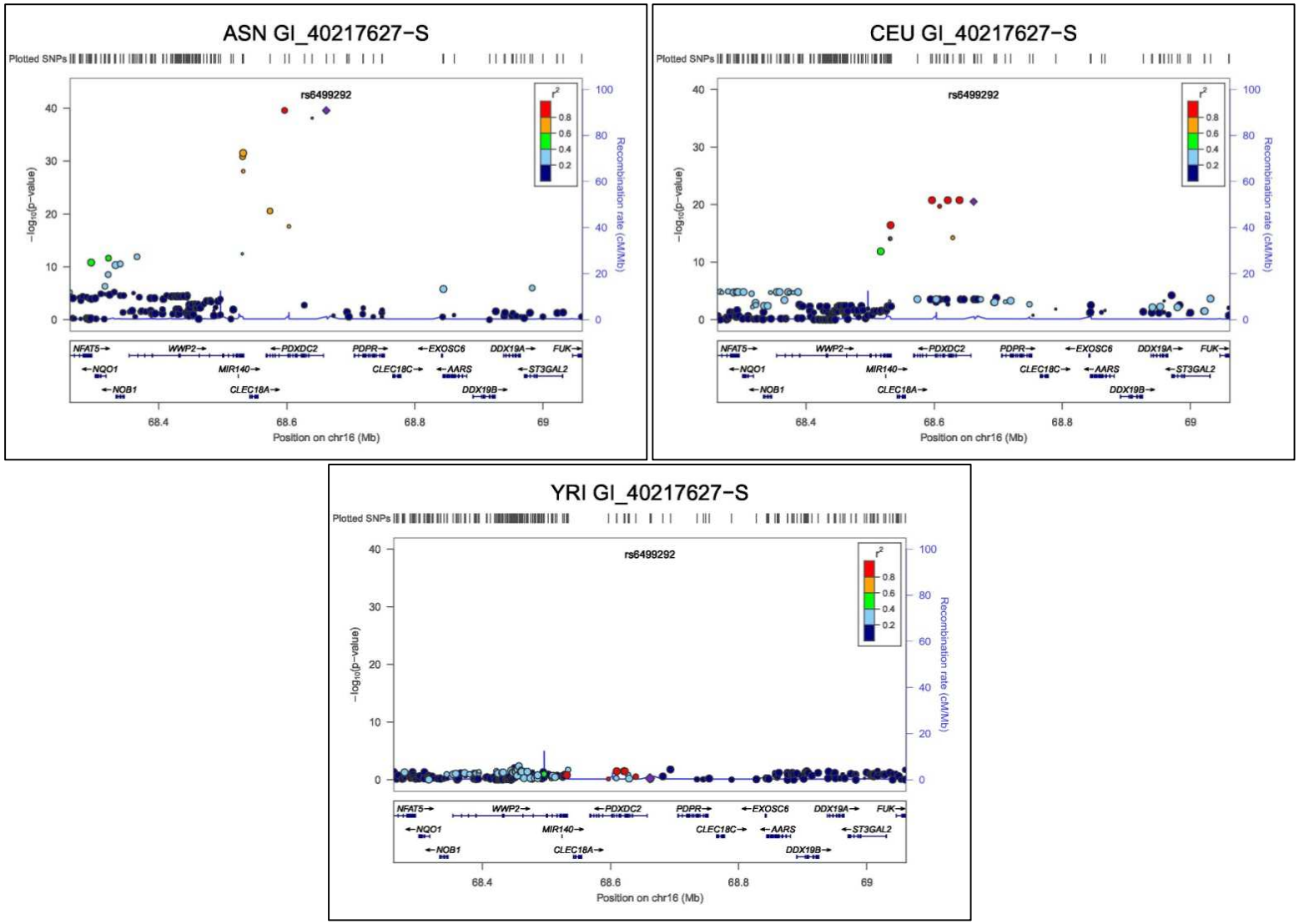


Figure 3.10: Signal plots for Phase II HapMap SNPs for probe GI_40217627-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data June 2010).

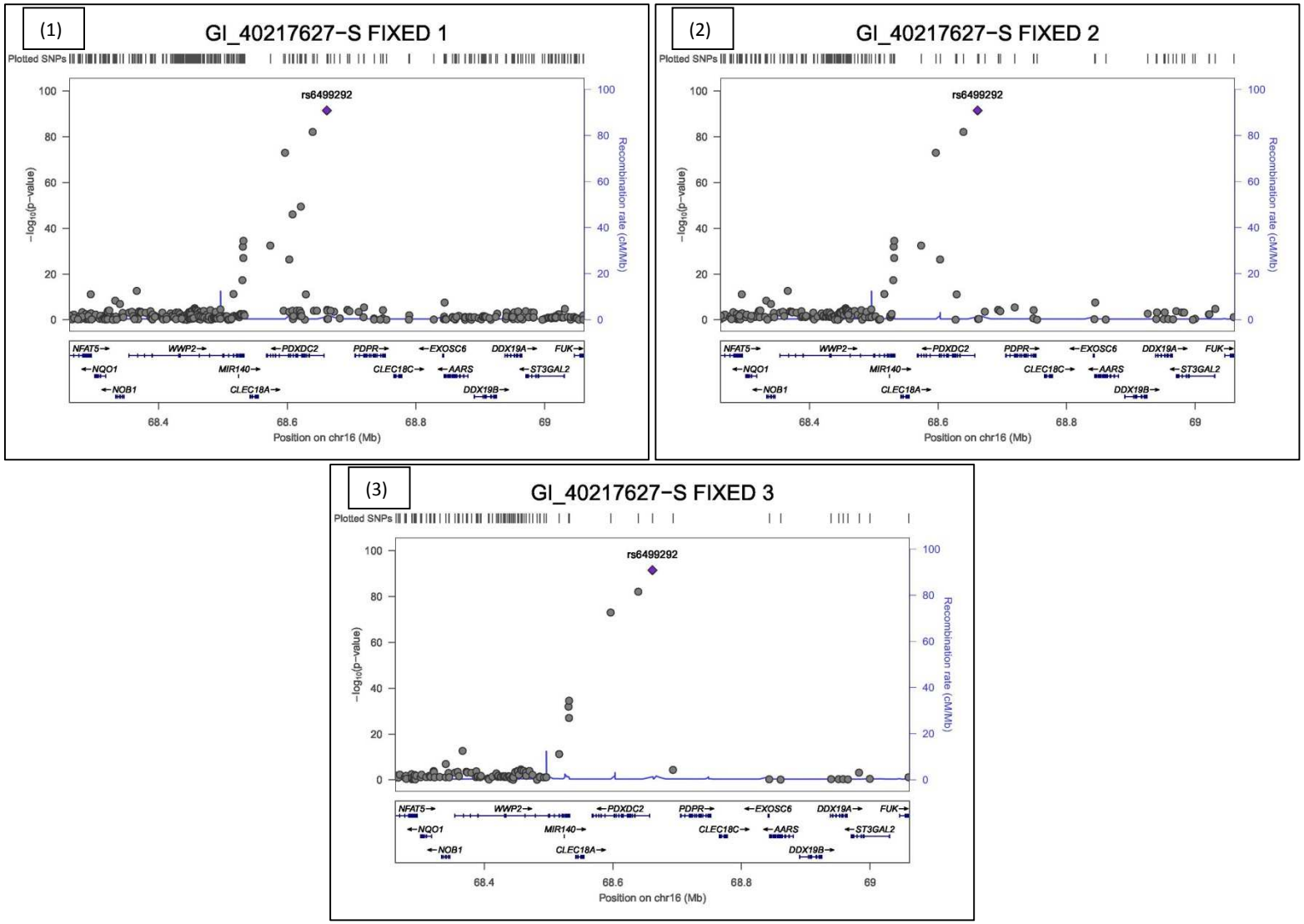


Figure 3.11: Signal plot of fixed effect meta-analysis for the probe GI_40217627-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) two or more SNPs not missing, (3) No missing SNPs: 3 SNPs present.

3.6.3 GI_14249375-S (ENSG00000173915, USMG5)

The probe GI_14249375-S has been selected as an example as it has the second most significant heterogeneity for the peak eSNP detected in the fixed effect meta-analysis for the SNP rs7831. (p-value = 6.85×10^{-113}, Cochran's Q p-value < 2×10^{-16}).

Gene name and location

The probe's gene has the Ensembl identifer ENSG00000173915 and the HGNC symbol *USMG5*. The full name of the gene is Up-Regulated During Skeletal Muscle Growth 5, and the gene's start position is chr10:105138788. The SNP rs7831 position is chr10:105195292, and maps within an exon of the gene *PDCD11*. The functional consequence of this SNP is missense at Amino Acid 1871, Alanine[A] -> Aspartic Acid[D]. *USMG5* plays a role in the maintaining the Adenosine triphosphate (ATP) synthase population in mitochondria. ATP synthase catalyses the synthesis of ATP from ADP. *PDCD11* full name is Programmed Cell Death 11.

Figure 3.12 shows a forest plot for association and fixed effect meta-analysis for probe GI_14249375-S with eSNP rs7831. As can be seen, the effect in the YRI population does not significantly differ from zero (p-value=0.05).

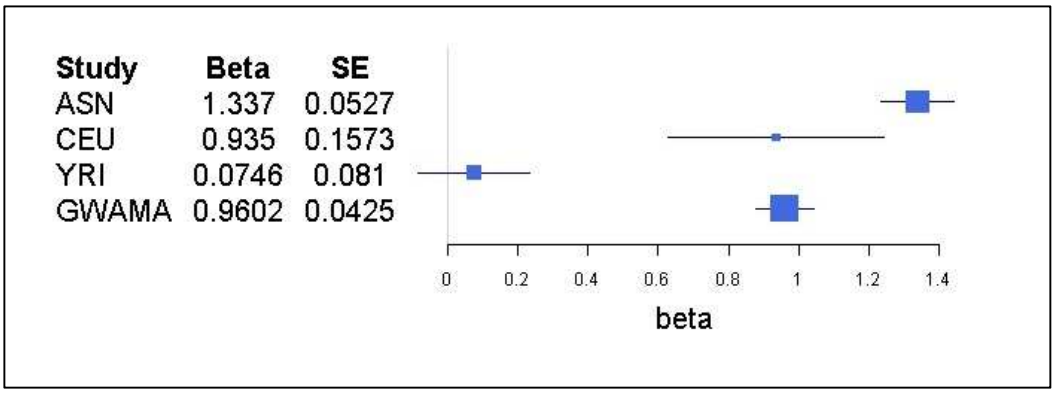


Figure 3.12: Forest plot of association and fixed effect meta-analyses for probe GI_14249375-S with phase II HapMap SNP rs7831.

Table 3.19 is a table of p-values, allele frequencies and peak SNP for probe GI_14249375-S with eSNP rs7831. In the ASN population, the SNP achieves GWS, whilst in the CEU population, it is

approaching GWS. Allele frequencies range from 0.24 – 0.39. None of the peak SNPs for the association analysis results is the same as that for the fixed effect meta-analysis on the basis of SNPs reported in all three populations. However, the peak signal in the fixed effect meta-analysis changes if we include SNPs that are not reported in all three populations. For this analysis, the peak SNP is rs7911488 with p-value 7.90×10^{-236}, which is not reported in the YRI population because it fails QC with a MAF of 0.017.

Population	Minor Allele (C/A)	Allele Frequencies (Allele C)	Beta (Allele C)	SE	p-value	Peak SNP
ASN	C	0.31	1.3370	0.0527	5.29×10^{-42}	N
CEU	C	0.39	0.9350	0.1573	1.69×10^{-7}	N
YRI	C	0.24	0.0746	0.0810	0.361	N
Fixed Effect Meta-analysis	--	--	0.9602	0.0424	6.85×10^{-113}	Y

Table 3.19: Table of allele frequencies and p-values for phase II HapMap eSNP rs7831 with probe GI_14249375-S.

The following pages show signal plots for the results of the association analysis (figure 3.13) and fixed effect meta-analysis (figure 3.14) for the probe GI_14249375-S. From the YRI signal plot, it can be observed that there is no significant association for the probe at this locus. From the results of the fixed effect meta-analysis, if the variants not reported in at least one population are included, a more significant SNP is observed. These data suggests that the heterogeneity in allelic effect sizes between populations is likely due to the lack of association signal in the YRI population, and because variants with strong signals of association are not reported across all three populations.

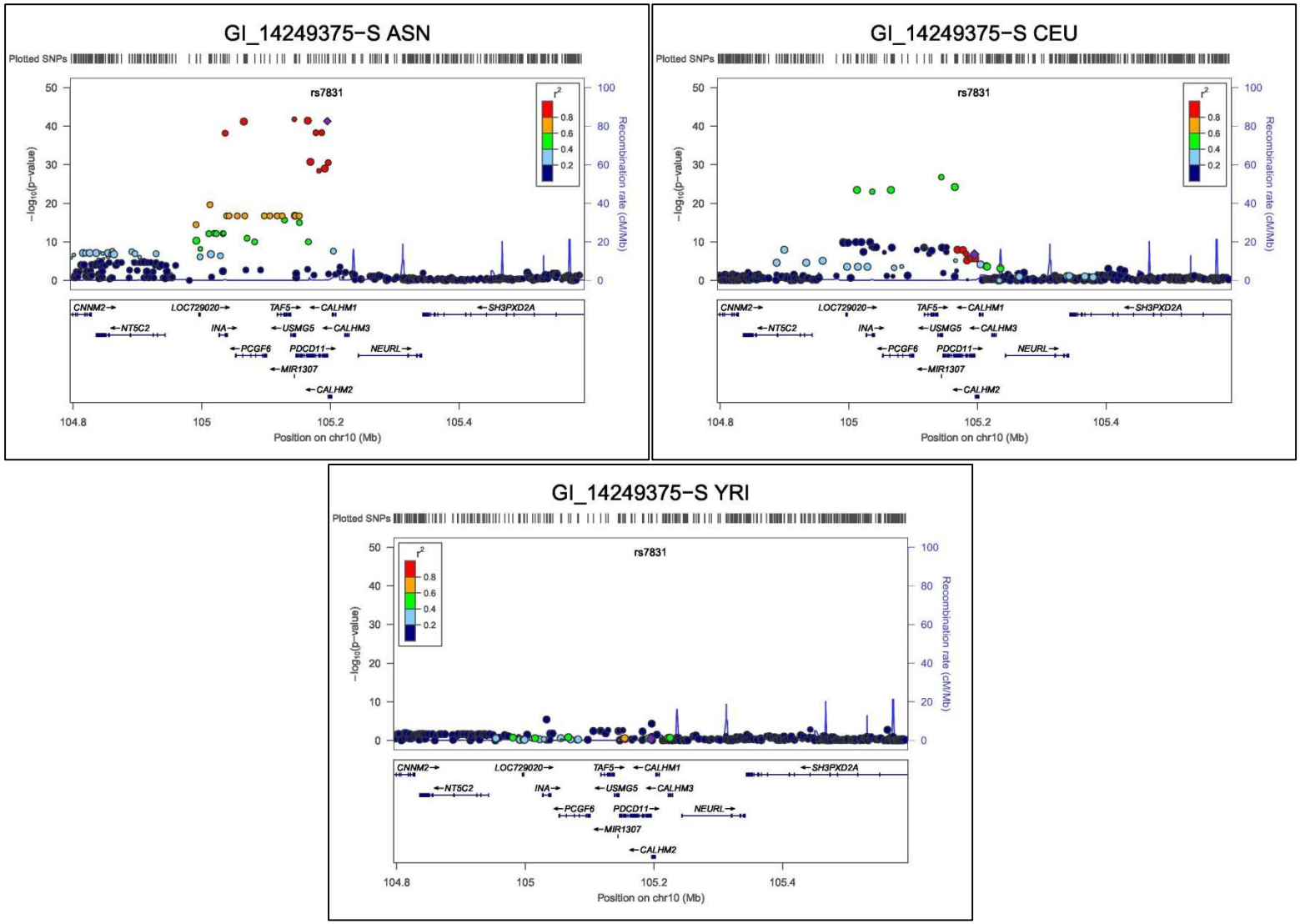


Figure 3.13: Signal plots for Phase II HapMap SNPs for probe GI_14249375-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data June 2010).

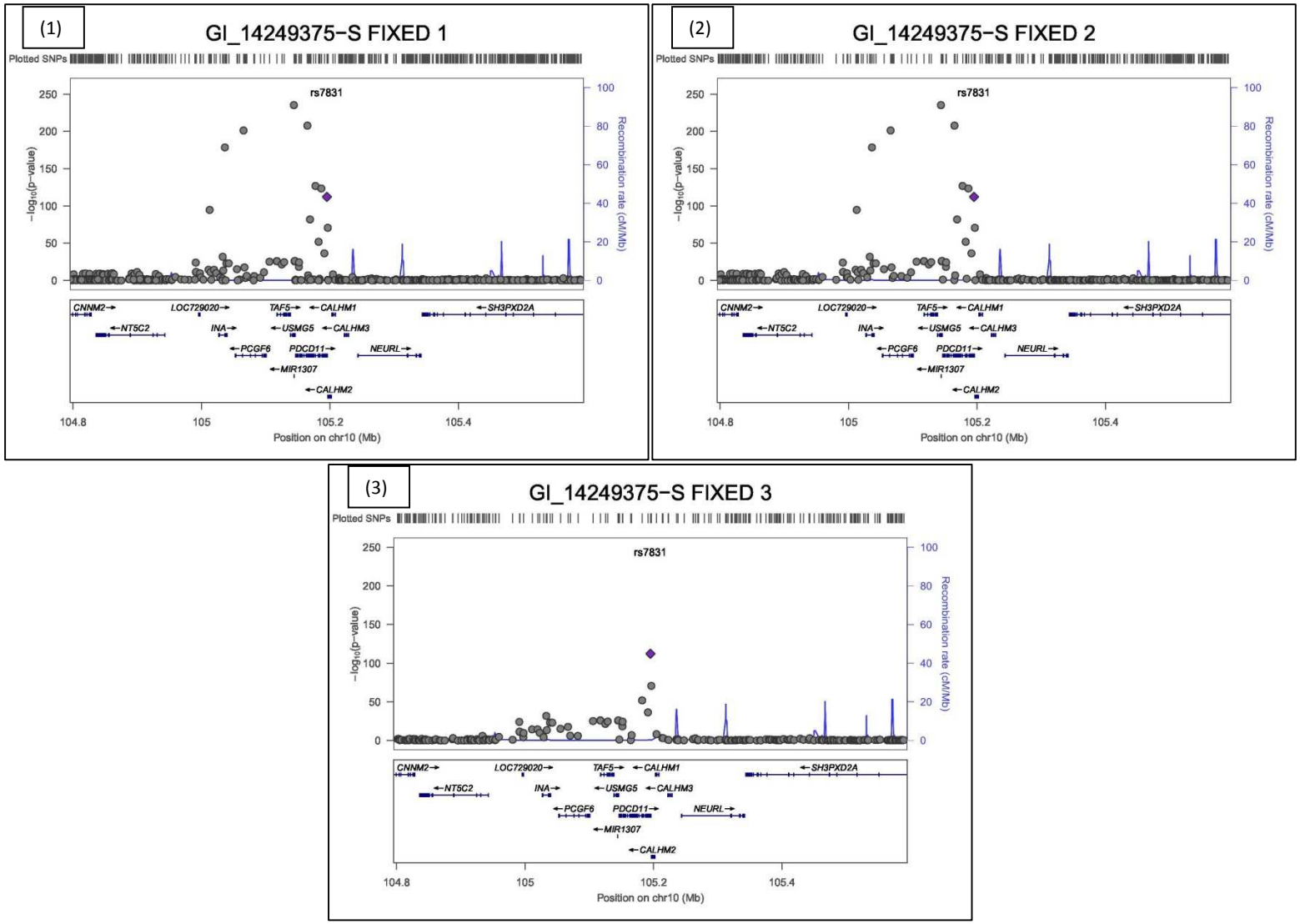


Figure 3.14: Signal plot of fixed effect meta-analysis for the probe GI_14249375-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) two or more SNPs not missing, (3) No missing SNPs: 3 SNPs present.

3.6.3 GI_37555934-S (ENSG00000196756, SNHG17).

This probe has been selected as an example of how LD differences in ancestry groups can be used to facilitate fine mapping. In the fixed effect meta-analysis, the probe GI_37555934-S has the *cis* eQTL eSNP (rs788344), which demonstrates highly significant association (p-value=1.38×10^{-76}), but no evidence of heterogeneity in allelic effects between populations (Cochran's Q p-value= 0.10).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000196756 and the HGNC symbol *SNHG17*. The full name of the gene is Small Nucleolar RNA Host Gene 17, and the gene's start position is chr20:36482658. The SNP rs788344 position is chr20:36492763, and maps within an intron of *SNHG17* (ncRNA Intronic). *SNHG17* is an RNA gene, and is part of the *processed transcript* RNA class. This RNA class is a transcript for which no open reading frame has been identified and for which no other function has been determined.

Figure 3.15 is a forest plot for the results of the association analysis and fixed effect meta-analysis for probe GI_37555934-S with eSNP rs788344. As can be seen, the effect sizes for all three populations differ significantly from zero.

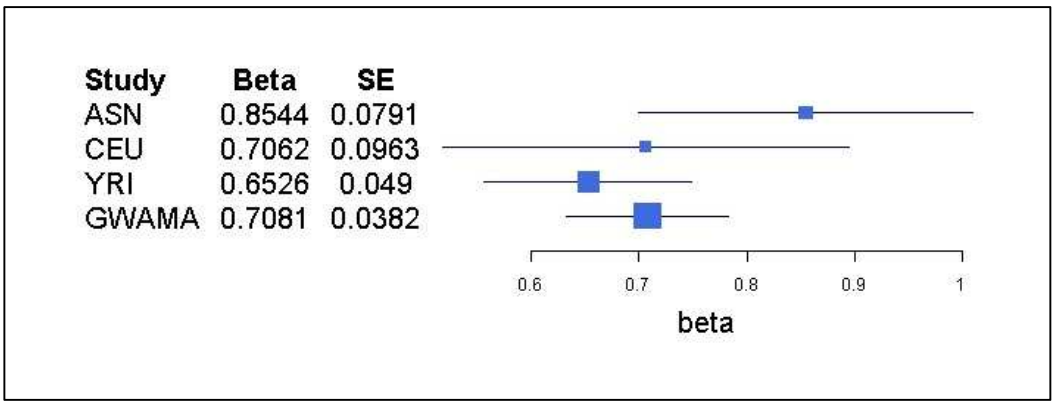


Figure 3.15: Forest plot of association and fixed effect meta-analyses for probe GI_37555934 with phase II HapMap SNP rs788344.

Table 3.20 shows the p-values, allele frequencies and peak SNPs for GI_37555934 at eSNP rs788344. Allele frequencies range from 0.09 to 0.48. The association signals in all three populations achieve GWS. A smaller standard error is detected in YRI, which is likely to be due to the larger allele frequency in this population. None of the population-specific peak SNPs is the same as that for the fixed effect meta-analysis.

Population	Minor Allele (C/T)	Allele Frequencies (Allele C)	Beta (Allele C)	SE	p-value	Peak SNP
ASN	C	0.12	0.8544	0.0791	8.28×10^{-18}	N
CEU	C	0.09	0.7062	0.0963	8.14×10^{-10}	N
YRI	C	0.48	0.6526	0.0490	2.64×10^{-19}	N
Fixed effect Meta-analysis	--	--	0.7081	0.0382	1.38×10^{-76}	Y

Table 3.20: Table of allele frequencies and p-values for phase II HapMap SNP rs788344 with probe GI_37555934-S.

The following pages show signal plots for association analysis (figure 3.16) and fixed effect meta-analysis (figure 3.17) for the probe GI_14249375-S. The following observations can be made. First, there are visual differences in LD structure between the three populations. The ASN population has the largest LD block, followed by CEU and YRI. In the fixed effect meta-analysis, rs788344 is still the peak SNP when variants not reported in all three populations are included. Second, the peak SNPs from the fixed effect meta-analysis corresponds to the overlap of variants in the LD block in the three populations. This is an example where the different LD structure between the populations can be used to fine map the causal variant.

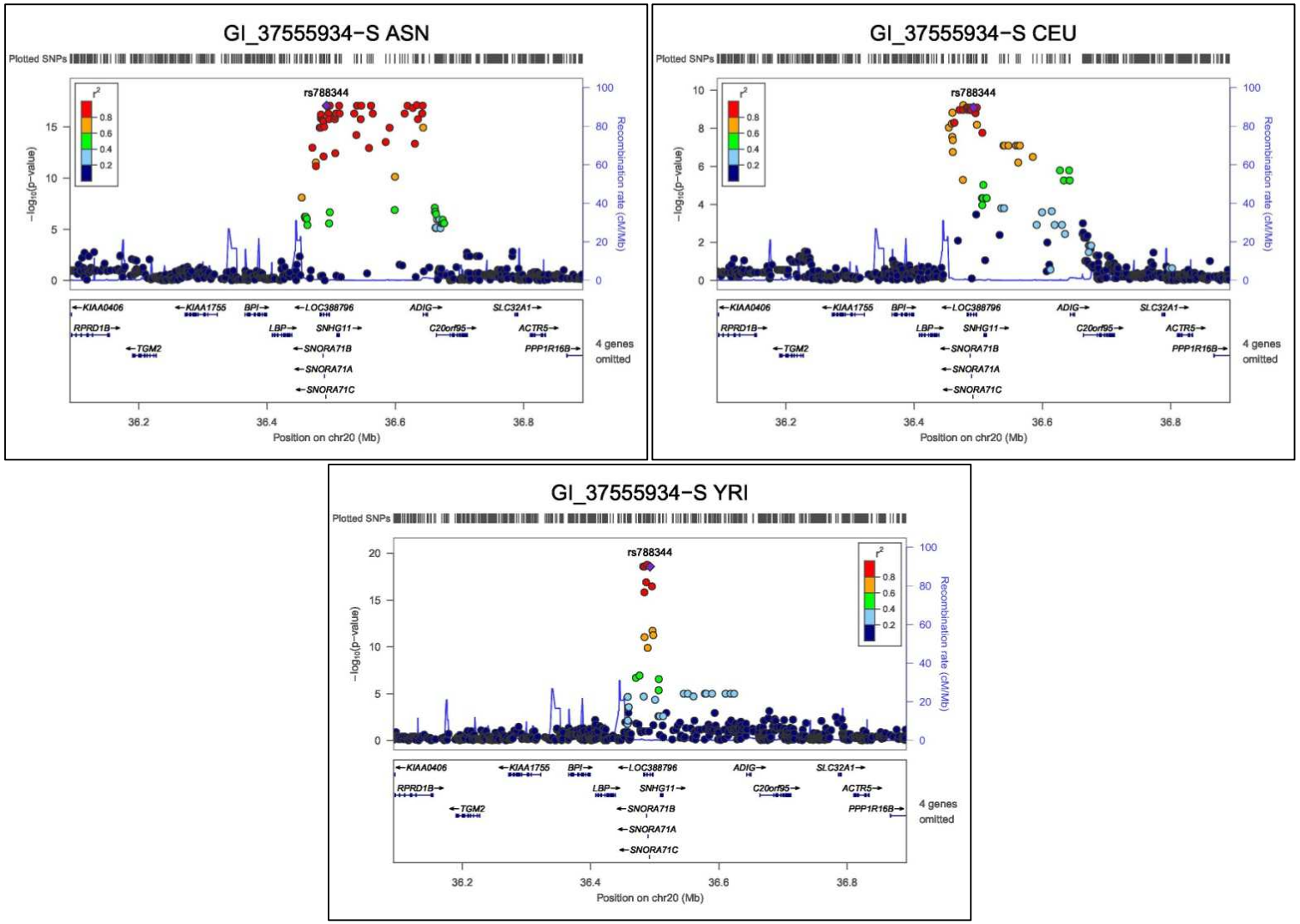


Figure 3.16: Signal plots association analysis for probe GI_37555934-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data June 2010).

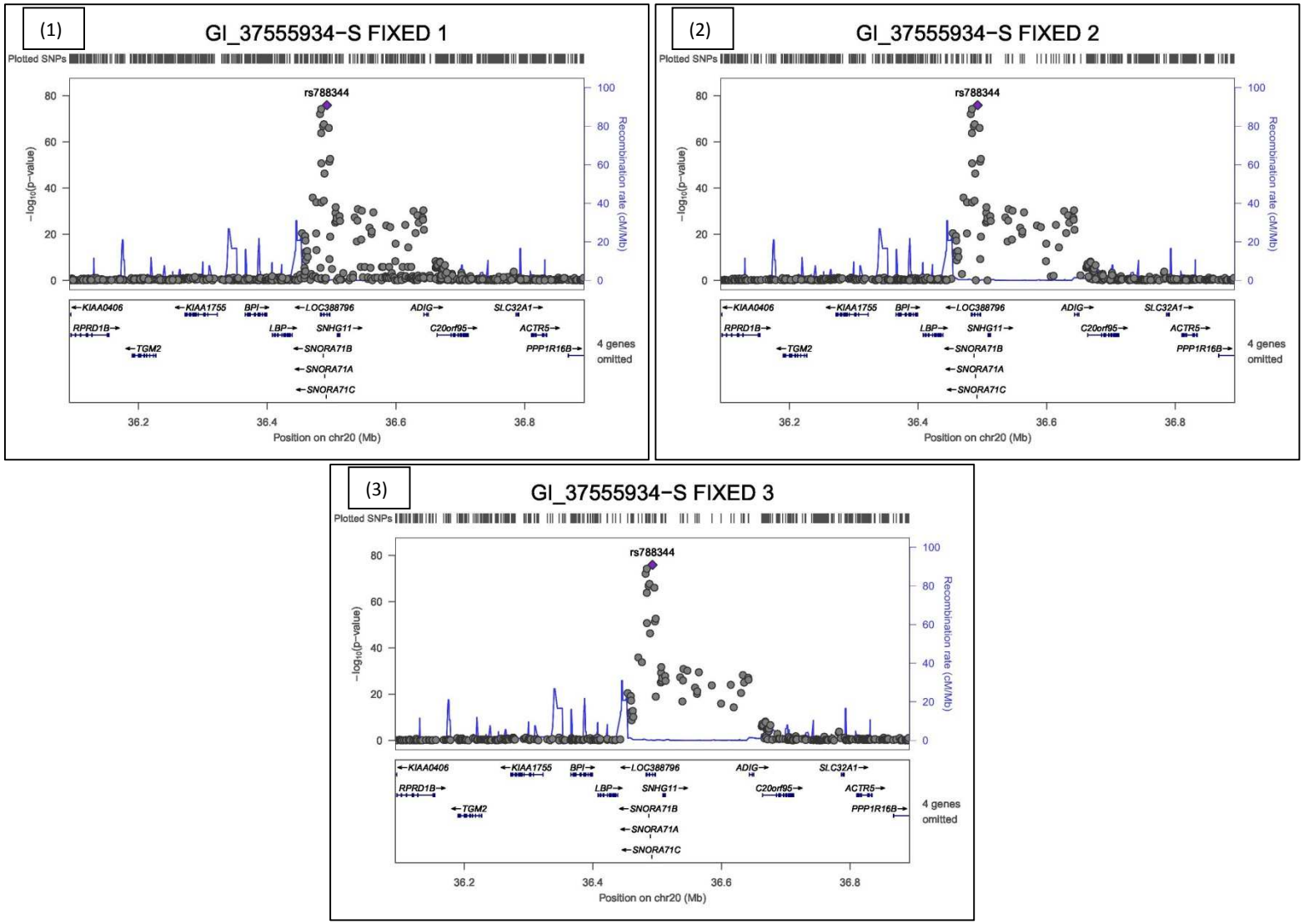


Figure 3.17: Signal plot of fixed effect meta-analysis for the probe GI_37555934-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) two or more SNPs not missing, (3) No missing SNPs: 3 SNPs present.

3.6.4 GI_29735811-S (ENSG00000221947, XKR9).

This probe has been selected as an example of how LD differences between populations can be used to facilitate fine mapping. In the fixed effect meta-analysis, the probe GI_29735811-S has the peak *cis* eQTL eSNP rs6998786, which has a strong signal of association (p-value = 4.23×10^{-89}), but no evidence of heterogeneity in allelic effects between populations (Cochran's Q p-value= 0.22).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000221947 and the HGNC symbol *XKR9*. The full name of the gene is XK, Kell Blood Group Complex Subunit-Related Family, Member 9, and the gene's start position is chr8: 71755848. The SNP rs6998786 position is chr8:71857184, and is intergenic, mapping 46,453 base pairs from the gene *XKR9*. The function of *XKR9* is unknown.

Figure 3.18 presents a forest plot for the results of the association analysis and fixed effect meta-analysis for the probe GI_29735811-S with the eSNP rs6998786. As indicated by the low level of heterogeneity, all three populations have similar effect sizes.

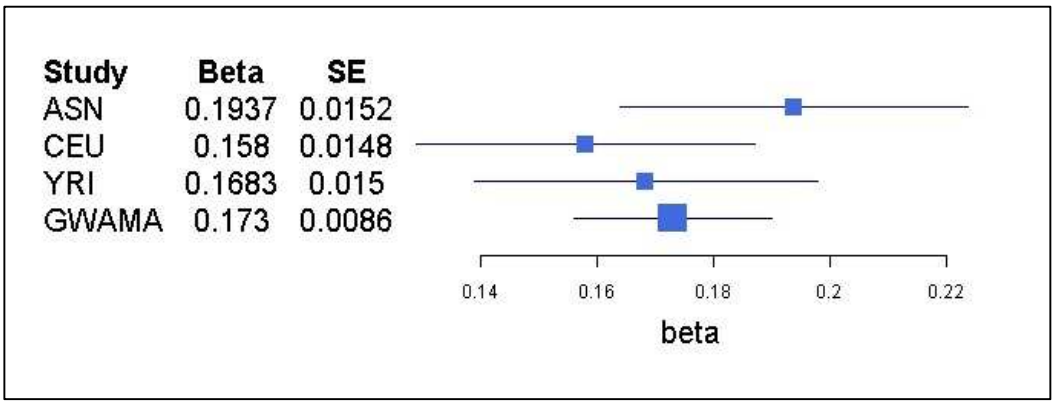


Figure 3.21: Forest plot of association and fixed effect meta-analyses for probe GI_29735811-S with phase II HapMap SNP rs6998786.

Table 3.21 presents p-values and allele frequencies for the probe GI_29735811-S with the eSNP rs6998786. Allele frequencies range from 0.11 to 0.53. The association signals in all three

populations achieve GWS. None of the population-specific peak SNPs matches that of the fixed effect meta-analysis.

Population	Minor Allele (G/A)	Allele Frequencies (Allele G)	Beta (Allele G)	SE	p-value	Peak SNP
ASN	G	0.11	0.1937	0.0152	$8.74x10^{-22}$	N
CEU	A	0.53	0.1580	0.0148	$3.72x10^{-15}$	N
YRI	G	0.15	0.1683	0.0150	$4.61x10^{-16}$	N
Fixed effect Meta-analysis	--	--	0.1730	0.0086	$4.23x10^{-89}$	Y

Table 3.21: Table of p-values and allele frequencies for the probe GI_29735811-S the eSNP rs6998786.

The following pages show signal plots for association analysis (figure 3.19) and fixed effect meta-analysis (figure 3.20) for the probe *GI_29735811-S*. In the three population-specific signal plots, the peak signals achieve GWS. When variants not reported in all three populations are included in the fixed effect meta-analysis, the peak SNP rs6998786 remains the same. The lengths of the peak signal LD blocks differ between the populations with YRI having the smallest LD block. From inspecting the signal plots for the population-specific association analyses and the fixed effect meta-analysis, it seems likely that the YRI population is contributing the most to fine mapping resolution.

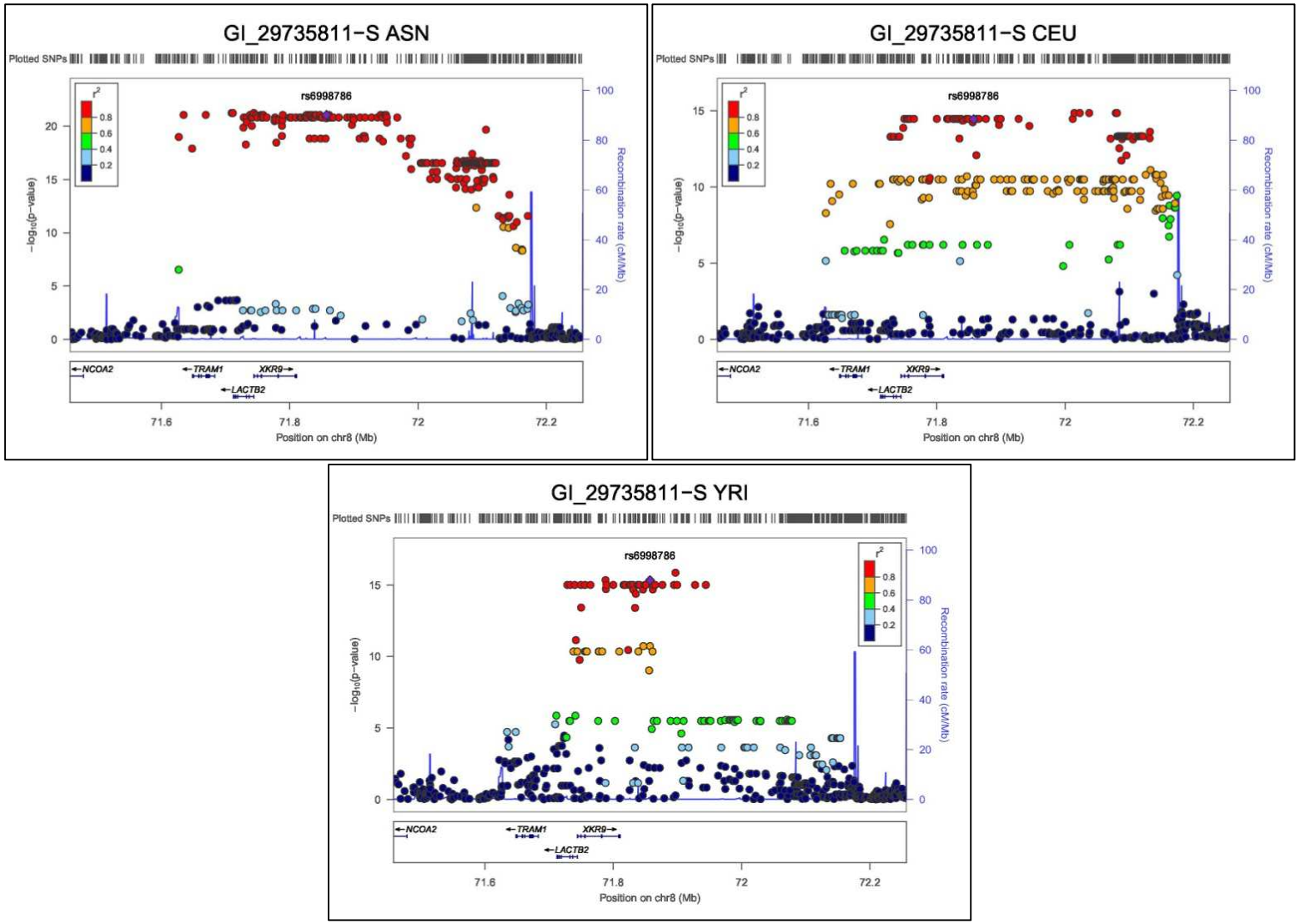


Figure 3.19: Signal plots for association analysis, for the probe GI_29735811-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data June 2010).

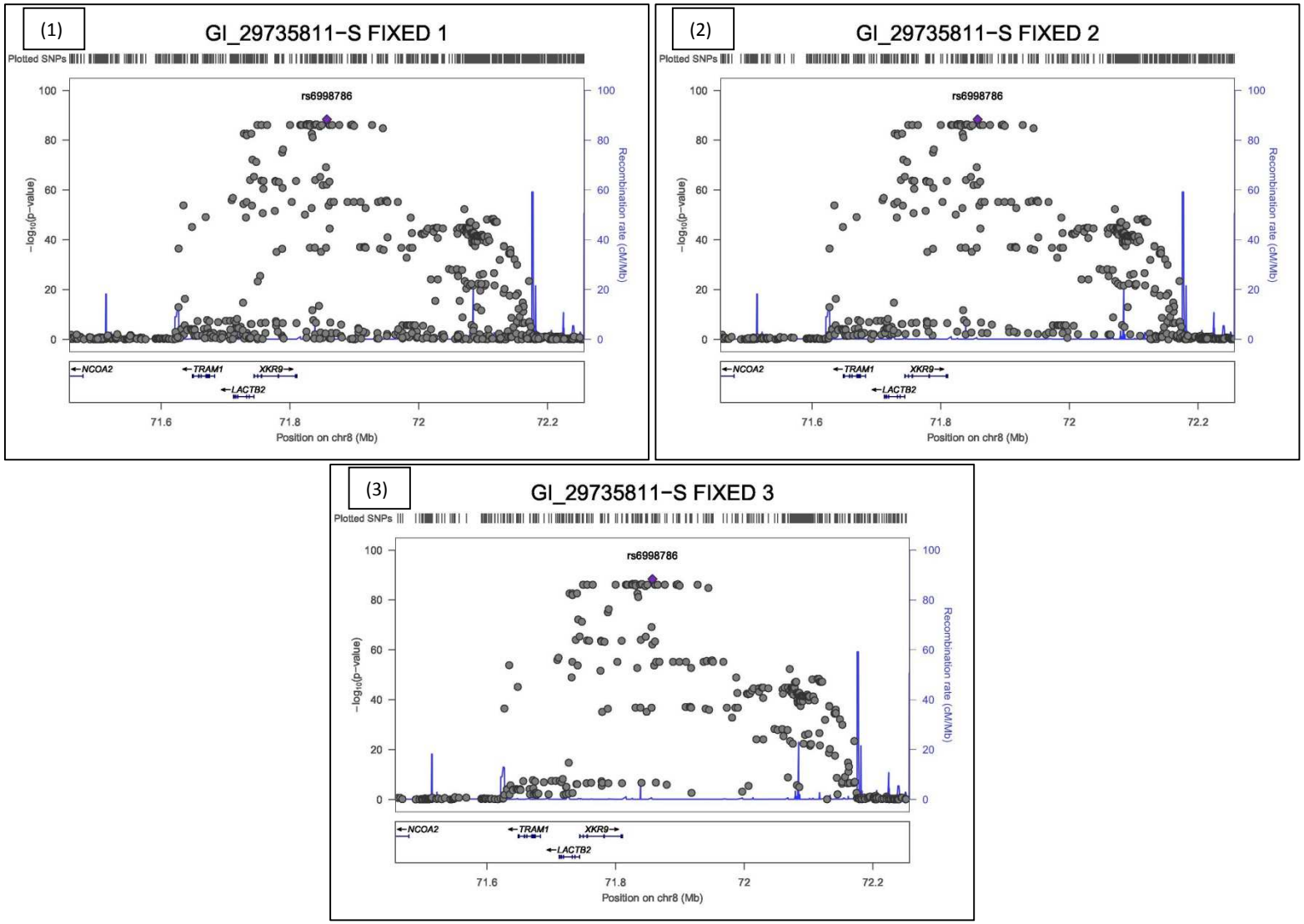


Figure 3.20: Signal plot of fixed effect meta-analysis for the probe GI_29735811-S. Each circle represents a Phase II HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) two or more SNPs not missing, (3) No missing SNPs: 3 SNPs present.

3.8 Summary

This chapter has presented the results of an association analysis and fixed effect meta-analysis to detect *cis* eQTLs in Phase II HapMap microarray expression and genotype datasets. The primary aims of this analysis were to: 1) improve power to detect *cis* eQTLs through increased sample size using meta-analysis; 2) identify and characterize heterogeneity of *cis* eQTL effect sizes between diverse populations; and 3) review the potential improvement in fine mapping by taking advantage of differing LD structure between ethnic groups.

Results in this chapter were compared to those reported in the Stranger *et al* 2007 study which used the same genetic and expression data. Results observed at FDR 5% differed between these studies, with larger numbers of eQTLs detected in this study. One possible reason for this is the different methods applied to assess statistical significance between this study and the Stranger 2007 study. However, a similar range of R2 values was observed between the two studies.

Two examples of heterogeneity in allelic effects between populations were selected for further analysis. In the first example, for the gene *PDXDC2P,* heterogeneity in effect sizes seemed likely to be due to: 1) no association signal in the YRI population and 2) differences in effect sizes at the peak SNP between ASN and CEU populations, which could arise because of differences in LD with a causal variant that is not present in Phase II HapMap. In the second example, for the gene *USMG5,* heterogeneity was again likely due to the lack of association signal in the YRI population, but also because the peak SNP from the fixed-effects meta-analysis was not reported in the association analysis for YRI.

Two examples have also been presented where a lack of heterogeneity in allelic effects between populations, but visual differences in the structure of LD between the populations can be used as the basis to fine map *cis* eQTLs. In the examples, for gene *XKR9* and *SNHG17*, the improved resolution in fine-mapping is primarily driven by a smaller LD block in the YRI population.

The primary limitations of the Phase II HapMap analysis for discovery and fine-mapping of eQTLs are its sample size and lack of diversity between the populations. In the next chapter, association analyses will be performed using the Phase III HapMap microarray expression and genotype data which have a larger overall sample size (709 individuals, compared with just 210 in Phase II) and greater ethnic diversity (eight populations, including Hispanic and South Asian ancestry, as well as greater diversity in African ancestry). The Phase III HapMap analysis will thus allow a more thorough investigation of heterogeneity and potential for trans-ethnic fine mapping.

CHAPTER 4 PHASE III HAPMAP ANALYSIS

4.1 Overview

In this chapter I present the results of the fixed-effect meta-analysis to detect eQTLs on genotype and expression data from eight Phase III HapMap populations (CEPH European ancestry: CEU, Chinese: CHB, Gujarati Indian: GIH, Japanese: JPT, Luhya, Kenya: LWK, Mexican: MEX, Maasai, Kenya: MKK and Yoruba, Nigeria: YRI). Full details of the genotype and expression data are in the methods sections 2.21 and 2.22.

The analysis aims to detect *cis* eQTLs, defined as 1 Mb upstream or downstream of the annotated start site of the transcript. The rationale is that performing meta-analysis across the eight populations will increase power, due to larger sample size, assuming that the causal eQTL variants are shared across ancestry groups. As compared with the Phase II HapMap analysis reported in the last chapter, the sample size and population diversity have been increased, although fewer SNPs have been directly genotyped (up to 1,599,206, compared with 2,907,520 in Phase II). I also evaluate the extent of heterogeneity of allelic effects at *cis* eQTLs across populations, and assess the potential for fine mapping of causal variants due to differing LD structure between the populations.

In this chapter, I begin by summarising quality control of the genotype data. I then discuss evidence for structure within each population. Next, I report the results of association analysis for each population, followed by the results of the fixed effect meta-analysis to combine association summary statistics across the populations. I then investigate the extent of heterogeneity in allelic effects between the populations, and present examples to illustrate the variable patterns of association across ancestry groups.

Please note that in this section, descriptions of gene function are taken directly from the GeneCard online resource, and therefore do not have citations. Also, positions within the genome use human genome version NCBI B37.

4.2 Quality Control (QC)

Poor quality SNPs have been removed from the Phase III HapMap data using the thresholds defined in section 2.3.1. Briefly QC exclusion measures applied are a MAF ≤ 0.05, SNP missing rate ≥ 0.05 and deviation from Hardy Weinberg Equilibrium p-value ≤ $1x10^{-3}$.

Table 4.1 gives an overview of polymorphic variant counts in each of the eight Phase III HapMap populations. The columns labelled "**Before QC"** and "**After QC"** specifies the number of variants present before and after QC has been applied. The row labelled **"Intersection"** is the number of polymorphic variants that are reported in all eight populations. The numbers of polymorphic variants are higher in the African populations (LWK, MKK, YRI), which would be expected as these populations are "older" than the others and therefore we expect to see more diversity in genetic variation. The lowest numbers of variants after QC are detected in the East Asian populations (CHB, JPT). The MKK has the highest population size (135) and the MEX has the lowest population size (41).

Population	SNPs Before QC	SNPs After QC	Individuals
LWK	1,309,721	1,283,373	83
MKK	1,316,017	1,302,287	135
YRI	1,295,378	1,281,127	108
CHB	1,103,523	1,083,499	79
JPT	1,095,758	1,080,301	81
CEU	1,201,053	1,198,993	107
GIH	1,214,205	1,190,891	75
MEX	1,193,660	1,160,400	41
Intersection	738,093	708,270	709

Table 4.1: Results of quality control performed on the Phase III HapMap SNPs.

4.3 Population Structure

In order to assess population structure, PCA of genetic data for each of the populations was performed. See methods section 2.4.3 for further details. PCA was carried out on each population separately (to assess fine-scale structure) and all populations together (to highlight ethnic outliers within each population). All variants that passed QC were used as input of SmartPCA of EIGENSTRAT (Price *et al* 2006) to generate principal components that reflect the genetic relatedness of the populations. SmartPCA takes account of LD between SNPs and therefore no LD pruning was undertaken prior to this analysis.

Figure 4.1 shows a PCA plot with all populations together. The populations cluster in three clear ancestry groups: African ancestry (LWK, MKK, YRI); East Asian ancestry (CHB, JPT); and Eurasian-Hispanic ancestry (CEU, GIH, MEX). Within each population group, it can be seen that there are no clear ethnic outliers. There is greatest spread in the MKK, GIH, and MEX, three populations which are known to be admixed.

PCA plots of the eight populations, separately, are shown in figure 4.2 and figure 4.3. The four populations also in the Phase II HapMap (CHB, JPT, CEU and YRI: figure 4.2) do not exhibit any structure (although some outliers are detected in the CEU and YRI populations as reported in section 3.2, and in Phase III, CHB and JPT are treated as separate populations). The four remaining populations (GIH, LWK, MEX, MKK, figure 4.3) have evidence of structure (with greater spread) and these populations are known to have admixture.

Since from the “All populations” PCA plot there are no clear ethnic outliers in populations, the outliers detected in the CEU and YRI populations were retained in the analysis. These outliers were also present in the Phase II HapMap populations and correcting the results for population structure using the first two principal components as covariates did not lead to any significant reduction in genomic control inflation factors (as reported in section 3.2).

Of the eight populations, the four known to have admixture (GIH, LWK, MEX, MKK) have been corrected for population structure by adjusting gene expression for eigenvectors obtained using EIGENSTRAT, as described in methods section 2.4.4. In the remaining four populations, no correction has been applied. Rank normalization has then been applied to expression data from all eight populations, after correction for population structure in the four admixed populations, to remove the impact of outliers.

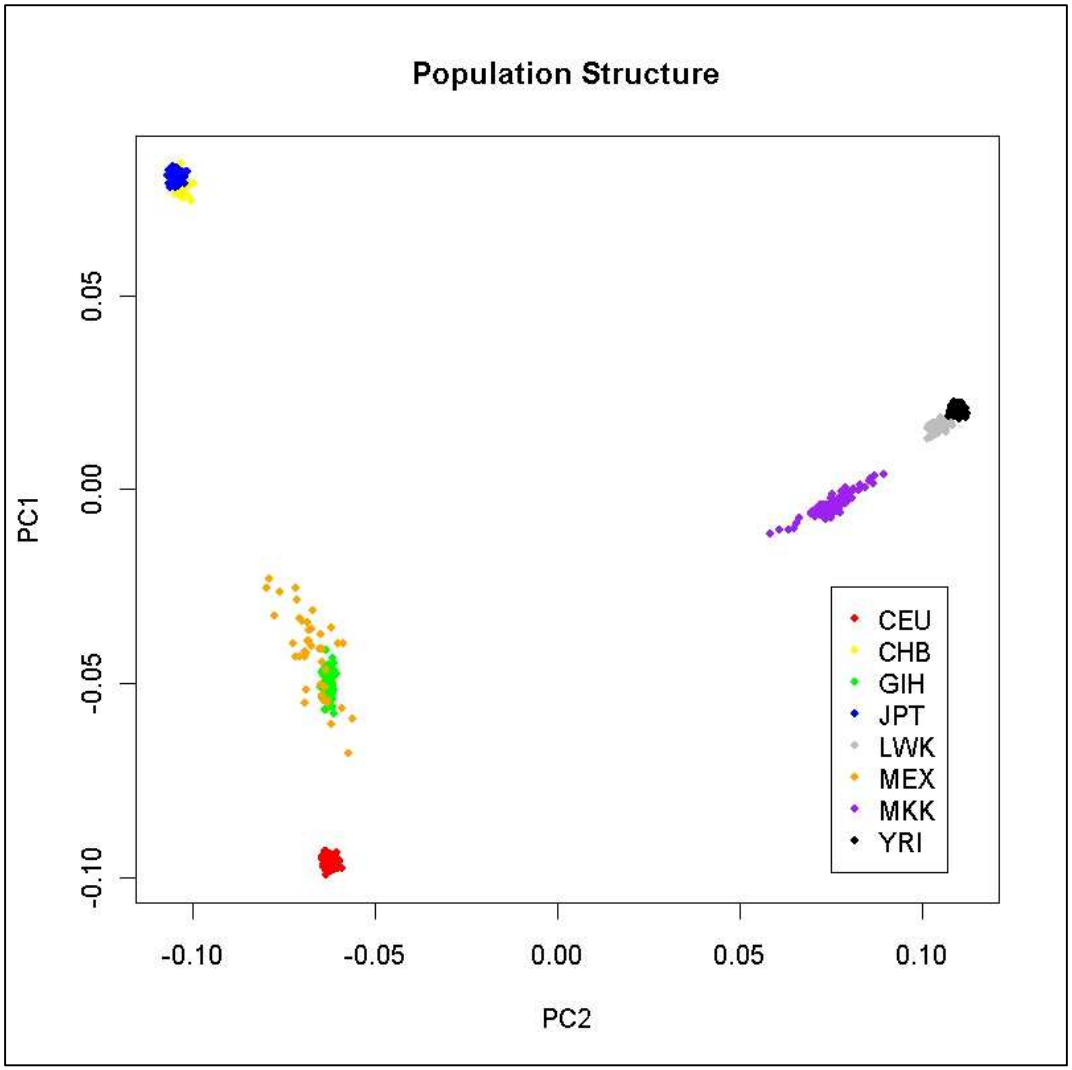


Figure 4.1: PCA plot for all eight Phase III HapMap populations. Each point specifies an individual in one of the eight populations. The legend specifies the colours of dots in each population. As can be seen the eight populations cluster into three ancestry groups: African: (LWK, MKK, YRI), East Asian (CHB, JPT) and Eurasian-Hispanic (CEU, GIH, MEX). PC: Principal Component.

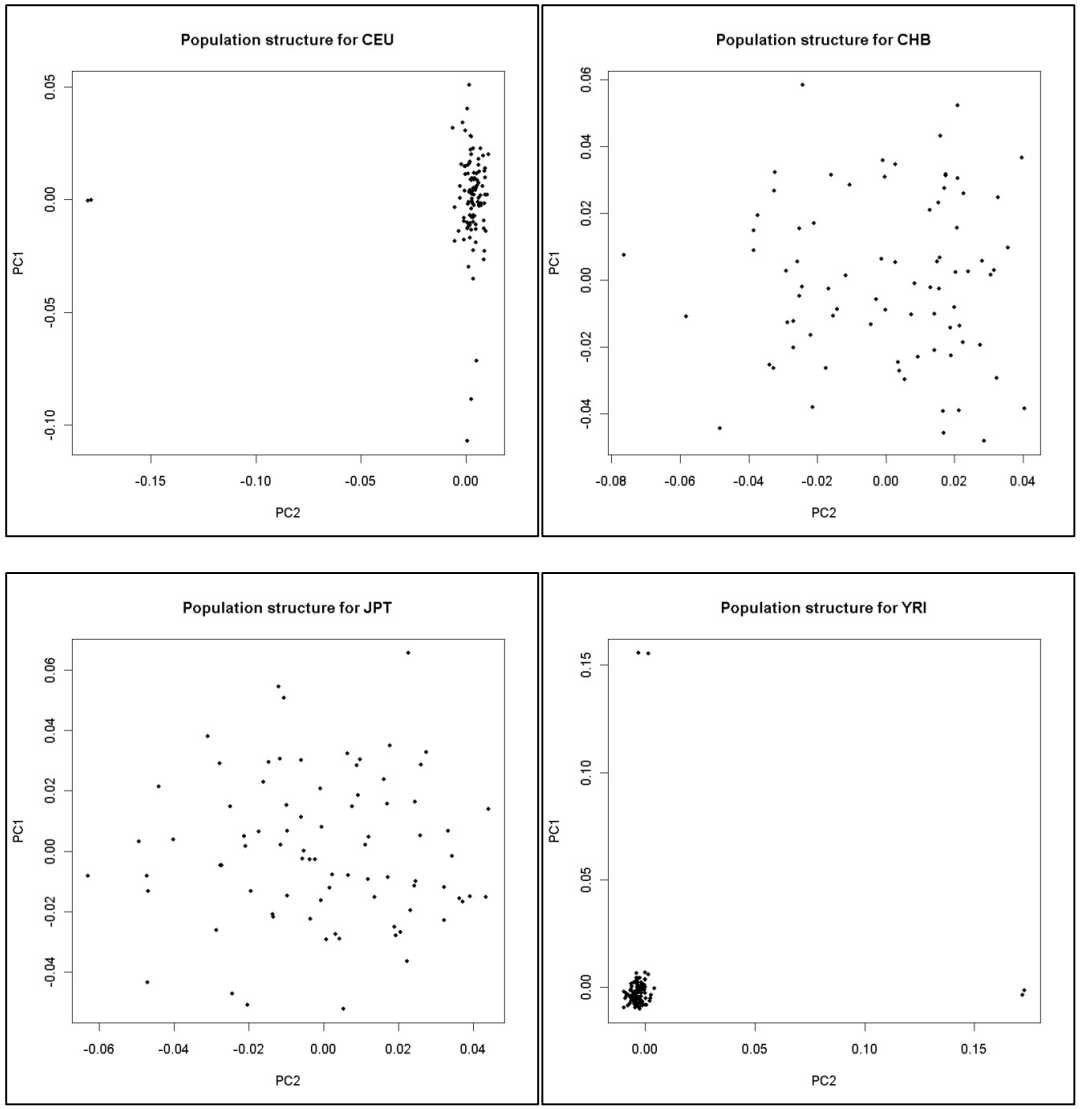


Figure 4.2: PCA plots for the Phase III HapMap Genetic data using populations also present in the Phase II HapMap with little structure (CEU, CHB, JPT, YRI). Each dot specified an individual from the specified population. PC: Principal Component.

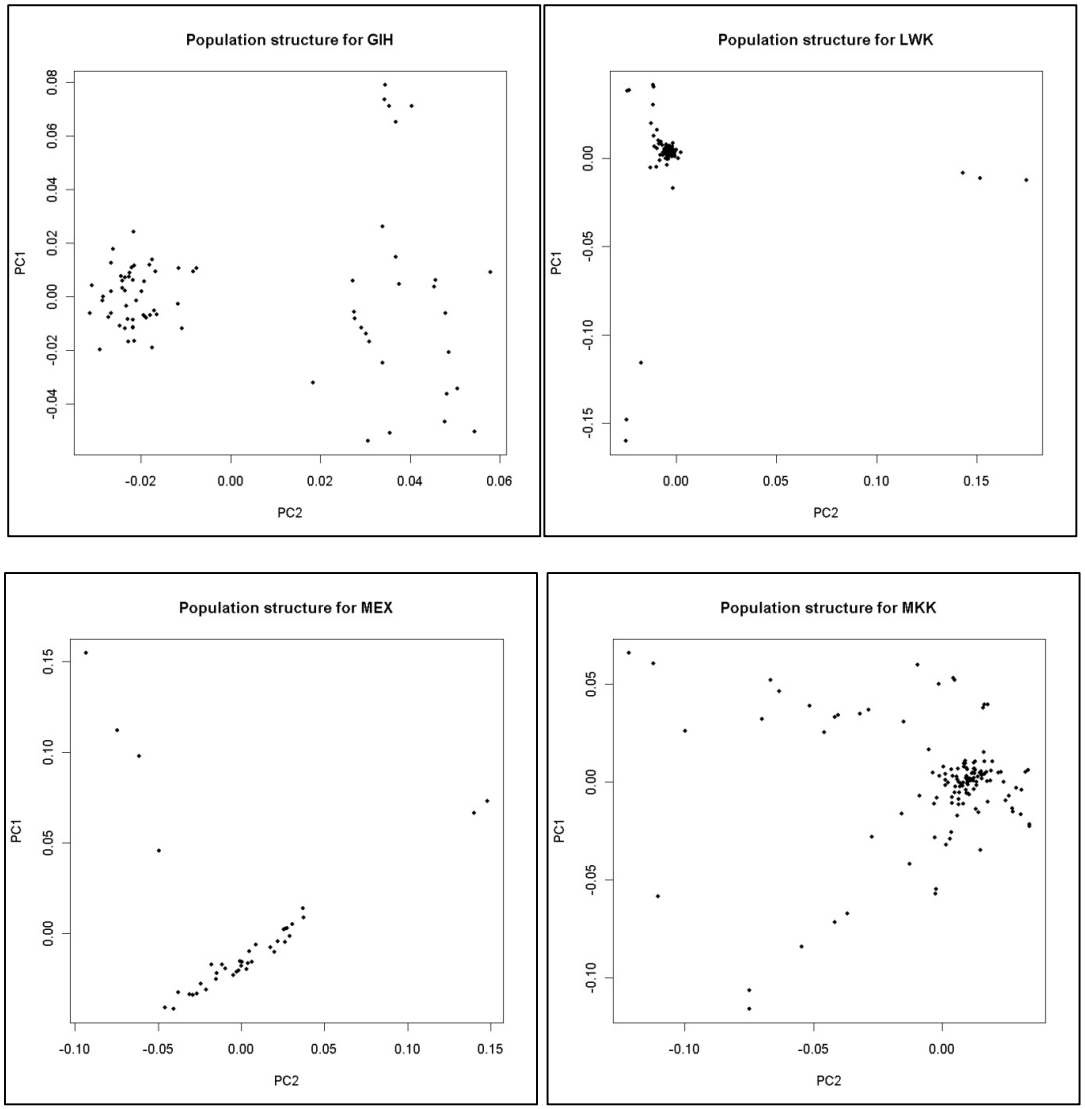


Figure 4.3: PCA plots for the Phase III HapMap Genetic data using populations known to have admixture (GIH, LWK, MEX, MKK). Each dot specified an individual from the specified population.

4.4 Multiple testing

To take into account the large number of tests being performed, the results of the association analysis and fixed effect meta-analysis were assessed at two significance thresholds: GWS (p-value ≤ $5x10^{-8}$) and FDR 5% significance determined using the Benjamini-Hochberg procedure. For background information regarding multiple testing correction, see methods section 2.10.

Table 4.2 shows the significance thresholds for FDR 5% determined using the Benjamini-Hochberg procedure. See methods section 2.10.2 for more information. The FDR threshold is calculated using the p-values generated for all SNPs in the dataset, including those specific to one or more population. The meta-analysis includes p-values with non-reported variant data; i.e. the number of populations with information at the SNP is less than eight. As can be seen the thresholds are broadly similar, with the MEX having the lowest cut-off, and the fixed effect meta-analysis having the largest.

The table also includes the number of test performed in the analysis for SNPs that pass QC (MAF ≥ 0.05), determined from the number of valid p-values generated. For the association analysis, this includes all the SNPs in each population. The fixed effect meta-analysis test number includes those variants that are reported in at least one population, effectively the union of the SNPs reported in any of the eight populations.

Population	Number of tests	FDR 5%
LWK	19,906,495	$1.26x10^{-5}$
MKK	19,989,543	$1.73x10^{-5}$
YRI	19,660,432	$1.24x10^{-5}$
CHB	16,830,135	$3.70x10^{-5}$
JPT	16,703,315	$4.07x10^{-5}$
CEU	18,297,853	$2.70x10^{-5}$
GIH	18,439,235	$1.92x10^{-5}$
MEX	18,292,111	$7.76x10^{-6}$
Meta-analysis	23,986,726	$2.26x10^{-4}$

Table 4.2: FDR 5% calculated using Benjamini-Hochberg for association analysis and fixed effect meta-analysis.

4.5 Association Analysis

This section presents the results of the Phase III HapMap association analysis to detect *cis* eQTLs. Quantitative trait association analysis using linear regression was performed using the *–linear* option of PLINK (Purcell *et al* 2007), including sex as a covariate. To detect *cis* eQTLs, only SNPs located 1Mb upstream and downstream of the probe's gene start site were included in the analysis. See methods section 2.5 for full details of this analysis. The results are presented at the two significance thresholds: GWS and FDR 5%.

4.5.1 QQ Plots

QQ plots were generated by plotting observed p-values from the association analysis against those expected from the null distribution. See Methods section 2.4.1 for more details on QQ plots. See figure 4.4 and 4.5 for QQ plots of association analysis results in each of the eight populations. As for the Phase II analysis, the p-values differ from the null distribution around the value $-\log_{10}(p) = 3$. Visually, there is no clear evidence of inflation indicative of population structure that has not been taken into account in the association analysis.

4.5.2 Genomic Control

To determine whether any general inflation of p-values has occurred in any of the populations, the GC inflation factor (λ) was calculated. See Methods Section 2.4.2 for definition and calculation of GC (λ).

Table 4.3 shows the GC (λ) inflation factors for the association analysis for each of the eight populations. The strongest GC (λ) inflation was observed in the MKK population (1.072). Population structure has already been taken into account in the MKK population using EIGENSTRAT (as described in methods section 2.4.4) because it is one of the four populations with admixture. Overall, there is no clear evidence of structure within any population that has not been accounted for in the association analysis. Furthermore, unlike the traditional application of genomic control in GWAS, for *cis* eQTLs we have a strong prior of association in the region, and

thus expect some inflation in association test statistics above the null distribution. Consequently, we have not corrected association summary statistics by genomic control in the presentation of results.

The table also shows the peak p-value for the association across all SNPs and probes. MEX has the weakest peak p-value, likely to be caused by the smaller population size. MKK has the strongest peak signal, likely to be due to being the largest population.

Population	Genomic Control	Peak p-value
LWK	1.023	2.15×10^{-29}
MKK	1.072	5.00×10^{-36}
YRI	1.024	1.93×10^{-32}
CHB	1.029	9.28×10^{-29}
JPT	1.036	2.51×10^{-29}
CEU	1.030	6.87×10^{-31}
GIH	1.030	1.50×10^{-24}
MEX	1.028	7.28×10^{-16}

Table 4.3: Genomic control for Phase III HapMap populations.

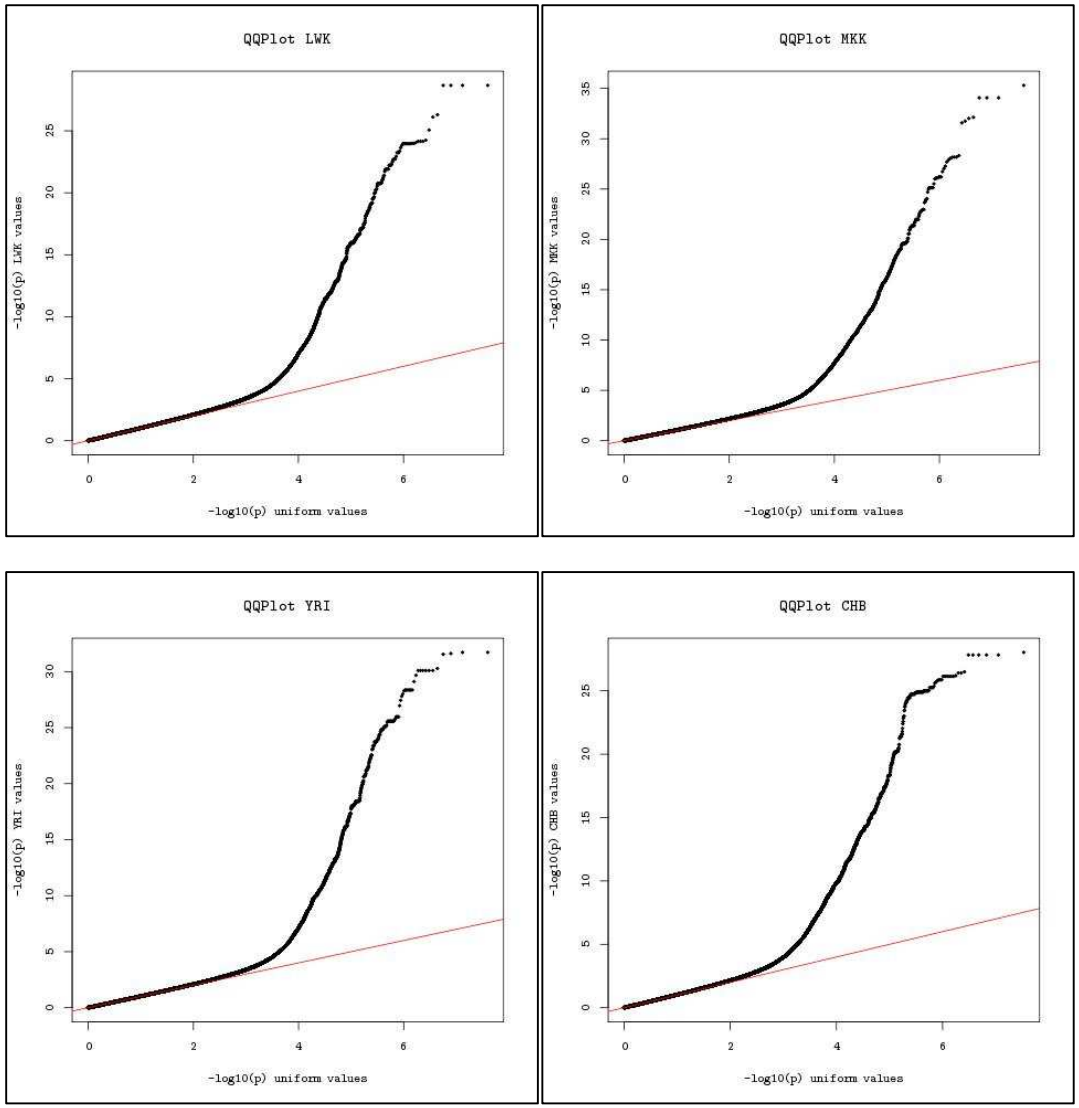


Figure 4.4: QQ plots of association analysis p-values against the uniform distribution

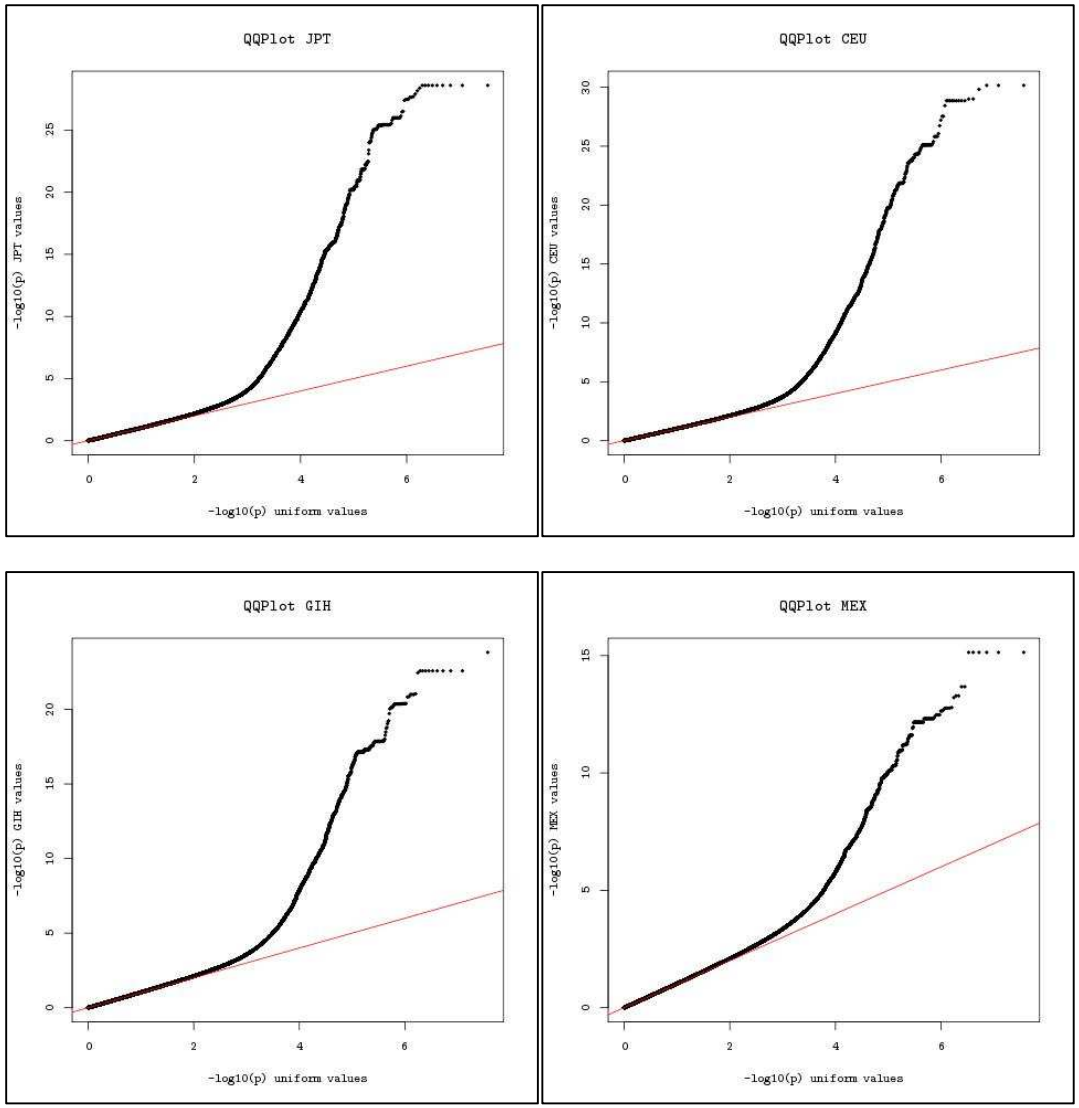


Figure 4.5: QQ plots of association analysis p-values against the uniform distribution

4.5.3 GWS ($p = 5x10^{-8}$)

Table 4.4 presents the counts of *cis* eQTLs signals detected at GWS using **Probes** and **Peak SNPs** see section 3.4.3 for description of these counts. JPT (305) has the largest number of *cis* eQTLs probes detected; MEX (81) has the smallest number. The East Asian ancestry group (CHB, JPT) have larger number of peak SNPs in comparison with the other groups; this could indicate greater extent of LD within these populations.

Population	Probes	Peak SNPs
LWK	244	447
MKK	294	427
YRI	245	427
CHB	276	743
JPT	305	827
CEU	215	409
GIH	185	426
MEX	81	245

Table 4.4: Summary of *cis*-eQTLs for Phase III HapMap with GWS ($p < 5x10^{-8}$).

In total, there is a nonredundant set of 736 probes showing a significant association in one or more populations. In total, 409 probes have significant associations in 2 or more populations and 23 probes have significant association for all eight populations.

4.5.4 FDR = 5%

Table 4.5 presents the counts of *cis* eQTL signals detected at FDR 5% significance. Once again, the largest number of *cis* eQTL probes detected is in the JPT (1088) population; MEX (311) has the smallest number. As in table 4.4, the East Asian ancestry group has a larger numbers of peak SNPs.

Population	Probes	Peak SNPs
LWK	770	1156
MKK	1074	1317
YRI	735	1040
CHB	1009	2051
JPT	1088	2105
CEU	784	1253
GIH	710	1317
MEX	311	751

Table 4.5: Summary of *cis* signals for individual Phase III HapMap populations with FDR 5%.

In total, there is a non-redundant set of 3445 probes showing a significant association in one or more populations. A total of 1155 probes have significant associations in 2 or more populations and 68 probes have significant associations across all eight populations.

At both significance thresholds, the MEX population has a lower number of *cis* eQTLs detected in comparison with the others; this may reflect the smaller sample size of MEX (41).

R2 data was not available for the association analysis results and so are not reported here.

4.5.5 Stranger 2012 results

The Stranger 2012 study used the same expression and genotype data analysed in this chapter. To compare results table 4.6 presents the results of the Stranger 2007 analysis at a 0.001 permutation threshold, which was reported to be equivalent to a FDR of approximately 5%.

Population	Gene Count
LWK	311
MKK	411
YRI	328
CHB	378
JPT	386
CEU	313
GIH	300
MEX	165

Table 4.6: Counts of eQTLs detected in the Stranger *et al* 2012 study at 0.001 permutation threshold.

In total, there is a nonredundant set of 1132 genes showing a significant association in one or more populations. A total of 547 genes have significant associations in 2 or more populations and 28 genes have significant associations across all three populations.

As can be seen the numbers of eQTLs detected in the Stranger 2012 study are smaller than those detected at FDR 5% in this study, this could be due to the differences in assessing significance between these studies. Also the Stranger results report gene counts rather than probe counts.

The Spearman Rho for eQTLs detected at the 0.001 permutation threshold reported in the Stranger 2012 paper ranged from 0.338 to 0.919.

4.6 Fixed effect meta-analysis

This section presents the results of fixed effect meta-analysis of association summary statistics across each of the Phase III HapMap populations. Fixed effect meta-analysis was performed using GWAMA (Magi *et al* 2010) using inverse variance weighting of effect sizes. For more information regarding fixed effect meta-analysis see methods section 2.8.1.

4.6.1 QQ Plot and Genomic Control

Figure 4.6 shows QQ plots for the fixed effect meta-analysis. The plots show a departure from the null distribution for each of the association results at approximately -log10 p-value = 3. The genomic control inflation factor is 1.119, the peak p-values is 1.40×10^{-305}. As described in the

previous section, for cis eQTL mapping, some deviation from the null distribution (and consequently some inflation in λ) is expected, and no correction for structure between populations after meta-analysis was performed.

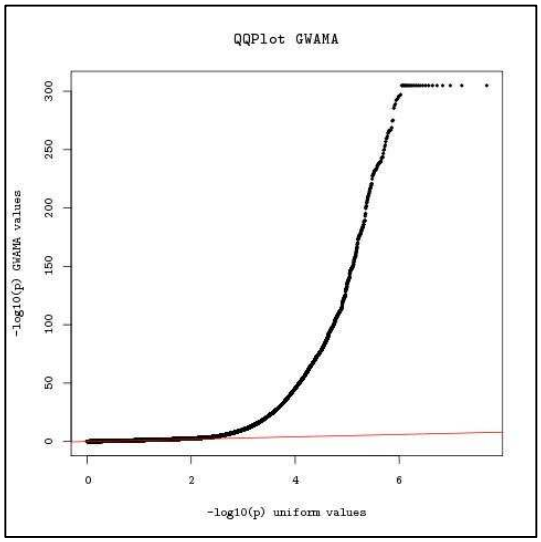


Figure 4.6: QQ plots of association analysis p-values against the uniform distribution

4.6.2 Signal Counts

The results of the meta-analysis are presented in the same way as the association analysis results (**Probes** and **Peak SNPs**).

Table 4.7 presents the counts of *cis* eQTLs detected with the fixed effect meta-analysis at GWS and FDR 5%. In addition to this, the number of populations within which the variant is reported (i.e. not monomorphic and passing QC) is also presented. A range of values are considered: SNP reported in all 8 populations; SNP reported in 6 or more populations; and SNP reported in at least one population.

Significance Level	Min population count	Probes	Peak SNPs
GWS	8	1846	1913
	6	2073	2167
	1	2305	2434
FDR 5%	8	5764	5888
	6	6935	7110
	1	8536	8860

Table 4.7: Results for fixed effect meta-analysis, at GWS and FDR. Probes and Peak SNPs explained in section 2.4.3. Min SNP count the minimum count of SNPs reported, i.e. 8: all eight reported, 1: one or more reported.

In order to assess the overlap of the HapMap phase 2 eQTLs identified in chapter 3 with the HapMap phase 3 eQTLs identified in this chapter, eQTL probes were mapped to their HGNC symbol and the overlap of the HGNC symbols were assessed. A total of 991 unique symbols were identified for the 1024 probes identified in the HapMap phase 2 eQTLs at GWS, and 1732 unique symbols were identified in the 1846 HapMap phase 3 probes at GWS. An overlap of 403 HGNC symbols was found between the HapMap phase 2 and phase 3 datasets.

4.6.3 Top results from fixed effect meta-analysis

Table 4.8 presents the ten strongest signals detected in the fixed effect meta-analysis, with SNPs reported in all eight populations. Effect sizes and standard errors of the eight population-specific association analyses are presented in the Appendix table A.4. As in the previous, chapter due to the extremely low p-values the validity of the signals was assessed by identifying whether each

probe contains one or more 1000 Genomes SNPs. It was found that four of the ten probes had one or more 1000 Genomes SNPs which mapped to their region and so maybe false-positives. R2 data was not available in the results generated by PLINK in this analysis and so have not been reported.

The nine SNPs detected in this table have been annotated using ANNOVAR (Wang *et al* 2010) (see section 2.12). Five of these are within introns and the remaining four are not within transcribed regions. Again as with the HapMap phase II peak SNPs none of these are within exons.

Probe	HGNC	SNP	Beta	SE	Z-Score	p-value	Cochran's Q p-value	1000 Genomes SNPs	ANNOVAR Annotation
ILMN_24844_2100600	*IRF5*	rs6965542	1.15	0.027	42.73	$p < 1.40x10^{-305}$	0.025	N	Intronic
ILMN_22173_5690095	*ERAP2*	rs2548533	1.70	0.041	41.13	$p < 1.40x10^{-305}$	0.022	N	Intronic
ILMN_27265_6330037 (*)	*CHURC1*	rs10141986	1.34	0.035	38.01	$p < 1.40x10^{-305}$	0.626	N	Intronic
ILMN_2170_6860347	*WBSCR27*	rs13228435	0.71	0.019	36.80	$8.41x10^{-298}$	0.236	Y	Intergenic
ILMN_7731_3520685	*C17orf97*	rs11150882	0.85	0.024	35.50	$6.40x10^{-276}$	$1.74x10^{-9}$	N	Upstream
ILMN_7731_2570703	*C17orf97*	rs11150882	1.23	0.035	35.45	$3.28x10^{-275}$	$4.98x10^{-10}$	Y	Upstream
ILMN_20550_7330093	*HLA-DRB1*	rs9271170	3.62	0.103	34.96	$1.26x10^{-267}$	0.590	Y	Intergenic
ILMN_15237_5860053	*IPO8*	rs11834524	-0.63	0.018	-34.93	$3.11x10^{-267}$	0.011	Y	Intergenic
ILMN_22465_2350368	*PTER*	rs7909832	-0.90	0.026	-34.01	$1.90x10^{-253}$	$1.22x10^{-15}$	N	Intronic
ILMN_33941_1570382 (*)	*HMSD*	rs9945924	-0.98	0.030	-32.57	$1.25x10^{-232}$	0.038	N	Intronic

Table 4.8: Ten probes with largest absolute z-score from the Phase III HapMap fixed effect meta-analysis. (*) deprecated in NCBI Build 37.

4.6.4 Heterogeneity

When performing fixed effect meta-analysis, heterogeneity of effect sizes across populations can be determined by calculating the Cochran's Q-statistic. This section investigates the extent of heterogeneity detected in the fixed effect meta-analysis of the Phase III HapMap populations. For more information regarding calculation of heterogeneity and the Cochran's Q statistics, see section 2.8.2.

Table 4.9 presents a summary of the number of observed association signals (at GWS and FDR 5%) that demonstrate evidence of heterogeneity for Cochran's Q p-value < 10^{-3}. For a full description of fields in this table, refer to section 3.53.

Significance Level	Min population Count	Peak Signals	Heterogeneous Signals	Expected Signals	Binomial Test p-value
GWS	8	1913	190	~2	< 2.2×10^{-16}
	6	2167	166	~2	< 2.2×10^{-16}
	1	2434	138	~3	< 2.2×10^{-16}
FDR 5%	8	5888	248	~6	< 2.2×10^{-16}
	6	7110	220	~7	< 2.2×10^{-16}
	1	8860	171	~9	< 2.2×10^{-16}

Table 4.9: Summary of heterogeneity results with GWS ($p < 5 \times 10^{-8}$) and FDR 5%. Heterogeneity is assessed with Cochran's Q ($P<1 \times 10^{-3}$). Shared indicates the number of populations the signal appeared within.

More heterogeneous *cis* eQTLs are observed than would be by chance. This could occur for the following reasons:

1. The peak SNP may not be the causal variant itself, but may be in strong LD with the causal variant. Different effect sizes may be observed between populations due differences in LD between the peak SNP and the causal variant. In some populations, if there is only weak LD between the causal variant and genotyped SNPs, there may be no association signal observed in those populations.

2. The causal variant or the closest tag SNP may be monomorphic or failed QC in one or more populations. As a result, the peak SNP from the meta-analysis might not be the best tag for the causal variant.
3. Gene-gene interaction with SNP that differs in frequency between populations, meaning that the marginal effect size of the causal variant will differ between populations
4. Interaction with some other factor that differs between populations (e.g. population structure, batch effects) again meaning that the marginal effect size of the causal variant will differ between populations.

It can also be observed from table 4.8 that the amount of heterogeneity is reduced when SNPs not reported in all eight populations are included. This indicates that some of the peak SNPs reported in all populations is not necessarily the best tag SNP for the causal variant.

In the same way as for Phase II HapMap (section 3.5.3) we have further investigated potential sources of heterogeneity using two criteria:

Zero Effects: Do any of the populations have an effect size not significantly different from zero?

Opposite Effects: Do populations have allelic effects in opposite directions?

Table 4.10 shows the results of this investigation of the heterogeneous association signals from the fixed effect meta-analysis. As can be seen approximately 70% of the heterogeneous probes have zero effects. This could be due to a number of reasons, for example, there may not be a good tag for the causal variant in some populations, with the result that an association signal is not observed. Approximately 1% of the heterogeneous probes have opposite effects. The reason for this lack of opposite direction effect signals is that they will tend to cancel each other out in the fixed-effects meta-analysis, and thus will not achieve GWS/FDR. Opposite directions of effect are less biologically plausible, and may be indicative of an interaction with a non-genetic factor influencing expression, for example.

Significance Level	Min population Count	Zero Effects	Opposite Effects
GWS	8	140 / 190 (74%)	2 / 190 (1%)
	6	113 / 166 (68%)	1 / 166 (1%)
	1	85 / 138 (62%)	1 / 138 (1%)
FDR 5%	8	198 / 248 (80%)	5 / 248 (2%)
	6	167 / 220 (76%)	2 / 220 (1%)
	1	118 / 171 (69%)	2 / 171 (1%)

Table 4.10: Summary of heterogeneity analysis.

4.6.5 Top Heterogeneous Results

Table 4.11 presents ten significant results from the fixed effect meta-analysis with largest heterogeneity observed. The two probes with the strongest heterogeneity are both for the gene *CAV2*. The first entry is used as an example of high heterogeneity in the examples section. See appendix A.5. It was found that five of these probes contain one or more 1000 Genome SNPs, and so maybe false positives.

Probe	HGNC	SNP	Beta	SE	z-score	p-value	Cochran's Q	Cochran's Q p-value	1000 Genomes SNP
ILMN_5108_2030195	*CAV2*	rs17138749	0.24	0.023	10.42	2.09×10^{-25}	217.12	$p < 2 \times 10^{-16}$	Y
ILMN_5108_2260136	*CAV2*	rs17138749	0.22	0.023	9.83	8.58×10^{-23}	195.44	$p < 2 \times 10^{-16}$	Y
ILMN_128814_5960180	*TMED4*	rs217378	0.34	0.026	13.11	3.03×10^{-39}	179.44	$p < 2 \times 10^{-16}$	Y
ILMN_1950_6220333	*MYRFL*	rs4512898	0.09	0.007	12.00	3.94×10^{-33}	167.15	$p < 2 \times 10^{-16}$	N
ILMN_30273_4860327	*UTS2*	rs2493214	0.12	0.016	7.22	5.44×10^{-13}	140.32	$p < 2 \times 10^{-16}$	Y
ILMN_4793_2490209	*PPIL3*	rs6435069	0.14	0.020	-7.17	7.99×10^{-13}	123.01	$p < 2 \times 10^{-16}$	N
ILMN_28044_4830575	*UTS2*	rs2493214	0.26	0.043	6.03	1.72×10^{-9}	122.32	$p < 2 \times 10^{-16}$	N
ILMN_4387_7210035	*FAM154B*	rs1877242	0.07	0.011	6.79	1.18×10^{-11}	100.28	$p < 2 \times 10^{-16}$	N
ILMN_17555_5560494	*ZP3*	rs1754183	0.50	0.040	12.55	4.30×10^{-36}	97.34	$p < 2 \times 10^{-16}$	NA
ILMN_13285_4210136	*FANCA*	rs2239360	0.28	0.014	-19.98	9.23×10^{-89}	92.34	$p < 2 \times 10^{-16}$	Y

Table 4.11: Probes with most significant heterogeneity detected using the Cochran's Q statistic

4.7 Examples

In this section I present four examples of *cis* eQTLs detected in fixed-effects meta-analysis:

1. Most significant meta-analysis association signal, with no evidence of heterogeneity in allelic effects between populations: *WBSCR27 (*ILMN_2170_6860347*)* (section 4.7.1). The probe ILMN_2170_6860347 contains a 1000 Genome SNP rs73369956 with an allele frequency of 0.028.
2. Two association signals with highly significant evidence of heterogeneity in allelic effects between populations: *CAV2 (*ILMN_5108_2030195*)* (section 4.7.2) and *FANCA* (ILMN_13285_4210136) (section 4.7.3). Both probes ILMN_5108_2030195 and ILMN_13285_4210136 contains several 1000 Genomes SNPs within their regions.
3. An example from the previous chapter, with highly significance evidence of heterogeneity: *USMG5 (*ILMN_10409_6860670*)* (section 4.7.4). The probe ILMN_10409_6860670 does not contain any 1000 Genomes SNPs within its region.

4.7.1 ILMN_2170_6860347 (ENSG00000165171, WBSCR27)

The probe ILMN_2170_6860347 has been chosen for further analysis due to it having the most significant association signal in the fixed effect analysis ($p=8.41x10^{-298}$) with no evidence of heterogeneity in allelic effect sizes between populations (Cochran's Q p-value= 0.236) (See table 4.9).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000165171 and the HGNC symbol *WBSCR27*. The full name of the gene is Williams Beuren Syndrome Chromosome Region 27, and the gene start site position is chr7:73256865. The SNP rs13228435 position is chr7:73239172, and is intergenic between the genes *CLDN3* (54572 bp) and *CLDN4* (6021 bp). The gene start site is 17,693 base

pairs from the eSNP. The gene encodes a protein belonging to the ubiE/COQ5 methyltransferase family. The gene is deleted in Williams Syndrome, a rare neurodevelopment disorder.

Figure 4.7 shows a forest plot for the results of the association analysis in each population and the fixed effect meta-analysis for probe ILMN_2170_6860347 with eSNP rs13228435. It can be observed, as expected from the low Cochran Q statistic, that effect sizes across the eight populations are very similar.

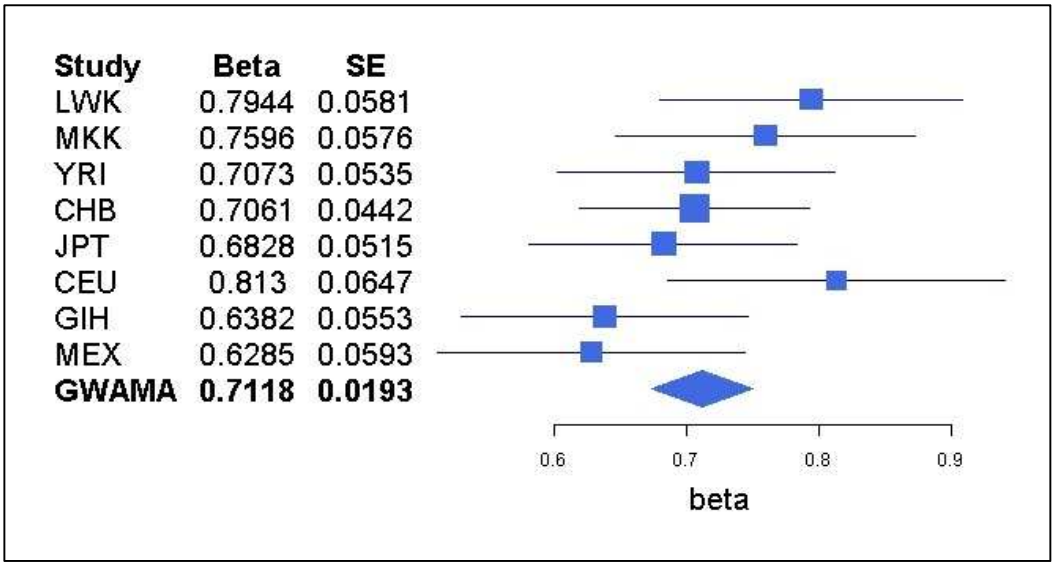


Figure 4.7: Forest plot of association and fixed effect meta-analyses for probe ILMN_2170_6860347 with Phase III HapMap SNP rs13228435. Blue squares specify the effect size for each analysis, lines specify the 95% confidence limits calculated from the standard error.

Table 4.12 presents a table of p-values and allele frequencies for association analysis and fixed effect meta-analysis for probe ILMN_2170_6860347 with eSNP rs13228435. Association signals in all eight populations achieve GWS. The allele frequencies range from 0.45 to 0.79. The fixed effect peak SNP rs13228435 is also the peak SNP in four of the population-specific association analyses (LWK, MKK, CHB and JPT).

Population	Minor Allele (C/T)	Allele Frequency (Allele C)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	C	0.45	0.7944	0.0581	1.21×10^{-22}	Y
MKK	T	0.66	0.7596	0.0576	7.35×10^{-26}	Y
YRI	C	0.47	0.7073	0.0535	4.30×10^{-24}	N
CHB	T	0.54	0.7061	0.0442	5.58×10^{-26}	Y
JPT	T	0.60	0.6828	0.0515	1.11×10^{-21}	Y
CEU	T	0.79	0.813	0.0647	1.38×10^{-22}	N
GIH	T	0.59	0.6382	0.0553	4.91×10^{-18}	N
MEX	T	0.57	0.6285	0.0593	6.85×10^{-13}	N
Fixed-effect Meta-analysis	--	--	0.7118	0.0193	8.41×10^{-298}	Y

Table 4.12: Table of allele frequencies and p-values for Phase III HapMap SNP rs13228435 with probe ILMN_2170_6860347. The column **Peak SNP** specifies whether the peak SNP for the association analysis is the same as the peak signal in the fixed effect meta-analysis.

Figures on the following pages show signal plots for the results of the association analysis in each of the eight populations (figure 4.8 and figure 4.9) and the fixed effect meta-analysis (figure 4.10).

The eSNP rs13228435 remains the peak eSNP in the fixed effect meta-analysis, when variants not reported in all eight populations are included. From the eight signal plots for the association analysis results the eSNP rs13228435 is the peak SNP, or in strong LD with the peak SNP, in each of the population-specific analyses. Since rs13228435 is not the peak SNP in all eight populations, this indicates that it may not be the causal SNP, but a tag SNP in strong LD with the causal SNP. The causal variant may not be genotyped or not reported in all eight populations. The peak signal LD blocks are very similar across populations, suggesting that the peak SNPs in each population is tagging the same causal variant. The peak signal LD block covers the smallest region in the African ancestry group populations.

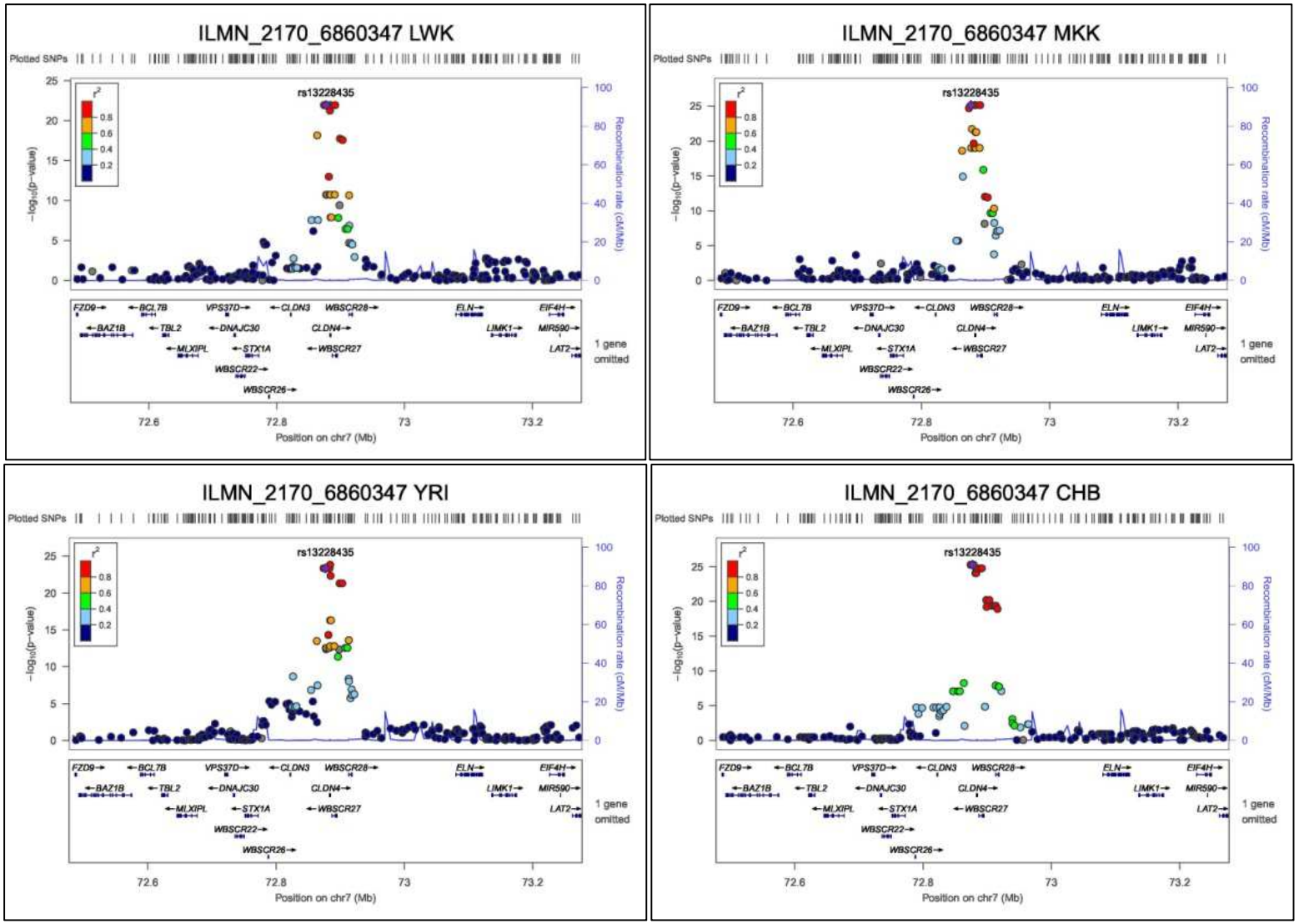


Figure 4.8: plots for peak Phase III HapMap SNPs for probe ILMN_2170_6860347. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

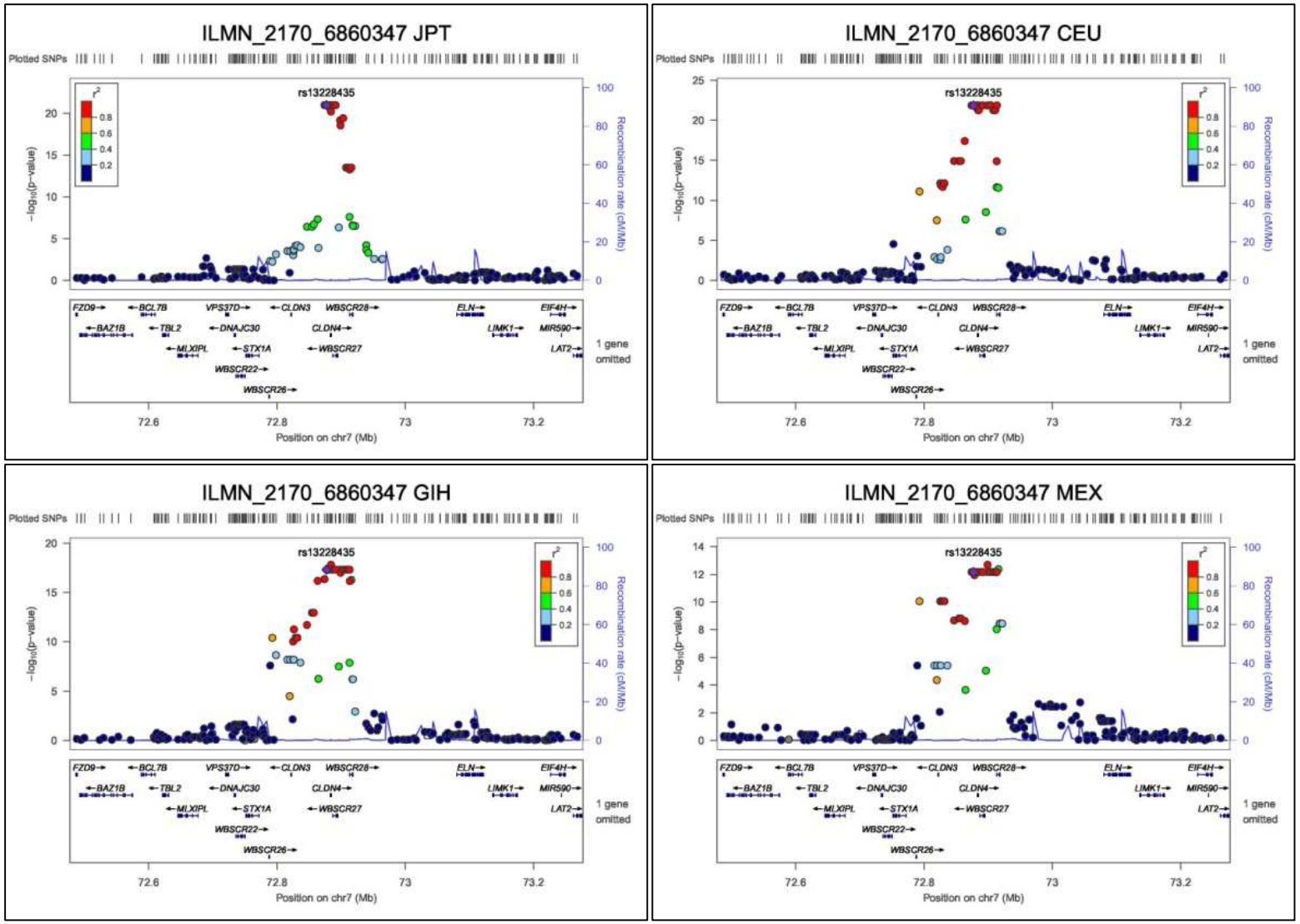


Figure 4.9: plots for peak Phase III HapMap SNPs for probe ILMN_2170_6860347. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

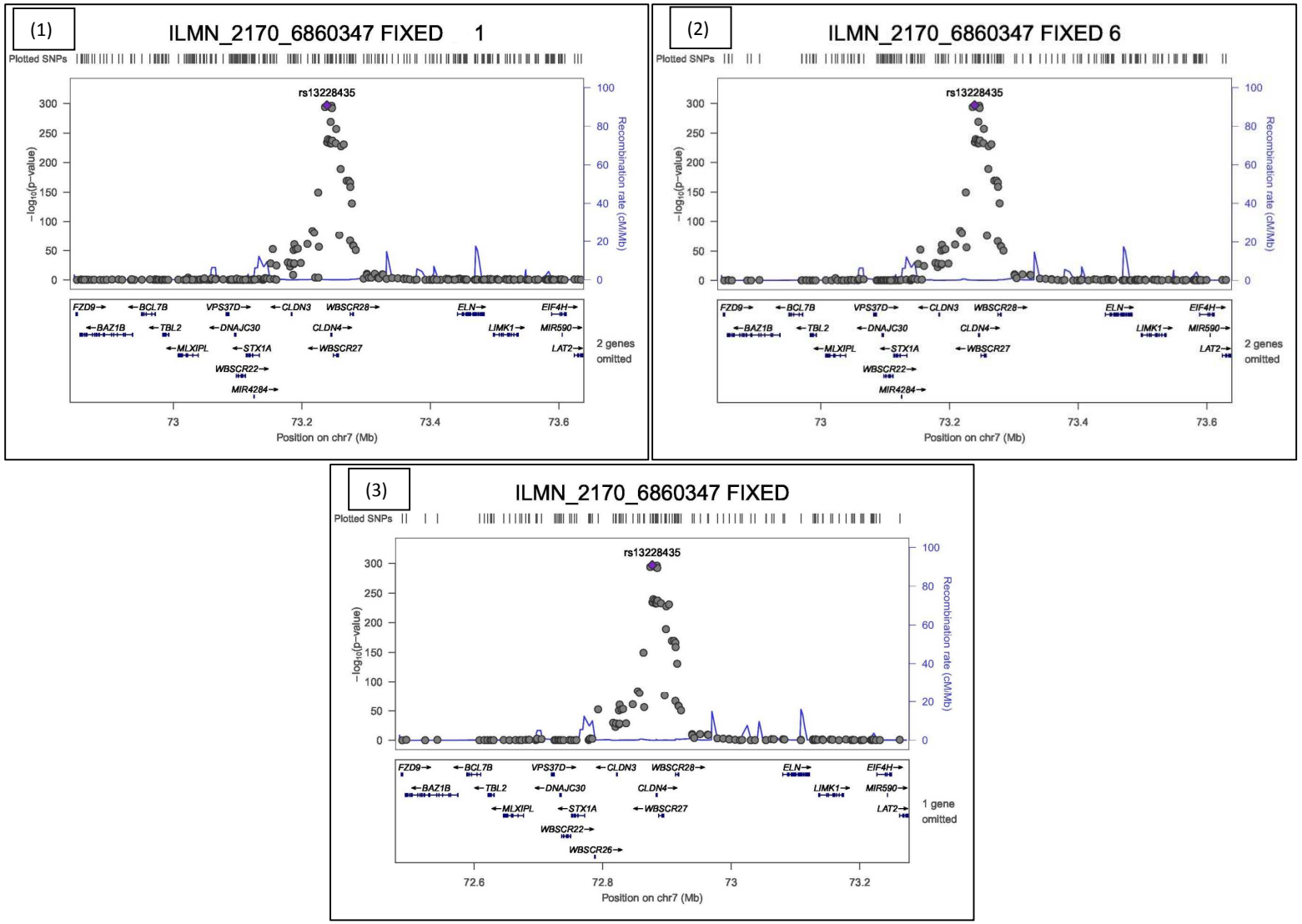


Figure 4.10: Signal plot for fixed effect meta-analysis for probe ILMN_2170_6860347. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) six or more SNPs not missing, (3) No missing SNPs: 8 SNPs present.

4.7.2 ILMN_5108_2030195 (ENSG00000105971, CAV2)

This example has been chosen because it has the greatest heterogeneity in allelic effect sizes between populations in the fixed-effects meta-analysis, as measured using the Cochran's Q statistic (Q=217.12, p-value < $2x10^{-16}$).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000105971 and the HGNC symbol *CAV2*. The full name of the gene is Caveolin 2, and the gene's start position is chr7:115927434. The eSNP rs17138749 is located at chr7:116133098 and is intergenic, mapping closest to *CAV2*. Caveolin 2 is a major component of the inner surface of caveolae, small invaginations of the plasma membrane and is involved in essential cell functions such as signal transduction, lipid metabolism, cellular growth control and apoptosis.

Figure 4.11 shows a forest plot for the results of the association analysis in each population and the fixed effect meta-analysis for probe ILMN_5108_2030195 with the eSNP rs17138749. From the plot it can be seen that the effect sizes of populations in the African ancestry cluster together, close to an effect size of zero. The effect sizes of two African populations (LWK and YRI) do not differ significantly from zero (p-value ≤ 0.05).

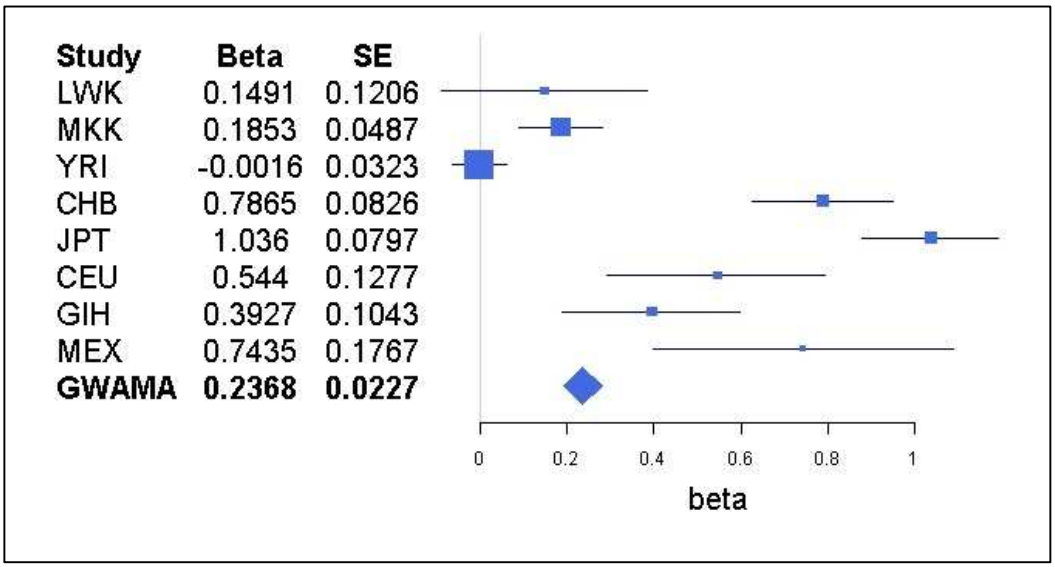


Figure 4.11: Forest plot of association and fixed effect meta-analyses for probe ILMN_5108_2030195 with Phase III HapMap eSNP rs17138749.

Table 4.13 shows p-values and allele frequencies for the association analysis and fixed effect for probe ILMN_5108_2030195 with eSNP rs17138749. The East Asian populations have the most significant effect sizes, both achieving GWS. In two populations (MKK and JPT), rs17138749 is the peak SNP. The allele frequencies range from 0.12 to 0.25.

Population	Minor Allele (C/A)	Allele Frequency (Allele C)	Beta (Allele C)	SE	p-value	Peak SNP
LWK	C	0.25	0.1491	0.1206	0.220	N
MKK	C	0.13	0.1853	0.0487	2.13×10^{-4}	Y
YRI	C	0.17	-0.0016	0.0323	0.962	N
CHB	C	0.16	0.7865	0.0826	1.33×10^{-14}	N
JPT	C	0.24	1.0360	0.0797	3.30×10^{-21}	Y
CEU	C	0.17	0.5440	0.1277	4.49×10^{-5}	N
GIH	C	0.12	0.3927	0.1043	3.38×10^{-4}	N
MEX	C	0.23	0.7435	0.1767	1.52×10^{-4}	N
Fixed effect Meta-analysis	--	--	0.2368	0.0227	2.09×10^{-25}	Y

Table 4.13: Table of allele frequencies and p-values for Phase III HapMap SNP rs17138749 with probe ILMN_5108_2030195

Figures on the following pages show the signal plots generated for each of the population-specific association analyses (figure 4.12 and 4.13) and fixed effect meta-analysis (figure 4.14).

Heterogeneity is being driven by effect size differences in the African ancestry groups: In the LWK association signal plot, there is a strong signal of association, which is not in LD with the peak SNP from the fixed-effects meta-analysis. In the MKK signal plot the eSNP (rs17138749) is the peak signal, but is quite weak ($p = 2.13x10^{-4}$). There is no evidence of any signal at all in the YRI signal plot. Signal plots for CHB and JPT show similar LD structure, with the eSNP rs17138749 located within the LD block for both populations. The eSNP rs17138749 is the peak signal in JPT, but not the peak SNP for CHB, but are in strong LD with each other and within the same LD block. The eSNP rs17138749 is not the peak signal in any of the Eurasian-Hispanic populations. From the fixed effect association plot it can be seen that the peak signal with SNPs reported in all eight populations is rs17138749. However, if SNPs not reported in at least one population are included, the peak SNP is different (rs12672236). Together, this indicates that the heterogeneity is in part due to the peak SNPs from population-specific association analyses having different LD with that from the fixed-effects meta-analysis. Furthermore additional heterogeneity in allelic effect sizes occurs because there is no signal of association in the YRI population.

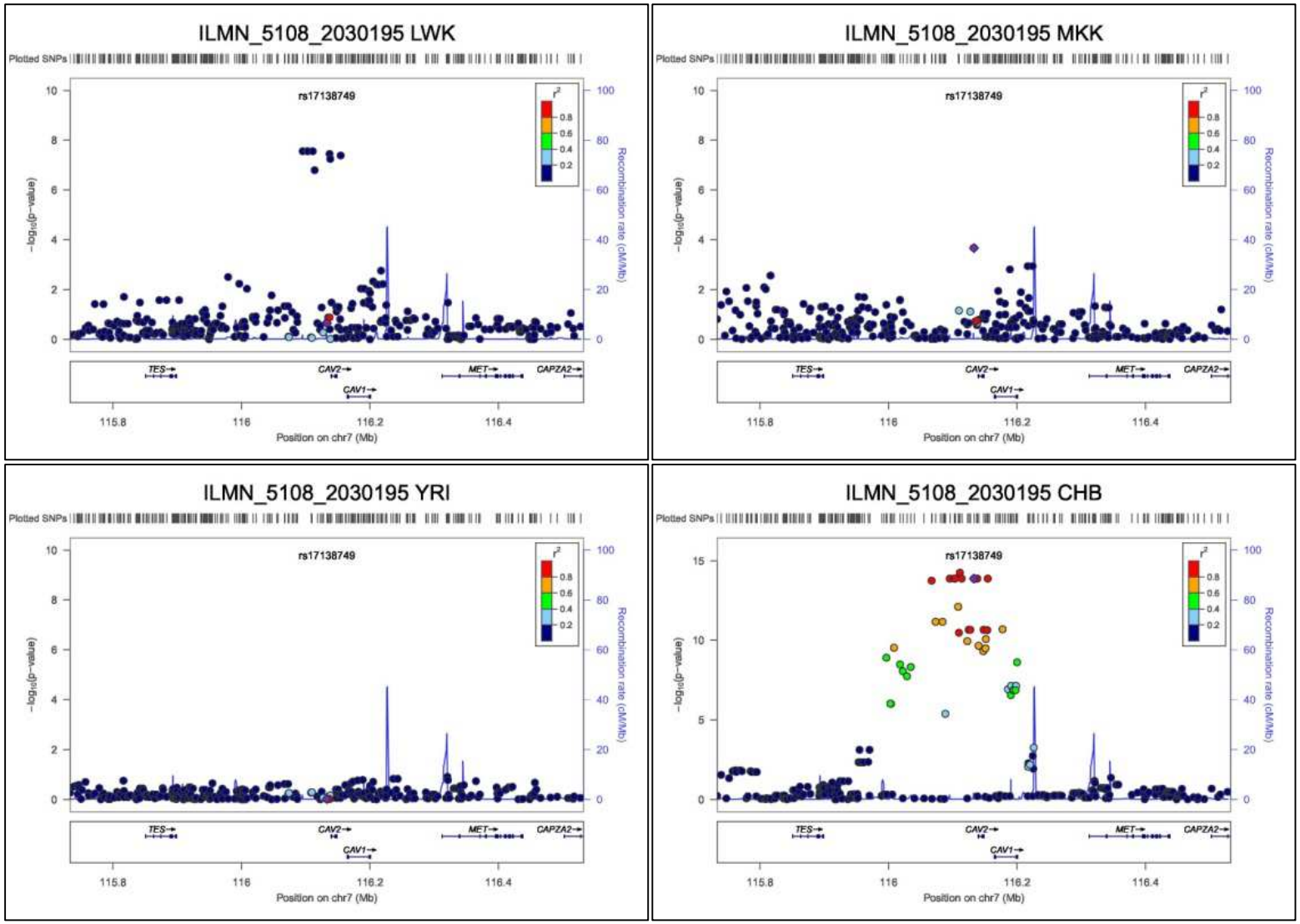


Figure 4.12: plots for peak Phase III HapMap SNPs for probe ILMN_5108_2030195. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

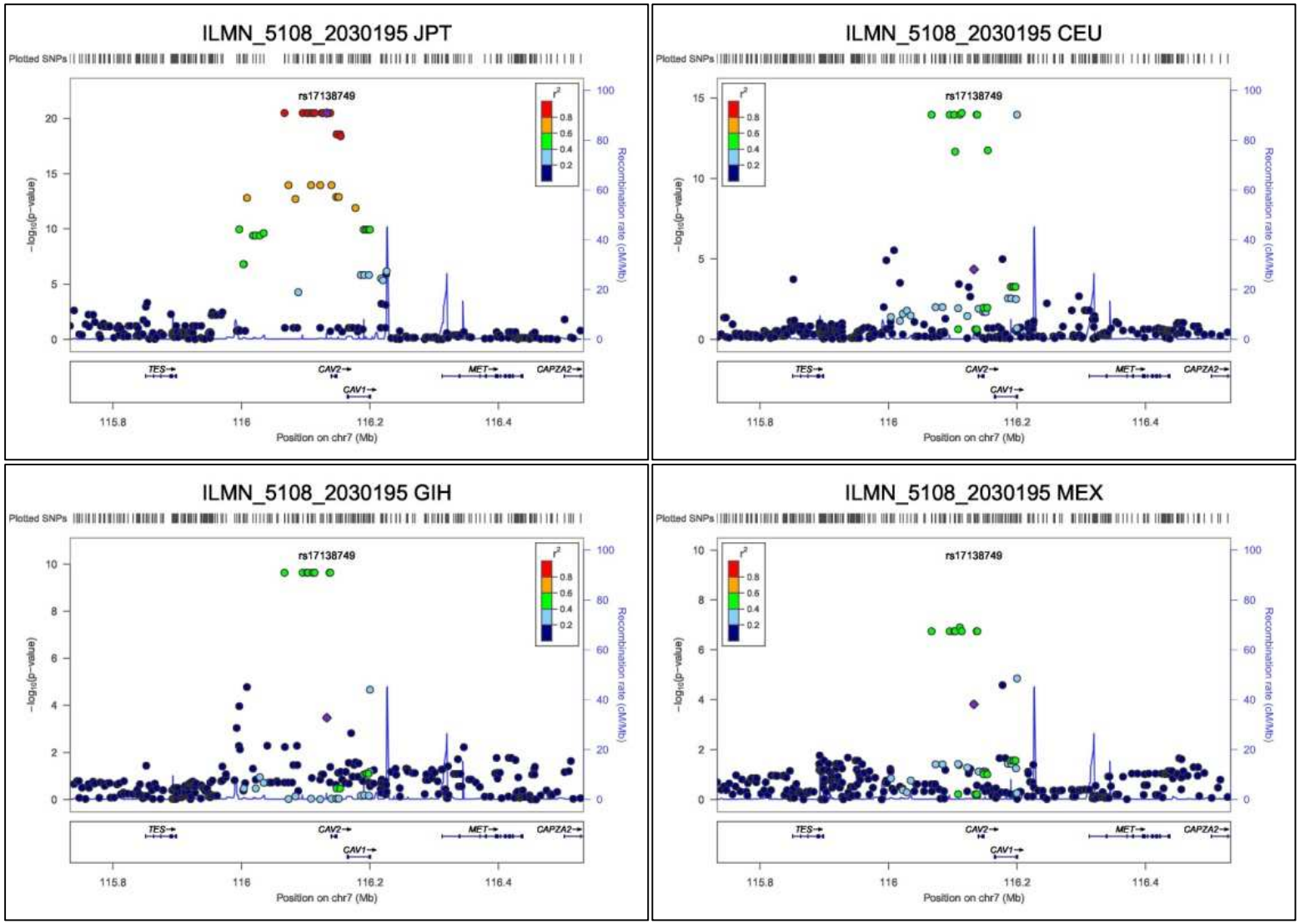


Figure 4.13: plots for peak Phase III HapMap SNPs for probe ILMN_5108_2030195. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

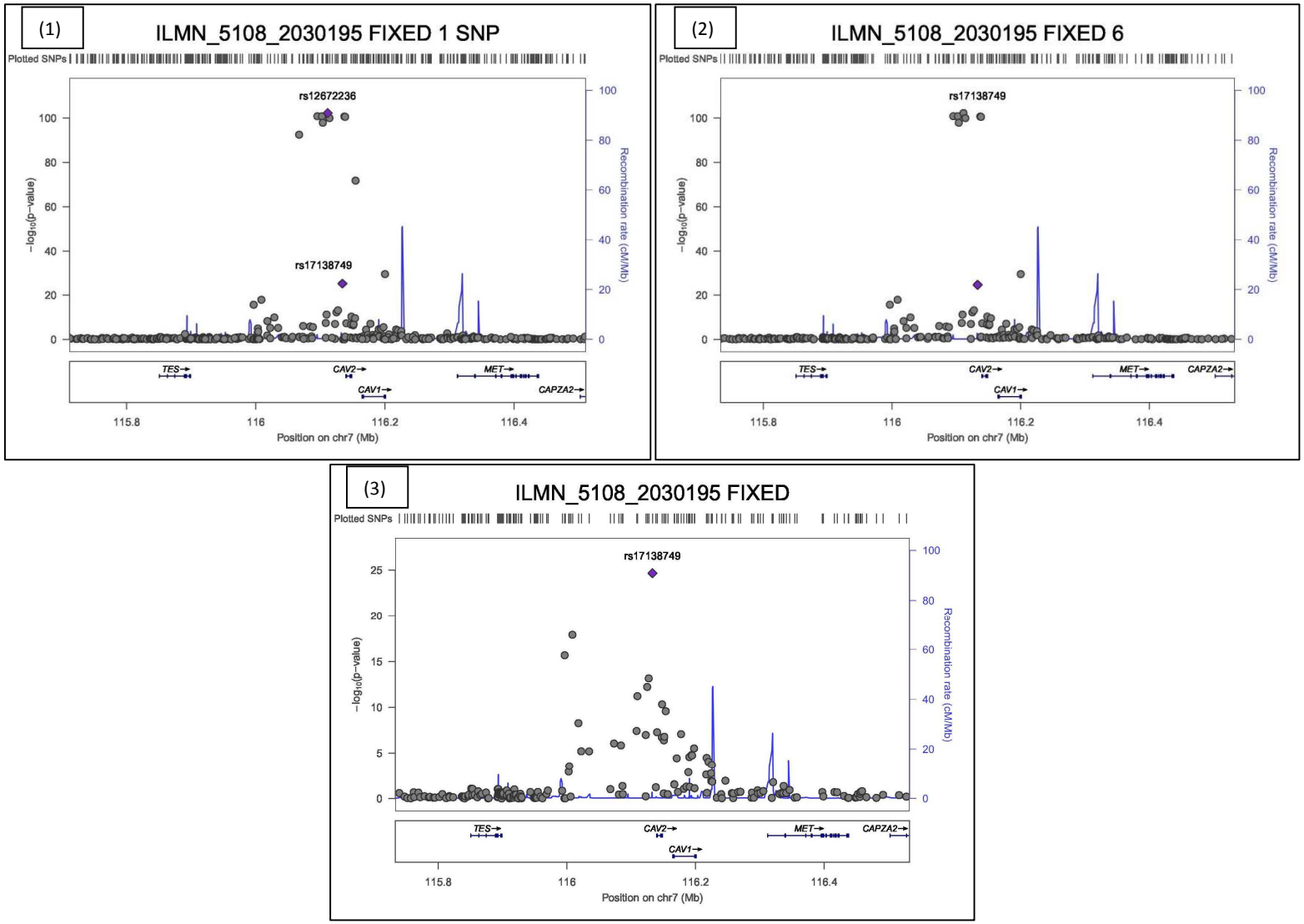


Figure 4.14: Signal plot for fixed effect meta-analysis for probe ILMN_5108_2030195. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) six or more SNPs not missing, (3) No missing SNPs: 8 SNPs present.

4.7.3 ILMN_13285_4210136 (ENSG00000187741, FANCA)

This probe has been selected as an example because it has strong evidence of heterogeneity in allelic effects between populations (Cochran's Q=92.34, $p < 2x10^{-16}$), and also a strong signal of association in the fixed effect meta-analysis (p-value = $9.23x10^{-89}$).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000187741 and the HGNC name *FANCA*. The full name of the gene is Fanconi Anaemia, Complementation Group A, and the gene's start position is chr16:89883065. The eSNP rs2239360 is located at chr16:89849583, and maps within an intron of *FANCA*. *FANCA* is a DNA repair protein that may operate in a post replication repair or cell cycle checkpoint function. The gene is a member of the Fanconi Anaemia complementation group (FANC). It may be involved in interstrand DNA cross link repair and in the maintenance of normal chromosome stability.

Figure 4.15 presents a forest plot for the results of the association analysis for the eSNP rs2239360 with the probe ILMN_13285_4210136 for *FANCA*. Some visual evidence of ancestry group clustering can be observed, with similar effect sizes within ancestry groups. The East Asian populations do not differ from zero at $p \leq 0.05$.

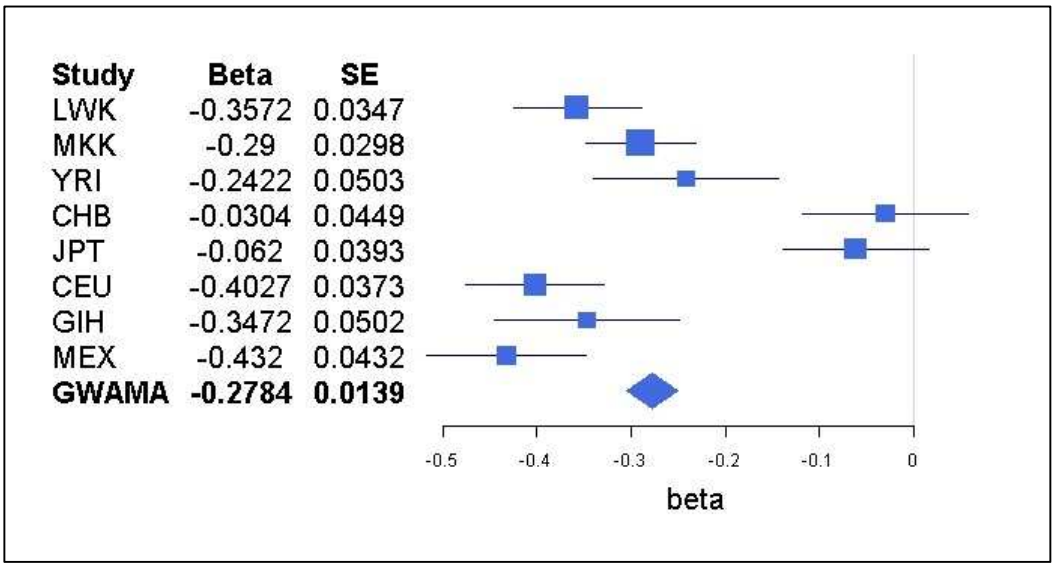


Figure 4.15: Forest plot of association and fixed effect meta-analyses for probe ILMN_13285_4210136 with SNP rs2239360.

Table 4.14 presents p-values and allele frequencies for the probe ILMN_13285_4210136 with SNP rs2239360. Allele frequencies range from 0.13 to 0.64. The association signals for the LWK, MKK, CEU, GIH and MEX populations achieve GWS. The fixed effect meta-analysis eSNP (rs2239360) is not the peak SNP in any of the population-specific association analyses.

Population	Minor Allele (C / T)	Allele Frequencies (Allele C)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	C	0.35	-0.3572	0.0347	2.56×10^{-16}	N
MKK	C	0.38	-0.2900	0.0298	3.32×10^{-17}	N
YRI	C	0.21	-0.2422	0.0503	4.98×10^{-6}	N
CHB	C	0.22	-0.0304	0.0449	0.500	N
JPT	C	0.13	-0.0620	0.0393	0.119	N
CEU	T	0.64	-0.4027	0.0373	1.20×10^{-18}	N
GIH	C	0.37	-0.3472	0.0502	1.60×10^{-9}	N
MEX	C	0.42	-0.4320	0.0432	3.32×10^{-12}	N
Fixed effect Meta-analysis	--	--	-0.2784	0.0139	9.23×10^{-89}	Y

Table 4.14: Table of allele frequencies for SNP rs2239360 with probe ILMN_13285_4210136.

Figures on the following pages show signal plots for population-specific association analyses (figure 4.16 and 4.17) and the fixed effect meta-analysis (figure 4.18). There is no evidence of any

association signal in the East Asian populations (CHB and JPT). Signals are observed at the six other populations. The eSNP rs2239360 is not the peak SNP in any of the population-specific analysis. When variants not reported in all eight populations are included in the fixed effect meta-analysis, the eSNP rs2239360 is no longer the peak SNP. Heterogeneity in allelic effects between populations may be due to: 1) the causal variant has not been genotyped and the East Asian populations do not have any tag SNPs for the causal variant; or 2) the peak SNP from the fixed effect meta-analysis is not the best tag SNP for the causal variant because it is not reported in all eight populations.

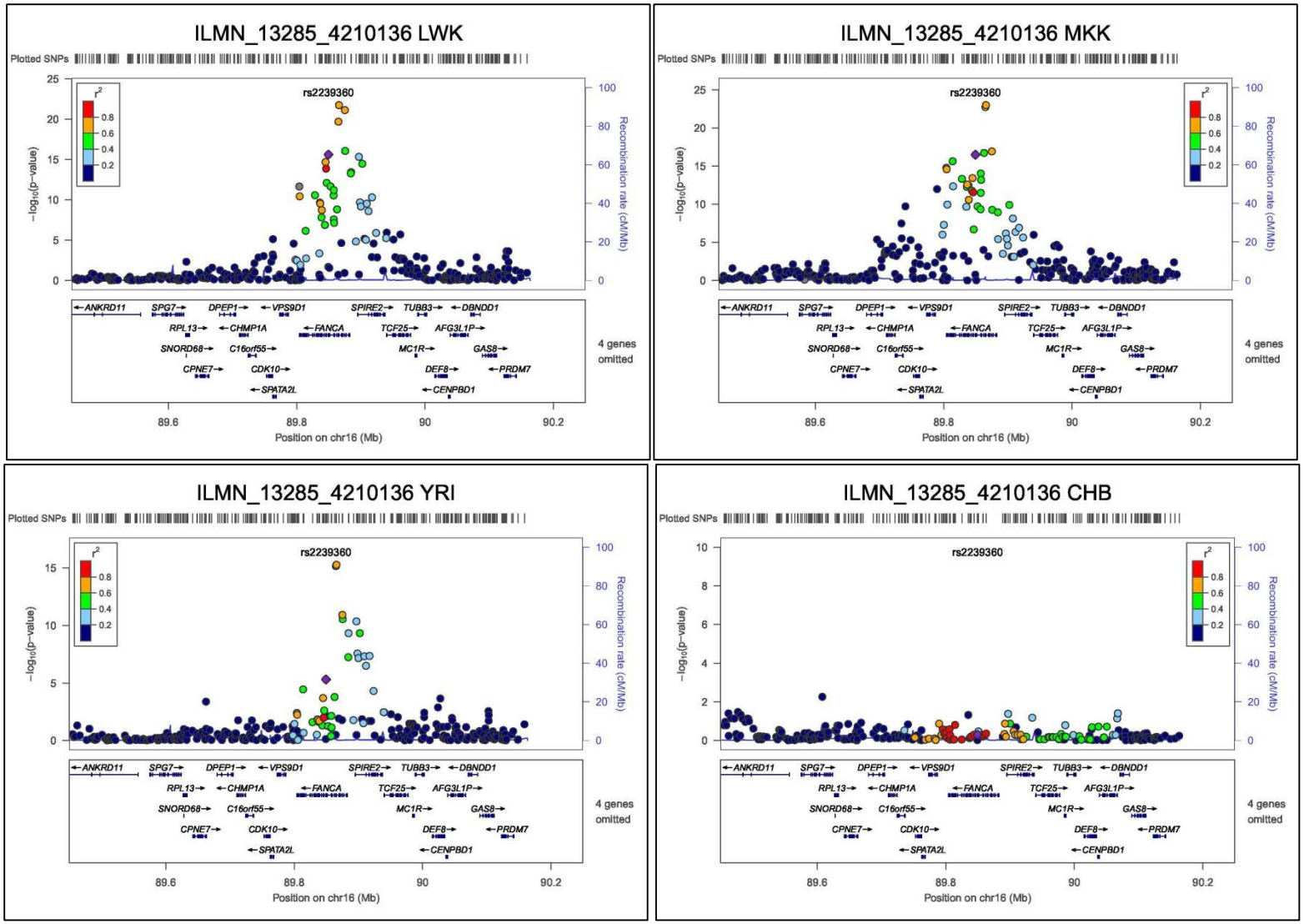


Figure 4.16: plots for peak Phase III HapMap SNPs for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

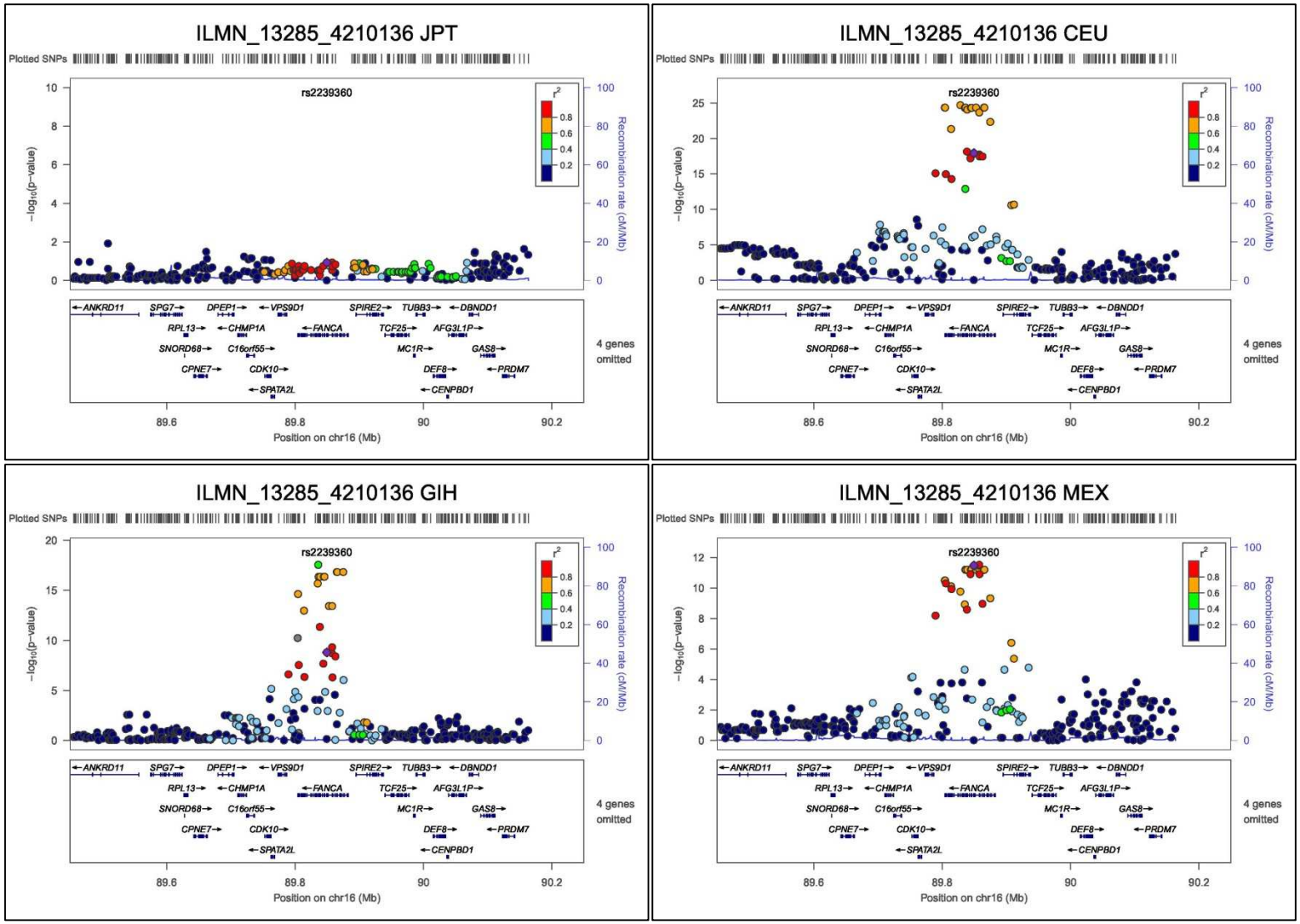


Figure 4.17: plots for peak Phase III HapMap SNPs for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

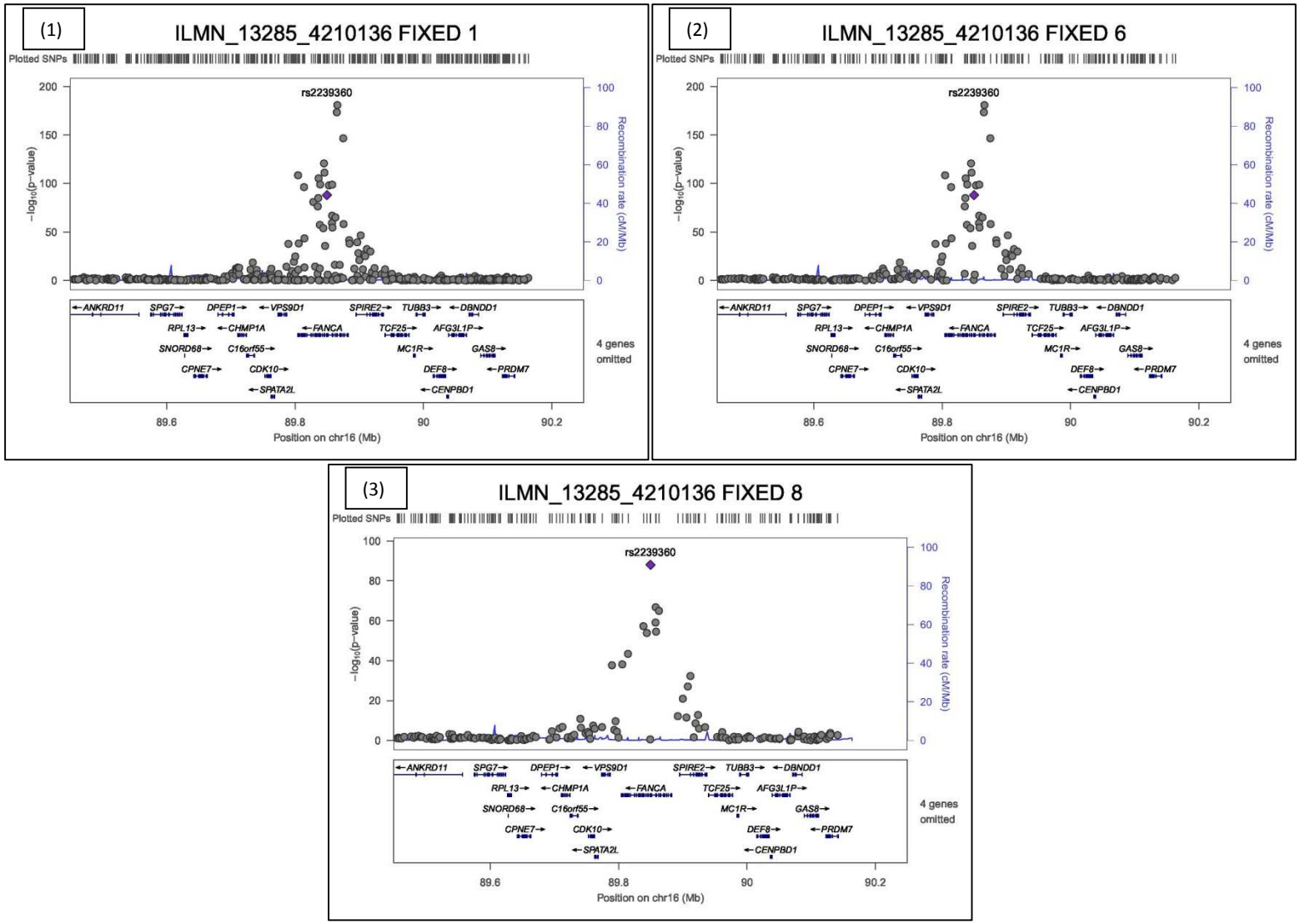


Figure 4.18: Signal plot for fixed effect meta-analysis for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) six or more SNPs not missing, (3) No missing SNPs: 8 SNPs present

4.7.4 ILMN_10409_6860670 (ENSG00000173915, USMG5)

This probe has been selected as an example because it has strong evidence of heterogeneity in allelic effects between populations (Cochran's Q= 75.89, p< 9.46×10^{-14}), and also a strong signal of association in the fixed effect meta-analysis (p-value= 2.00×10^{-101}). This probe is also in Phase II HapMap examples see section 3.6.3.

Gene name and location

The probe's gene has the Ensembl ID ENSG00000173915 and the HGNC name *USMG5*. The full name of the gene is Up-Regulated During Skeletal Muscle Growth 5, and the gene's start position is chr10: 105156223. The eSNP rs6580 is located at chr10:105206874, and maps within an exon of *CALHM2*. See section 3.6.3 for functional information for this gene.

Figure 4.19 presents a forest plot of association analysis and fixed effect meta-analysis results for the probe ILMN_10409_6860670 with the SNP rs6580. Visually there is evidence of ancestry group clustering. African effect sizes are weaker than the other ancestry groups.

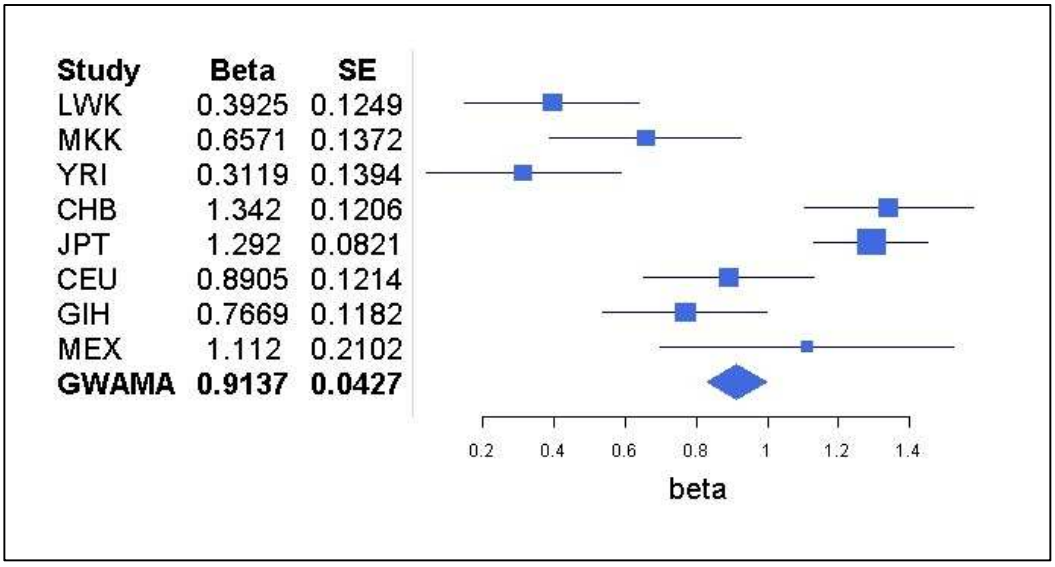


Figure 4.19: Forest plot of association and fixed effect meta-analyses for probe ILMN_10409_6860670 with SNP rs6580.

Table 4.15 presents allele frequencies, association analysis and fixed effect meta-analysis results for the probe ILMN_10409_6860670 with the SNP rs6580. Allele frequencies range from 0.11 to 0.56. All effect sizes are significant at $p \leq 0.05$. CHB, JPT, CEU and GIH are significant at GWS

Population	Minor Allele (C / T)	Allele Frequencies (Allele C)	Beta (Allele C)	SE	p-value
LWK	C	0.19	0.3925	0.1249	2.35×10^{-3}
MKK	C	0.22	0.6571	0.1372	4.41×10^{-6}
YRI	C	0.11	0.3119	0.1394	0.027
CHB	C	0.31	1.3420	0.1206	1.25×10^{-17}
JPT	C	0.39	1.2920	0.0821	6.36×10^{-26}
CEU	C	0.41	0.8905	0.1214	5.05×10^{-11}
GIH	T	0.56	0.7669	0.1182	9.64×10^{-9}
MEX	C	0.32	1.1120	0.2102	5.36×10^{-6}
Fixed effect Meta-analysis	--	--	0.9137	0.0427	2.00×10^{-101}

Table 4.15: Table of allele frequencies for SNP rs6580 with probe ILMN_10409_6860670.

Figures on the following pages show signal plots for population-specific association analyses (figure 4.20 and 4.21) and fixed effect meta-analysis (figure 4.22) for the probe ILMN_10409_6860670. The results are similar to the Phase II HapMap probe GI_14249375-S (see section 3.6.3), in both signals is very weak in the YRI population. When SNPs not reported are included in the fixed effect meta-analysis rs6580 is no longer the peak SNP (also seen in Phase II HapMap). Heterogeneity is due to difference in effect sizes in the African populations, due to no signal in YRI and weak signal in LWK and MKK. LWK and MKK do have more significant signals, but these are not detected in fixed effect meta-analysis due to SNPs not reported.

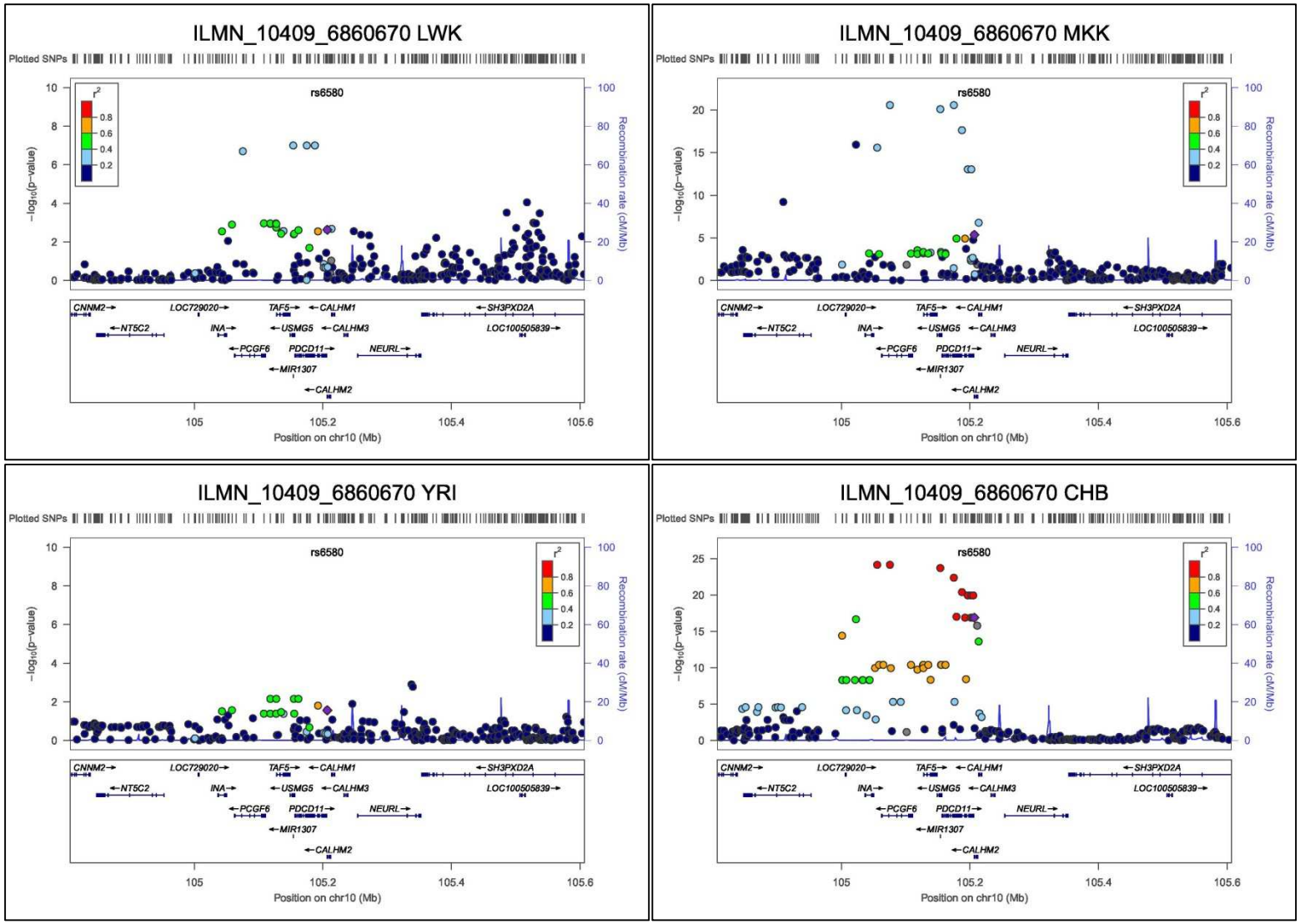
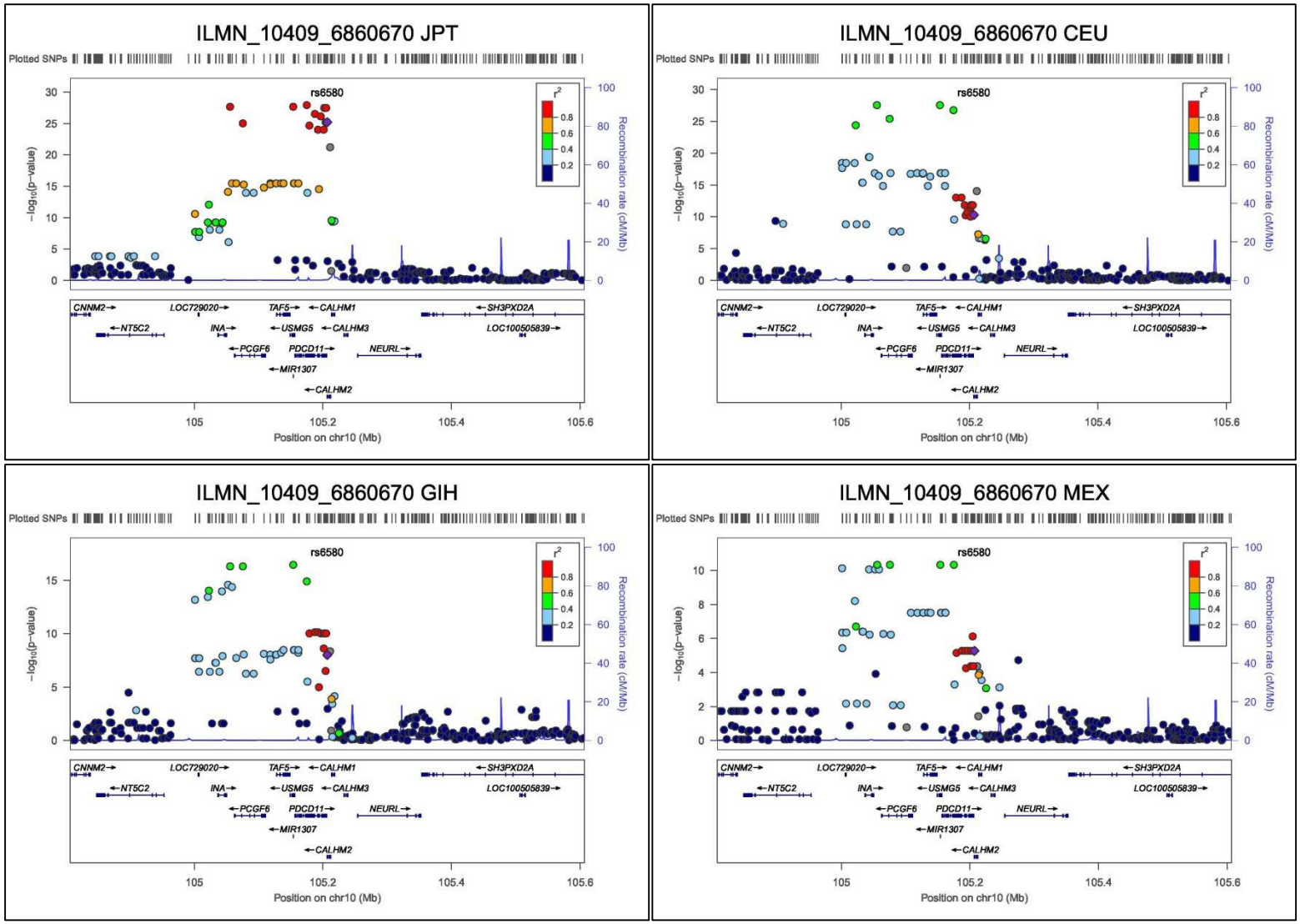


Figure 4.20: plots for peak Phase III HapMap SNPs for probe ILMN_10409_6860670. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).



4.21: plots for peak Phase III HapMap SNPs for probe ILMN_10409_6860670. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

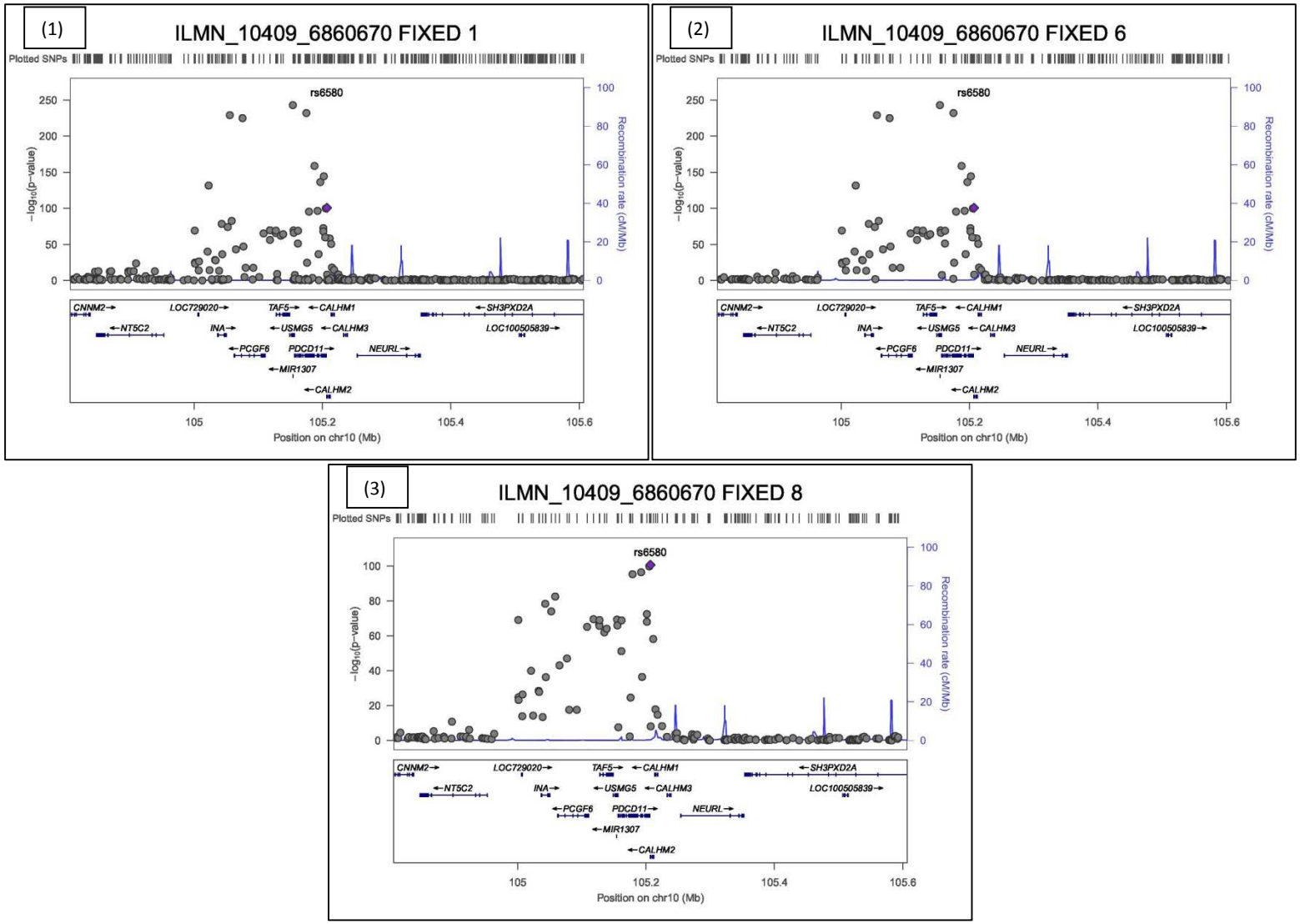


Figure 4.22: Signal plot for fixed effect meta-analysis for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap SNP. For each locus, the lead SNP is represented as a purple diamond. Numbering: (1) One or more SNPs not missing, (2) six or more SNPs not missing, (3) No missing SNPs: 8 SNPs present

4.8 Summary

This chapter presented the results of an association analysis and fixed effect meta-analysis to detect *cis* eQTLs in Phase III HapMap microarray expression and genotype data.

The aims of this analysis were to: 1) Improve power to detect *cis* eQTLs using fixed effect meta-analysis with the increased sample count in the Phase III HapMap dataset and 2) Detect and characterize effect size heterogeneity across populations.

In this chapter I have shown that merging association results from individual populations with fixed effect meta-analysis can improve power to detect *cis* eQTLs. I have also identified considerable heterogeneity of *cis* eQTL effect sizes between populations.

Results in this chapter were compared to those reported in the Stranger *et al* 2012 study which used the same genetic and expression data. Again, as observed in chapter 3, results between these studies differed when considering eQTLs detected using a FDR of 5%. It seems likely that this may be due to the differing methods to assess significance between these studies. Stranger 2012 uses permutations to calculate significance, whilst this study used the Benjamini-Hochberg method to account for multiple testing.

Three examples of heterogeneity in allelic effect between populations have been selected. The first for the gene CAV2 were heterogeneity is being driven by 1) no association signal in the YRI population. 2) Non-reported SNPs. The second example was for *FANCA* and is being driven by 1) no association signal in the CHB or JPT populations. 2) the peak SNP from the fixed effect meta-analysis is not the best tag SNP for the causal variant because it is not reported in all eight populations. The third example was for *USMG5*, which was also analysed in chapter 3, section 3.6.3. The peak SNP is different between the two datasets.

Taken together, these results suggest that much of the cause of heterogeneity in allelic effects between populations is likely to be due to differing LD structure between them. However,

examples where lack of any signal in YRI needs to be investigated in more detail (using 1000 Genomes variants) to rule out a biological cause for the lack of signal in this population.

One of the disadvantages of performing the analysis on the Phase III HapMap genotype data is that it is relatively sparse. To overcome this problem in the next chapter I present the results of imputing the phase III HapMap data with the March 2012 “All ancestries” 1000 Genomes Project reference panel.

CHAPTER 5 PHASE III HAPMAP IMPUTATION ANALYSIS

5.1 Overview

The previous chapter presented the results of association analysis and fixed effect meta-analysis on Phase III HapMap genotype and microarray expression data to detect *cis* eQTLs.

This chapter extends the analysis by introducing additional genetic variants to the Phase III HapMap genotypes through imputation up to the "all ancestries" March 2012 1000 Genomes Project Phase I reference panel (1000 Genomes Project Consortium *et al* 2010). The rationale behind this analysis is that imputation of the Phase III HapMap populations up to 1000 Genomes density will increase power to detect new *cis* eQTLs and improve fine mapping because the reference panel is more likely to include the causal variant (or a better tag for the causal variant) than HapMap.

As in the previous section, quantitative trait association analysis and fixed effect meta-analysis are performed, but here on the imputed genotypes and expression data, to detect *cis* eQTLs 1Mb upstream and downstream of each probe's gene transcription start site. Annotation of peak SNPs for the *cis* eQTLs detected are also presented.

Examples of *cis* eQTLs are selected for further interrogation, based on the following criteria:

1. Strongest association signal. The strongest overall *cis* eQTL meta-analysis p-value.
2. Heterogeneous signal. Using Cochran's Q statistic, eQTLs that show the strongest evidence of heterogeneity in allelic effects between populations.
3. Improvement in strongest association signal compared to Phase III HapMap when imputed SNPs are included in the analysis.
4. Change in evidence for heterogeneity in allelic effects between populations compared to Phase III HapMap when imputed SNPs are included in the analysis.

Positions of SNPs and genes use NCBI Build 37. Gene functions are not cited as the information has been taken from GeneCards and so is assumed to be public knowledge.

5.2 Imputation

5.2.1 Overview

The Phase III HapMap genotype data was imputed up to the 1000 Genomes "all ancestries" March 2012 reference panel using the software IMPUTE V2 (Howie *et al* 2009). For more information regarding IMPUTE V2, see methods section 2.6.1. A subset of the Phase III HapMap individuals used in this analysis has already been whole-genome sequenced as part of the 1000 Genomes Project reference panel (see table 2.2). Therefore imputation is only necessary for individuals present only in the Phase III HapMap dataset. Subsequently the imputed and directly genotyped subsets of individuals are merged for downstream association analysis.

IMPUTE requires a genotype "scaffold" as an input, which are the known genotypes present in both the imputation reference panel and the genotype data to be imputed. The scaffold was generated by starting with the QC filtered Phase III HapMap genotypes and applying a filter of MAF ≥ 1% instead of 5% and updating genotypes from NCBI Build 36 to Build 37. For more information regarding the imputation scaffold pipeline, see Methods section 2.3.2. Each population was imputed separately.

On completion of imputation, each imputed SNP for an individual is assigned a posterior distribution of possible genotypes. Across all individuals in a population, this distribution is used to calculate an expected allele frequency and a QC metric named the INFO score. As with most traditional GWAS, variants have been discarded from downstream analysis if the INFO ≤ 0.4 and, because of the small sample size, MAF ≤ 5%. See Winkler *et al* 2014 for rationale of INFO score threshold. For more information regarding the INFO score, see Methods section 2.6.2.

5.2.2 Imputation results

This section presents the results of the imputation which are shown in table 5.1. The table has the following columns:

1. **Total:** This is the total number of variants in the 1000 Genomes reference panel for which imputation has been performed.
2. **Total INFO ≥ 0.4:** The total number of imputed variants with an INFO score ≥ 0.4
3. **MAF ≥ 5%, INFO ≥ 0.4:** The total number of imputed variants with MAF ≥ 5% and an INFO score ≥ 0.4

In addition, the row "**Intersection"** specifies the number of variants reported in all eight populations. **"All variants"** specifies the counts of all types of genetic variants in the 1000 Genomes datasets: including SNPs, indels and CNVs.

The numbers of variants imputed are similar between members of ancestry groups: (African ~$8x10^6$, East Asian: ~$5.5x10^6$, Eurasian-Hispanic ~$5.5x10^6$, at MAF ≥ 5%, INFO ≥ 0.4). The MEX population has a lower number of imputed genotypes which passed QC than the other populations, which is likely due to sample size, which is lower than the other populations (41). As expected, African ancestry populations have greater diversity in genetic variation because they are more ancient. The number of SNPs which can be used in the fixed effect meta-analysis that are reported in all eight populations is 2,831,770.

Population	All Variants			SNPs only		
	Total	Total INFO ≥ 0.4	MAF ≥ 5% INFO ≥ 0.4	Total	Total INFO ≥ 0.4	MAF ≥ 5% INFO ≥ 0.4
LWK	37,969,374	13,762,955	8,988,265	36,331,747	12,738,665	8,198,828
MKK	38,047,942	16,726,806	8,682,327	36,407,286	15,592,688	7,912,230
YRI	38,047,573	15,555,610	9,069,395	36,406,914	14,471,477	8,271,243
CHB	38,046,817	7,550,870	5,777,953	36,406,150	6,920,542	5,268,239
JPT	38,046,793	7,269,622	5,676,991	36,406,124	6,655,741	5,172,250
CEU	38,047,111	8,735,642	6,346,281	36,406,444	8,007,176	5,782,964
GIH	38,047,322	8,964,707	6,462,792	36,406,661	8,231,753	5,895,132
MEX	36,061,698	5,857,893	5,814,299	34,506,384	5,349,404	5,308,757
Intersection	--	4,047,027	3,110,069	--	3,693,201	2,831,770

Table 5.1: Results of Phase III HapMap 1000 Genomes imputation. MAF: Minor allele frequency, INFO: IMPUTE V2 quality metric.

5.2.3 Merging dosage data

As described above, some individuals within the Phase III HapMap individuals are also present in the 1000 Genomes reference panel. For an overview of individuals in both datasets, see Methods section 2.2.1. For these individuals, it is not necessary to perform imputation. The 1000 Genomes genotype data was converted to IMPUTE GEN file format, where each SNP for each individual is represented by three probabilities, one for each possible genotype. The imputed GEN files and 1000 Genomes GEN files were merged using the software GTOOL. See methods section 2.6.4 for full details of merging datasets.

Table 5.2 shows the results of the merged genotype data. The row **"union"** is the number of variants which are reported in at least one population. Variants are filtered for INFO score ≥ 0.4 and MAF ≥ 0.05. The intersection of SNPs which can be used in the fixed effect meta-analysis that are reported in all eight populations is 2,841,022. This is this is a substantial improvement over Phase III HapMap genotype data (708,270) and Phase II HapMap genotype data (1,271,775).

Populations	All Variants	SNPs
LWK	8,556,217	7,960,451
MKK	8,632,085	7,865,973
YRI	9,051,461	8,270,990
CHB	5,659,426	5,261,079
JPT	5,692,003	5,199,500
CEU	6,318,163	5,769,589
GIH	6,418,352	5,854,010
MEX	4,972,856	5,854,010
Intersection	3,016,504	2,841,022
Union	12,345,574	11,283,547

Table 5.2: Results of merged genotype data for MAF 5% INFO 0.4, all variants and SNPs only.

All variants other than SNPs have been removed from the downstream analyses.

5.3 Multiple Testing

As with previous analyses, we have considered two approaches to adjusting for multiple testing: GWS and FDR of 5% using the Benjamini-Hochberg procedure. See Methods section 2.10 for full description of multiple testing procedures used.

5.3.1 FDR – Benjamini-Hochberg

Table 5.3 shows the p-values for FDR 5% determined using the Benjamini-Hochberg procedure. See 2.10.2 for full description of the Benjamini-Hochberg procedure. For the fixed effect analysis results, the FDR thresholds is determined by using the union which is the total number of tests performed when including all variants reported in at least one population.

Population	Tests performed	FDR 5% p-value
LWK	115,913,466	1.29×10^{-5}
MKK	116,120,257	1.92×10^{-5}
YRI	122,099,253	1.27×10^{-5}
CHB	75,321,283	4.19×10^{-5}
JPT	75,776,024	4.88×10^{-5}
CEU	84,635,719	3.62×10^{-5}
GIH	85,347,207	2.45×10^{-5}
MEX	65,094,032	1.21×10^{-5}
Fixed effect Meta-analysis	166,850,680	1.69×10^{-4}

Table 5.3: Phase III HapMap FDR 5% p-value threshold levels, after 1000 Genomes Project imputation, generated by Benjamini-Hochberg procedure

5.4 Association Analysis

Correction for population structure proceeded in the same way as in the Phase III HapMap only analysis, where the expression data from the admixed populations (GIH, LWK, MEX and MKK) were corrected for eigenvectors from EIGENSTRAT. See section 4.3 for more information on population structure in the Phase III HapMap populations.

Quantitative trait association analysis was performed on the merged GEN files generated from imputation with the microarray expression intensity data. The imputed genotypes were treated as dosages with the command –dosage if the software PLINK (Purcell *et al* (2007). The analysis included sex as a covariate. See Methods section 2.7.1 for more information on imputed association analysis.

5.4.1 QQ Plots

QQ plots were generated using the imputed variants p-values from association analysis against the null distribution. See Methods section 2.4.1 for more details on QQ plots. Figure 5.1 and 5.2 shows QQ plots for the association analyses. The plots show a departure from the uniform distribution for each of the association results at approximately $-\log_{10}$ p-value = 3.

5.4.2 Genomic Control (λ)

To determine whether there is any substantial population structure within any of the populations that has not been accounted for in the association analysis, the GC inflation factor (λ) was calculated. See Methods Section 2.4.2 for definition and calculation of GC (λ).

Table 5.4 shows the GC inflation factors (λ) for each population. As in the analysis of these data before imputation, the strongest GC (λ) inflation was observed in the MKK (Maasai) population (1.072). As described in section 4.5.2, we expect some inflation, due to a strong prior of association of SNPs in cis eQTL studies. GC (λ) between non-imputed and imputed variants is very similar.

The table also shows the peak p-value for the association. The MEX population has the weakest peak association signal, likely to be due to the small population size (41).

Population	GC (λ) before imputation	GC (λ) after imputation	Peak p-value
LWK	1.023	1.027	1.94×10^{-28}
MKK	1.072	1.072	8.66×10^{-31}
YRI	1.024	1.029	6.73×10^{-33}
CHB	1.029	1.034	1.27×10^{-28}
JPT	1.036	1.037	2.11×10^{-29}
CEU	1.030	1.043	4.56×10^{-32}
GIH	1.030	1.032	2.46×10^{-25}
MEX	1.028	1.037	7.25×10^{-16}

Table 5.4: GC (λ) values for population-specific association analyses. Also shown is the lowest p-value detected in each analysis.

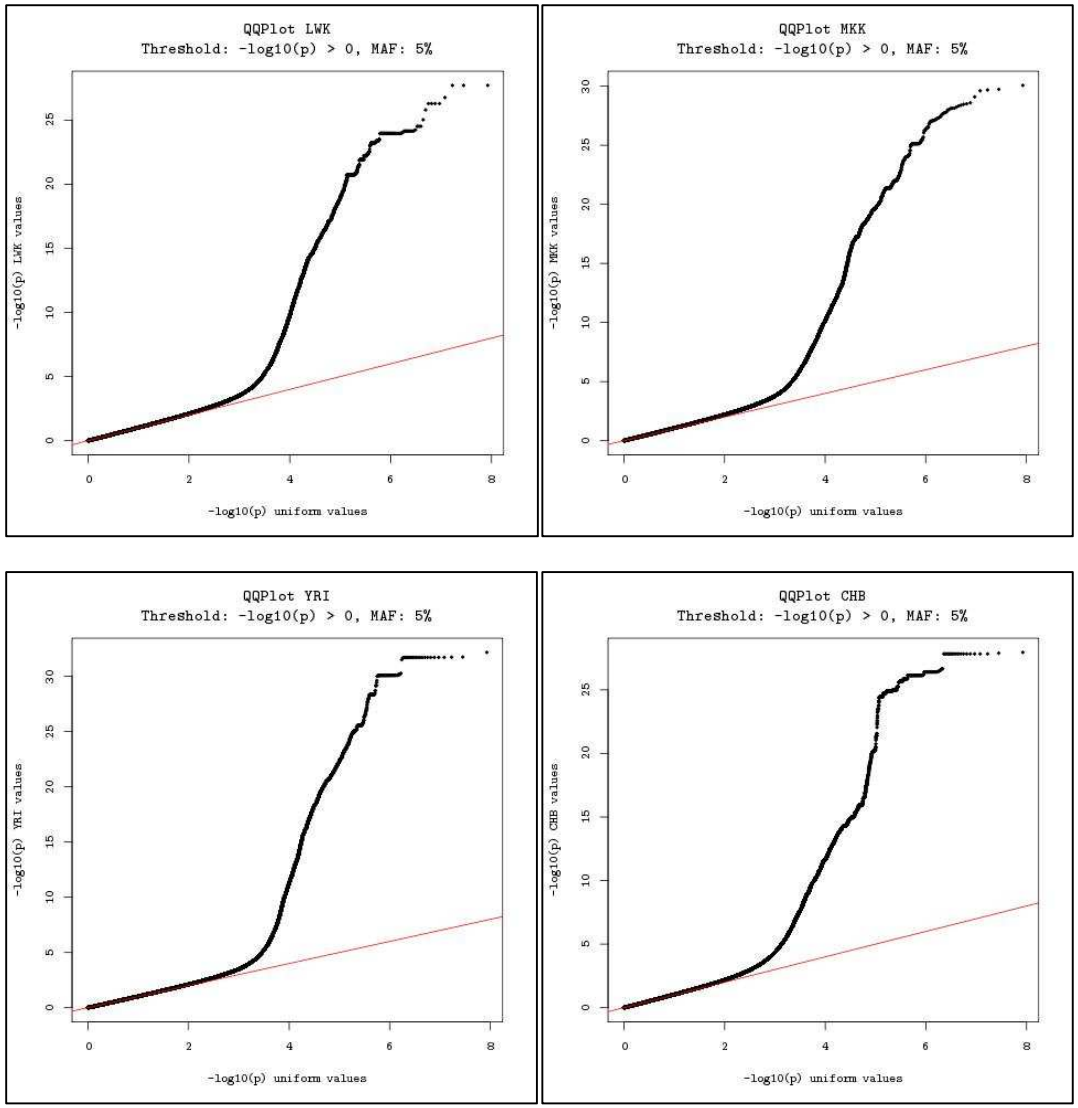


Figure 5.1: QQ plots of association analysis p-values against the uniform distribution

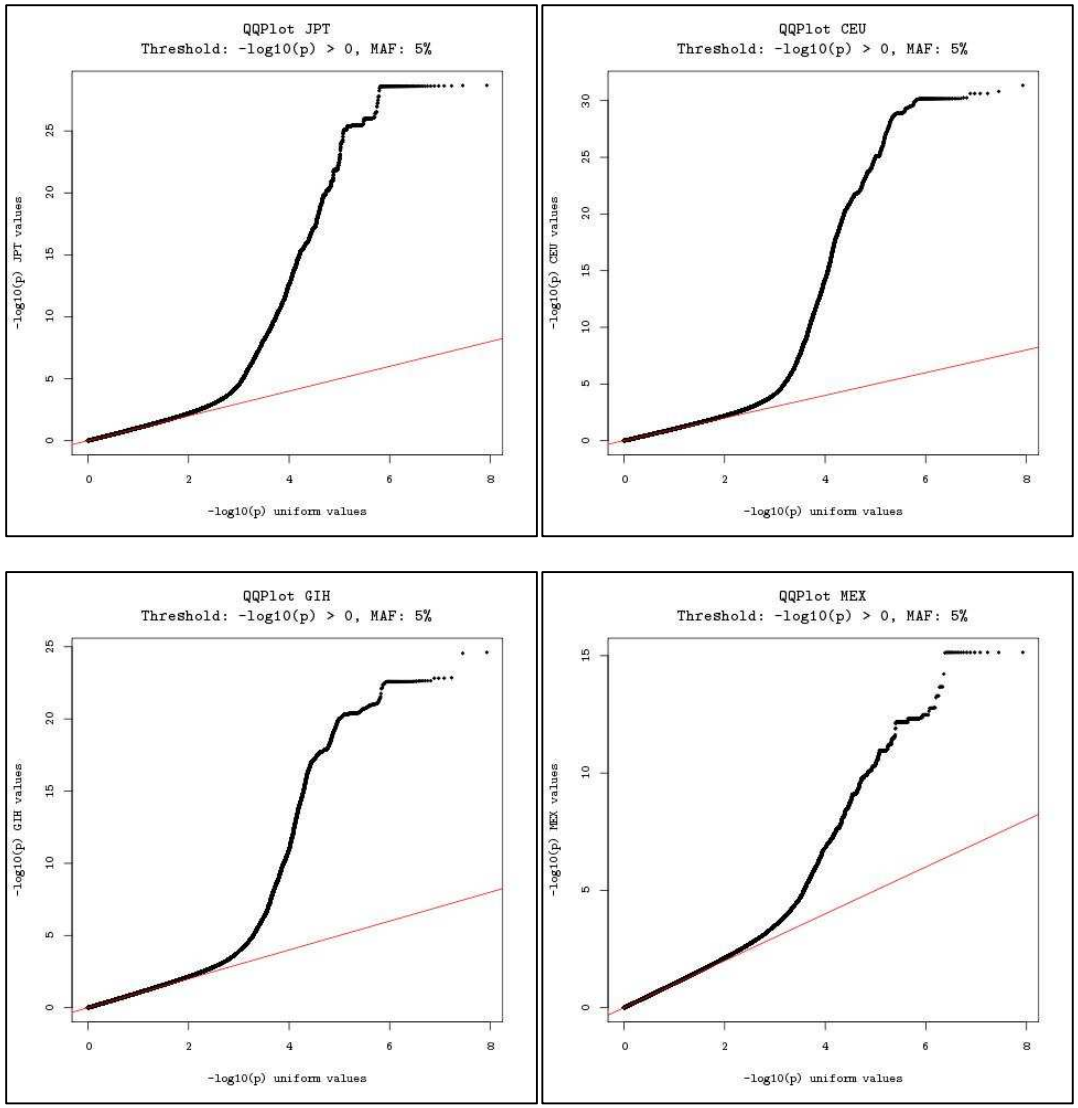


Figure 5.2: QQ plots of association analysis p-values against the uniform distribution

5.4.3 GWS ($p \leq 5x10^{-8}$)

Table 5.5 summarises the results of quantitative trait association scans for *cis* eQTLs at GWS ($p \leq 5x10^{-8}$). The counts of probes range from 322 (JPT) to 91 (MEX), where the low number in MEX is likely to be due to the low population size (41) in comparison with the other groups.

Population	Probes	Peak SNPs
LWK	297	571
MKK	334	338
YRI	304	335
CHB	303	740
JPT	322	468
CEU	227	255
GIH	199	208
MEX	91	142

Table 5.5: Summary of *cis*-eQTLs for imputed Phase III HapMap at GWS ($p < 5x10^{-8}$).

In total, there is a non-redundant set of 843 probes showing a significant association in one or more populations. A total of 477 probes have significant associations detected in 2 or more populations and 22 probes have significant associations across all eight populations.

5.4.4 False Discovery Rate (FDR) ≤ 5%

Table 5.6 summarizes the results with FDR of 5%, determined using the Benjamini-Hochberg method (see table 5.3 for thresholds). The counts of probes range from 1640 (JPT) to 467 (MEX). As in the case of using GWS, the MEX ancestry group has the lowest number of *cis* eQTLs probes.

Population	Probes	Peak SNPs
LWK	1323	1908
MKK	1873	1934
YRI	1323	1386
CHB	1583	2756
JPT	1640	2125
CEU	1390	1540
GIH	989	1026
MEX	467	770

Table 5.6: *cis* eQTL signals at 5% FDR (determined using Benjamini Hochberg method).

In total, there is a non-redundant set of 6275 probes showing a significant association in one or more populations. At total of 1905 probes have significant associations detected in 2 or more populations and 85 probes have significant associations across all eight populations.

5.5 Fixed Effect Meta-Analysis

Fixed effect meta-analysis was performed using the software GWAMA (Magi *et al* 2010). Summary statistics (effect sizes and standard errors) from the population-specific association analyses were used as input. See Methods section 2.8 for background on fixed effect meta-analysis.

5.5.1 QQ Plots

QQ plots were generated by plotting the observed p-values from the fixed effect meta-analysis against those expected from the null distribution. See Methods section 2.5.1 for more details on QQ plots. Figure 5.3 shows QQ plots for the fixed effect meta-analysis. The plots show a departure from the null distribution at approximately -$\log_{10}$ p-value = 2. As described previously inflation is expected in *cis* eQTL studies because there is a stronger prior for association than in GWAS.

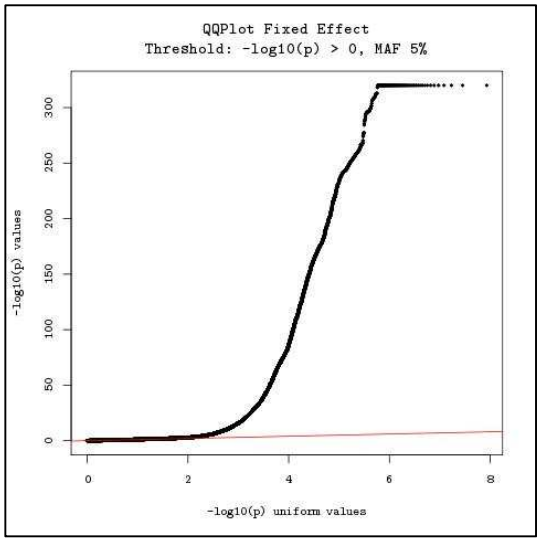


Figure 5.3: QQ plot of fixed effect meta-analysis p-values against the null distribution

The genomic control inflation factor was calculated for the fixed effect meta-analysis using the union (1.116) and intersection (1.145) of the SNPs in the association analysis. These values are

similar to GC calculated with the non-imputed Phase III HapMap dataset (1.119). The peak signal from the meta-analysis has p-value = 1.69×10^{-312}.

5.5.2 Signal counts

Table 5.7 shows the results of the fixed effect meta-analysis at GWS ($p \leq 5 \times 10^{-8}$) and FDR 5%. Results with differing numbers of not reported variants are reported. For calculating the p-value for FDR of 5%, the union of the SNPs across the eight populations was used (see Table 5.3).

Significance Level	Minimum Population Count	Probes	Peak SNPs
GWS	8	1811	1816
	6	2176	2190
	1	2558	2580
FDR 5%	8	6226	6247
	6	8146	8181
	1	13,045	13,348

Table 5.7: Results for fixed effect meta-analysis at GWS ($p < 5 \times 10^{-8}$) and FDR 5%.

5.5.3 Comparison of peak signals before and after imputation

Out of the 1811 probes detected at GWS with eSNPs reported in all eight populations, 1238 had an improvement in signal with the imputed SNPs. Figure 5.4 shows a comparison of peak signals before and after imputation. To generate this plot the peak SNP from the fixed effect meta-analysis was plotted against the peak Phase III HapMap SNP. 573 of the probes had a Phase III HapMap SNP as the peak signal.

An example where the p-value is substantially better after imputation is the probe ILMN_20456_4590554 for the gene *CORO2A*. The imputed SNP rs28703824 has a z-score of 20.68 and p-value of 6.01×10^{-95}. The non-imputed SNP rs10818610 has a z-score of 9.32 and a p-value of 1.26×10^{-20}. This has the largest change in z-score of all the probes. (This example is investigated in more detail in section 5.7.4).

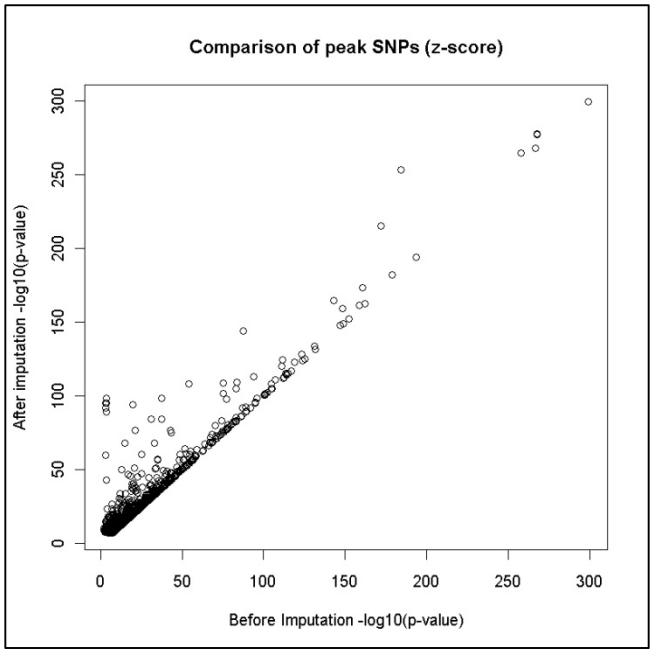


Figure 5.4 Comparison of peak SNPs, before and after imputation.

5.5.4 Top results from fixed effect meta-analysis

Table 5.8 shows the ten *cis* eQTL probes with the strongest signals of association after 1000 Genomes imputation. See appendix A.6. It was found that six of the probes also contained one or more 1000 Genomes SNPs and so maybe false-positives.

Probe	**HGNC**	**SNP**	**Beta**	**SE**	**z-score**	**p-value**	**Q p-value**	**1000 Genomes SNP**
ILMN_24844_2100600	*IRF5*	rs10954213	-1.16	0.027	-43.34	$p < 1.69x10^{-312}$	0.03	N
ILMN_22173_5690095	*ERAP2*	rs2549782	1.70	0.041	41.25	$p < 1.69x10^{-312}$	0.02	N
ILMN_2170_6860347	*WBSCR27*	rs10225983	-0.72	0.019	-37.01	$6.99x10^{-300}$	0.23	Y
ILMN_7731_3520685	*C17orf97*	rs7502594	0.85	0.024	35.65	$2.13x10^{-278}$	$1.29x10^{-8}$	N
ILMN_7731_2570703	*C17orf97*	rs7502594	1.24	0.035	35.62	$6.35x10^{-278}$	$5.00x10^{-9}$	Y
ILMN_15237_5860053	*IPO8*	rs28661513	-0.63	0.018	-35.02	$1.23x10^{-268}$	0.01	Y
ILMN_20550_7330093	*HLA-DRB1*	rs9271073	3.61	0.104	34.79	$3.10x10^{-265}$	0.67	Y
ILMN_25790_1110273	*SLFN5*	rs11080327	0.77	0.023	34.04	$6.52x10^{-254}$	0.02	N
ILMN_3178_4390692	*HLA-DRB5*	rs115466555	-4.40	0.140	-31.38	$4.10x10^{-216}$	0.68	Y
ILMN_18858_160242	*PKHD1L1*	rs1783166	-0.58	0.020	-29.77	$1.13x10^{-194}$	$2.35x10^{-6}$	Y

Table 5.8: Ten probes from fixed effect meta-analysis with largest absolute Z score in 1000 Genomes.

Table 5.9 presents the peak Phase III HapMap association results for the same ten SNPs. Association signals are of the same order of magnitude or stronger after imputation, with similar effect sizes and similar extent of heterogeneity in allelic effects between populations.

Probe	**HGNC**	**SNP**	**Beta**	**SE**	**z-score**	**p-value**	**Q p-value**
ILMN_24844_2100600	*IRF5*	rs6965542	1.15	0.027	42.90	$p < 1.69x10^{-312}$	0.028
ILMN_22173_5690095	*ERAP2*	rs2549782	1.70	0.041	41.26	$p < 1.69x10^{-312}$	0.017
ILMN_2170_6860347	*WBSCR27*	rs13223593	0.72	0.019	37.01	$9.30x10^{-300}$	0.229
ILMN_7731_3520685	*C17orf97*	rs11150882	0.86	0.025	35.02	$1.38x10^{-268}$	$1.46x10^{-8}$
ILMN_7731_2570703	*C17orf97*	rs11150882	1.25	0.036	35.00	$2.37x10^{-268}$	$6.56x10^{-9}$
ILMN_15237_5860053	*IPO8*	rs7327	-0.63	0.018	-34.94	$2.01x10^{-267}$	0.011
ILMN_20550_7330093	*HLA-DRB1*	rs9271170	-3.59	0.104	-34.36	$1.11x10^{-258}$	0.616
ILMN_25790_1110273	*SLFN5*	rs883416	0.72	0.025	28.99	$1.05x10^{-184}$	$5.37x10^{-3}$
ILMN_3178_4390692	*HLA-DRB5*	rs9271170	-3.66	0.131	-28.01	$1.50x10^{-172}$	$7.77x10^{-3}$
ILMN_18858_160242	*PKHD1L1*	rs1783170	0.58	0.020	29.73	$4.28x10^{-194}$	$2.19x10^{-6}$

Table 5.9: Ten probes from fixed effect meta-analysis with largest absolute z-score in Phase III HapMap

5.5.5 Heterogeneity

To investigate heterogeneity in allelic effects between populations, Cochran's Q statistic has been calculated at each of the detected *cis* eQTLs at GWS and FDR 5%. Table 5.10 shows the number of SNPs detected at the two significance thresholds that show evidence of heterogeneity in allelic effects between populations at Cochran's Q p-value < $1x10^{-3}$. For more information about the Cochran's Q statistic, see Methods section 2.8.2. As can be seen, the number of heterogeneous signals is greater than that expected by chance (see expected signals and binomial test p-value).

When fixed effect meta-analysis results are considered for variants not reported in all eight populations, the number of heterogeneous signals decreases. This would be expected since fewer populations are being considered, and therefore heterogeneity in allelic effects between populations is less likely to occur (for example, you cannot observe heterogeneity at a SNP reported in just one population, and you are less likely to observe heterogeneity at a SNP reported only in populations from one ancestry group).

Significance Level	Min Population Count	Peak	Heterogeneous Signals	Expected	Binomial Test p-value
GWS	8	1816	180	~2	< $2.2x10^{-16}$
	6	2190	149	~2	< $2.2x10^{-16}$
	1	2580	118	~2.5	< $2.2x10^{-16}$
FDR 5%	8	6247	233	~6	< $2.2x10^{-16}$
	6	8181	200	~8	< $2.2x10^{-16}$
	1	13,348	156	~13	< $2.2x10^{-16}$

Table 5.10: Imputed Dataset: Summary of heterogeneity results with GWS (p-value < $5x10^{-8}$) and FDR 5%. Heterogeneity is assessed with Cochran's Q (p-value < $1x10^{-3}$).

5.5.6 Top heterogeneous results

Table 5.11 shows the ten *cis* eQTLs detected in the fixed effect meta-analysis with the largest Cochran's Q-statistic. Table 5.12 shows the corresponding Phase III HapMap SNP for the ten probes in table 5.11. Results are consistent before and after imputation in terms of heterogeneity. See Appendix A.7. It was found that nine of the probes also contained one or more 1000 Genomes SNPs and so maybe false positives.

Probe	HGNC	SNP	Beta	SE	z-score	p-value	Cochran's Q	Q p-value	1000 Genomes SNP
ILMN_5108_2030195	*CAV2*	rs13235183	-0.45	0.031	-14.57	4.72×10^{-48}	182.12	$p < 2 \times 10^{-16}$	Y
ILMN_128814_5960180	*TMED4*	rs217378	0.34	0.026	13.12	2.62×10^{-39}	178.92	$p < 2 \times 10^{-16}$	Y
ILMN_5108_2260136	*CAV2*	rs13235183	-0.43	0.030	-14.24	5.69×10^{-46}	168.51	$p < 2 \times 10^{-16}$	Y
ILMN_4387_7210035	*FAM154B*	rs11635460	0.07	0.011	6.84	7.95×10^{-12}	101.97	$p < 2 \times 10^{-16}$	N
ILMN_13285_4210136	*FANCA*	rs8051231	-0.27	0.013	-20.49	2.52×10^{-93}	101.34	$p < 2 \times 10^{-16}$	Y
ILMN_2228_6250228	*STOX1*	rs4746792	-0.05	0.006	-8.85	9.10×10^{-19}	94.27	$p < 2 \times 10^{-16}$	Y
ILMN_27741_6650402	*SERF2*	rs2467402	0.11	0.007	14.25	4.95×10^{-46}	94.15	$p < 2 \times 10^{-16}$	Y
ILMN_2021_2350243	*UGT2B17*	rs139440909	1.78	0.086	20.79	5.94×10^{-96}	88.21	3.33×10^{-16}	Y
ILMN_22479_2900379	*UGT2B7*	rs139440909	1.58	0.075	21.15	2.77×10^{-99}	85.76	8.88×10^{-16}	Y
ILMN_21130_5090184	*HEBP2*	rs6919872	1.36	0.050	27.22	4.42×10^{-163}	85.21	1.22×10^{-15}	Y

Table 5.11: Ten probes with greatest heterogeneity (largest Cochran's Q-statistic) for imputed SNPs.

Probe	HGNC	SNP	Beta	SE	z-score	p-value	Cochran's Q	Q p-value
ILMN_5108_2030195	*CAV2*	rs17138749	-0.24	0.0227	-10.53	7.04×10^{-26}	212.75	$p < 2 \times 10^{-16}$
ILMN_128814_5960180	*TMED4*	rs217378	0.34	0.0257	13.12	2.62×10^{-39}	178.92	$p < 2 \times 10^{-16}$
ILMN_5108_2260136	*CAV2*	rs17138749	-0.23	0.0226	-10.03	1.20×10^{-23}	191.32	$p < 2 \times 10^{-16}$
ILMN_4387_7210035	*FAM154B*	rs11635460	0.07	0.0109	6.84	7.95×10^{-12}	101.97	$p < 2 \times 10^{-16}$
ILMN_13285_4210136	*FANCA*	rs2239360	0.28	0.0139	20.08	1.27×10^{-89}	94.76	$p < 2 \times 10^{-16}$
ILMN_2228_6250228	*STOX1*	rs7081960	-0.05	0.0062	-8.35	7.44×10^{-17}	79.75	1.55×10^{-14}
ILMN_27741_6650402	*SERF2*	rs2467402	0.11	0.0075	14.25	4.95×10^{-46}	94.15	$p < 2 \times 10^{-16}$
ILMN_2021_2350243	*UGT2B17*	rs34572119	0.40	0.1198	3.38	7.41×10^{-4}	18.03	0.012
ILMN_22479_2900379	*UGT2B7*	rs34572119	0.37	0.1049	3.56	3.79×10^{-4}	19.44	6.90×10^{-3}
ILMN_21130_5090184	*HEBP2*	rs6919872	1.36	0.0500	27.22	4.42×10^{-163}	85.21	1.22×10^{-15}

Table 5.12: Ten probes with greatest heterogeneity (largest Cochran's Q-statistic) for Phase III HapMap SNPs.

5.5.7 Comparison of Cochran's Q p-values before and after imputation.

In order to determine whether heterogeneity is increasing or decreasing between the peak 1000 Genomes and Phase III HapMap SNPs, the peak SNP Cochran's Q -$\log_{10}$ p-value has been plotted against the peak Phase III HapMap SNP Q –$\log_{10}$ p-value. Figure 5.5 shows a scatter plot of –log10 Cochran's Q p-value for the imputed and non-imputed datasets. Out of 1238 probes with an improvement in association signal after imputation, 657 had a larger Cochran's Q statistic in the Phase III HapMap SNP.

It may have been expected that there would be less heterogeneity in allelic effects after imputation because: (i) we are more likely to include the causal variant in the 1000 Genomes Project reference panel; and (ii) if the causal variant is not present in the reference panel, there should be a better tag for it, and for these better tags we would expect LD to be less variable between diverse populations. However, this is not seen in this analysis. This could suggest that heterogeneity is not due to LD differences between populations at some cis eQTLs, but a result of causal variants that are specific to one population group, or which interact with factors that differ between populations.

An example with a large reduction in heterogeneity between imputed and non-imputed SNPs is for the probe ILMN_1295_2340291 which encodes the gene *CHI3L2*. The peak non-imputed SNP has a Cochran's Q score of 46.47 ($p=7.07 \times 10^{-8}$) the peak imputed SNP has a Q score of 11.36 (p=0.124). This example is looked at in more detail in section 5.7.5.

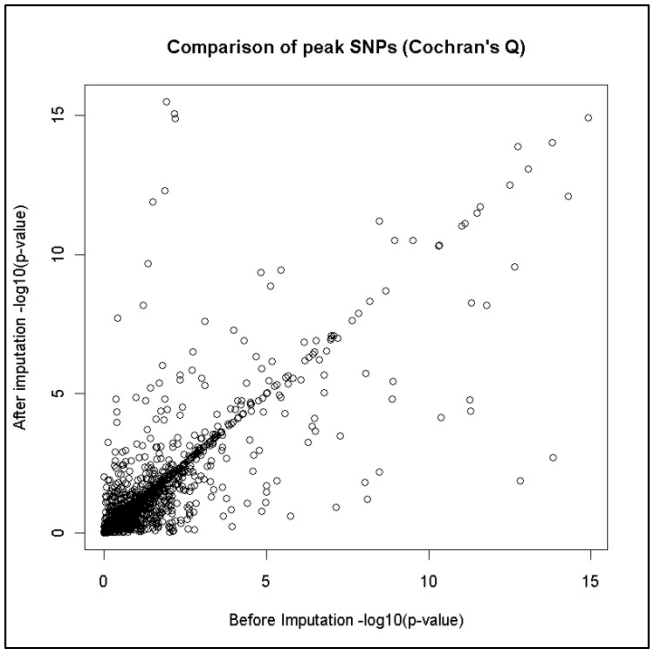


Figure 5.5: Scatter plot of Cochran's Q p-value for lead SNP before and after imputation

5.6 Annotation

In total, 1752 unique eSNPs were identified from the 1811 probes detected with the March 2012 imputed Phase III HapMap data with fixed effect meta-analysis at GWS. In order to gain insight into the mechanisms through which these eSNPs might impact expression, we investigated their annotation. For the 1752 eSNPs, annotations of 1738 were available in the NCBI B37 ENSEMBL Biomart. See methods section 2.2 for more information on ANNOVAR.

5.6.1 Gene-Based Annotation

Table 5.13 gives overview of gene based annotation for peak SNPs detected in the fixed effect meta-analysis at GWS.

Type	**Count**
Intronic	887
Intergenic	447
UTR3	121
Upstream	67
Exonic	64
Downstream	52
UTR5	43
ncRNA Intronic	36
ncRNA Exonic	15
Splicing	2
Upstream; Downstream	2
Exonic; Splicing	1
ncRNA UTR3	1
Total	1738

Table 5.13: Functional annotation for peak SNPs from fixed effect meta-analysis with GWS ($p < 5x10^{-8}$) and MAF 5%

Approximately 51% of the SNPs (887 / 1738) are located within introns, suggesting that factors such as transcript stability and RNA splicing could be causal in eQTLs. Approximately 26% (447 / 1738) are intergenic, which suggests variants in *cis* elements, such as promoters and enhancers, are also causing differences in expression. Only 64 SNPs are within exons.

5.6.2 Region-Based Annotation

Region-based annotation investigates the overlap of the eSNP variant and functional regions in the genome. The functional regions can be *cis* acting elements, such as transcription factor binding sites. Table 5.14 shows the region-based annotation for SNPs detected with the fixed effect meta-analysis at GWS. Descriptions of the annotation used in this analysis are presented in the Methods Section 2.12.2. A total of 22% to 31% of SNPs are located within regions of histone modification, which indicate active promoters and enhancers. Only 3 SNPs are within regions of CpG Methylation.

ENCODE Annotation Type	Cell Line	Count
chromHMM, chromatin state predictions	GM12878	1712
Histone Modification H3K4me1 (active enhancer)	GM12878	547
Histone Modification H3K4me3 (active promoter)	GM12878	472
DNase I Hypersensitivity Sites	NA	401
Histone Modification H3K27Ac (active enhancer)	GM12878	401
Transcription Factor ChIP-Seq	NA	396
Histone Modification H3K9Ac (active promoter)	GM12878	383
Methylation 450 Bead Array	GM12878	3

Table 5.14: Region-based annotation for peak SNPs from fixed effect meta-analysis at GWS. Cell line is specified if required in the assay, all use the GM12878 CEU LCL cell line.

Table 5.15 shows the counts of each category detected. Almost all the SNPs (1722) have corresponding chromHMM chromatin state prediction annotation. Approximately 25% of SNPs are within a region of heterochromatin which could indicate a region of low gene expression. In addition to this approximately 22% are within a region of weak transcription.

Category	Count
Heterochromatin/low signal	450
Weak Transcription	380
Transcription Elongation	285
Active Promoter	122
Strong Enhancer	84
Weak Enhancer	81
Weak Enhancer	78
Weak Promoter	70
Strong Enhancer	55
Repressed	44
Transcription Transition	31
Poised Promoter	16
Insulator	13
Repetitive/CNV	2
Repetitive/CNV	1
Total	1712

Table 5.15: Categories from the chromHMM annotation.

5.6.2 Overlap of Gene-Based and Region-based Annotation

Table 5.16 shows the overlap between the gene based and region based annotation. As can be seen the majority of the region-based annotation are located within intronic or intergenic regions. All of the SNPs within methylation regions are within introns.

		ENCODE Region Annotation							
		DNase I	TF ChiP-Seq	Methylation	H3k4me1	H3k4me3	H3K27Ac	H3K9Ac	Total
ENCODE Gene Annotation	Intronic	185	176	3	280	238	200	184	887
	Intergenic	76	72	0	136	83	72	59	448
	UTR3	23	17	0	29	19	20	20	121
	Upstream	45	51	0	26	52	40	47	67
	Exonic	19	21	0	22	20	17	19	64
	Downstream	6	9	0	12	6	6	5	52
	UTR5	35	40	0	22	36	29	33	43
	ncRNA Intronic	5	6	0	15	10	12	10	36
	ncRNA Exonic	5	3	0	4	6	3	4	15
	Splicing	0	0	0	0	0	0	0	2
	Upstream; Downstream	2	1	0	0	1	1	1	2
	Exonic; Splicing	0	0	0	1	1	1	1	1
	ncRNA UTR3	0	0	0	0	0	0	0	1
	Total	401	396	3	547	472	401	383	1738

Table 5.16: Overlap of Gene based and Region based annotation.

5.7 Examples

This section presents examples for a series of *cis* eQTL detected in the fixed effect meta-analysis. Examples have been selected with the following criteria:

1. Most significant association signal from the fixed effect meta-analysis together with no evidence of heterogeneity in allelic effects between populations: *WBSCR27* (Section 5.7.1). This cis eQTL was also studied in the Phase III HapMap analyses (Section 4.7.1).
2. Greatest extent of heterogeneity in allelic effects between populations detected: *CAV2* (Section 5.7.2) and *FANCA* (Section 5.7.3). These cis eQTLs were also studied in the Phase III HapMap analyses (Section 4.7.2 and 4.7.3).
3. Greatest improvement in association signals between the imputed and non-imputed peak SNPs: *CORO2A* (Section 5.7.4).
4. Greatest reduction in heterogeneity in allelic effects between populations observed between the imputed and non-imputed peak SNPs: *CHI3L2* (Section 5.7.5). This example has non-significant heterogeneity in the 1000 Genomes eSNP.

5.7.1 ILMN_2170_6860347 (ENSG00000165171, WBSCR27)

This probe was previously analysed in the Phase III HapMap non-imputed analysis (Section 4.7.1).

As in the previous chapter, this *cis* eQTL has the strongest signal of association, with no evidence of heterogeneity in allelic effects between populations. In the fixed effect meta-analysis after imputation, the probe ILMN_2170_6860347 has the 1000 Genomes eSNP rs10225983 (p-value=6.99×10^{-300}), with no evidence of heterogeneity (Cochran's Q p-value= 0.23) (See table 5.8). The signal in the 1000 Genomes dataset is more significant than that identified in the Phase III HapMap dataset before imputation (eSNP=rs13228435, p-value = 8.41×10^{-298}, Q p-value= 0.24).

The eSNP rs10225983 position is chr7:73258750, the SNP is intergenic between genes *WBSCR27* and *WBSCR28*. The distance is 1,895 bp from *WBSCR27* and 16,739 bp from *WBSCR28*. Additional

details regarding the gene *WBSCR27* are summarized in Section 4.7.1. The position of the Phase III HapMap eSNP (rs13228435) is chr7:73239172, which is 19,578 base pairs from WBSCR27, and thus further away from the gene.

Figure 5.6 shows the forest plot for the eSNP rs10225983 at the *cis* eQTL for *WBSCR27*. As expected by the low heterogeneity, effect sizes and standard errors are similar across the populations. The results are very similar to those before imputation (Figure 4.6).

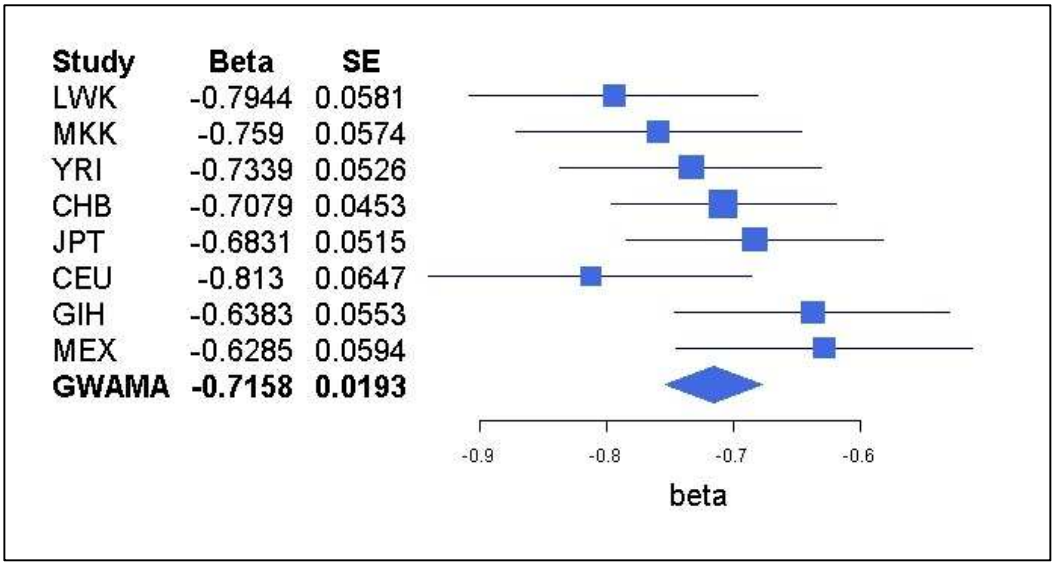


Figure 5.6: Forest plot of association and fixed effect meta-analyses for probe ILMN_2170_6860347 with 1000 Genomes SNP rs10225983. Blue squares are the value of the effect sizes, lines specify the 95% confidence intervals calculated from standard error.

Table 5.17 shows p-values, allele frequencies and INFO scores for the eSNP rs10225983 with probe ILMN_2170_6860347. As can be seen all eight populations have association signals achieving GWS. Allele frequencies ranges from 0.47 – 0.79. The peak SNP in the fixed effect meta-analysis is not the same as the peak SNP in any of the population-specific association analyses. These results are very similar to the peak Phase III HapMap SNP (table 4.11).

Population	Minor Allele (A/G)	Allele Frequencies (Allele A)	Beta (Allele A)	SE	p-value	INFO	Peak SNP
LWK	A	0.47	-0.7944	0.0581	1.21×10^{-22}	1.000	N
MKK	G	0.67	-0.7590	0.0574	6.05×10^{-26}	0.983	N
YRI	A	0.47	-0.7339	0.0526	1.11×10^{-25}	0.990	N
CHB	G	0.55	-0.7079	0.0453	1.82×10^{-25}	0.965	N
JPT	G	0.56	-0.6831	0.0515	1.13×10^{-21}	0.997	N
CEU	G	0.79	-0.8130	0.0647	1.38×10^{-22}	1.000	N
GIH	G	0.54	-0.6383	0.0553	4.84×10^{-18}	1.000	N
MEX	G	0.52	-0.6285	0.0594	6.85×10^{-13}	0.999	N
Fixed effect Meta-analysis	--	--	-0.7158	0.0193	6.99×10^{-300}	--	Y

Table 5.17: Table of allele frequencies and p-values for 1000 Genomes SNP rs10225983 with probe ILMN_2170_6860347. The column **Peak SNP** specifies whether the peak SNP for the association analysis is the same as the peak signal in the fixed effect meta-analysis.

Figures on the following pages show signal plots for population-specific association analyses (figures 5.7 and 5.8) and for the fixed effect meta-analysis (figure 5.9). The SNP rs10225983 remains the peak eSNP when variants not reported in all eight populations are included in the fixed effect meta-analysis. This indicates that the SNP identified is optimal with the imputed dataset.

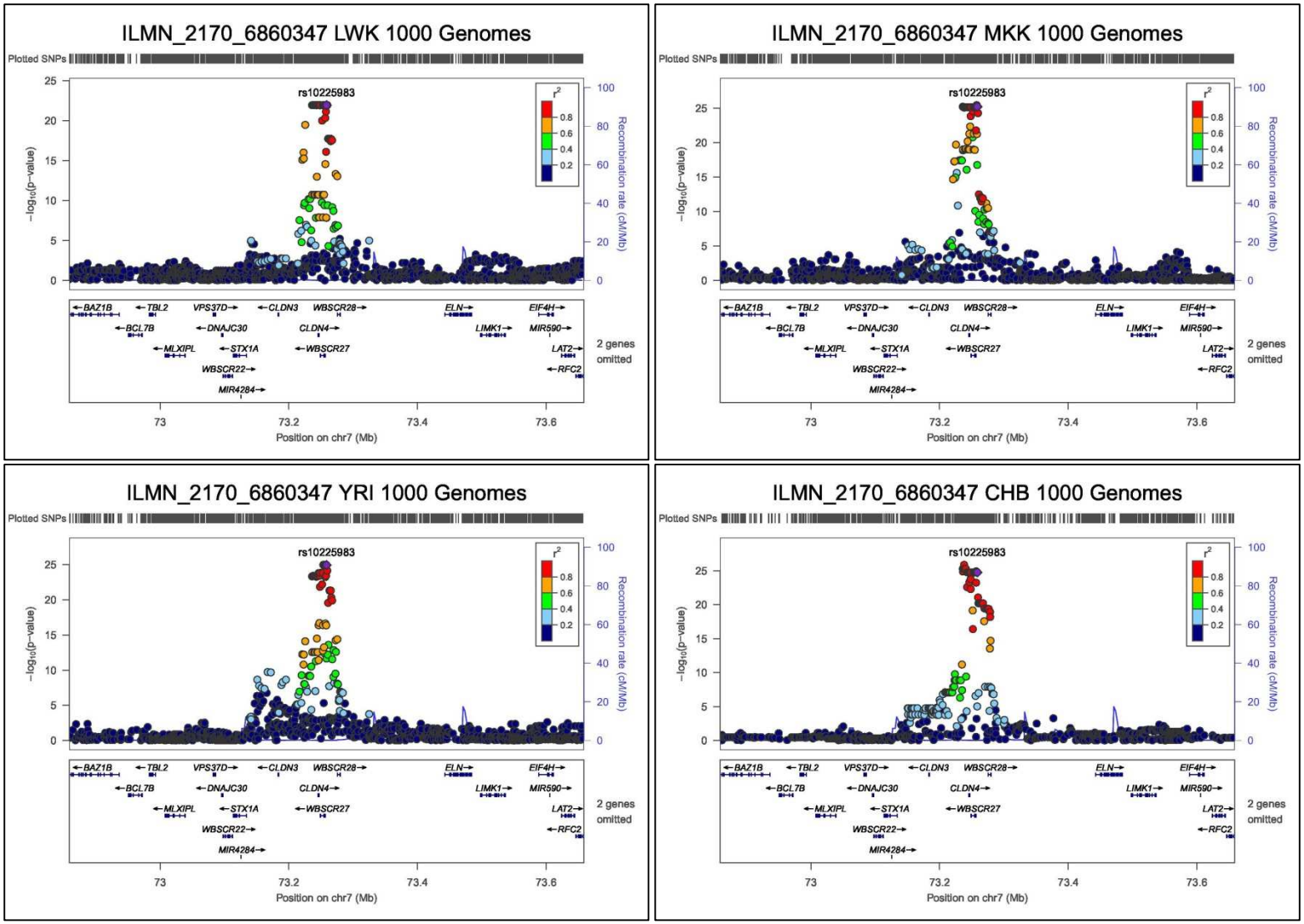


Figure 5.7: Signal plots for peak 1000 Genomes and Phase III HapMap eSNPs for the probe ILMN_2170_6860347. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

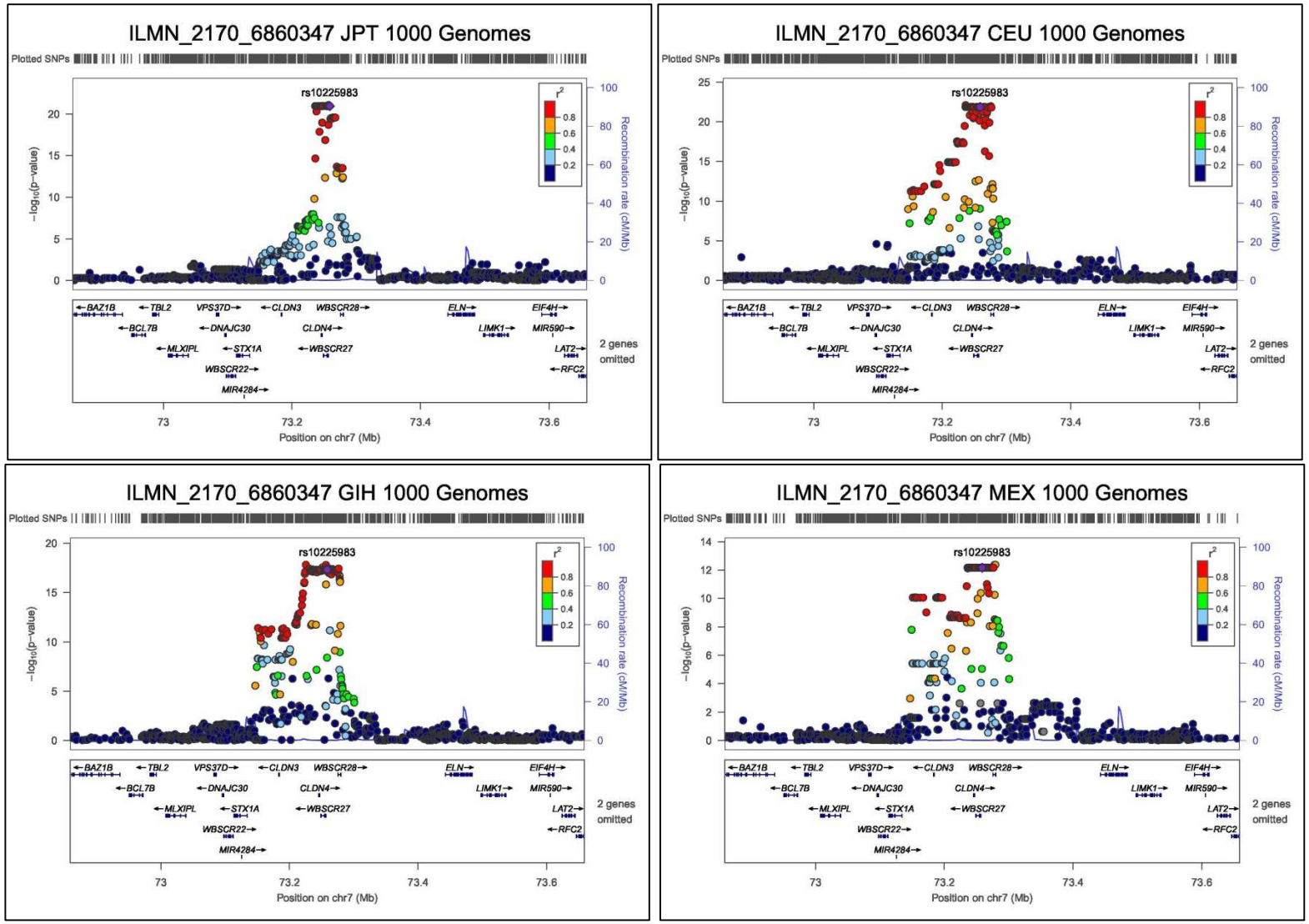


Figure 5.8: Signal plots for peak 1000 Genomes and Phase III HapMap SNPs for probe ILMN_2170_6860347. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

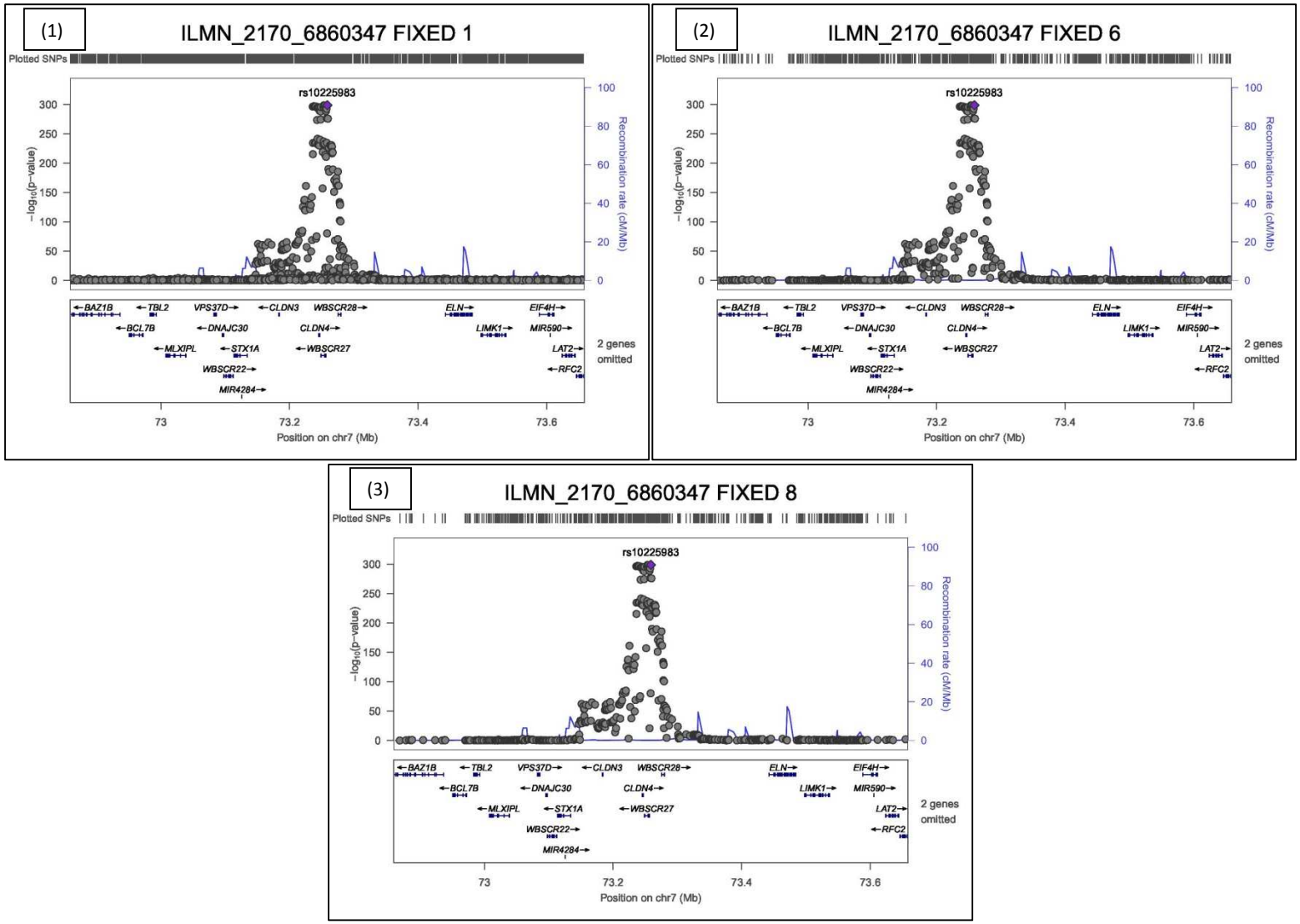


Figure 5.9: Signal plot of SNP $-\log_{10}$ p-values from fixed effect meta-analysis for probe ILMN_2170_6860347. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond.

5.7.2 ILMN_5108_2030195 (ENSG00000105971, CAV2)

This probe was previously analysed in the Phase III HapMap non-imputed analysis (Section 4.7.2).

This probe has the strongest signal of heterogeneity in allelic effects between populations across all the *cis* eQTLs. In the fixed effect meta-analysis, this probe has the largest Cochran's Q-statistic (Q=182.13, p-value < $2x10^{-16}$). The peak SNP is rs13235183, which is not present in Phase III HapMap (p-value=$4.72x10^{-48}$). The association signal after imputation is stronger than that identified in the Phase III HapMap dataset, but with less evidence of heterogeneity (SNP = rs17138749, p-value = $2.09x10^{-25}$, Cochran's Q = 217.12, Q p-value < $2x10^{-16}$).

The function of this gene is summarized in Section 4.7.2. The 1000 Genomes eSNP rs13235183 is located at chr7:116140149. The eSNP is within the five prime untranslated region of the gene *CAV2*. The Phase III HapMap SNP is rs17138749 is located at chr7:116133098 and is intergenic, mapping closest to *CAV2*.

Figure 5.10 shows the forest plot for the SNP rs13235183 with the probe ILMN_5108_2030195. As can be seen, the African ancestry group effect sizes cluster together and have a lower magnitude than the other ancestry groups, with LWK and YRI not being significantly different from zero. The forest plot is very similar to the plot for Phase III HapMap eSNP (See figure 4.10).

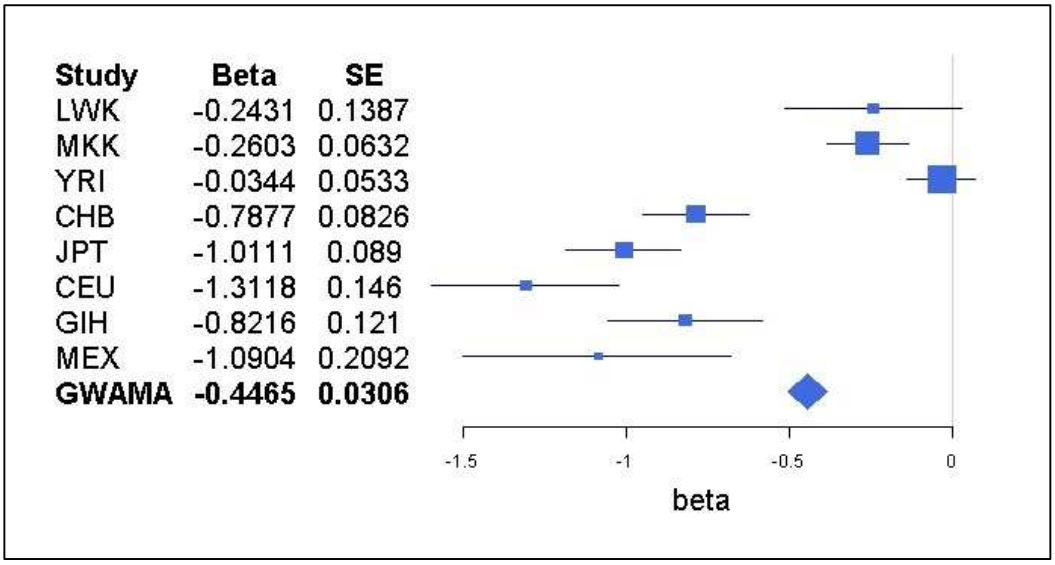


Figure 5.10: Forest plot of association and fixed effect meta-analyses for probe ILMN_5108_2030195 with SNP rs13235183.

Table 5.18 is a table of the p-values, allele frequencies and INFO score for probe ILMN_5108_2030195 with SNP rs13235183. Four populations (CHB, JPT, CEU and GIH) have association signals achieving GWS. Allele frequencies range from 0.07 to 0.21. The association p-values of Eurasian-Hispanic populations are more significant than when using the Phase III HapMap eSNP see table 4.12.

Population	Minor Allele (T / G)	Allele Frequency (Allele T)	Beta (Allele G)	SE	p-value	INFO	Peak SNP
LWK	T	0.16	-0.2431	0.1387	0.083	0.906	N
MKK	T	0.07	-0.2603	0.0632	6.66×10^{-5}	0.908	N
YRI	T	0.07	-0.0344	0.0533	0.521	0.898	N
CHB	T	0.15	-0.7877	0.0826	1.29×10^{-14}	0.895	N
JPT	T	0.21	-1.0111	0.0890	3.33×10^{-18}	0.916	N
CEU	T	0.08	-1.3118	0.1460	1.23×10^{-14}	0.980	N
GIH	T	0.07	-0.8216	0.1210	2.70×10^{-9}	0.965	N
MEX	T	0.09	-1.0904	0.2092	6.85×10^{-6}	0.873	N
Fixed effect Meta-analysis	--	--	-0.4465	0.0306	4.72×10^{-48}	--	Y

Table 5.18: Table of allele frequencies for SNP rs13235183 for probe ILMN_5108_2030195.

Figures on the following pages show signal plots for the population-specific association analysis (figures 5.11 and 5.12) and the fixed effect meta-analysis (figure 5.13). It is clear from the forest plot that the heterogeneity is being caused by smaller effect sizes in the African populations.

From the signal plots for the African populations, it can be seen that the YRI population has no evidence of any association signal at this locus. This is the same as was observed in the Phase III HapMap analysis. For LWK, although the peak SNP does not have a significant association signal, it can be seen from the signal plot that significant signals for this population do exist for other SNPs in this region. However, they are not in LD with the peak SNP detected in the fixed effect meta-analysis. The SNP with the strongest signal of association for LWK (rs36011826) has a p-value of 4.60×10^{-8}. This was also observed in the LWK population with the Phase III HapMap eSNP (rs17138749) also. Reviewing signal plots for the Non-African ancestry groups, significant association signals are detected at the fixed effect meta-analysis peak SNP. For the fixed effect meta-analysis, if SNPs not reported in all eight populations are included in the analysis, the peak SNP changes, indicating there is an issue due to variable reporting of SNPs across populations

Taken together, these data suggest that the heterogeneity is being mainly caused by difference in effect size of the African ancestry populations compared to the others. In the case of LWK and MKK, the heterogeneity of effect size may be due to different LD with the causal variant. However in the case of YRI, there is no evident of signal whatsoever, even after 1000 Genomes imputation. If the causal variant is shared across populations, this indicates that it is not present in (or well tagged by) the 1000 Genomes Project reference panel (with MAF>5%).

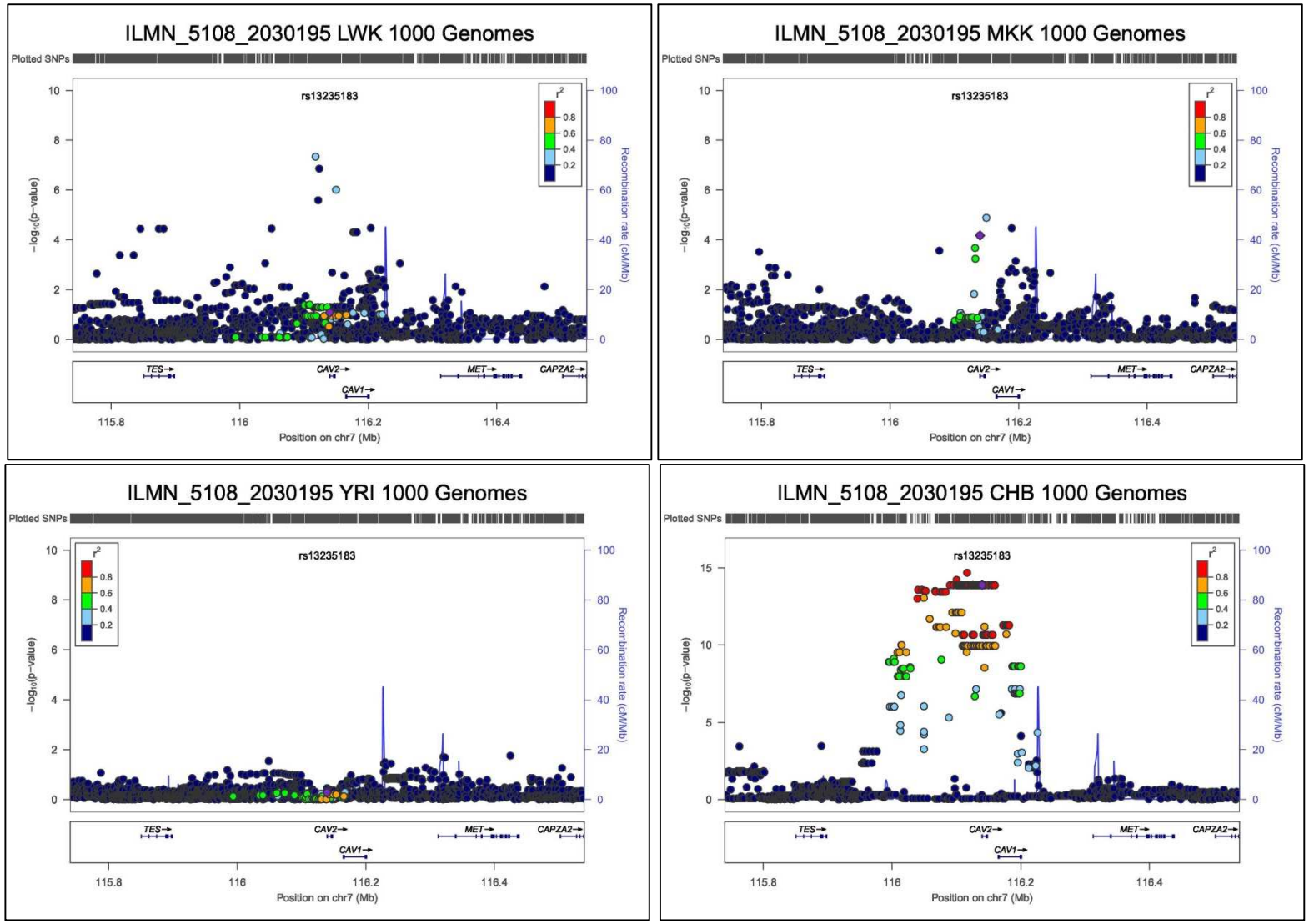


Figure 5.11: plots for peak 1000 Genomes and Phase III HapMap SNPs for probe ILMN_5108_2030195. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

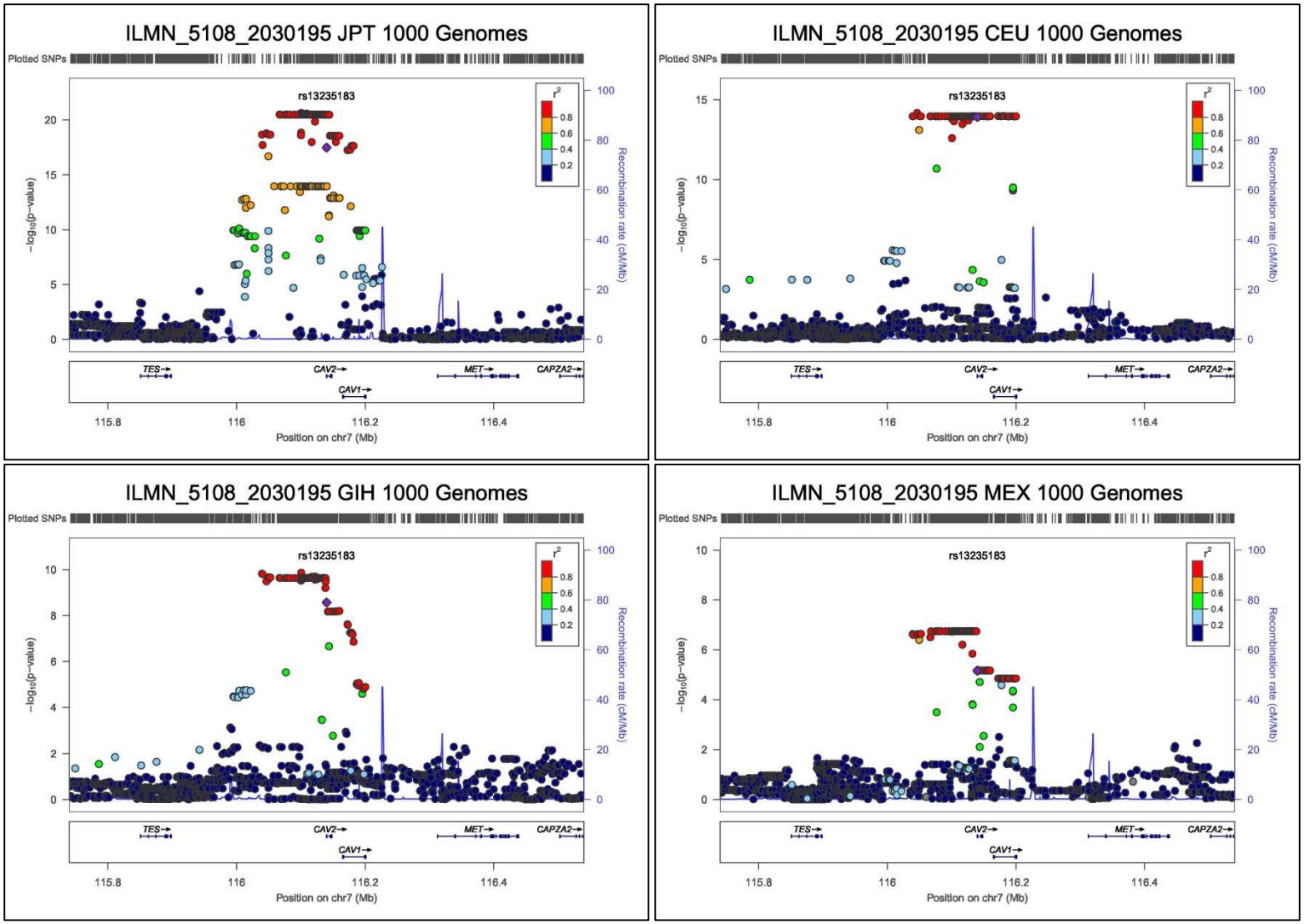


Figure 5.12: plots for peak 1000 Genomes and Phase III HapMap SNPs for probe ILMN_5108_2030195. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

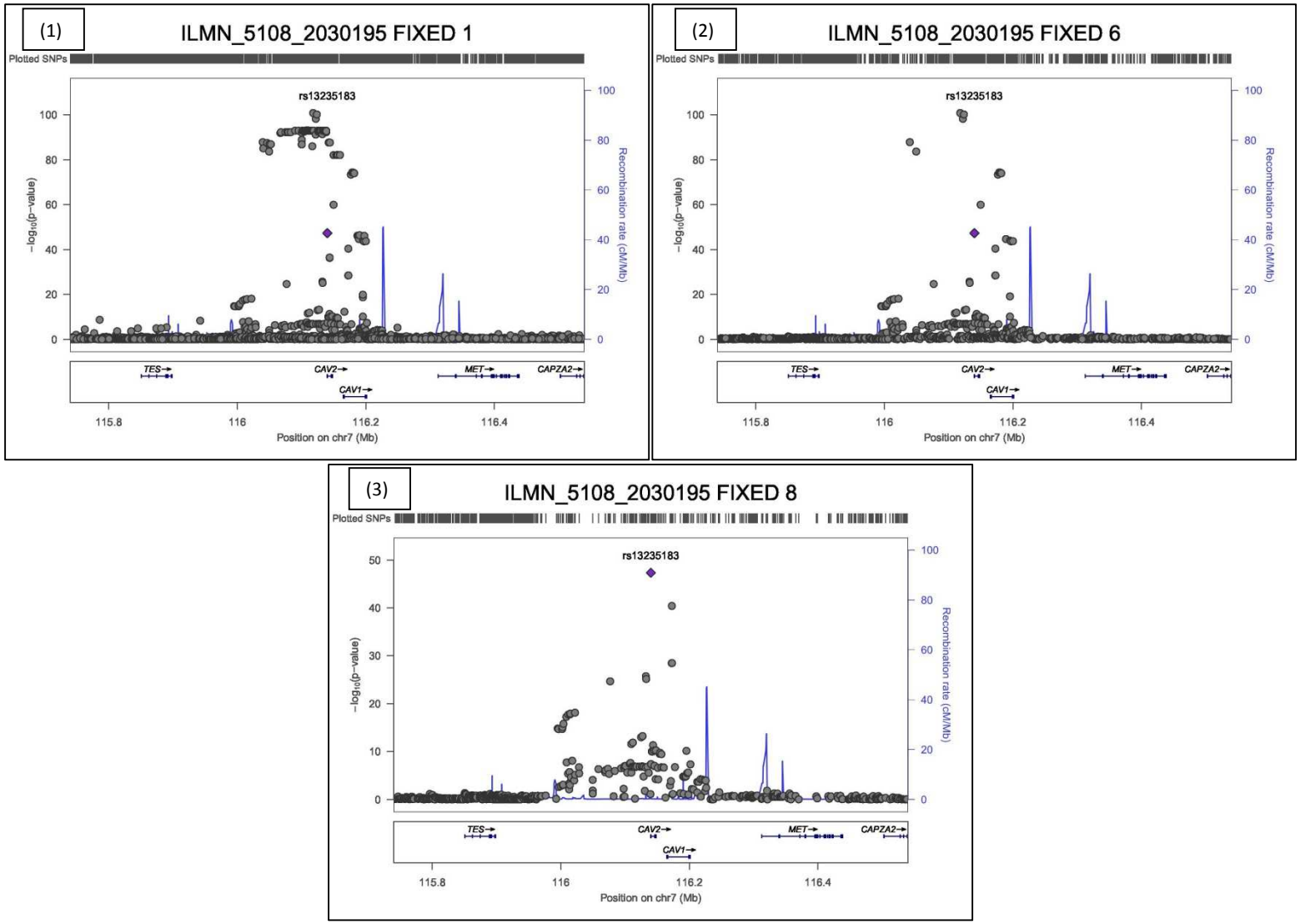


Figure 5.13: Signal plot for fixed effect meta-analysis for probe ILMN_5108_2030195. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond.

5.7.3 ILMN_13285_4210136 (ENSG00000187741, FANCA)

This probe was previously analysed in the Phase III HapMap non-imputed analysis (Section 4.7.3).

This probe has been selected because it has strong evidence of heterogeneity in allelic effects between populations and a strong signal of association from the fixed effect meta-analysis. This probe has the fifth largest Cochran's Q-statistic value (Q=101.34, Q p-value < $2x10^{-16}$). The peak SNP is rs8051231, and is not present in Phase III HapMap (p-value=$2.52x10^{-93}$).

The Phase III HapMap peak eSNP rs2239360 has less heterogeneity and a weaker association signal than after imputation (SNP=rs2239360, p-value = $9.23x10^{-89}$, Cochran's Q = 92.34, Q p-value < $2x10^{-16}$). The 1000 Genomes SNP rs8051231 is located at chr16:89862132. It is within an intron of the gene *FANCA*. The function of this gene is summarized in Section 4.7.3. The eSNP rs2239360 is located at chr16: 89849583, and maps within an intron of *FANCA*.

Figure 5.14 shows the forest plot for SNP rs8051231 with probe ILMN_13285_4210136. It can be seen that the effect sizes for populations within the same ancestry groups cluster together. The largest effect sizes are observed in the Eurasian-Hispanic ancestry group, followed by African and East Asian. Effect sizes in the East Asian ancestry group do not differ significantly from zero at $p \leq 0.05$. These results are very similar to those with the Phase III HapMap eSNP rs2239360 (See Chapter 4, Figure 4.14).

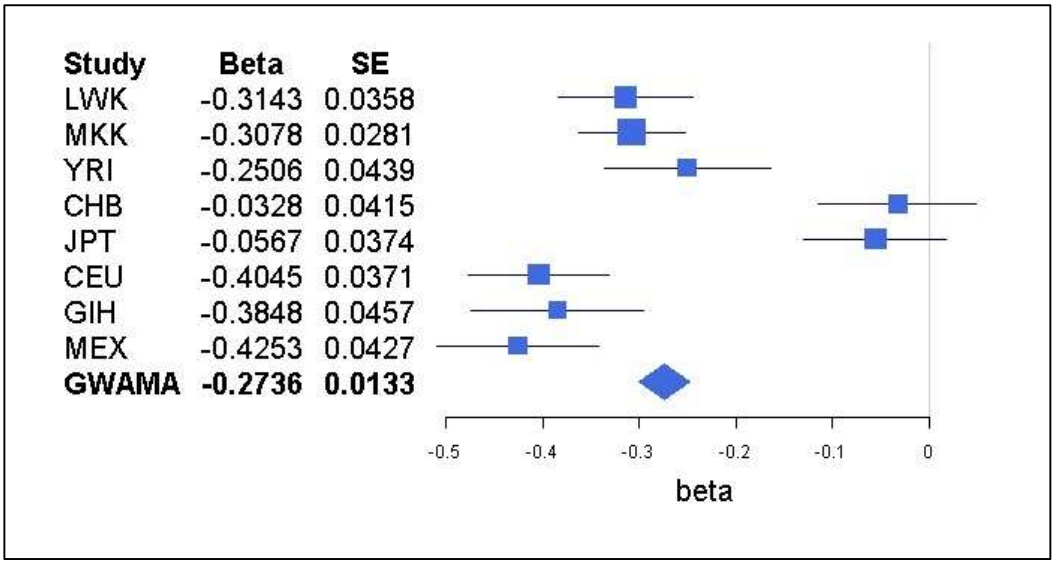


Figure 5.14: Forest plot of association and fixed effect meta-analyses for probe ILMN_13285_4210136 with SNP rs8051231.

Table 5.19 shows a table for p-values, allele frequencies and INFO scores for the probe ILMN_13285_4210136 with SNP rs8051231. Allele frequencies range from 0.12 to 0.43. Five of the populations have association signals achieving GWS. The eSNP from the fixed-effects meta-analysis is not the peak SNP for any of the population-specific association analyses.

Population	Minor Allele (G / C)	Allele Frequency (Allele G)	Beta (Allele C)	SE	p-value	INFO	Peak SNP
LWK	G	0.42	-0.3143	0.0358	2.39×10^{-13}	0.982	N
MKK	G	0.35	-0.3078	0.0281	3.07×10^{-20}	0.943	N
YRI	G	0.29	-0.2506	0.0439	1.07×10^{-7}	0.983	N
CHB	G	0.25	-0.0328	0.0415	0.43	0.966	N
JPT	G	0.12	-0.0567	0.0374	0.13	0.929	N
CEU	C	0.65	-0.4045	0.0371	6.88×10^{-19}	0.996	N
GIH	G	0.33	-0.3848	0.0457	2.43×10^{-12}	0.975	N
MEX	G	0.43	-0.4253	0.0427	3.83×10^{-12}	0.977	N
Fixed effect Meta-analysis	--	--	-0.2736	0.0133	2.52×10^{-93}	--	Y

Table 5.19: Table of allele frequencies for SNP rs8051231 with probe ILMN_13285_4210136.

Figures on the following pages show signal plots for population-specific association analysis (figures 5.15 and 5.16) and the fixed effect meta-analysis (figure 5.17). From the plots the following observations can be made. First, the SNP rs8051231 is the peak eSNP when the intersection of variants are considered in the fixed effect meta-analysis. However, when variants

not reported in all eight populations are included in the meta-analysis, rs8051231 is no longer the peak SNP. The East Asian ancestry populations, CHB and JPT, have little evidence of any association signal. This is particularly the case in JPT, where no significant signal can be detected. There is some evidence for a weak signal in CHB population. Visually the length of the peak LD block in the African populations is narrower than that in the Eurasian-Hispanic populations. The signal plots suggest that heterogeneity has two possible causes. The first is that the East Asian populations do not have any tag SNPs for the *cis* eQTL for *FANCA*. The second is that peak signals may be in variants which are not reported/present in a subset of populations, and are therefore not detected in the fixed effect meta-analysis with the intersection of variants. This causes differences in effect sizes between populations due to differing LD with the causal variant.

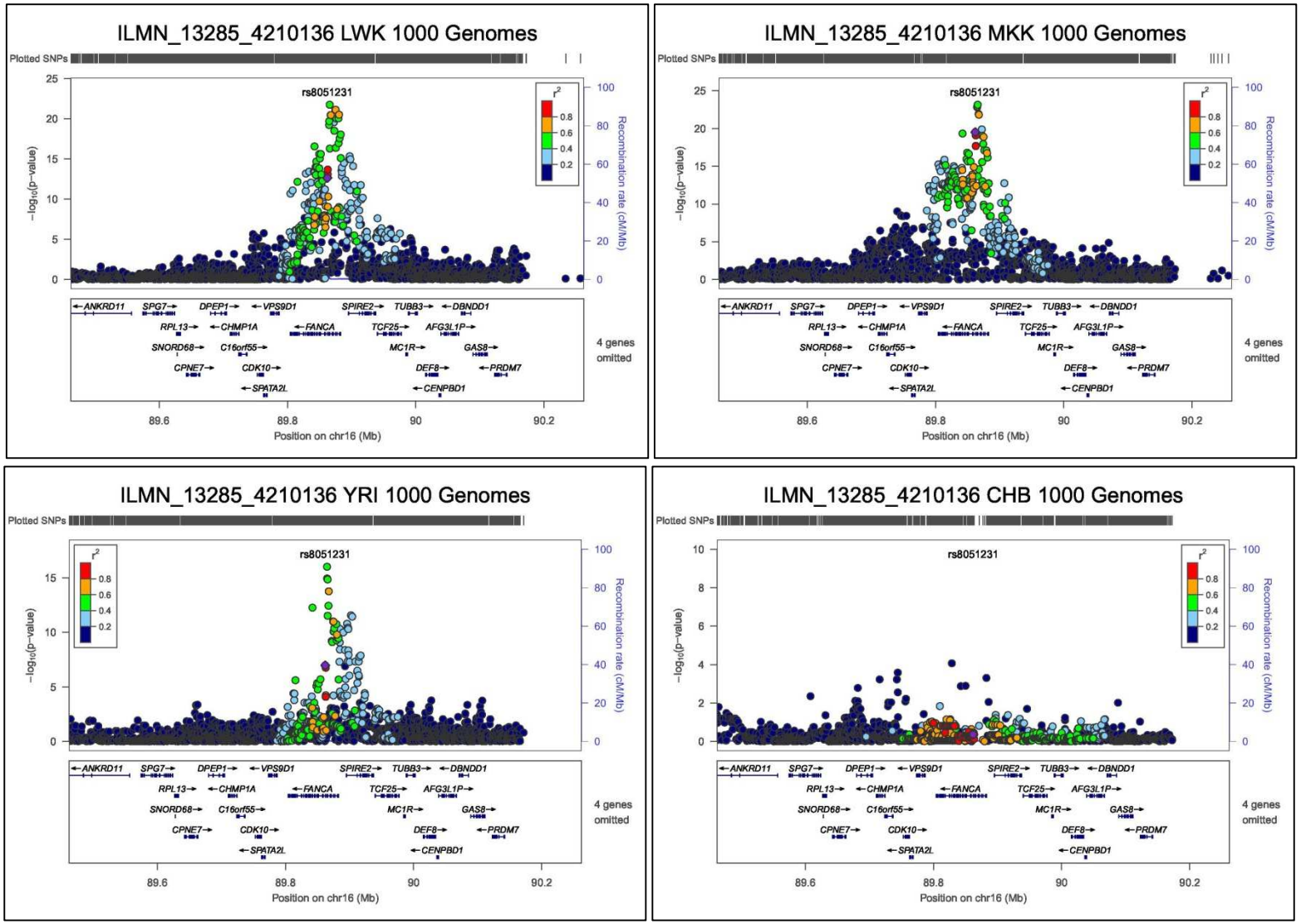


Figure 5.15: plots for peak 1000 Genomes and Phase III HapMap SNPs for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

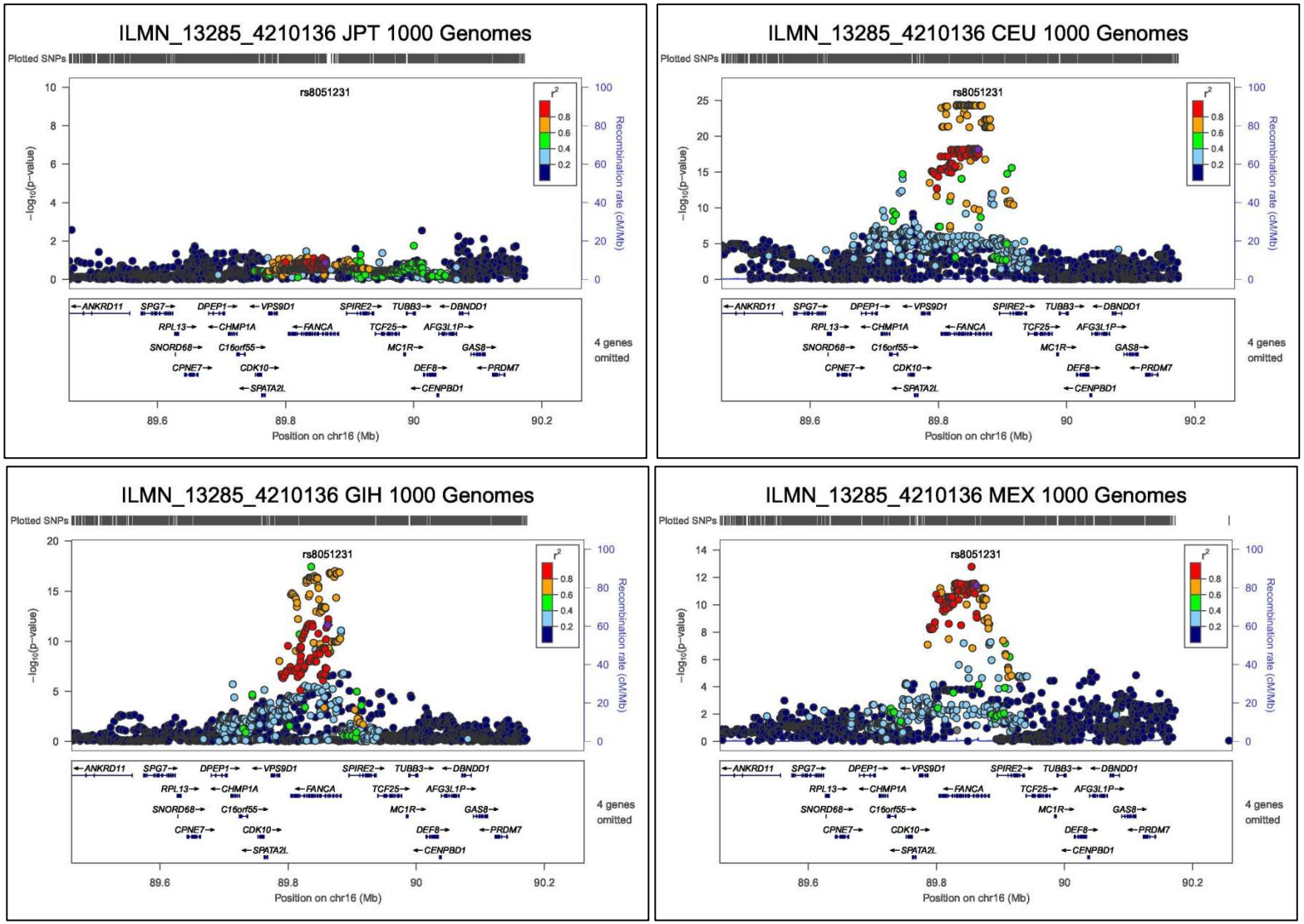


Figure 5.16: plots for peak 1000 Genomes and Phase III HapMap SNPs for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

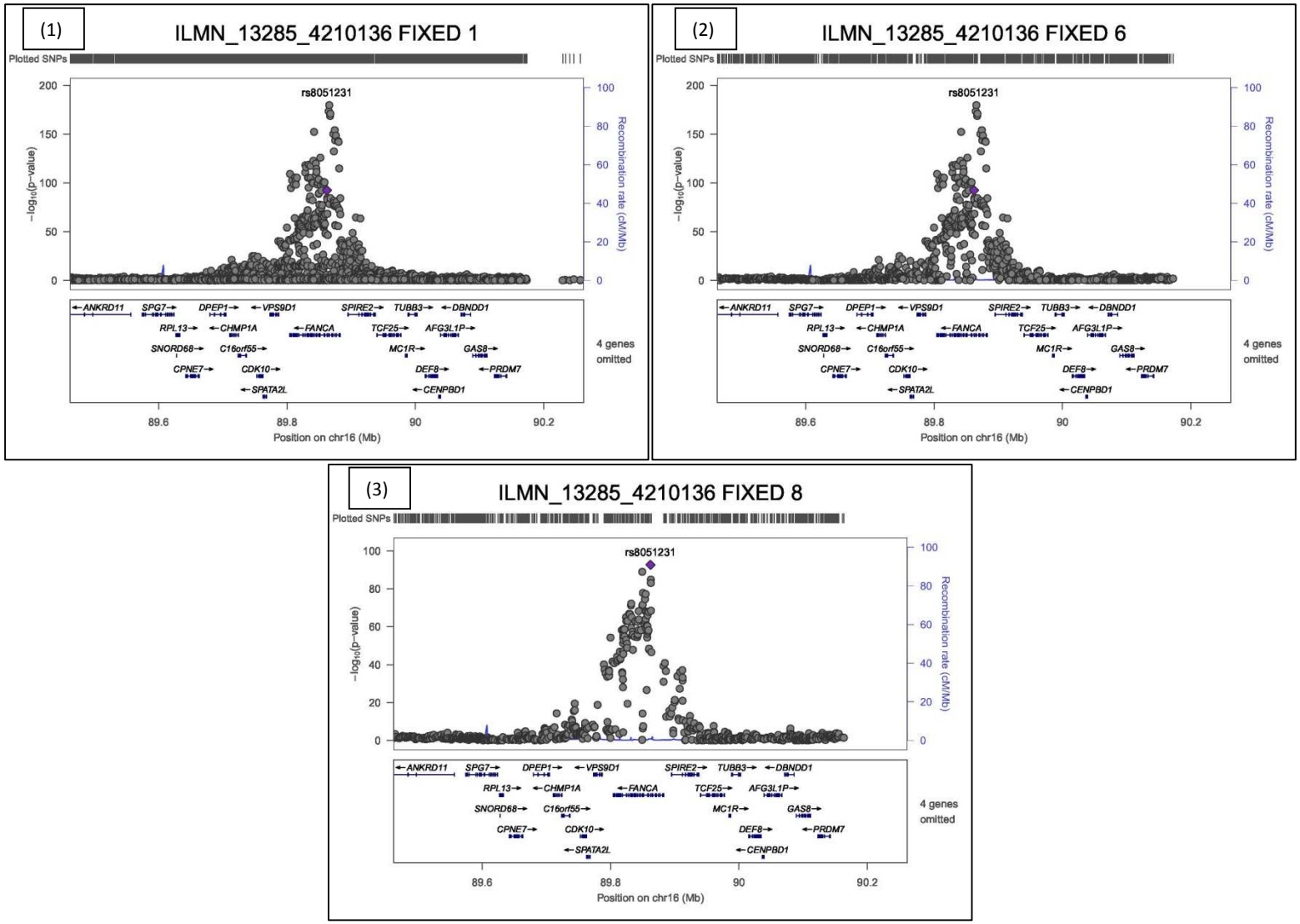


Figure 5.17: Signal plot for fixed effect meta-analysis for probe ILMN_13285_4210136. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond.

5.7.4 ILMN_20456_4590554 (ENSG00000106789, CORO2A)

In the fixed effect meta-analysis, this probe has the first overall largest improvement in association signal between the peak 1000 Genomes eSNP rs28703824 (p-value=$6.01x10^{-95}$, Cochran's Q p-value=$7.50x10^{-3}$) and peak Phase III HapMap eSNP rs10818610 (p-value=$1.26x10^{-20}$, Cochran's Q p-value=0.032): See table 5.13.

Gene name and location

The probe ILMN_20456_4590554 gene has the Ensembl ID ENSG00000106789 and the HGNC symbol *CORO2A*. The full name of the gene is Coronin, Actin Binding Protein, 2A, and the gene's start position is chr9:100954922. This gene encodes a member of the WD repeat protein family, which are involved in a variety of cellular processes including cell cycle progression, signal transduction, apoptosis and gene regulation. WD repeats are short structural motifs of approximately 40 amino acids, often terminating in a tryptophan-aspartic acid (W-D) dipeptide. The protein contains 5 WD repeats and has structural similarity with actin binding proteins.

The 1000 Genomes peak eSNP is rs28703824, which is located at chr9:100954429 and is within an intron in the gene *CORO2A*. The Phase III HapMap eSNP is rs10818610, which is located at chr9:100969348 and is within an intronic in the gene TBC1D2. The distance between the eSNPs is 14,919 base pairs.

Linkage Disequilbrium

Table 5.20 shows LD between the eSNPs before and after imputation, rs28703824 and rs10818610 respectively, using the Broad institute's tool SNAP (Johnson *et al* 2008). The SNPs are only in moderate LD with each other across ancestry groups. The strongest LD between the SNPs is in the East Asian ancestry group, which can also be seen in the signal plots in figure 5.20, where the strong signals at both the Phase III HapMap and 1000 Genomes SNPs are detected.

Population (Ancestry Group)	r^2
CEU (Eurasian-Hispanic)	0.197
CHB + JPT (EastAsian)	0.307
YRI (African)	0.168

Figure 5.20: LD between rs28703824 and rs10818610 using Broad SNAP.

Figure 5.18 is a forest plot of the Phase III HapMap peak eSNP rs10818610 for probe ILMN_20456_4590554 displaying results from the population-specific association analyses and fixed effect meta-analysis. Visually there appears to be some effect size clustering within ancestry groups. For example the East Asian populations (CHB and JPT) have very similar effect sizes (0.7053 and 0.7306, respectively). The effect sizes for the LWK and GIH populations do not significantly differ from zero at $p \leq 0.05$.

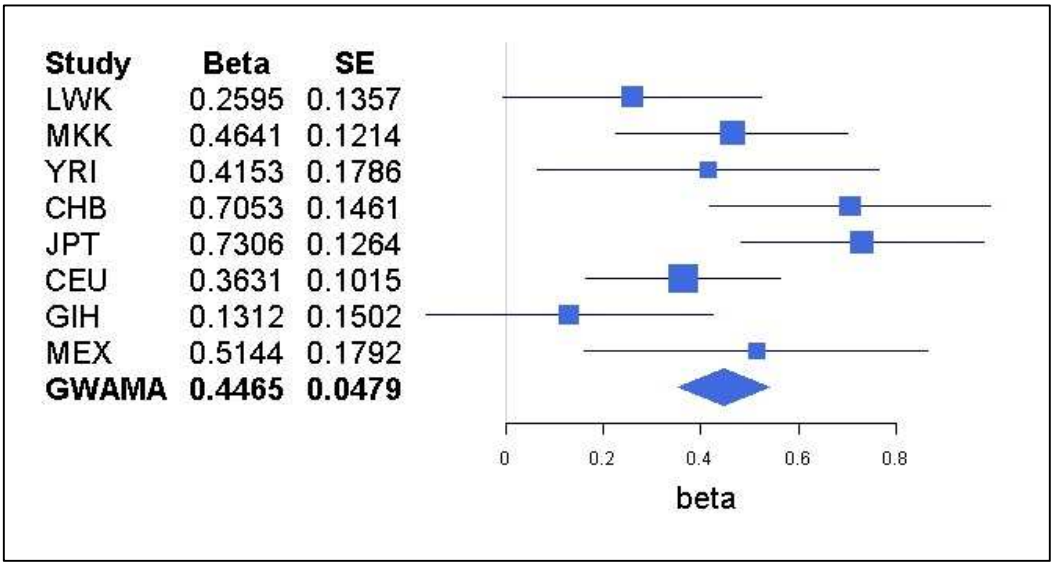


Figure 5.18: Forest plot of association and fixed effect meta-analyses for probe ILMN_20456_4590554 with Phase III HapMap SNP rs10818610

Table 5.21 shows p-values and allele frequencies for the Phase III HapMap peak eSNP rs10818610. Allele frequencies range from 0.24-0.75. None of the populations have signals with GWS. None of the peak SNPs for the population-specific association analyses are the same as that for the fixed effect meta-analysis.

Population	Minor Allele (G / A)	Allele Frequencies (Allele G)	Beta (Allele A)	SE	p-value	INFO	Peak SNP
LWK	G	0.40	0.2595	0.1357	0.059	1.000	N
MKK	G	0.44	0.4641	0.1214	2.01×10^{-4}	0.974	N
YRI	G	0.24	0.4153	0.1786	0.022	0.990	N
CHB	A	0.65	0.7053	0.1461	6.95×10^{-6}	1.000	N
JPT	A	0.63	0.7306	0.1264	1.47×10^{-7}	1.000	N
CEU	A	0.66	0.3631	0.1015	5.28×10^{-4}	0.994	N
GIH	A	0.75	0.1312	0.1502	0.386	0.989	N
MEX	A	0.69	0.5144	0.1792	6.65×10^{-3}	0.999	N
FIXED	--	--	0.4465	0.0479	1.26×10^{-20}		Y

Table 5.21: Table of allele frequencies for Phase III HapMap SNP rs10818610 with probe ILMN_20456_4590554.

Figure 5.19 is a forest plot of the 1000 Genomes peak eSNP (rs28703824) displaying results from the association analysis and fixed effect meta-analysis. As can be seen, the effect sizes for all populations are now significantly different from zero at $p < 0.05$. Some clustering of effect sizes is observed, for example, as before, the JPT and CHB populations of East Asian ancestry have similar effect sizes (1.2661, 1.2442 respectively).

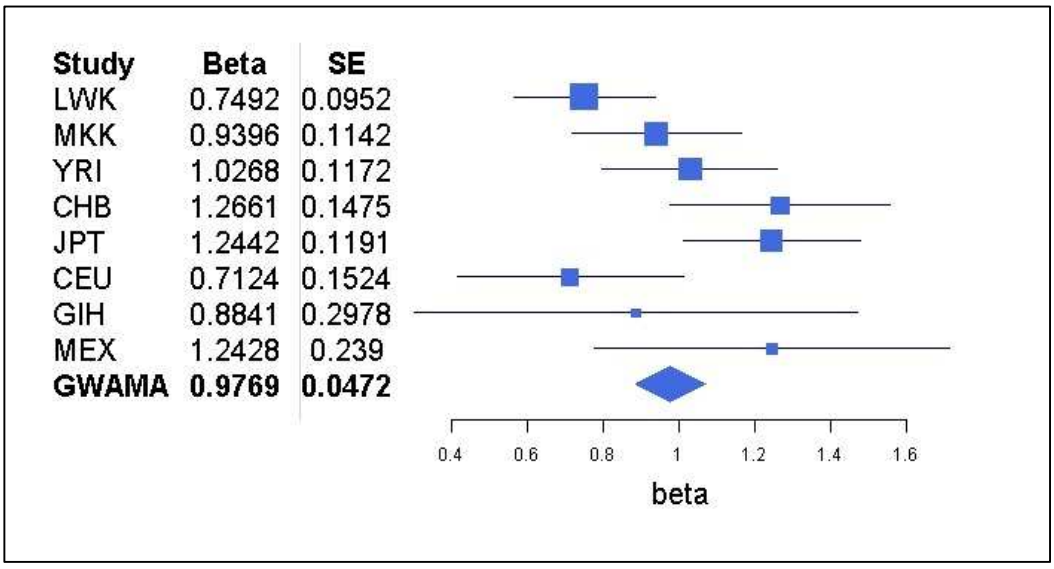


Figure 5.19: Forest plot of association and fixed effect meta-analyses for probe ILMN_20456_4590554 with 1000 Genomes SNP rs28703824.

Table 5.22 shows p-values, allele frequencies and INFO scores at the 1000 Genomes eSNP rs28703824. All members of the African and East Asian populations have p-values at GWS. The weaker signals in the Eurasian-Hispanic ancestry group are due to a larger standard error in the GIH and MEX populations and a weaker effect size in CEU. The Eurasian-Hispanic ancestry

populations have smaller MAF than the other ancestry groups, which could be the cause of the large standard errors in this group. Allele frequencies range from 0.05 to 0.38. Seven of the populations (with the exception of GIH) share the same peak eSNP as that of the fixed effect meta-analysis.

Population	Minor Allele (A/G)	Allele Frequencies (Allele A)	Beta (Allele A)	SE	p-value	INFO	Peak SNP
LWK	A	0.37	0.7492	0.0952	$1.45x10^{-11}$	0.83	Y
MKK	A	0.35	0.9396	0.1142	$1.56x10^{-13}$	0.868	Y
YRI	A	0.38	1.0268	0.1172	$3.62x10^{-14}$	0.941	Y
CHB	A	0.17	1.2661	0.1475	$8.43x10^{-13}$	0.856	Y
JPT	A	0.19	1.2442	0.1191	$1.78x10^{-16}$	0.946	Y
CEU	A	0.11	0.7124	0.1524	$8.89x10^{-6}$	0.897	Y
GIH	A	0.05	0.9941	0.2978	$4.10x10^{-3}$	0.709	N
MEX	A	0.12	1.2428	0.2390	$7.09x10^{-6}$	0.993	Y
FIXED	--	--	0.9769	0.0472	$6.01x10^{-95}$	--	Y

Table 5.22: Table of allele frequencies and p-values for 1000 Genomes SNP rs28703824 with probe ILMN_20456_4590554

Figures on the following pages show signal plots for population-specific association analyses (Figure 5.20 and 5.21) and the fixed effect meta-analysis (figure 5.22). From the plots the following observations can be made. First, for all eight populations, the 1000 Genomes peak eSNP signal is more significant than the Phase III HapMap peak eSNP signal, which can also be seen by comparing tables 5.21 and 5.22). The Phase III HapMap signal plot for the LWK and YRI populations show that the SNP reported in the fixed effect meta-analysis is not the peak signal for these populations. This suggests that variants not reported in all eight populations and LD structure is causing the African peak SNPs to be missed. Further evidence that SNPs not reported in all eight populations are causing signals to be missed is shown by the fixed effect meta-analysis signal plot for Phase III HapMap SNPs (figure 5.21). In this analysis, a stronger signal is observed when considering the union of all variants, rather than the intersection. The peak SNP for the 1000 Genomes fixed effect meta-analysis (rs28703824) is also the peak SNP for seven of the eight populations (apart from GIH). There is no signal for the association analysis in the GIH population for Phase III HapMap SNPs, but a moderately significant SNP is observed after 1000 Genomes imputation.

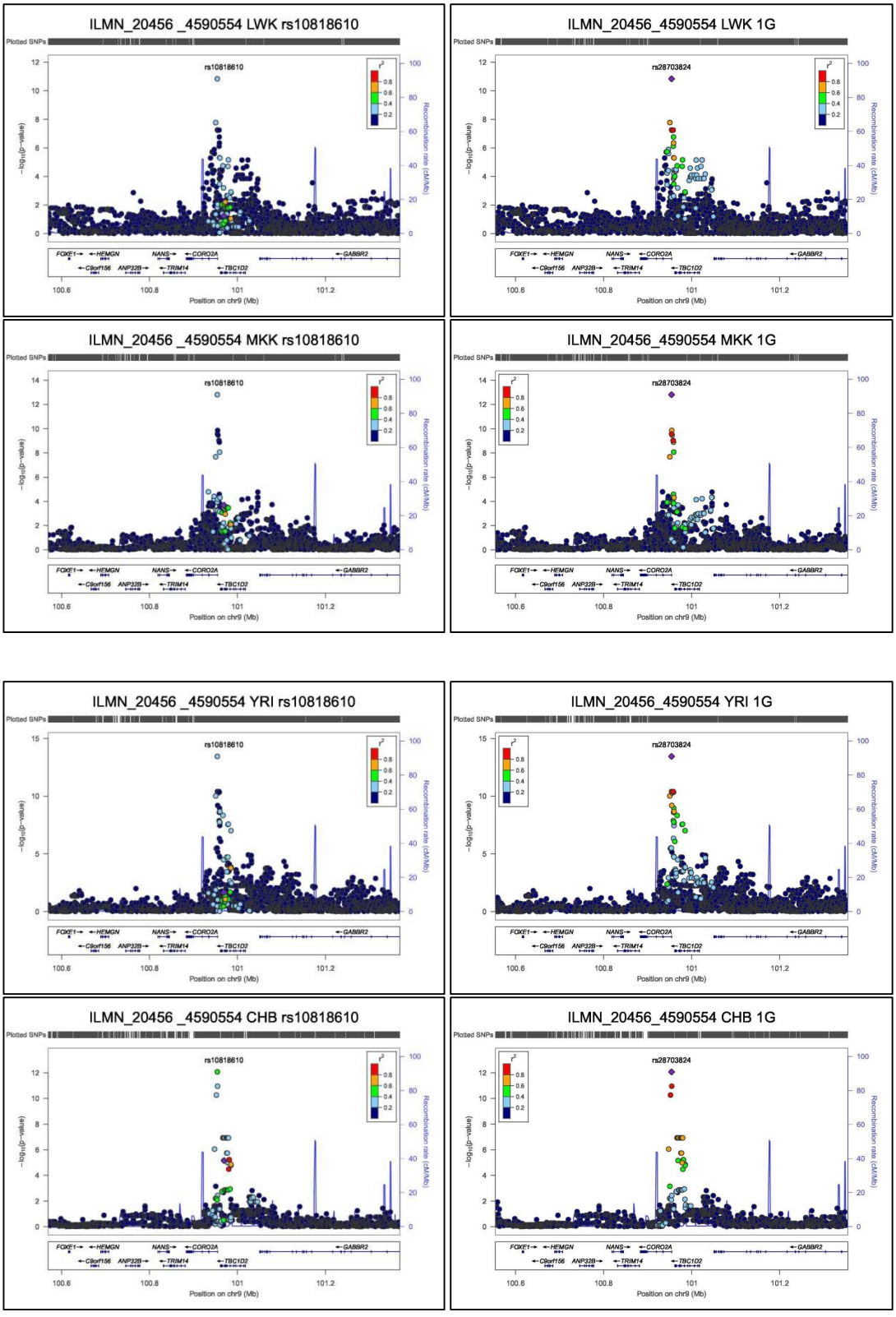


Figure 5.20: Signal plots for probe ILMN_20456_4590554 for Phase III HapMap and 1000 Genomes SNPs. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012)

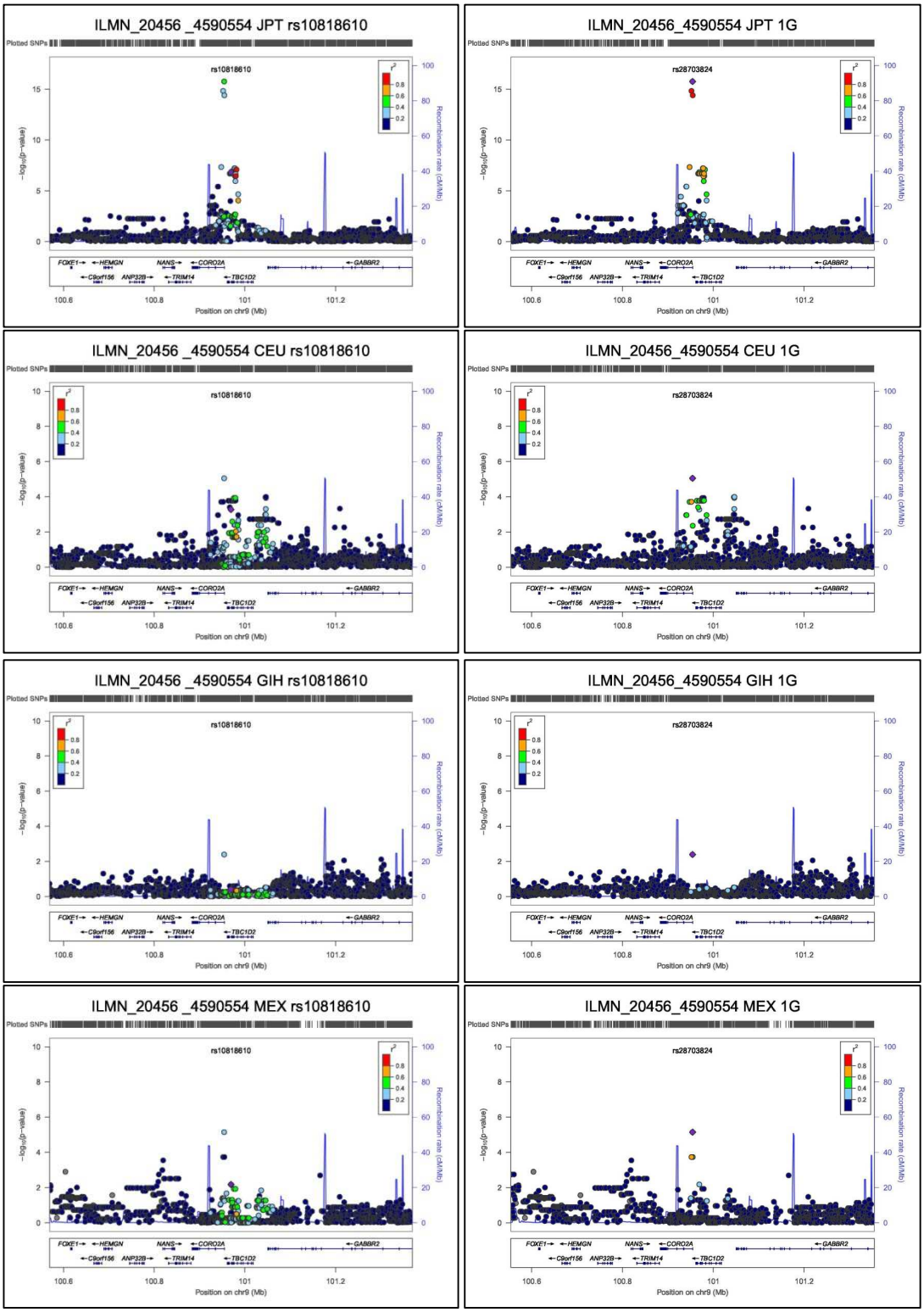


Figure 5.21: Signal plots for probe ILMN_20456_4590554 for Phase III HapMap and 1000 Genomes SNPs. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

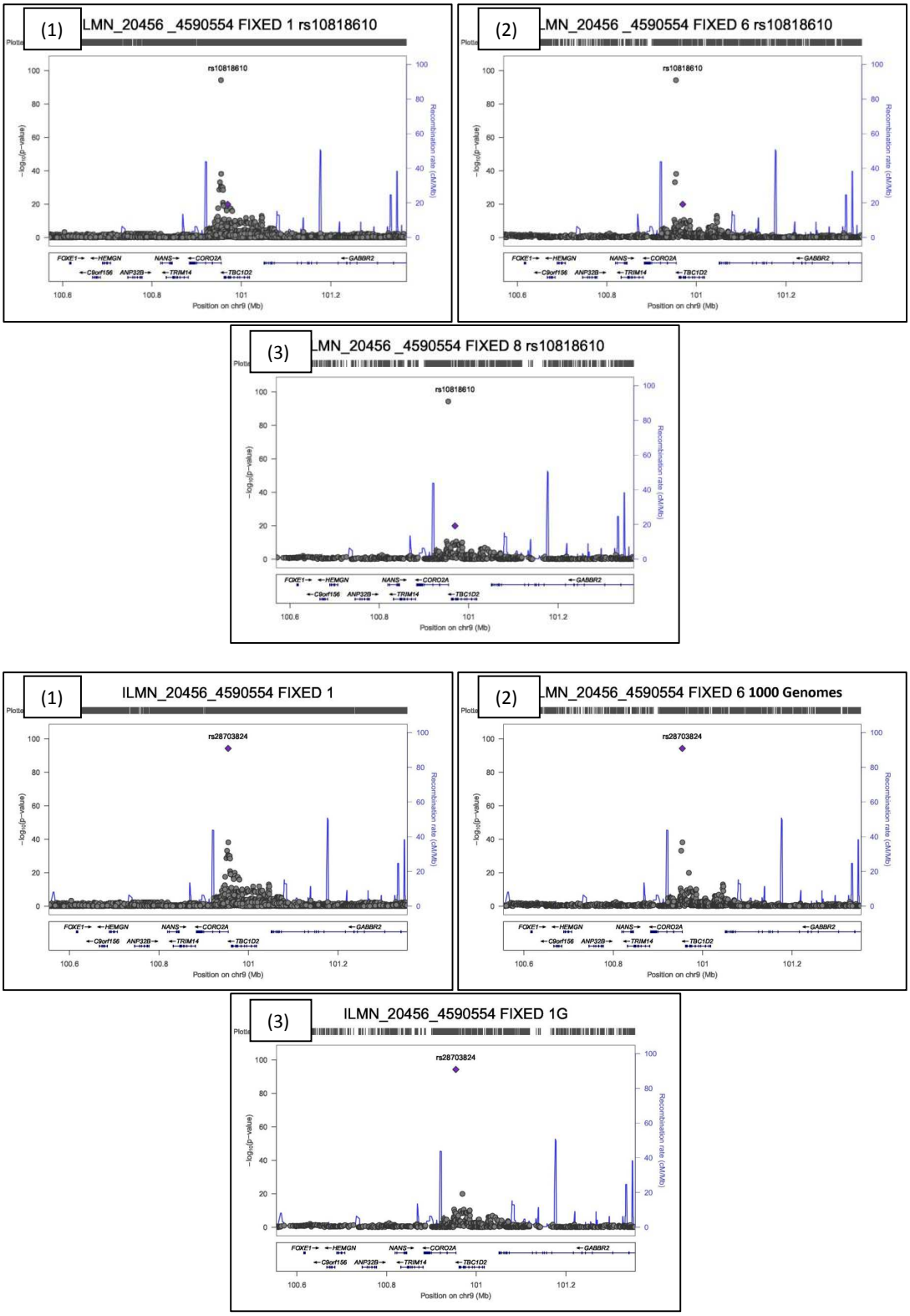


Figure 5.22: Signal plots for probe ILMN_20456_4590554 for Phase III HapMap and 1000 Genomes SNPs. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond.

5.7.5 ILMN_1295_2340291 (ENSG00000064886, CHI3L2)

In the fixed effect meta-analysis, this probe has a large reduction in heterogeneity between the peak 1000 Genomes eSNP rs3118207 (Z-score=-18.62, p-value=$2.56x10^{-77}$, Cochran's Q p-value = 0.124) and peak Phase III HapMap eSNP rs11102223 (Z-score= -13.81, p-value= $2.51x10^{-43}$, Cochran's Q p-value = $7.07x10^{-8}$) (See table 5.14).

Gene Name and Location

The probe ILMN_1295_2340291 gene has the Ensembl ID ENSG00000064886 and the HGNC name *CHI3L2*. The full name of the gene is Chitinase 3-Like 2, and the start position is chr1:111743393. This gene is similar in structure to bacterial chitinases, but does not have this activity. The encoded protein is involved in cartilage biogenesis. The 1000 Genomes peak SNP is rs3118207, which is located at chr1:111768576, and is intergenic between the genes DENND2D (21416 bp) and CHI3L2 (1705 bp). The Phase III HapMap SNP is rs11102223, which is located at chr1:111791134.

Linkage Disequilibrium

Table 5.23 shows LD between the two SNPs rs3118207 and rs11102223. SNPs in the Eurasian-Hispanic and East Asian ancestry groups are moderately correlated with each other. No data is available for the African ancestry group.

Population (Ancestry Group)	**r^2**
CEU (Eurasian-Hispanic)	0.444
CHB + JPT (EastAsian)	0.334
YRI (African)	--

Table 5.23: LD between rs11102223 and rs3118207 using Broad SNAP.

Figure 5.23 shows a forest plot for the population-specific association analyses and fixed effect meta-analysis results for the probe ILMN_1295_2340291 with the Phase III HapMap SNP rs11102223. A can be seen, effect sizes for populations of African ancestry cluster together, with LWK and YRI not significantly different from zero at $p \leq 0.05$.

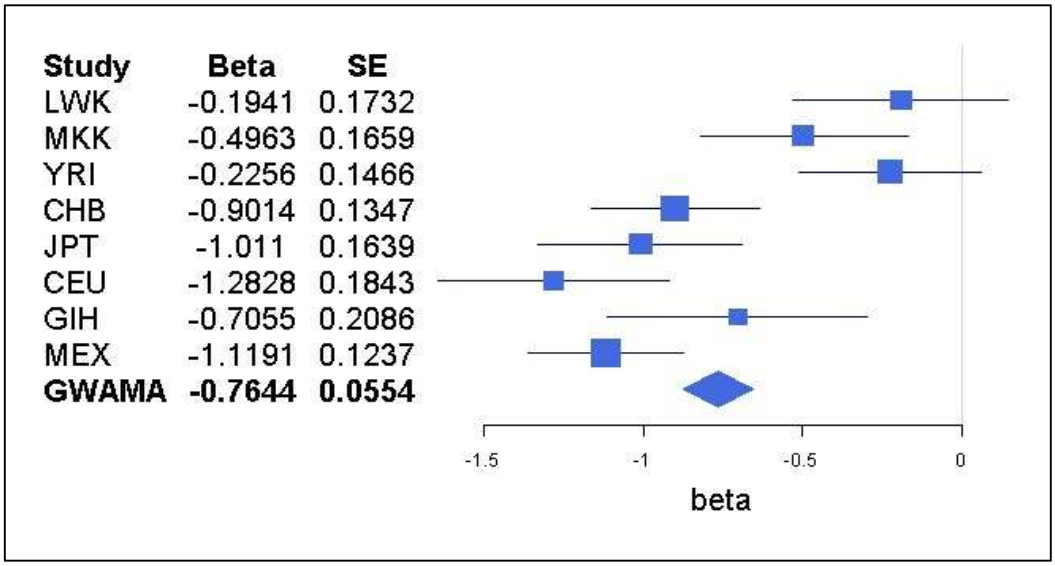


Figure 5.23: Forest Plot for probe ILMN_1295_2340291 with Phase III HapMap SNP rs11102223

Table 5.24 presents allele frequencies, INFO scores and association analysis and fixed effect meta-analysis results for probe ILMN_1295_2340291 with the Phase III HapMap SNP rs11102223. Allele frequencies range from 0.25 to 0.57. The association signals for CHB, JPT, CEU and MEX are significant at GWS.

Population	Minor Allele (G / T)	Allele Frequencies (Allele G)	Beta	SE	p-value	INFO
LWK	G	0.39	-0.1941	0.1732	0.266	0.999
MKK	G	0.40	-0.4963	0.1659	3.31×10^{-3}	0.996
YRI	G	0.49	-0.2256	0.1466	0.127	0.984
CHB	G	0.31	-0.9014	0.1347	3.35×10^{-9}	0.972
JPT	G	0.25	-1.011	0.1639	2.86×10^{-8}	0.997
CEU	G	0.49	-1.2828	0.1843	3.14×10^{-10}	1.000
GIH	G	0.41	-0.7055	0.2086	1.17×10^{-3}	0.994
MEX	T	0.57	-1.1191	0.1237	5.17×10^{-11}	1.000
Fixed effect Meta-analysis	--	--	-0.7644	0.0554	2.51×10^{-43}	--

Table 5.24: Table of allele frequencies for Phase III HapMap SNP rs11102223 with probe ILMN_1295_2340291

Figure 5.24 is a forest plot of the population-specific association analyses and fixed effect meta-analysis results for the probe ILMN_2195_2340291 with the 1000 Genomes SNP rs3118207. As can be seen the heterogeneity has reduced, African ancestry groups form a cluster with the other populations.

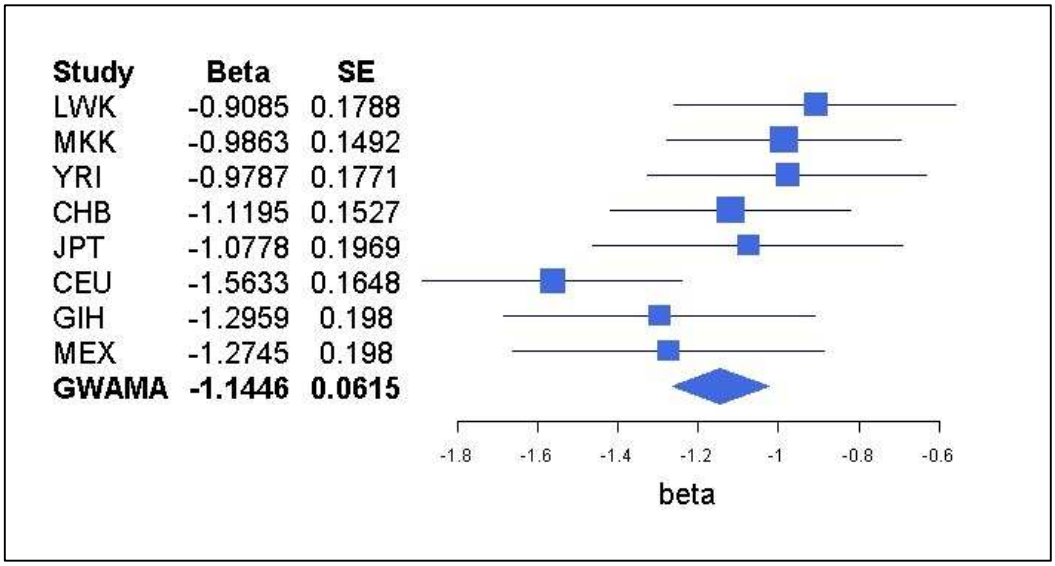


Figure 5.24: Forest Plot for probe ILMN_1295_2340291 with 1000 Genomes SNP rs3118207

Table 5.25 presents results for the association analysis and fixed effect meta-analysis together with allele frequencies and INFO scores for the probe ILMN_1295_2340291 with the 1000 Genomes SNP rs3118207. Allele frequencies range from 0.12 to 0.35. All association signals are significant at $p \leq 0.05$, and for MKK, CHB, CEU and GIH, achieve GWS.

Population	Minor Allele (A / G)	Allele Frequencies (Allele A)	Beta	SE	p-value	INFO
LWK	A	0.21	-0.9085	0.1788	2.40×10^{-6}	0.973
MKK	A	0.28	-0.9863	0.1492	8.56×10^{-10}	0.953
YRI	A	0.21	-0.9787	0.1771	2.39×10^{-7}	0.921
CHB	A	0.16	-1.1195	0.1527	2.09×10^{-10}	0.840
JPT	A	0.12	-1.0778	0.1969	5.15×10^{-7}	0.987
CEU	A	0.35	-1.5633	0.1648	9.41×10^{-16}	0.995
GIH	A	0.29	-1.2959	0.1980	7.58×10^{-9}	0.978
MEX	A	0.22	-1.2745	0.1980	1.44×10^{-7}	1.000
Fixed effect Meta-analysis	--	--	-1.1446	0.0615	2.56×10^{-77}	--

Table 5.25: Table of allele frequencies for 1000 Genomes SNP rs3118207 with probe ILMN_1295_2340291

Figure 5.25 and 5.26 presents signal plots for association analysis results and figure 5.27 presents fixed effect meta-analysis results. Peak Phase III HapMap and 1000 Genomes SNPs are marked in each of the plots.

The reduction in heterogeneity is primarily due to stronger effect sizes in the African populations after imputation. Significance levels before imputation (LWK=0.266, MKK=3.31×10^{-3}, YRI=0.127) are weaker than those after imputation (LWK=2.40×10^{-6}, MKK=8.56×10^{-10}, YRI=2.39×10^{-7}). This indicates that signals in the African populations are being uncovered by the imputed SNPs. The 1000 Genomes peak SNP remains peak, when non-reported SNPs are included, indicating that the peak SNP after imputation is optimal.

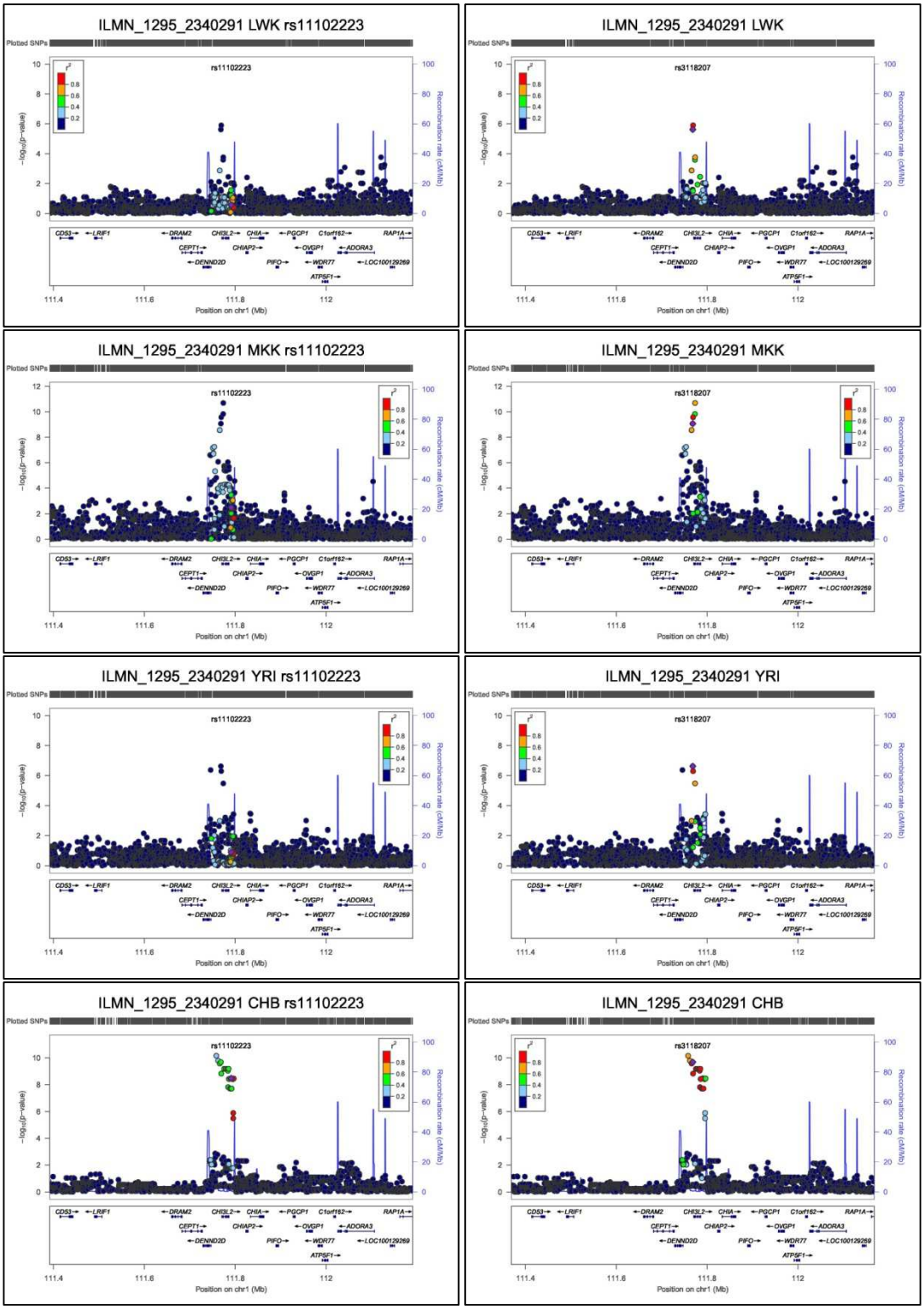


Figure 5.25: Signal plots for probe ILMN_1295_2340291 for Phase III HapMap and 1000 Genomes SNPs. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

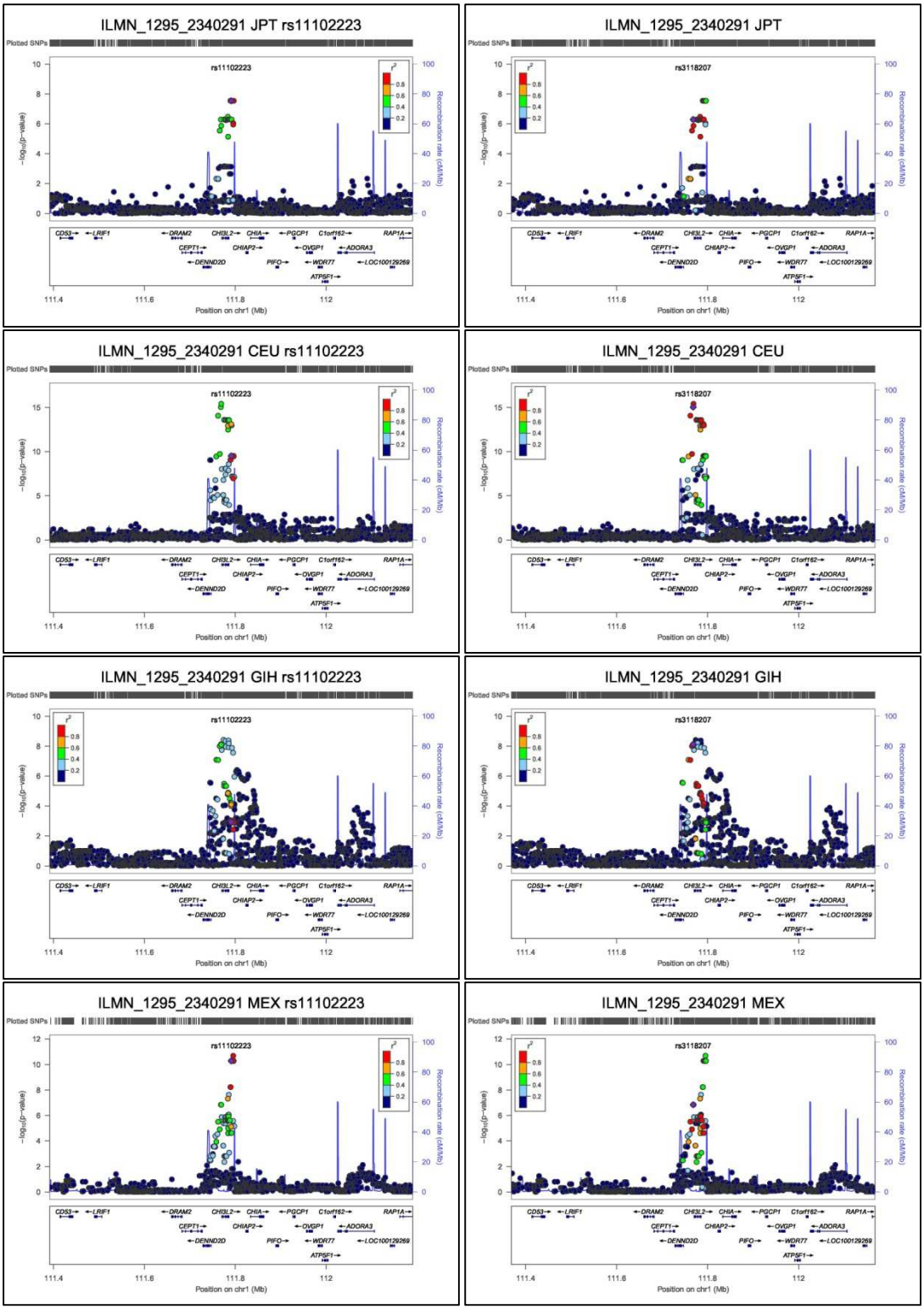


Figure 5.26: Signal plots for probe ILMN_1295_2340291 for Phase III HapMap and 1000 Genomes SNPs. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012).

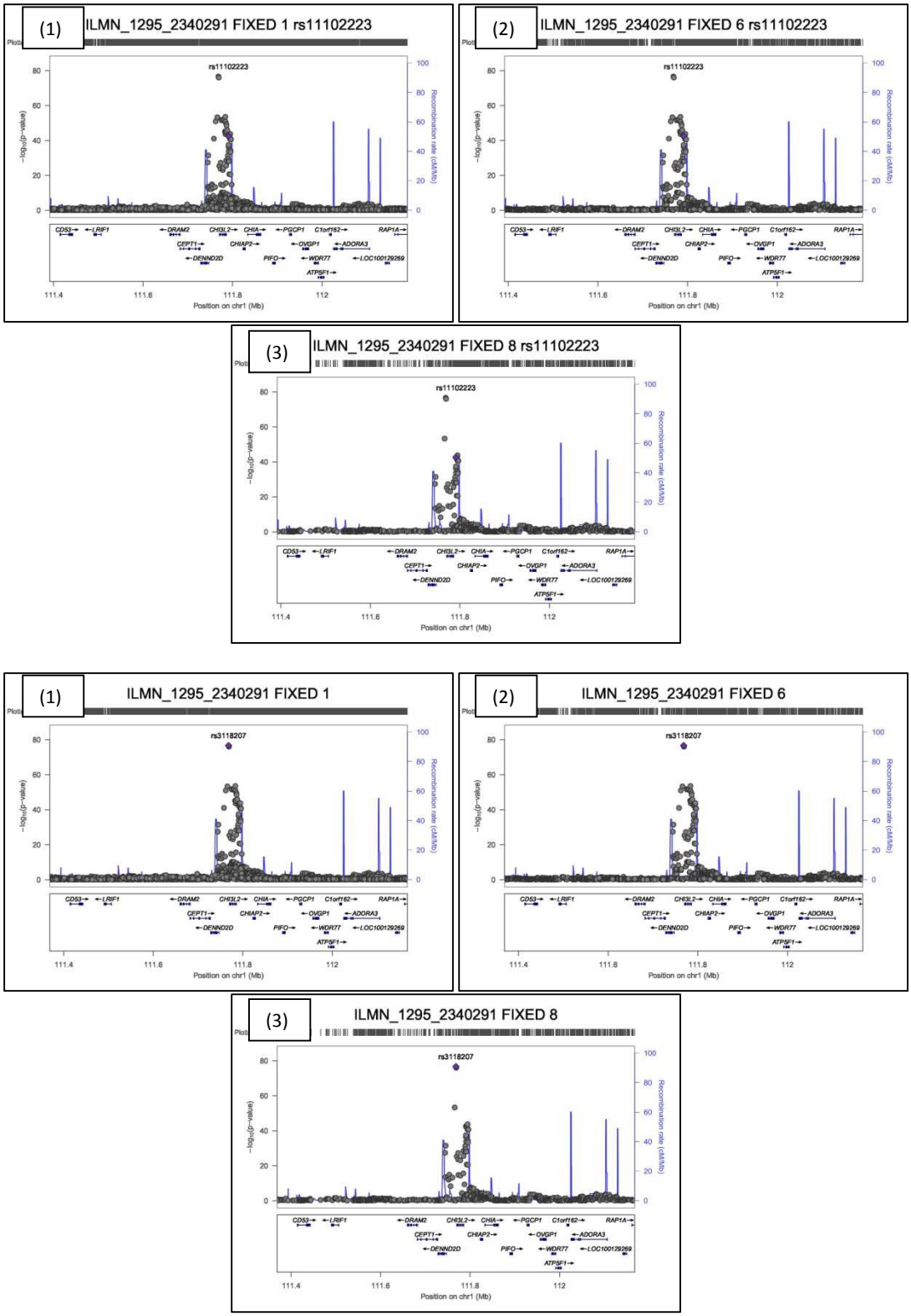


Figure 5.27: Signal plots for probe ILMN_1295_2340291 for Phase III HapMap and 1000 Genomes SNPs. Each circle represents a Phase III HapMap or 1000 Genome SNP. For each locus, the lead SNP is represented as a purple diamond.

5.8 Summary

This chapter has presented the results of imputation of Phase III HapMap genotypes with the 1000 Genomes "all ancestries" March 2012 panel. These imputed genotypes have been used in an association analysis and fixed effect meta-analysis with Phase III HapMap microarray expression data. The aims of this analysis were to: 1) Fine map *cis* eQTLs using additional variants from the 1000 Genomes, and 2) Look at the result of imputation on heterogeneity. It was found that out of the 1811 probes detected from 1000 Genome imputed SNPs, 1238 had an improvement in signal with 1000 Genomes imputed SNPs. Out of 1238 probes with an improvement in imputed SNPs, 657 had a larger Cochran's Q statistic in the Phase III HapMap SNP.

Annotation of probes detected in this analysis indicates that over half of the eSNPs are within introns, and approximately one quarter are intergenic. Region based annotation have revealed that approximately one quarter of these eSNPs are located within, transcription factor binding, DNase sensitivity or histone modification regions.

Examples from chapter 4 have been re-visited with 1000 Genomes imputed SNPs, these are: 1) *WBSCR27* which has the most significant z-score with no heterogeneity. 2) *CAV2* which has strong evidence of heterogeneity. 3) *FANCA* which also has strong evidence of heterogeneity. An example where using the 1000 Genomes imputed data has led to a large improvement in signal has been presented (*CORO2A).* An example where using the 1000 Genomes imputed data has led to a reduction in heterogeneity has been presented (*CHI3L2)* This is a good example were the 1000 Genomes is causing reduction in heterogeneity, due to tagging signal in African populations. One of the uses of eQTLs is in helping infer the mechanism in GWAS – so in the next section is presented the results of a GWAS Integration using the 1000 Genomes imputed data.

CHAPTER 6 GWAS INTEGRATION

6.1 Overview

The National Human Genome Research Institute (NHGRI) regularly publishes a catalogue of Genome Wide Association Study (GWAS) results. Many of the disease SNPs (dSNPs) identified in these studies are located within intergenic regions, sometimes within regions with no candidate genes, and have no obvious functional impact.

A possible hypothesis for the function of these intergenic dSNPs could be through variants within elements that control transcription such as promoter, enhancers, silencers and transcription factor binding sites. In order to test this hypothesis, dSNPs from the catalogue can be integrated with peak SNPs (eSNPs) from a cis eQTL study to detect candidate genes through which the observed association signal influences disease.

This section presents the results of integrating the dSNPs from the NHGRI GWAS catalog (accessed 05-03-2014) with the eSNPs detected with the 1000 Genomes imputed Phase III HapMap fixed effect meta-analysis data described in chapter 5. Briefly, reciprocal conditional analysis has been used to determine whether the dSNP and eSNP represent the same *cis* eQTL signal.

Positions of SNPs and genes in this section use NCBI Build 37. Gene functions are not cited as the information has been taken from GeneCards and so is assumed to be public knowledge.

6.2 Pipeline

I have designed an analysis pipeline to identify dSNP – eSNP pairs that are the same variant or tag each other (and therefore represent the same underlying association signal). The workflow uses reciprocal conditional analysis, as described in methods section 2.13.1.

The analysis has the following workflow. First, filter the dSNP catalogue for those which are significant at GWS ($p \leq 5x10^{-8}$). Following this, generate a list of (filtered) dSNP and eSNPs pairs

that map within 1 Mb of each other. Then, perform a reciprocal conditional analysis with the eSNP and dSNP pairs. Finally, filter results using the detection criteria specified below. Full details of the GWAS integration pipeline are presented in methods section 2.13.

6.2.1 Detection criteria

Three detection criteria are used to select dSNP – eSNP pairs for follow up are as follows:

Criterion 1: The dSNP and eSNP are the same variant.

Criterion 2: Performing reciprocal analysis removes the signal in both the eSNP and the dSNP. The association signal of both the dSNP and eSNP are extinguished after including the other as a covariate in conditional analysis (p-value ≥ 0.05). Although the eSNP and dSNP are not the same, they likely represent the same underlying association signal.

Criterion 3: Performing reciprocal conditional analysis removes part of the signal in both the eSNP and the dSNP (p-value ≥ $1x10^{-3}$), and the disease phenotype is relevant to the cell type studied (here, LCLs are relevant to immunological phenotypes). In this category, the dSNP and eSNP may reflect the same underlying association signal, but the potential importance of the gene for which the eSNP was detected is bolstered by the relevance of the cell type to the disease studied. This is important as eQTLs can operate in a cell-type specific manner, as shown in the study Dimas *et al* 2009, which looked at the overlap of eQTLs in primary fibroblasts, LCLs and T-Cells in 75 individuals. It was found that 69 – 80% of eQTL regulatory elements operated in a cell-type specific manner.

In order to determine whether the eSNP – dSNP pairs had been detected previously, the GTEX database (GTEx Consortium 2013) was used to detect whether an existing eQTL is reported at the eSNP for whole blood. If the eSNP exists in the GTEX database then the eSNP – dSNP pair is reported as known. This information is included in each of the tables in the column eSNP Reported (GTEX Whole Blood).

6.3 Results

6.3.1 Criterion 1

Table 6.1 shows the results for criterion 1 where the dSNP and eSNP are the same variant. In total 12 probes were identified that fulfilled this criteria. The following general observations can be made about these results. First, there are four probes that have eSNPs which are also dSNPs for cervical cancer. Two the probes are for the same gene (*ZBPB2*). All four probes have the same eSNP, rs8067378, identified in the study Shi *et al* 2013. Second, for the twelve probes detected, there are four dSNPs that are associated with five immunological phenotypes: Ulcerative Colitis (UC), Inflammatory Bowel Disease (IBD), Bechet's Disease, Sarcoidosis and IgG glycosylation.

6.3.2 Criterion 2

Table 6.2 shows the results for criterion 2, where the dSNP and eSNP extinguish the association signal in reciprocal conditional analysis. In total 12 dSNP-eSNP pairs were identified which fulfilled this criteria. The following general observations can be made. First, of the twelve dSNP – eSNP pairs, six are associated with immunological function phenotypes: UC, self-reported allergy, Crohn's Disease, IBD, Grave's Disease and Primary Biliary Cirrhosis (PBC). Second, there are two dSNP - eSNP pairs for multiple sclerosis, and both are cis eQTLs for the same gene, *EOMES.* Third, there are two dSNP – eSNP pairs for Crohn's Disease, and are cis eQTLs for the genes *CARD9* and *GPX4*.

6.3.3 Criterion 3

Table 6.3 shows the results for criterion 3, where the dSNP and eSNP partially explain the association signal of the other in reciprocal conditional analysis, and the disease phenotype has an immunological function. In total 18 dSNP-eSNP pairs were identified that fulfilled this criteria. The following general observations can be made. First there are five dSNP – eSNP pairs for IBD, all of which were identified in the study Jostins *et al* 2012, each mapping to different regions of the genome. Second, there are two dSNP – eSNP pairs for UC, but both of which are cis eQTLs for the

same gene, *TNFRSF14*. Third, there are two dSNP - eSNP pairs for Bechet's Disease, and are cis eQTLs for *STAT4* and *KLRK1*.

Gene				SNP		p-value	Disease(s) / Phenotype(s)	Publication	eSNP Reported (GTEX Whole Blood)
Probe	Ensembl	HGNC	Start Site	ID	Position (B37)				
ILMN_22365_5870093	ENSG00000187796	*CARD9*	chr9:139268133	rs10781499	chr9:139266405	$7.35x10^{-21}$	UC	Anderson *et al* 2011 Nat Genet	Y
							IBD	Jostins *et al* 2012 Nature	Y
ILMN_1521_1990669	ENSG00000142224	*IL19*	chr1:206972215	rs1518111	chr1:206944645	$9.52x10^{-11}$	Behcet's disease	Remmers *et al* 2010 Nat Genet	N
ILMN_137122_1440324	ENSG00000168071	*CCDC88B*	chr11:64107695	rs479777	chr11:64107477	$1.01x10^{-15}$	Sarcoidosis	Fischer *et al* 2012 Am J Respir Crit Care Med	Y
ILMN_29393_6520095	ENSG00000139197	*PEX5*	chr12:7341281	rs7973719	chr12:7334813	$5.59x10^{-22}$	IgG glycosylation	Lauc *et al* 2013 PLoS Genet	N
ILMN_12329_4040132	ENSG00000172057	*ORMDL3*	chr17:38083854	rs8067378	chr17:38051348	$4.21x10^{-35}$	Cervical cancer	Shi *et al* 2013 Nat Genet	Y
ILMN_137680_2710228	ENSG00000186075	*ZPBP2*	chr17:38024417	rs8067378	chr17:38051348	$2.24x10^{-19}$	Cervical cancer	Shi *et al* 2013 Nat Genet	N
ILMN_138844_6040674	ENSG00000186075	*ZPBP2*	chr17:38024417	rs8067378	chr17:38051348	$2.62x10^{-12}$	Cervical cancer	Shi *et al* 2013 Nat Genet	N
ILMN_19619_6620170	ENSG00000073605	*GSDMB*	chr17:38077313	rs8067378	chr17:38051348	$6.04x10^{-11}$	Cervical cancer	Shi *et al* 2013 Nat Genet	Y
ILMN_13293_5290674	ENSG00000128928	*IVD*	chr15:40697686	rs10518693	chr15:40700022	$4.73x10^{-24}$	Metabolic traits	--	--
ILMN_6495_4570739	ENSG00000164761	*TNFRSF11B*	chr8:119964439	rs2062377	chr8:120007420	$2.24x10^{-8}$	Bone mineral density	--	--
ILMN_6244_2760707	ENSG00000106366	*SERPINE1*	chr7:100770370	rs2227631	chr7:100769538	$2.51x10^{-8}$	Plasminogen activator inhibitor type 1 levels (PAI-1)	--	--
ILMN_12727_5720647	ENSG00000135074	*ADAM19*	chr5:157002783	rs2277027	chr5:156932376	$3.14x10^{-17}$	Pulmonary function	--	--

Table 6.1: *Criterion 1* matches: dSNP and eSNP matches with fixed effect meta-analysis p-value ≤ $5x10^{-8}$ only including SNPs with MAF ≥ 5% and INFO ≥ 0.4

Gene				Disease(s) / Phenotype(s)	Publication(s)	dSNP			eSNP			Distance	LD (r_2) (ASN,CEU,YRI)	eSNP Reported (GTEX – Whole Blood)
Probe ID	Ensembl	HGNC	Start Site			Name	p-value	Conditional	Name	p-value	Conditional			
ILMN_22365_5870093	ENSG00000187796	*CARD9*	chr9:139268133	Crohn's disease	Franke et al 2010 Nat Genet	rs4077515	$1.33x10^{-18}$	0.647	rs10781499	$1.25x10^{-19}$	0.490	91	(1.0,1.0,1.0)	Y
				UC	McGovern *et al* 2010 Nat Genet									Y
ILMN_13328_5050541	ENSG00000171522	*PTGER4*	chr5:4069600	Self-reported allergy	Hinds *et al* 2013 Nat Genet	rs7720838	$2.05x10^{-17}$	0.251	rs10440635	$1.62x10^{-18}$	0.109	3894	(1.0,1.0,0.449)	N
ILMN_16320_1570414	ENSG00000166501	*PRKCB*	chr16:23847322	IBD	**Jostins *et al* 2012 Nature**	rs7404095	$1.51x10^{-14}$	0.616	rs7193632	$1.65x10^{-15}$	0.174	580	(0.966,0.965,1.0)	N
ILMN_7307_4590646	ENSG00000160856	*FCRL3*	chr1:157670775	Graves' disease	Chu *et al* 2011 Nat Genet	rs3761959	$8.66x10^{-11}$	0.552	rs2210913	$1.35x10^{-14}$	0.105	285	(1.0,1.0,0.167)	Y
ILMN_138844_6040674	ENSG00000186075	*ZPBP2*	chr17:38024417	PBC	Nakamura *et al* 2012 Am J Hum Genet	rs9303277	$2.69x10^{-12}$	0.148	rs8067378	$4.27x10^{-13}$	0.111	74879	(0.836,1.0,0.837)	Y
				PBC	Liu *et al* 2010 Nat Genet									Y
ILMN_19998_6760075	ENSG00000163508	*EOMES*	chr3:27764206	Multiple sclerosis	IMSGC *et al* 2011 Nature	rs669607	$3.88x10^{-10}$	0.080	rs413544	$7.72x10^{-11}$	0.092	2640	(0.967,0.967,1.0)	N
ILMN_19998_6760075	ENSG00000163508	*EOMES*	chr3:27764206	Multiple sclerosis	Patsopoulos *et al* 2011 Ann Neuro	rs170934	$2.97x10^{-10}$	0.747	rs413544	$7.72x10^{-11}$	0.659	5001	(1.0,1.0,0.526)	N
ILMN_137100_5270672	ENSG00000167468	*GPX4*	chr19:1103936	Crohn's Disease	**Jostins *et al* 2012 Nature**	rs2024092	$7.70x10^{-7}$	0.641	rs8178977	$1.47x10^{-8}$	0.071	17554	(0.656,0.744,0.538)	N
ILMN_8696_3930754	ENSG00000016864	*GLT8D1*	chr3:52740048	Adiponectin levels	--	rs2590838	$1.33x10^{-7}$	0.206	rs4435633	$3.37x10^{-8}$	0.051	65476	(NA,NA,NA)	--
ILMN_138428_3130066	ENSG00000185808	*PIGP*	chr21:38445470	Eye color traits	--	rs1003719	$4.55x10^{-20}$	0.980	rs2154536	$4.74x10^{-20}$	0.374	1052	(1.0,1.0,1.0)	--
ILMN_5873_6480368	ENSG00000173821	*RNF213*	chr17:78234665	Moyamoya disease	--	rs6565681	$1.62x10^{-17}$	0.512	rs4890020	$5.87x10^{-18}$	0.088	24340	(NA,NA,NA)	--
ILMN_5798_1580669	ENSG00000139428	*MIMAB*	chr12:110011679	Triglycerides	--	rs7134594	$1.03x10^{-9}$	0.640	rs7296044	$1.79x10^{-10}$	0.079	16754	(1.0,0.791,0.802)	--

Table 6.2: *Criterion 2 matches*: Results of conditional analysis between dSNP and eSNPs, where conditional p-value for eSNP ≥ 0.05. All SNPs have MAF ≥ 5% and INFO score ≥ 0.4.

Gene				Disease(s) / Phenotype(s)	Publication(s)	dSNP			eSNP			Distance	LD (r_2) (ASN,CEU,YRI)	eSNP Reported (GTEX – Whole Blood)
Probe ID	Ensembl	HGNC	Start Site			Name	p-value	Conditional	Name	p-value	Conditional			
ILMN_12426_4540301	ENSG00000123427	*METT21B*	chr12:58165275	Multiple sclerosis	ANZgene 2009 Nat Genet	rs703842	$8.88x10^{-40}$	0.259	rs923829	$5.01x10^{-43}$	0.041	11567	(1.0,0.963,0.669)	Y
ILMN_8937_1770468	ENSG00000138378	*STAT4*	chr2:192016322	Behcet's disease	Hou *et al* 2012 Arthritis Rheum	rs897200	$6.82x10^{-33}$	0.605	rs1031507	$9.44x10^{-35}$	0.012	2847	(1.0,0.898,0.775)	N
ILMN_29874_6480543	ENSG00000213809	*KLRK1*	chr12:10544473	Behcet's disease	Kirino *et al* 2013 Nat Genet	rs2617170	$5.08x10^{-24}$	0.186	rs1841957	$1.81x10^{-27}$	0.004	1068	(0.851,0.961,0.381)	N
ILMN_23059_6650537	ENSG00000030110	*BAK1*	chr6:33548019	Chronic lymphocytic leukemia	Berndt *et al* 2013 Nat Genet	rs210142	$4.77x10^{-24}$	0.187	rs210143	$9.45x10^{-27}$	0.001	93	(NA,NA,NA)	Y
				Chronic lymphocytic leukemia	Slager *et al* 2012 Blood									Y
ILMN_137680_2710228	ENSG00000186075	*ZPBP2*	chr17:38024417	PBC	Nakamura *et al* 2012 Am J Hum Genet	rs9303277	$1.72x10^{-16}$	0.473	rs8067378	$7.39x10^{-20}$	0.006	74879	(0.836,1.0,0.837)	Y
				PBC	Liu *et al* 2010 Nat Genet									Y
ILMN_11334_3870670	ENSG00000149311	*ATM*	chr11:108093211	Response to metformin	Florez *et al* 2011 Curr Diab Rep	rs11212617	$3.17x10^{-17}$	0.384	rs676729	$1.63x10^{-18}$	0.009	98320	(NA,NA,NA)	N
				Response to metformin in type 2 diabetes	GoDarts... *et al* 2011 Nat Genet									N
ILMN_14543_4590767	ENSG00000135926	*TIMBIM1*	chr2:219157309	IBD	**Jostins *et al* 2012 Nature**	rs2382817	$3.57x10^{-15}$	0.509	rs12987219	$2.14x10^{-18}$	0.006	7069	(1.0,0.964,0.539)	Y
ILMN_12900_4070215	ENSG00000166501	*PRKCB*	chr16:23847322	IBD	**Jostins *et al* 2012 Nature**	rs7404095	$2.77x10^{-12}$	0.359	rs3785403	$6.60x10^{-14}$	0.003	2433	(NA,NA,NA)	N
ILMN_3329_3190551	ENSG00000157873	*TNFRSF14*	chr1:2487078	UC	**Jostins *et al* 2012 Nature**	rs10797432	$9.73x10^{-11}$	0.770	rs2281852	$4.57x10^{-12}$	0.012	10396	(NA,NA,NA)	Y
ILMN_3329_3190551	ENSG00000157873	*TNFRSF14*	chr1:2487078	UC	Anderson *et al* 2011 Nat Genet	rs734999	$1.02x10^{-9}$	0.742	rs2281852	$4.57x10^{-12}$	0.002	22274	(NA,NA,NA)	Y
ILMN_26929_6380088	ENSG00000138031	*ADCY3*	chr2:25142708	IBD	**Jostins *et al* 2012 Nature**	rs6545800	$5.57x10^{-10}$	0.400	rs13387729	$2.77x10^{-11}$	0.032	42759	(NA,NA,NA)	Y
ILMN_19619_6620170	ENSG00000073605	*GSDMB*	chr17:38077313	Type 1 diabetes	Barrett *et al* 2009 Nat Genet	rs2290400	$2.60x10^{-10}$	0.0823	rs8067378	$3.37x10^{-11}$	0.032	14892	(1.0,1.0,0.410)	Y
ILMN_17345_510437	ENSG00000131507	*NDFIP1*	chr5:141488070	IBD	**Jostins *et al* 2012 Nature**	rs6863411	$5.57x10^{-9}$	0.174	rs12515668	$3.21x10^{-10}$	0.008	17114	(NA,NA,NA)	N
ILMN_12900_7150095	ENSG00000166501	*PRKCB*	chr16:23847322	IBD	**Jostins *et al* 2012 Nature**	rs7404095	$4.52x10^{-8}$	0.585	rs8056879	$1.70x10^{-9}$	0.004	1633	(NA,NA,NA)	N
ILMN_5266_6840672	ENSG00000197712	*FAM114A1*	chr4:38869298	Allergic sensitization	Bønnelykke *et al* 2013 Nat Genet	rs17616434	$9.23x10^{-9}$	0.682	rs6815814	$2.12x10^{-9}$	0.031	3462	(NA,NA,NA)	N
ILMN_5266_6840672	ENSG00000197712	*FAM114A1*	chr4:38869298	Self-reported allergy	Hinds *et al* 2013 Nat Genet	rs2101521	$9.38x10^{-8}$	0.211	rs6815814	$2.12x10^{-9}$	0.027	4787	(NA,NA,NA)	N
ILMN_7544_6270026	ENSG00000160856	*FCRL3*	chr1:157670775	Graves' disease	Chu *et al* 2011 Nat Genet	rs3761959	$4.36x10^{-5}$	0.207	rs945635	$1.19x10^{-8}$	0.032	1012	(NA,NA,NA)	Y
ILMN_8696_3930754	ENSG00000016864	*GLT8D1*	chr3:52740048	Osteoarthritis	arcOGEN *et al* 2012 Lancet	rs11177	$2.52x10^{-7}$	0.022	rs4435633	$3.37x10^{-8}$	0.005	33743	(NA,NA,NA)	N

Table 6.3: *Criterion 3 matches:* Results of conditional analysis between dSNP and eSNPs, where conditional p-value for eSNP ≥ $1x10^{-3}$. All SNPs have MAF ≥ 5% and INFO score ≥ 0.4. Only dSNPs with disease with plausible function in LCLs have been selected.

6.4 Gene Function

This section expands on findings from the GWAS integration pipeline.

6.4.1 CARD9

The gene *CARD9* has the full name Caspase Recruitment Domain-Containing Protein 9 and is represented by the probe ILMN_22365_5870093. *CARD9* is a member of the Caspace Recruitment Domain (CARD) family. It is thought to function as a molecular scaffold for the assembly of the BCL10 signalling complex that activates NF-Kappa-B. Diseases known to be associated with this gene are deep dermatophytosis and candidiasis, familial 2, auto recessive.

CARD9 has an eSNP (rs10781499) that can explain the association signal for dSNPs (rs10781499, rs4077515) with the diseases UC, IBD and Crohn's disease. UC and Crohn's disease are both forms of IBD. The eSNP rs10781499 is positioned at chr9:139266405, and is exonic within *CARD9*. It is a variant for a synonymous codon, amino acid position 42, residue Proline.

The eSNP rs10781499 for *CARD9* is also a dSNP for UC and IBD in two studies. The dSNP rs10781499 for UC was identified in the study Anderson *et al* 2011. The dSNP rs10781499 for IBD was identified in the study Jostins *et al* 2012. *CARD9* was selected as the candidate gene in these studies because the dSNP rs10781499 is within an exon of this gene.

The eSNP rs10781499 also passes *criterion 2* conditional analysis with the dSNP rs4077515 for UC and Crohn's disease in two studies. The dSNP rs4077515 for UC was identified in the study McGovern *et al* 2010. The dSNP rs4077515 for Crohn's Disease was identified in the study Franke *et al* 2010. Again in these studies *CARD9* was selected as the candidate gene in these studies because the dSNP rs4077515 is within an exon of this gene.

CARD9 was identified as the relevant functional gene at this locus in each of the four studies, so the findings of the GWAS integration are confirmatory of previous reports.

6.4.2 IL19

The gene *IL19* has the full name Interleukin 19 and is represented by the probe ILMN_1521_1990669. *IL19* encodes a cytokine that belongs to the *IL10* cytokine subfamily. It can bind to *IL20* receptor complex and lead to the activation of the signal transduction and activator of transcription 3 (*STAT3*). The gene has a suggested role in inflammatory responses. Diseases known to be associated with *IL19* are melanoma and Richter's Syndrome.

IL19 has an eSNP (rs1518111) that is also reported as a dSNP with Bechet's Disease. Bechet's Disease is a rare immune-mediated, small-vessel systemic vasculitis. The dSNP rs1518111 was identified in the study Remmers *et al* 2010. The SNP rs1518111 is positioned at chr1:206944645 within the second intron in the gene *IL10 (*which was identified as the relevant functional gene at this locus on the basis of position). The findings of the GWAS integration contradict the identified functional gene in this report, and suggest that *IL19* is a better candidate gene through which the disease association signal is mediated.

6.4.3 CCDC88B

The gene *CCDC88B* has the full name Coiled-Coil Domain Containing Protein 88B and is represented by the probe ILMN_137122_1440324. The protein for this gene may be involved in linking organelles to micro-tubules. *CCDC88B* has an eSNP (rs479777) that is also reported as a dSNP for the auto-immune disease Sarcoidosis. Sarcoidosis is involved with the abnormal collection of inflammatory cells (granulomas) that can form nodules in multiple organs. The dSNP rs479777 was identified in the study Fischer *et al* 2012. The SNP rs479777 is positioned at chr11:64107477, upstream of gene *CCDC88B*. The gene *CCDC88B* was identified as the relevant functional gene at this locus because rs479777 is upstream of this gene, so the findings of the GWAS integration are again confirmatory of previous reports.

6.4.4 ZPBP2

The gene *ZPBP2* has the full name Zona Pellucida-Binding Protein 2 and is represented by the probes ILMN_137680_2710228 and ILMN_138844_6040674. *ZPBP2* is known to be associated with primary billiary cirrhosis. It may be implicated in gamete interaction during fertilization.

ZPBP2 has an eSNP (rs8067378) that is also reported as a dSNP for cervical cancer. The dSNP (rs8067378) for cervical cancer was identified in the study Shi *et al* 2013. The SNP rs8067378 is located at chr17:38051348, intergenic between the genes *ZPBP2* (17199 bases) and *GSDMB* (9500 bases).The gene GSDMB was identified as the relevant functional gene at this locus due to rs8067378 being located downstream of this gene, so the findings of the GWAS integration are contradictory of previous reports.

The ZPBP2 eSNP (rs8067378) also passes *criterion 2* conditional analysis with the dSNP rs9303277 for the disease Primary Billiary Cirrhosis (PBC). PBC is an autoimmune disease of the liver marked by the slow destruction of the small bile ducts of the liver. The dSNP (rs9303277) for PBC was identified in two studies: Nakamura *et al* 2012 and Liu *et al* 2010. The dSNP, rs9303277, is located at chr17:37976469, and maps to an intron of *IKZF3*. In both these studies, *IKZF3* was identified as

most likely functional gene at this locus due to the position within an intron. Therefore the findings of the GWAS integration contradict the identified functional gene in this report.

6.4.5 PRKCB

The gene *PRKCB* has the full name Protein Kinase C, Beta-1. The gene is represented by three probes: ILMN_16320_1570414, ILMN_12900_4070215 and ILMN_12900_7150095. The protein encoded by this gene is a member of the protein kinase C (PKC) family. It has been reported to be involved in B cell activation, apoptosis induction, endothelial cell proliferation and intestinal sugar absorption. The gene is associated with the diabetic macular edema.

The eSNP rs7193632 for the probe ILMN_16320_1570414 passes *criterion 2* conditional analysis with the dSNP rs7404095 for IBD. The eSNP rs3785403 for the probe ILMN_12900_4070215, and the eSNP rs8056879 for the probe ILMN_12900_7150095, both pass *criterion 3* conditional analysis with the dSNP rs7404095 for IBD. The dSNP rs7404095 for IBD was identified in the study Jostins *et al* 2012, and is located at chr16:23864590, within an intron of *PRKCB*. The eSNPs rs3785403 (chr16:23867023), rs7193632 (chr16:23865170) and rs8056879 (chr16:23866223) are all within introns of the gene *PRKCB*. The likely functional gene at this locus was reported to be *PRKBC* due to the SNP being within an intron of this gene, so the GWAS integration is confirmatory of these findings.

6.4.6 FCRL3

The gene *FCRL3* has the full name FC Receptor Like Protein 3. The gene is represented by two probes, namely ILMN_7307_4590646 and ILMN_7544_6270026. *FCRL3* contains immunoreceptor domains and may play role in regulation of the immune system. The gene has been found to be associated with rheumatoid arthritis, auto-immune thyroid disease and systemic lupus erythematosus. This gene is also associated with auto-immune pancreatic and juvenile rheumatoid arthritis. The gene is associated with a dSNP for Graves' Disease. Graves' Disease is an auto-immune disease, which affects the thyroid causing it to enlarge in size.

The eSNP rs2210913 for the probe ILMN_7307_4590646 passes criteria 2 with the dSNP rs3761959. The eSNP rs2210913 is intronic within the gene *FCRL3*. The eSNP rs945635 for the probe ILMN_7544_6270026 passes criteria 3 with the dSNP rs3761959. The eSNP rs945635 is within the UTR3 of *FCRL3*.The dSNP rs3761959 was identified in the study Chu *et al* 2011. This study identified *FCRL3* as the candidate gene as rs3761959 is within an intron of FCRL3, so the GWAS integration is confirmatory of these findings.

6.4.7 GPX4

The gene *GPX4* has the full name Glutathione Peroxidase 4. The gene is represented by the probe ILMN_137100_5270672. The gene is a member of the gluthathione peroxidase family, which is involved in the reduction of hydrogen peroxide, organic hydroperoxide and lipid peroxides by reduced gluthathione, and functions in the protection of cells against oxidative damage. Diseases associated with this gene are spondylometaphyseal dysplasia sedaghatian type and keshan disease.

GPX4 has an eSNP (rs8178977) which passes *criterion 2* conditional analysis with the dSNP rs2024092 for Crohn's disease. The dSNP was identified in the study Jostins *et al* 2012. The eSNP rs8178977 has the position chr19:1106477, and maps to an intron of the gene *GPX4*. *GPX4*, *HMHA1* and 20 other loci were identified as the most likely functional genes at this locus based on positional information. However, the GWAS integration would favour GPX4 as the more relevant gene on the basis of LCL expression.

6.4.8 STAT4

The gene *STAT4* has the full name Signal Transducer and Activator of Transcription 4. The gene is represented by the probe ILMN_8937_1770468. *STAT4* is a member of the Signal Transducer and Activator of Transcription (STAT) family of transcription factors. The protein is essential for mediating responses to Interleukin-12 (IL12) in lymphocytes, and for regulating the differentiation

of T helper cells. *STAT4* is associated with the disease Systemic Lupus Erythematosus and Rheumatoid Arthritis.

STAT4 has an eSNP (rs1031507) which passes *criterion 3* conditional analysis with the dSNP rs897200 for Bechet's Disease. The dSNP (rs897200) was identified in the study Hou *et al* 2012. The eSNP rs1031507 has the position chr2:192020618 and is within an intron of the gene *STAT4*. The dSNP rs897200 has the position chr2:192017771 and is intronic within STAT4: *STAT4* was identified as the likely functional gene at this locus, and GWAS integration thus confirms previous reports.

6.4.9 EOMES

The gene *EOMES* has the full name Eomesodermin. The gene is represented by the probe ILMN_19998_6760075. *EOMES* is a transcription factor which is crucial for embryonic development of mesoderm and the central nervous system. It may also be necessary for the differentiation of effector CD8+ T cells which are involved in defence of viral infection.

EOMES has an eSNP (rs413544) which passes *criterion 2* conditional analysis with two dSNPs (rs669607 and rs170934) for Multiple Sclerosis (MS). MS is an inflammatory disease in which the insulating covers of nerve endings in the brain and spinal cord are damaged. The eSNP rs413544 has the position chr3:28074084 and is intergenic between the genes *EOMES* (310,299 bases) and CMC1 (209,040 bases).The dSNP rs669607 was identified in the study IMSGC *et al 2011*, but no likely functional gene at this locus was identified in this study. The dSNP rs170934 was identified in the study Patsopoulos *et al 2011*, and EOMES was identified as the relevant functional gene due to being located closest to the SNP. The findings of GWAS integration are thus confirmatory of previous reports.

6.5 Summary

This chapter has presented an analysis to detect GWAS dSNPs associated with *cis* eQTL eSNP from the imputed dataset introduced in Chapter 5. The analysis used a pipeline which identified eSNP – dSNP pairs which potentially both tagged the causal variant of an eQTL and GWAS variant. Three analyses were carried out using different criteria for detecting eSNP – dSNP pairs. The first analysis looked to identify exact matched between eSNPs and dSNPs, in total 12 were identified, 4 of which had a plausible immunological phenotype. The other two analyses used reciprocal conditional analysis to detect were both eSNP and dSNP cancelled each other out, in this analysis, in total 12 were identified, 8 of which had a plausible immunological phenotype.

The following eSNP – dSNP pairs all shared the same GWAS candidate gene and eQTL gene: *CARD9:* An eSNP for the gene *CARD9* was also found to pass conditional analysis with dSNPs for the diseases UC, IBD and Crohn's disease. *CCDC88B* was found to be the gene for an eSNP which is also a dSNP for Sarcoidosis. An eSNP for the gene *PRKCB* passed conditional analysis with dSNPs for the disease IBD. An eSNP for the gene *FCRL3* passes conditional analysis with a dSNP for the disease Graves' Disease. An eSNP for the gene *GPX4* passes conditional analysis with a dSNP for the disease Crohn's Disease. An eSNP for the gene *STAT4* passes conditional analysis with a dSNP for the disease Bechet's Disease. *STAT4.* An eSNP for the gene *EOMES* passes conditional analysis with a dSNP for the disease Multiple Sclerosis.

Two eSNP – dSNP pairs had contradictory GWAS candidate gene and eQTL gene. The eSNP for the gene *IL19* was found to be a dSNP for Bechet's Disease, however in this case the candidate gene (*IL10*) contradicted the eSNP gene (*IL19*). The gene *ZPBP2* has an eQTL which is also a dSNP for cervical cancer, candidate genes in this case contradicted one another.

It is noteworthy that, in general, the significance levels of eQTLs that have been identified in this section as potentially causal for GWAS have much weaker significance (3.37×10^{-8} to 5.01×10^{-43}) than those identified as the top signals in chapters 3, 4 and 5 (1.13×10^{-194} to 1.69×10^{-312}).

Although some of the top results may be false-positives due to SNPs within the corresponding probe region, the general weaker significance of the GWAS eQTLs indicates a complex genetic architecture for eQTLs involved in disease phenotypes as compared to a Mendelian genetic architecture identified in the top results. The reason why disease associated eSNPs are generally weaker could be due to the fact that the effect sizes of individual variants for complex diseases are generally modest in size, and the eQTL signals simply reflects this.

CHAPTER 7 MANTRA ANALYSIS

7.1 Overview

In this chapter, I present the results of performing a Bayesian trans-ethnic meta-analysis (implemented in the software MANTRA) using population-specific association summary statistics for cis eQTLs from the Phase III Hapmap after imputation up to the 1000 Genomes Project reference panel (described in chapter 5). This analysis has been run to facilitate the construction of "credible sets" of variants that are most likely to be causal for eQTLs to assess the benefits of trans-ethnic meta-analysis for improving the resolution of fine mapping within and between ancestry groups.

MANTRA uses a Bayesian partition model to cluster populations into ancestry groups (using differences in allele frequencies between populations). The populations within a cluster are modelled as having the same underlying effect size. In contrast, between populations from different clusters, the underlying effect size is allowed to vary. MANTRA effectively performs a Bayesian hybrid of fixed and random effect meta-analysis, allowing for the similarity between populations in terms of their ancestry. MANTRA provides a Bayes' factor (BF) in favour of association for each SNP. Using MANTRA enables calculation of credible set of variants that are most likely to be causal for association signals, allowing for heterogeneity in allelic effects between populations, due to differences in LD between ancestry groups, which can be leveraged for fine mapping eQTLs. See Methods section 2.9 for full description of MANTRA.

In order to assess the improvement in fine mapping offered by trans-ethnic meta-analysis, MANTRA has been run within and between ancestry groups. Following this, 99% credible sets have been constructed using association summary statistics from the MANTRA analysis within and between ancestry groups. The size of the 99% credible sets are then compared: smaller credible sets correspond to fine-mapping at higher resolution. On the basis of these results, I have selected four examples that show the contrasting results of trans-ethnic meta-analysis. Two of

the cis eQTLs show an improvement in fine mapping resolution after trans-ethnic meta-analysis, whilst two show less precise fine-mapping after trans-ethnic meta-analysis, compared to within one ancestry group.

7.2 MANTRA Analysis

MANTRA requires population-specific association analysis summary statistics (beta and standard error) be specified for each SNP, together with allele frequencies.

7.2.1 Ancestry Groups

MANTRA requires that a dendogram of the populations in the analysis be constructed using differences in allele frequencies between each pair. Figure 7.1 shows a dendogram of the three ancestry groups for the eight Phase III HapMap populations. The dendogram was generated using hierarchical clustering, based on distance matrix calculated as average pairwise difference in allele frequencies between populations. The ancestry groups determined are African (LWK, MKK, YRI), East Asian (CHB, JPT) and Eurasian-Hispanic (CEU, GIH, MKK). The dendogram is used as a prior model for clustering.

MANTRA was run separately for each of the ancestry groups, and then for all populations together in a trans-ethnic meta-analysis.

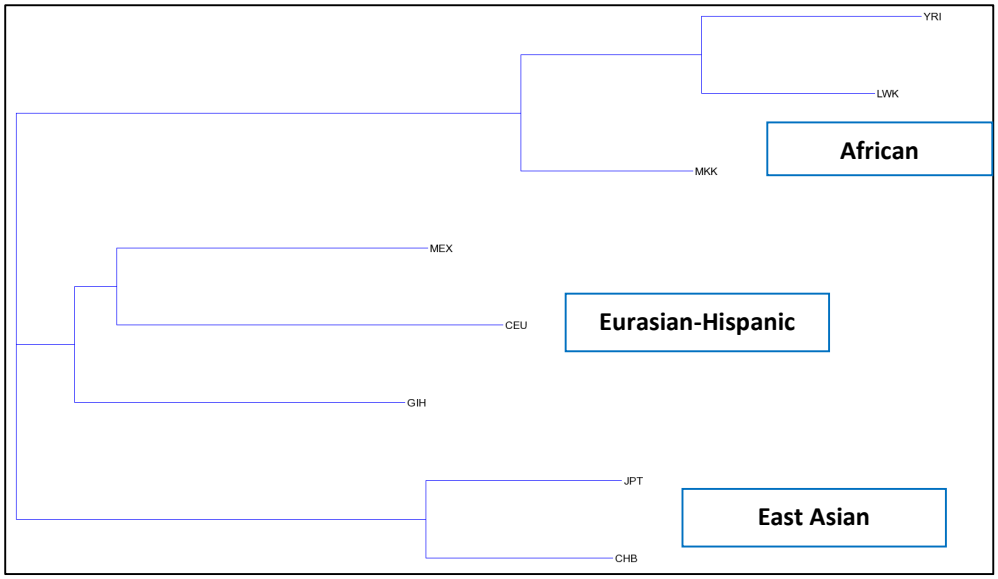


Figure 7.1: Dendogram of ancestry groups in Phase III HapMap populations.

7.2.2 Credible Set Analysis

In order to determine the extent to which the ancestry groups contribute to resolution of fine-mapping, 99% credible sets of variants have been determined for each of the *cis* eQTLs detected in the fixed effect meta-analysis at GWS. The number of variants at each *cis* eQTL can be compared between ancestry groups. Credible sets with fewer variants correspond to fine-mapping at greater resolution. The median credible set size can be used to determine how much each ancestry group contributes to fine mapping in the trans-ethnic meta-analysis. For more information on how to credible sets are calculated, see methods section 2.11. Also reported are the number of credible sets that include a single variant, and those that contain less than 10 variants. Credible sets of this size are tractable for further investigation.

The results of the credible set analysis are presented in table 7.1 for the 1811 probes detected in the fixed effect meta-analysis at GWS with SNPs reported in all eight populations. (See section 5.5.2 for results of fixed effect meta-analysis).The results indicate that trans-ethnic meta-analysis substantially reduces the size of credible sets, reflecting the improved resolution in fine-mapping offered by trans-ethnic meta-analysis by taking advantage of the differences in LD structure between diverse populations. Of the ancestry groups, African credible sets are the smallest –

reflecting the shortest range LD in these populations. The East Asian credible sets are the biggest, probably because there is less diversity between those populations (JPT and CHB) than the Eurasian-Hispanic populations (CEU, MEX, and GIH). Overall 1281 / 1811 probes achieved an improvement in resolution when using all three ancestry groups together.

Ancestry Groups	**SNP Count**			
	Median	**Min / Max**	**1 Variant**	**< 10 variants**
All	7	1 / 313	306	1091
African	90.5	1 / 4014	143	530
East Asian	149	1 / 4058	58	308
Eurasian-Hispanic	129.5	1 / 4007	47	317

Table 7.1: Results of 99% credible set analysis for SNPs with MAF ≥ 5%, using all 1811 probes identified in fixed effect meta-analysis at GWS.

Of the 1811 probes evaluated, 212 had a larger credible set after trans-ethnic meta-analysis than in one or more of the ancestry groups alone. For these cis eQTLs, using all three ancestry groups did not achieve an improvement in fine-mapping. There are several reasons why this might occur: (i) the eQTL might be specific to one or more ancestry group, and including those populations where there is no effect will likely introduce noise and increase the credible set size; (ii) there may be multiple eQTLs acting in different populations, and by performing trans-ethnic meta-analysis, the credible set will include variants for more than one eQTL signal, thus increasing in size.

7.3 Fine mapping examples

This section presents examples of fine mapping highlighting the contribution of the ancestry groups LD differences. Examples are selected based on the credible sets calculated using the output of MANTRA. Examples of eQTLs where trans-ethnic meta-analysis has improved fine mapping resolution are presented and discussed in sections 7.4.1 (*C12orf54*) and 7.4.2 (*GDE1)*. Examples of eQTLs where trans-ethnic meta-analysis has worsened fine mapping resolution are presented and discussed in sections 7.4.3 (*PAX8*) and 7.4.4 (*PEX6*)

7.3.1 ILMN_26449_2510523 (ENSG00000177627, C12orf54)

This probe is an example where fine mapping resolution improves when using all three ancestry groups compared with each ancestry group separately. The peak eSNP detected using Bayesian trans-ethnic meta-analysis was rs2731096 with $\log_{10}$ BF = 84.97. The peak eSNP detected using fixed effect meta-analysis was also rs2731096 (p-value = 1.15×10^{-86}), and there was minimal evidence of heterogeneity in allelic effect sizes between populations (Cochran's Q p-value = 3.50×10^{-3}).

Table 7.2 shows the peak SNPs and log10 BF from the MANTRA analysis for each ancestry groups.

Ancestry Group	**Peak SNP**	**$\log_{10}$ BF**	**rs2731096 $\log_{10}$ BF**
All	rs2731096	84.97	84.97
African	rs2731108	44.41	40.98
East Asian	rs2731063	19.11	18.96
Eurasian-Hispanic	rs2731063	26.16	25.55

Table 7.2: Peak SNPs from MANTRA analysis for each ancestry group.

Gene name and location

The probe ILMN_26449_2510523 gene has the Ensembl Identifier ENSG00000177627 and the HGNC symbol *C12orf54*. The full name of the gene is Chromosome 12 Open Reading Frame 54, and the gene's position is chr12:48876286. The peak eSNP is rs2731096 which is located at chr12:48883020, and is intronic within *C12orf54*. The function of this gene is unknown.

Credible sets

Table 7.3 presents SNP counts for the 99% credible sets for gene *C12orf54*. When MANTRA is run with all ancestry groups the 99% credible set includes a single variant (the eSNP rs2731096). However, when MANTRA is run for each ancestry group separately, the 99% credible sets include more than one variant, being smallest in the African ancestry group, whilst the East Asian and Eurasian-Hispanic ancestry groups have the same resolution.

Ancestry Group	**SNP Count**	**Rank** rs2731096
All	1	1
African	8	9
East Asian	22	17
Eurasian-Hispanic	22	16

Table 7.3: 99% Credible sets for signals in ancestry groups

Figure 7.2 shows the forest plot for probe ILMN_26449_2510523 with SNP rs2731096 in population-specific association analyses and fixed effect meta-analysis results. As indicated by the low Cochran's Q statistic, effect sizes are quite similar across the populations. The effect size for the MKK population is weakest, but not significantly different from others. The lack of heterogeneity in allelic effects between populations is an indication that the causal variant is shared across ancestry groups and is therefore amenable to fine-mapping through trans-ethnic meta-analysis.

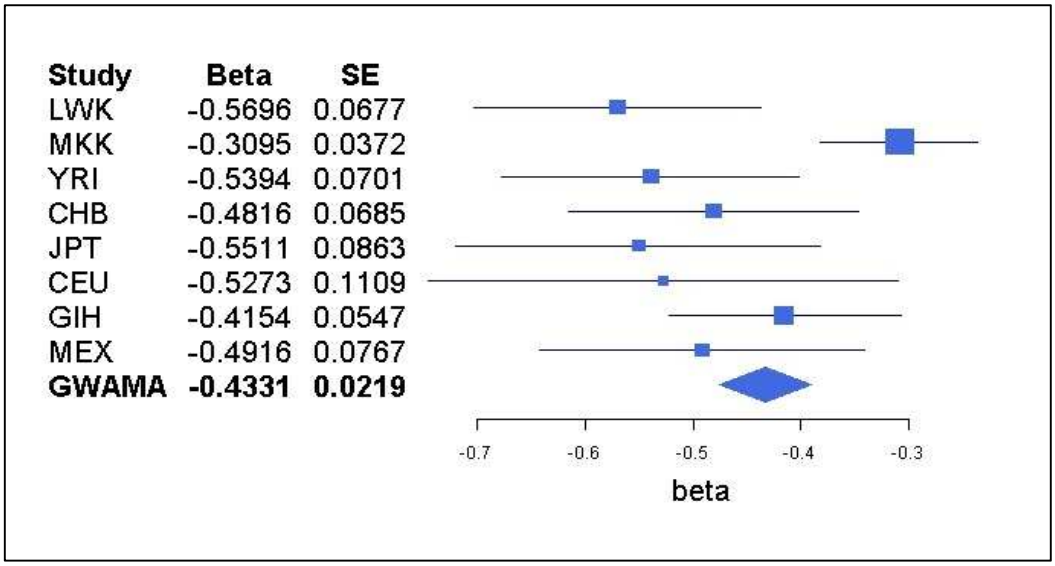


Figure 7.2: Forest plot for association analysis and fixed effect meta-analysis results for the probe ILMN_26449_2510523 with SNP rs2731096

Table 7.4 presents allele frequencies, p-values and INFO scores for probe ILMN_26449_2510523 with SNP rs2731096. Allele frequencies range from 0.11 to 0.43. Six populations have association signals achieving GWS, and in CEU and MEX approach GWS.

Population	Minor Allele (G/T)	Allele Frequency (Allele T)	Beta (Allele T)	SE	p-value	INFO
LWK	T	0.41	-0.5696	0.0677	1.23×10^{-12}	0.969
MKK	T	0.43	-0.3095	0.0372	1.02×10^{-13}	0.978
YRI	T	0.33	-0.5394	0.0701	8.02×10^{-12}	0.959
CHB	T	0.11	-0.4816	0.0685	7.58×10^{-10}	0.999
JPT	T	0.11	-0.5511	0.0863	1.15×10^{-8}	0.999
CEU	T	0.18	-0.5273	0.1109	6.41×10^{-6}	1.000
GIH	T	0.22	-0.4154	0.0547	8.82×10^{-11}	1.000
MEX	T	0.13	-0.4916	0.0767	1.59×10^{-7}	1.000
Fixed effect Meta-analysis	--	--	-0.4331	0.0219	1.15×10^{-86}	--
MANTRA	--	--	--	--	84.97 (Log_{10} BF)	--

Table 7.4: Table presents association analysis and meta-analysis results for probe ILMN_26449_2510523 with SNP rs2731096. Allele frequencies and INFO scores are also presented for SNP rs2731096.

Figure 7.3 presents the results of MANTRA trans-ethnic meta-analysis for individual ancestry groups and all ancestry groups together in the complete trans-ethnic meta-analysis. From the MANTRA trans-ethnic results it can be seen that the East Asian and Eurasian-Hispanic ancestry groups both have large LD blocks, whilst the African ancestry group has a more narrow block. The African ancestry group has stronger signals than the other groups. This suggests that improvement in resolution is mainly due to strong signals and a small block of LD in the African ancestry group. However, the reduction in the credible set size between the African ancestry analysis and the trans-ethnic meta-analysis suggests that the differential patterns of LD in the East Asian and Eurasian-Hispanic ancestry groups have also contributed to the improved fine-mapping resolution.

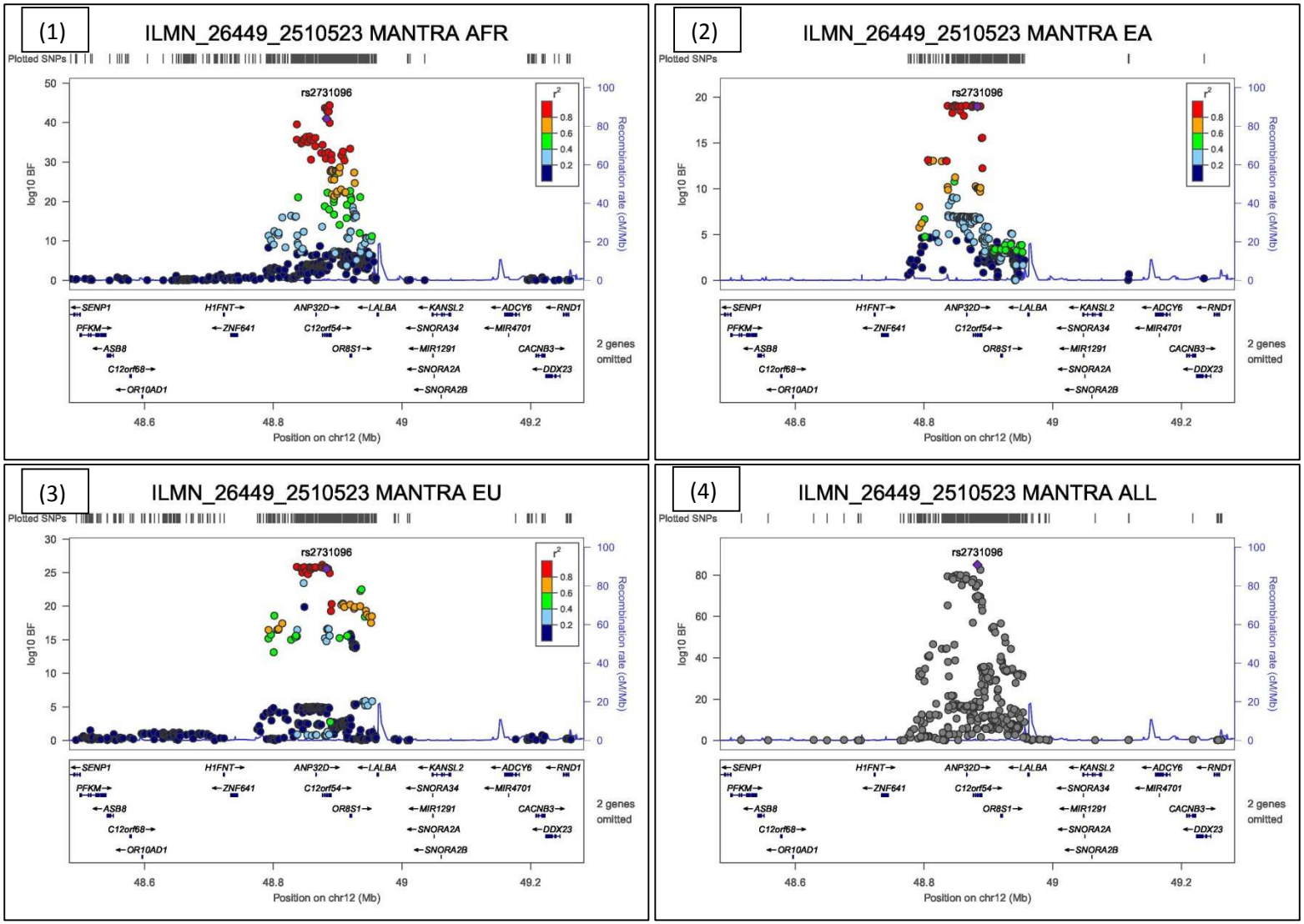


Figure 7.3: Signal plots for eSNPs for the probe ILMN_26449_2510523 detected using trans-ethnic Bayesian meta-analysis (MANTRA). Each circle represents a SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012). Plots are labelled with: (1) African ancestry group only. (2) EastAsian ancestry group only. (3) Eurasian-Hispanic ancestry group only. (4) All ancestry groups.

7.3.2 ILMN_3857_4810452 (ENSG00000006007, GDE1)

This section presents an example where an improvement in fine mapping occurs when all three ancestry groups are used together.

A *cis* eQTL was detected for the probe ILMN_3857_4810452 at GWS in the fixed-effects meta-analysis. The peak eSNP detected using Bayesian trans-ethnic meta-analysis was rs11865920 with $\log_{10}$ BF = 61.67. The same eSNP was detected with the fixed-effects meta-analysis (p-value = 3.50×10^{-64}), with no significant heterogeneity in allelic effect sizes observed between populations (p-value = 0.39).

Table 7.5 presents the peak SNPs detected in the MANTRA analysis.

Ancestry Group	**Peak SNP**	**$\log_{10}$ BF**	**rs11865920 $\log_{10}$ BF**
All	rs11865920	61.67	61.67
African	rs9888817	18.38	17.06
East Asian	rs1898676	31.96	30.99
Eurasian-Hispanic	rs11860735	13.24	13.01

Table 7.5: Peak SNPs from MANTRA analysis for each ancestry group.

Gene name and location

The probe ILMN_3857_4810452 gene has the Ensembl ID ENSG00000006007 and the HGNC symbol *GDE1*. The full name of the gene is Glycerophosphodiester Phosphodiesterase, and the gene's start position is chr16:19533467. The peak eSNP, rs11865920, is located at chr16:19571317, andis intronic within *C16orf62*. The gene has glycerophosphoinositol phosphodiesterase activity.

Credible sets

Table 7.6 presents SNP counts for the 99% credible sets for gene *GDE1*. When MANTRA is run with all ancestry groups together, the 99% credible set includes a single variant (the eSNP rs11865920). However, when MANTRA is run in each ancestry group, separately, the credible sets are larger, with the greatest resolution observed in the East Asian ancestry group.,

Ancestry Group	SNP Count	RANK rs11865920
All	1	1
African	12	12
East Asian	4	2
Eurasian-Hispanic	24	15

Table 7.6: 99% Credible sets for signals in ancestry groups

Figure 7.4 presents a forest plot for population-specific association analyses and fixed effect meta-analysis for probe ILMN_3857_4810452 with SNP rs11865920. Visually there seems to be little evidence of heterogeneity, as suggested by the Cochran's Q statistic at this SNP.

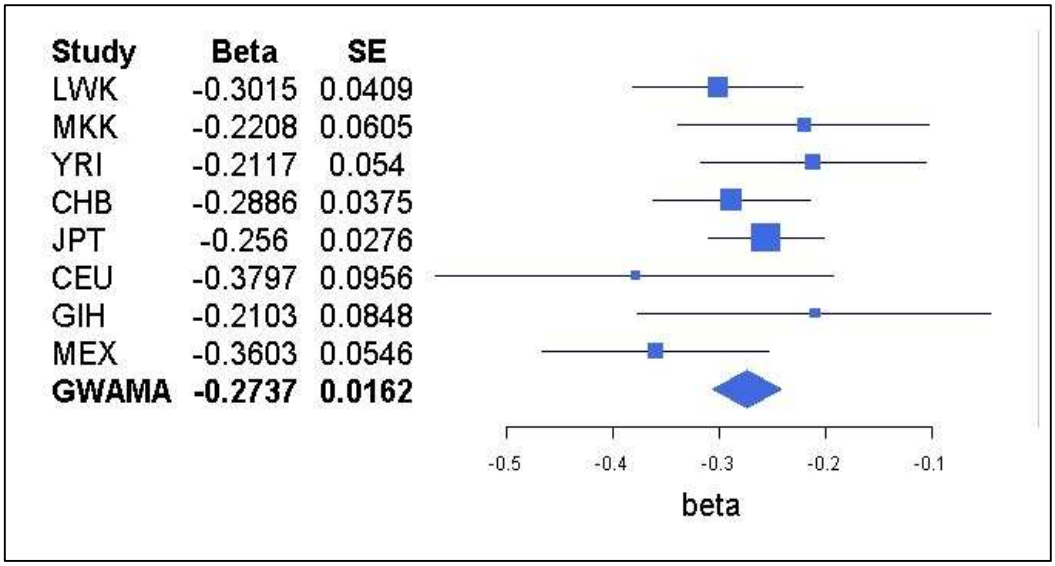


Figure 7.4: Forest plot for association analysis and fixed effect meta-analysis results for the probe ILMN_3857_4810452 with SNP rs11865920.

Table 7.7 shows allele frequencies, p-values and INFO scores for the SNP rs11865920 with probe ILMN_3857_4810452. Allele frequencies range from 0.06 to 0.38. The strongest signals of association are observed in the East Asian ancestry populations, where the allele frequency is greatest.

Population	Minor Allele (A / G)	Allele Frequency (Allele A)	Beta (Allele A)	SE	p-value	INFO
LWK	A	0.31	-0.3015	0.0409	$1.40x10^{-10}$	0.976
MKK	A	0.32	-0.2208	0.0605	$3.80x10^{-4}$	0.979
YRI	A	0.32	-0.2117	0.0540	$1.58x10^{-4}$	0.974
CHB	A	0.32	-0.2886	0.0375	$4.21x10^{-11}$	0.969
JPT	A	0.38	-0.2560	0.0276	$3.27x10^{-14}$	0.985
CEU	A	0.06	-0.3797	0.0956	$1.32x10^{-4}$	0.882
GIH	A	0.12	-0.2103	0.0848	0.02	0.831
MEX	A	0.17	-0.3603	0.0546	$8.56x10^{-8}$	0.811
Fixed effect Meta-analysis	--	--	-0.2737	0.0162	$3.50x10^{-64}$	--
MANTRA	--	--	--	--	61.67 (Log_{10} BF)	--

Table 7.7: Table presents association analysis and meta-analysis results for probe ILMN_3857_4810452 with SNP rs11865920. Allele frequencies and INFO scores are also presented for SNP rs11865920.

Figure 7.5 presents the results of MANTRA for individual ancestry groups and all ancestry groups together in trans-ethnic meta-analysis. From the MANTRA analysis signal plots it can be seen that both the African and Eurasian-Hispanic ancestry groups have LD blocks of similar size. However the East Asian ancestry group has a much narrower LD block than the other two groups, and shows the strongest signal of association. This suggests that for this cis eQTL, the strong signal and narrow LD block in the East Asian ancestry group is the major contributing factor to the fine-mapping resolution. However, differences in LD structure between East Asians and the other ancestry groups as also contributed to the improved fine-mapping resolution in the trans-ethnic meta-analysis.

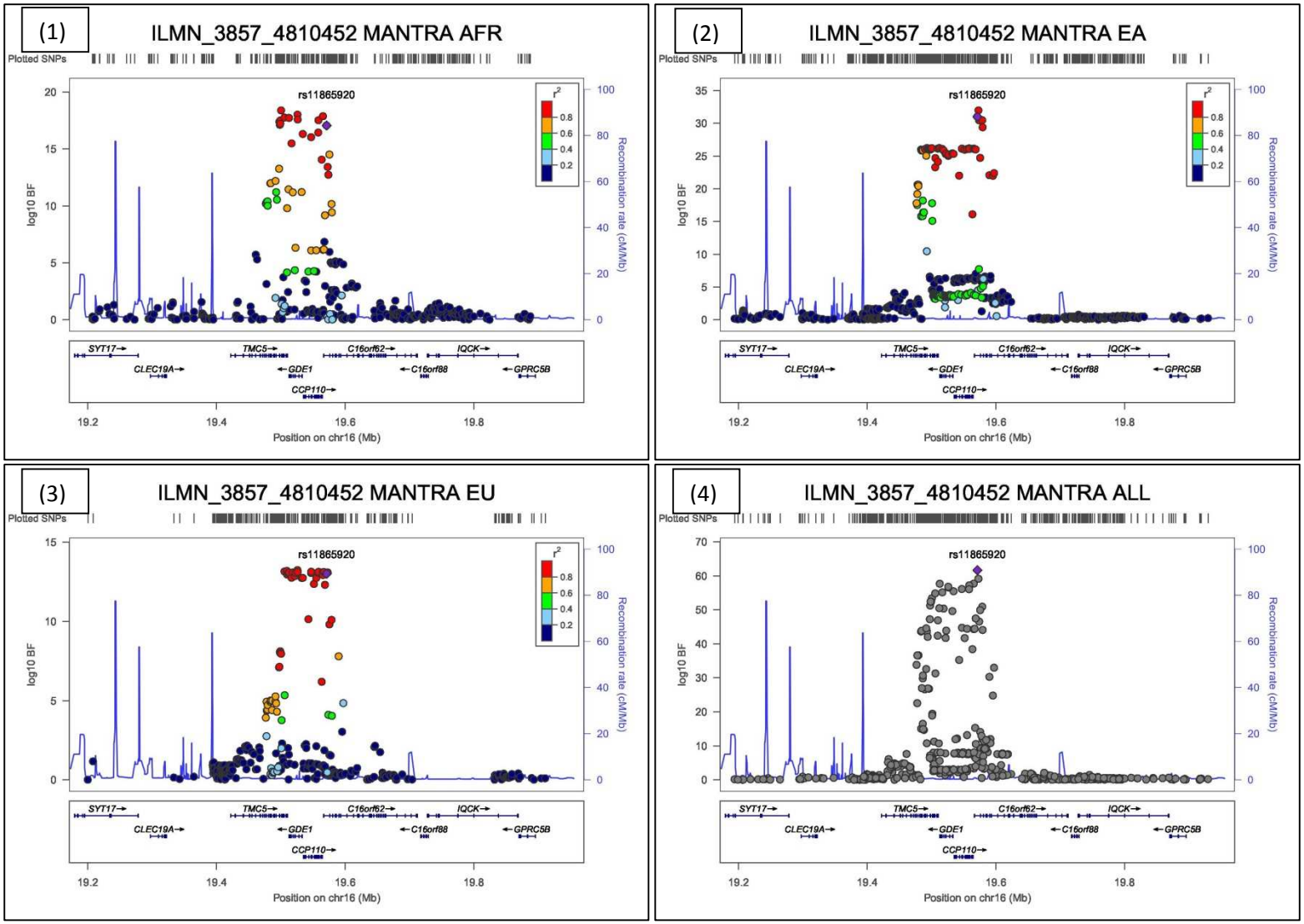


Figure 7.5: Signal plots for eSNPs for the probe ILMN_3857_4810452 detected using trans-ethnic Bayesian meta-analysis (MANTRA). Each circle represents a SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012). Plots are labelled with: (1) African ancestry group only. (2) EastAsian ancestry group only. (3) Eurasian-Hispanic ancestry group only. (4) All ancestry groups.

7.3.3 ILMN_137172_1070754 (ENSG00000125618, PAX8)

This probe is an example where fine mapping resolution gets worse when using all three ancestry groups together rather than each ancestry group separately.

The peak eSNP detected using Bayesian trans-ethnic meta-analysis was rs7421852 with $\log_{10}$ BF = 56.53. The same peak eSNP was detected using fixed effect meta-analysis (p-value = 4.93×10^{-57}), and some heterogeneity in allelic effects between populations was detected (Cochran's Q p-value = 1.62×10^{-3}).

Table 7.8 presents the peak SNPs detected in the MANTRA analysis.

Ancestry Group	**Peak SNP**	**$\log_{10}$ BF**	**rs7421852 $\log_{10}$ BF**
All	rs7421852	56.53	56.53
African	rs7421852	10.85	10.85
East Asian	rs11123170	23.78	23.76
Eurasian-Hispanic	rs7589901	23.84	23.12

Table 7.8: Peak SNPs from MANTRA analysis for each ancestry group.

Gene name and location

The probe ILMN_137172_1070754 gene has the Ensembl ID ENSG00000125618 and the HGNC symbol *PAX8*. The full name of the gene is Paired Box 8, and the gene's start position is chr2:114036527. The peak eSNP, rs7421852, is located at chr2:113990261, and maps to an intron of *PAX8*. *PAX8* is a transcription factor for expression of genes exclusively expressed in thyroid cells. Diseases associated with *PAX8* include endosalpingiosis, and hypothyroidism, congenital, due to thyroid dysgenesis or hypoplasia. *PAX8* candidate gene from GWAS for renal function related traits (p-value = 3.00×10^{-10}).

Credible sets

Table 7.9 presents SNP counts for the 99% credible sets for gene *PAX8*. The credible set after trans-ethnic meta-analysis is actually larger than for the African ancestry group, indicating poorer

fine-mapping resolution. However, the credible set after trans-ethnic meta-analysis is smaller than that in the East Asian or Eurasian Hispanic ancestry groups.

Ancestry Group	**SNP Count**	**RANK** rs7421852
All	10	1
African	7	1
East Asian	17	4
Eurasian-Hispanic	14	3

Table 7.9: 99% Credible sets for signals in ancestry groups

Figure 7.6 presents a forest plot of population-specific association analyses and fixed effect meta-analysis for the probe ILMN_13172_1070754 with the SNP rs7421852. All populations have significant effect size at p-value ≤ 0.05. There is some evidence of effect size clustering in the African populations, and also in the East Asian and Eurasian-Hispanic populations.

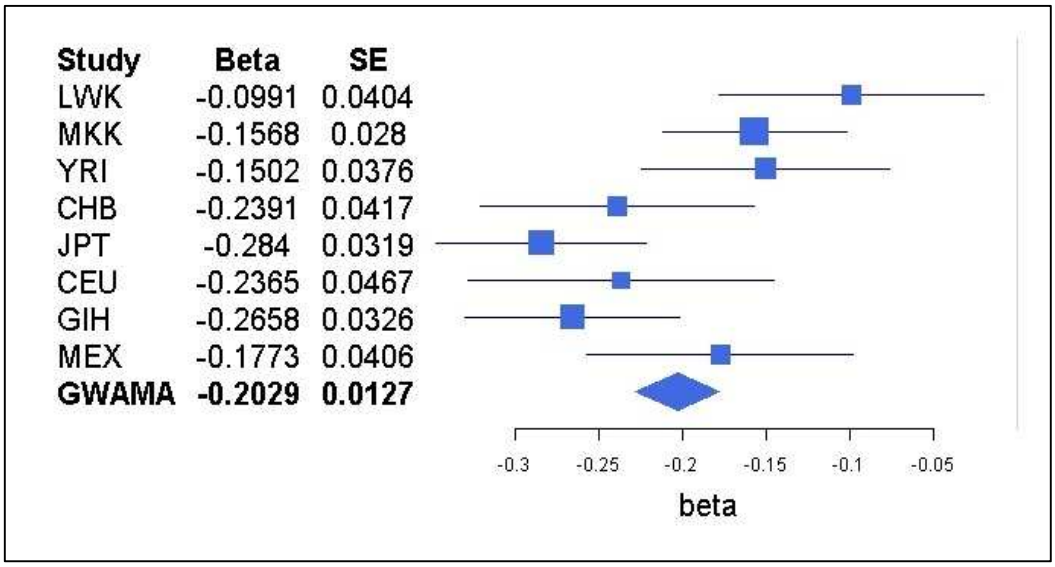


Figure 7.6: Forest plot for association analysis and fixed effect meta-analysis results for the probe ILMN_137172_1070754 with SNP rs7421852.

Table 7.10 presents allele frequencies, p-values and INFO score for the probe ILMN_137172_1070754 with the SNP rs7421852. Allele frequencies range from 0.28 to 0.54. Association signals in GIH and JPT achieve GWS, although all other populations also attain nominal significance at $p < 0.05$.

Population	Minor Allele (A/G)	Allele Frequency (Allele A)	Beta (Allele A)	SE	p-value	INFO
LWK	A	0.32	-0.0991	0.0404	0.02	0.881
MKK	A	0.29	-0.1568	0.0280	$1.20x10^{-7}$	0.935
YRI	A	0.28	-0.1502	0.0376	$1.19x10^{-4}$	0.954
CHB	A	0.36	-0.2391	0.0417	$1.86x10^{-7}$	0.991
JPT	A	0.31	-0.2840	0.0319	$1.68x10^{-13}$	0.996
CEU	A	0.41	-0.2365	0.0467	$1.80x10^{-6}$	0.998
GIH	G	0.54	-0.2658	0.0326	$7.77x10^{-12}$	0.990
MEX	A	0.34	-0.1773	0.0406	$9.54x10^{-5}$	0.999
Fixed effect Meta-analysis	--	--	-0.2029	0.0127	$4.93x10^{-57}$	--
MANTRA	--	--	--	--	56.53 (Log_{10} BF)	--

Table 7.10: Table presents association analysis and meta-analysis results for probe ILMN_137172_1070754 with SNP rs7421852. Allele frequencies and INFO scores are also presented for SNP rs7421852.

Figure 7.7 presents signal plots for the results of trans-ethnic meta-analysis using MANTRA. The reason for the increase in the credible set size after trans-ethnic meta-analysis compared to the African ancestry group appears to be due to the stronger signals of association in the other ethnic groups, coupled with the larger LD block sizes in East Asian and Eurasian-Hispanic populations. As a result, the narrow signal in the African ancestry group is being "swamped" by the other two ethnicities. Adding in the African ancestry group in the trans-ethnic meta-analysis has improved resolution over the Eurasian-Hispanic and East Asian ancestry groups, but it would be optimal, for this cis eQTL, to perform fine-mapping only in the African ancestry populations

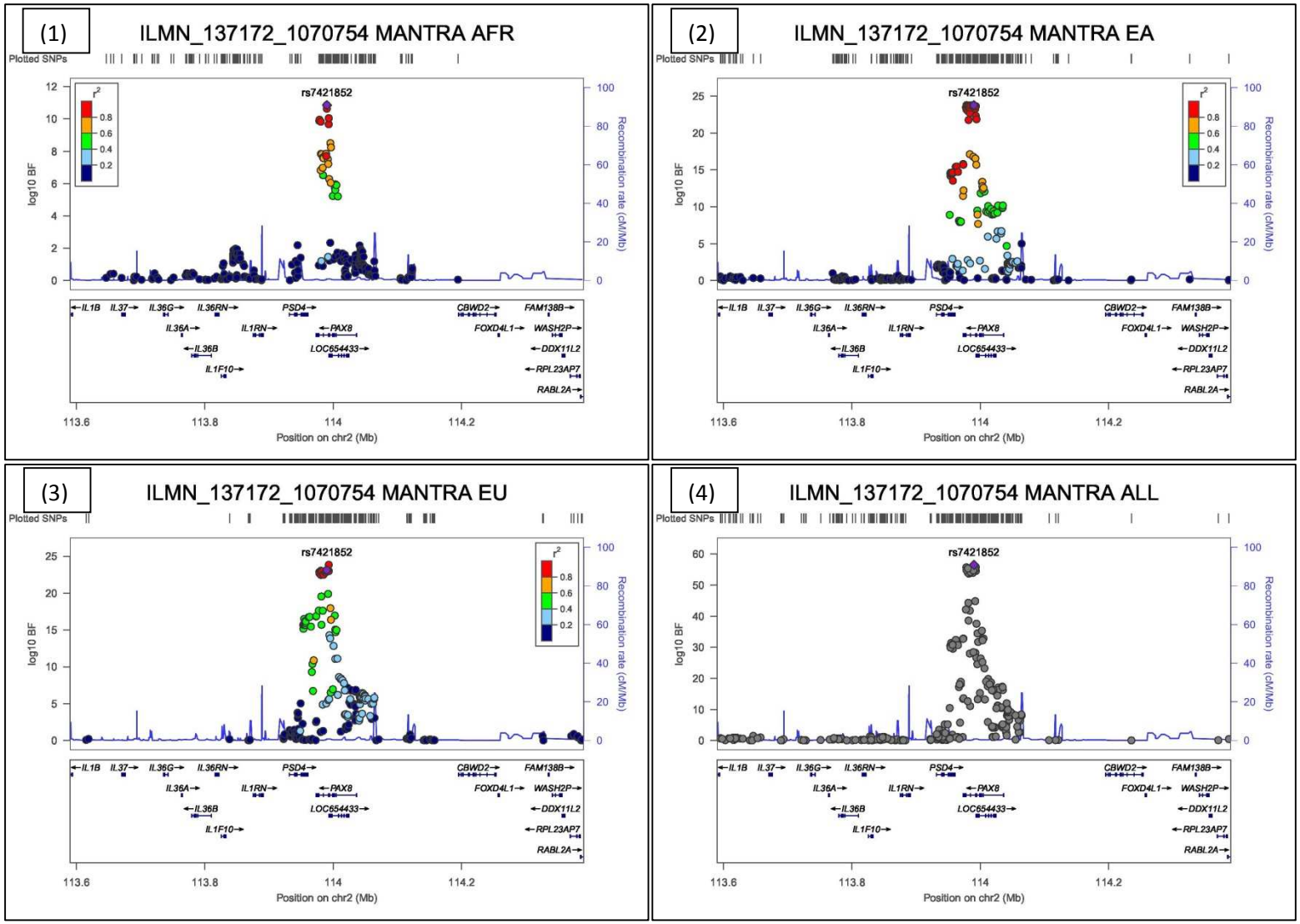


Figure 7.7: Signal plots for eSNPs for the probe ILMN_137172_1070754 detected using trans-ethnic Bayesian meta-analysis (MANTRA). Each circle represents a SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012). Plots are labelled with: (1) African ancestry group only. (2) EastAsian ancestry group only. (3) Eurasian-Hispanic ancestry group only. (4) All ancestry groups.

7.3.4 ILMN_3043_870301 (ENSG00000124587, PEX6)

This probe is an example where the resolution of fine mapping reduces when using all three ancestry groups together rather than each ancestry group separately.

The peak eSNP detected using Bayesian trans-ethnic meta-analysis was rs6458312 with $\log_{10}$ BF = 105.93. The peak eSNP detected using fixed effect meta-analysis was rs9986447 (p-value = 5.41×10^{-101}), and there was strong evidence of heterogeneity in allelic effects detected between populations (Cochran's Q p-value = 8.49×10^{-8}).

Table 7.11 presents the peak SNPs detected in the MANTRA analysis.

Ancestry Group	**Peak SNP**	**$\log_{10}$ BF**	rs6458312 **$\log_{10}$ BF**
All	rs6458312	105.93	105.93
African	rs9986447	30.42	6.75
East Asian	rs6458312	49.01	49.01
Eurasian-Hispanic	rs6907751	64.83	50.22

Table 7.11: Peak SNPs from MANTRA analysis for each ancestry group.

Gene name and location

The probe ILMN_3043_870301 gene has the Ensembl ID ENSG00000124587 and the HGNC symbol *PEX6*. The full name of the gene is Peroxisomal Biogenesis Factor 6, and the gene's start position is chr6:42946958. The peak eSNP using MANTRA, rs6458312, is located at chr6:42904274, and maps an intron of the gene *CNPY3*. The peak eSNP using fixed effect meta-analysis, rs9986447, is located at chr6:42942779, and maps to an intron of the gene *PEX6*. *PEX6* is involved in peroxisome biosynthesis. Diseases associated with *PEX6* include Zellweger Syndrome, and Peroxisome Biogenesis Disorder 4B.

Credible sets

Table 7.12 presents SNP counts for the 99% credible sets for gene PEX6. As can be seen the resolution using all ancestry groups together is not as good as using the any of the ancestry groups alone.

Ancestry Group	SNP Count	RANK rs6458312
All	3	1
African	1	61
East Asian	1	1
Eurasian-Hispanic	2	53

Table 7.12: 99% Credible sets for signals in ancestry groups

Figure 7.8 shows a forest plot for population-specific association analysis and fixed effect meta-analysis for probe ILMN_3043_870301 with SNP rs6458312 The effect sizes of the non-African populations are similar, whilst those in African ancestry populations are less significant.

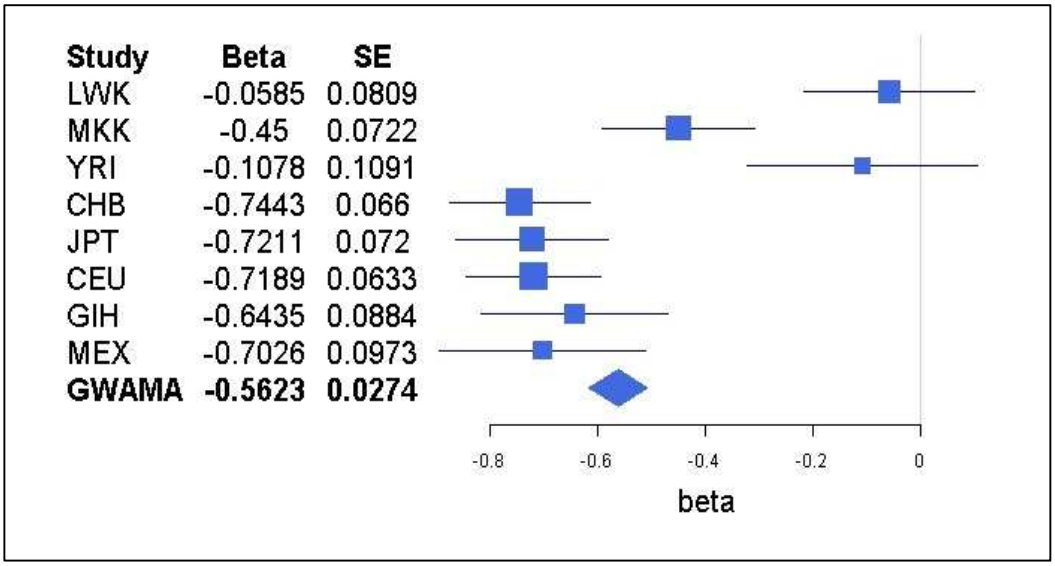


Figure 7.8: Forest plot for association analysis and fixed effect meta-analysis results for the probe ILMN_3043_870301 with SNP rs6458312.

Table 7.13 presents allele frequencies, p-values and INFO scores for the association analysis and meta-analysis for ILMN_3043_870301 with SNP rs6458312. Allele frequencies range from 0.11 to 0.72. Association signals for the eSNP in six of the population-specific association analysis achieve GWS, but are flat in LWK and YRI.

Population	Allele 1 (G/T)	Allele Frequency (Allele G)	Beta (Allele G)	SE	p-value	INFO
LWK	G	0.18	-0.5850	0.0809	0.47	0.977
MKK	G	0.20	-0.4500	0.0722	5.62×10^{-9}	0.936
YRI	G	0.11	-0.1078	0.1091	0.33	0.836
CHB	G	0.72	-0.7443	0.0660	6.82×10^{-18}	0.946
JPT	G	0.63	-0.7211	0.0720	1.18×10^{-15}	0.970
CEU	G	0.46	-0.7189	0.0633	6.58×10^{-20}	0.930
GIH	G	0.36	-0.6435	0.0884	3.39×10^{-10}	0.958
MEX	G	0.54	-0.7026	0.0973	1.25×10^{-8}	1.000
Fixed effect Meta-analysis	--	--	-0.5623	0.0274	1.09×10^{-93}	--
MANTRA	--	--	--	--	105.93 (Log_{10} BF)	--

Table 7.13: Table presents association analysis and meta-analysis results for probe ILMN_3043_870301 with SNP rs6458312. Allele frequencies and INFO scores are also presented for SNP rs6458312.

The following pages show signal plots for trans-ethnic Bayesian Meta-analysis results (MANTRA) (figure 7.9) for the probe ILMN_3043_870301. There appears to be (at least) two independent eQTL signals at this locus. Looking at the African ancestry group alone in the MANTRA analysis, the peak SNP from the MANTRA analysis using all three groups is not in LD with the peak in the African group. The MANTRA analysis identifies three variants in the credible set for this eQTL: These appear to correspond to different eSNPs in each ancestry group, all of which are in weak LD with each other, and thus potentially reflect independent eQTL signals in different populations.

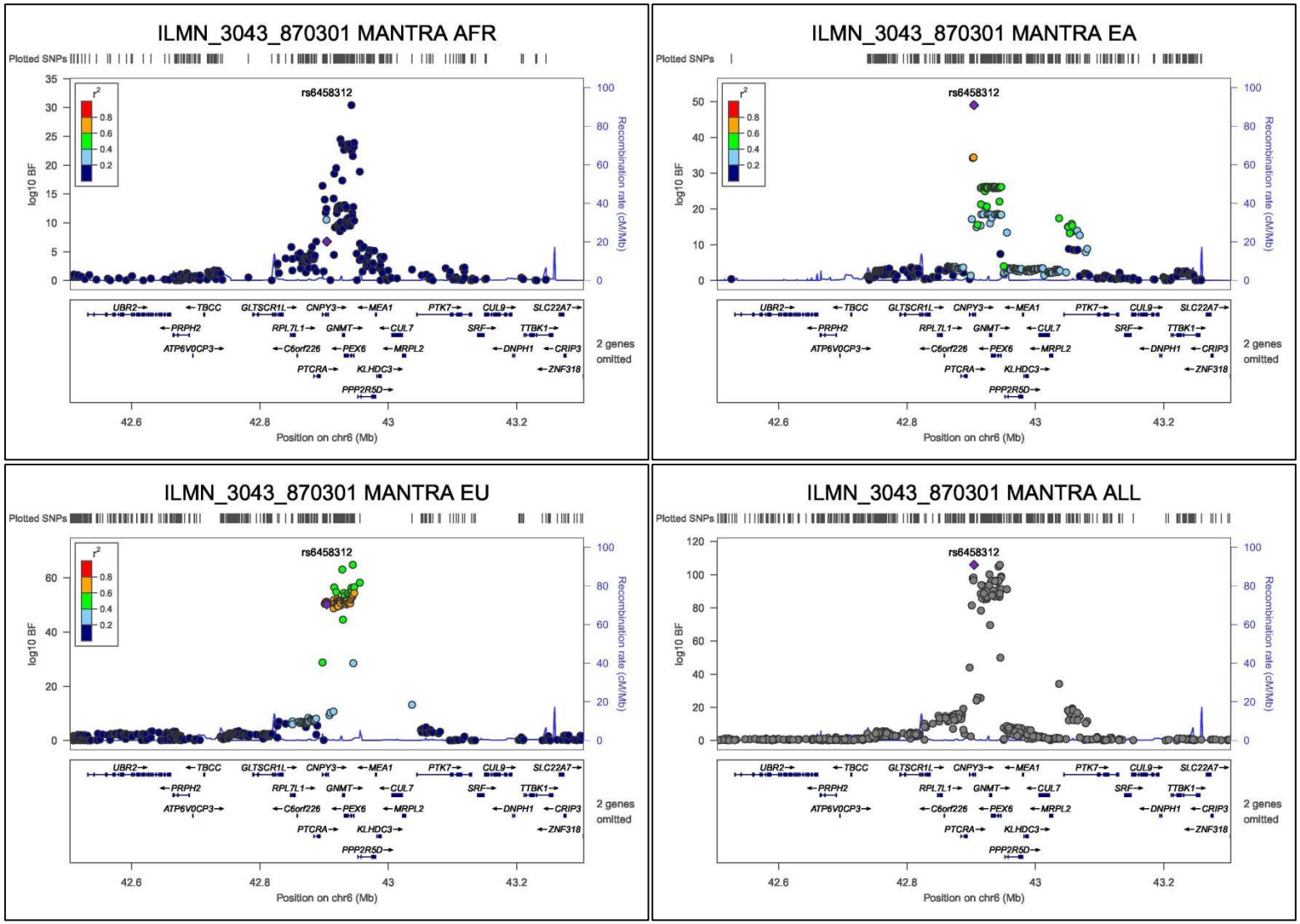


Figure 7.9: Signal plots for eSNPs for the probe ILMN_3043_8700301 detected using trans-ethnic Bayesian meta-analysis (MANTRA). Each circle represents a SNP. For each locus, the lead SNP is represented as a purple diamond. The colour of all other SNPs indicated LD with peak eSNP (estimated using relevant ancestry group (ASN, EUR, AFR) r^2 from 1000 Genomes data March 2012). are labelled with: (1) African ancestry group only. (2) EastAsian ancestry group only. (3) Eurasian-Hispanic ancestry group only. (4) All ancestry groups.

7.4 Summary

This chapter presents the results of a Bayesian trans-ethnic meta-analysis (MANTRA) using microarray expression and March 2012 1000 Genomes imputed genotype Phase III HapMap (Generated in Chapter 5). The aim of the analysis was to calculate credible sets for each of the probes used in the analysis which can be used to compare and characterize multi-ethnic fine mapping.

It has been shown that an improvement in fine-mapping resolution occurs across the majority of eQTLs through trans-ethnic meta-analysis compared with ancestry-specific analyses when there is no heterogeneity in effect sizes between populations or secondary signals. The African ancestry group has the smallest ancestry-specific credible set size across eQTLs, and thus offers greater fine-mapping resolution than either Eurasian-Hispanic or East Asian ancestry groups.

Two examples were presented where trans-ethnic meta-analysis demonstrated improved fine-mapping resolution compared to any ancestry group alone: *C12orf54* and *GDE1*. In both of these examples, the extent of heterogeneity in allelic effect sizes between populations was minimal, suggesting the the eQTL shares the same underlying causal variant across populations, and thus amenable to trans-ethnic fine-mapping.

However, improved fine-mapping resolution was not always obtained through trans-ethnic meta-analysis compared to ancestry-specific analyses: Two examples *PAX8* and *PEX6* resolution was worse when taking all ancestry groups together against each ancestry group individually. In the case of *PAX8*: the narrow signal in the African ancestry group is being “swamped” by the other two ethnicities, due to weak strength of African effect size. In the case of *PEX6*: There appears to be (at least) two independent eQTL signals at this locus.

CHAPTER 8 ADME EQTL ANALYSIS

8.1 Overview

This section presents an analysis of Phase III HapMap *cis* eQTLs within the "absorption, distribution, metabolism, and excretion" (ADME) gene list taken from the pharmaadme resource. Counts of *cis* eQTLs detected within this list are presented and an enrichment analysis is carried out using a Fisher exact test, finally heterogeneity of effect sizes are assessed in ADME eQTLs.

The analysis uses the Phase III HapMap *cis* eQTL dataset which was introduced in chapter 4. This dataset does not include the 1000 Genomes imputed genotypes because the analysis in this section was performed prior to the generation of the imputed data.

8.1.1 eQTLs in Drug Discovery and Development

The motivation for this analysis is for the usage of identified eQTL variants in drug discovery and development. eQTLs can be used within the field of drug discovery and development in the following ways:

1. **GWAS Integration:** Integration of eQTLs with GWAS dSNP data to discover dSNPs associated with expression regulation. An integration of imputed Phase III HapMap eQTLs with dSNPs from the NHGRI GWAS catalog is performed in chapter 6.
2. **Drug Target Validation:** Identification of eQTLs for drug targets and correlation of drug target eSNPs with the phenotype of interest. This involves first identifying eQTL variants for drug target genes and determining whether the variant is in correlation with phenotypes related to the drug target. This can be achieved in some cases through GWAS integration.
3. **Network construction:** ADME eQTLs can be used as a starting point of building drug metabolism networks and analyse the effect of differing expression of ADME genes on these networks.

4. **Personalized medicine:** Expression profile of drug targets or ADME genes can be used to determine best drug and dosage given a patient's genotype. Using population / ancestry group heterogeneity of eQTL effect sizes, personalized medicine can be generalized to specific ethnic groups, and give biological insight into differences between ethnic groups.

This chapter focuses on ADME genes and identifies *cis* eQTLs for these genes. ADME genes are important in drug discovery and development, as they may act as marker for differences in drug metabolism and therefore may indicate the best drug and dosage to administer, therefore forming the basis of personalized medicine. The next section presents the ADME gene list in detail.

8.1.2 ADME

ADME is an abbreviation for "absorption, distribution, metabolism, and excretion" and describes the disposition of a pharmaceutical compound within an organism. ADME gene lists were taken from the resource pharmaadme where there are two lists of ADME genes:

- Core List: The genes that are considered to be the most important in drug metabolism.
- Extended List: Remaining genes to be associated with drug metabolism.

ADME genes can be one of the following categories:

- Phase I Metabolism Enzymes: Responsible for the modification of functional groups.
- Phase II Metabolism Enzymes: Responsible for the conjugation with endogenous moieties.
- Transporter: Responsible for the uptake and excretion of drugs in and out of cells.
- Modifier: Can either alter the expression of other ADME genes or affect the biochemistry of ADME enzymes.

8.1.3 Analysis Pipeline

The analysis has the following steps:

Step one: Determine the criteria for selecting probes for the analysis. This uses the average intensity of the probes in each population as the criteria for inclusion. The threshold selected is ≥ 10, which is close to the lower bound of the upper quartile in each population.

Step two: Determine *cis* eQTL probe and gene counts that pass the average intensity threshold in the whole dataset. This is followed by filtering the counts with the ADME gene lists. Information for each *cis* eQTLs discovered is summarized.

Step three: Enrichment analysis of fixed effect meta-analysis results for genes in the ADME dataset. This uses GSK in-house enrichment tool Multiverse Fisher exact test.

Step four: An analysis of heterogeneity is performed on the ADME gene lists. Counts of heterogeneous probes are presented and summarized. Following this each of the heterogeneous probes are analysed in more detail, in order to gain insight on what is causing the heterogeneity. Finally an example where the *cis* eQTL peak SNP detected is also an ADME functional variant is presented (*DHRS1*).

In this chapter functions of genes are taken from the resource GeneCards and so the information is not referenced, as it is assumed to be public knowledge. Positions of SNPs and genes in this section use NCBI Build 37.

8.2 Microarray average intensity analysis

This section presents the criteria used to select probes for the analysis and the number of probes which are selected. Briefly probes with the highest intensities are selected by using the approximate lower bound of the upper quartile of the probe average intensity.

8.2.1 Inclusion threshold

To determine which probes to include in the analysis, the average intensity of each micro-array probe was calculated for each population. Average intensity is the average of each probe's expression intensities across individuals within a population.

Table 8.1 presents the mean, median, upper quartile and maximum values of the average intensities for each population. Values are very similar with range 8.28 to 8.41 for mean average intensity and 7.28 to 7.36 for the median average intensity. The lower bound of the upper quartile ranges from 9.41 to 9.65, the max value ranges from 15.77 to 16.04. When the lower bound of the upper quartile is rounded up to the nearest integer, each population has the value 10, this has been selected as the inclusion criteria of a probe in the analysis.

Population	Mean average intensity	Median average intensity	Lower Bound (upper quartile)	Max average intensity
LWK	8.32	7.28	9.45	15.92
MKK	8.28	7.28	9.41	16.04
YRI	8.34	7.30	9.51	15.77
CHB	8.41	7.36	9.66	15.85
JPT	8.39	7.33	9.63	15.84
CEU	8.36	7.32	9.54	15.78
GIH	8.38	7.33	9.60	15.92
MEX	8.41	7.33	9.65	15.86

Table 8.1: Table of average intensities for phase III HapMap microarray probes.

Figure 8.1 shows a box plot of the average intensities for each population. The inclusion criterion of ≤ 10 is shown with a red line in figure 8.1. The distribution of average intensities is very similar across populations.

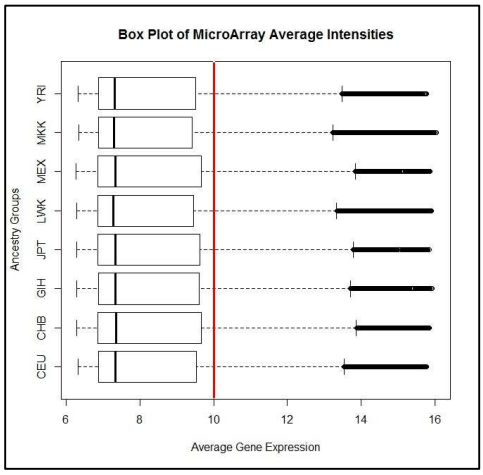


Figure 8.1: Box plots of Phase III HapMap Microarray average intensities. The lower bounds of upper quartile is close to 10 (red line), which has been selected as the inclusion criteria for probes in this analysis.

8.2.2 Probes: Passing inclusion threshold

Table 8.2 presents the number of probes and (corresponding) genes selected from each population for further analysis using the average intensity criteria of ≥ 10.

The following data is presented in this table:

- **Probes:** The number of probes out of the 21,800 in the original analysis, that pass the inclusion threshold
- **Genes:** The number of genes that these probes represent, some of the probes encodes the same gene.

Also included are the union and the intersection of the probes from the eight populations. These are included to indicate the number of probes that can be detected in the fixed effect meta-analysis. The intersection will show how many probes can be detected without any SNPs being not reported, the union will show how many probes can be detected with one or more SNPs not being reported.

Population	Probes	Genes
LWK	4405	4202
MKK	4160	3962
YRI	4474	4247
CHB	4738	4504
JPT	4695	4458
CEU	4508	4281
GIH	4610	4375
MEX	4742	4474
Union	5056	4776
Intersection	3929	3752

Table 8.2: Table of phase III HapMap probe and gene counts that pass the average intensity threshold ≥ 10.

8.3 Association analysis and fixed effect meta-analysis

This section presents counts of *cis* eQTLs detected with association analysis and fixed effect meta-analysis for probes which pass the average intensity inclusion criterion ≥ 10. These results are then filtered using the ADME gene list.

8.3.1 Results passing inclusion criteria.

Table 8.3 presents counts of *cis* eQTLs discovered at genome wide significance (GWS) (p-value ≤ $5x10^{-8}$), and passing the average intensity inclusion threshold. The table has the following columns:

- Probes: The total number of probes with a significant *cis* eQTL signal detected.
- Genes: The total number of genes corresponding to the probes detected.

The CHB and JPT populations of the East Asian ancestry group have the largest number of *cis* eQTLs detected. The MEX population has the lowest number of *cis* eQTLs detected, due to the smaller population size.

Population	Probes	Genes
LWK	63	62
MKK	55	53
YRI	51	49
CHB	76	73
JPT	86	83
CEU	57	54
GIH	49	48
MEX	25	24

Table 8.3: Number of *cis* eQTLs detected in phase III HapMap populations at GWS.

Table 8.4 presents the results of the fixed effect meta-analysis which passed the inclusion criteria. The table has the following additional column: Min Population Count which specifies the minimum number of SNPs which are reported in the populations. For example; 8 indicates that all 8 SNPs are reported, 1 indicates that 1 or more SNPs are reported.

Comparing the intersection of probes passing the inclusion threshold (3929) with the number of *cis* eQTLs detected with fixed effect meta-analysis with no missing variants indicates that approximately 14% of probes passing the inclusion threshold are also *cis* eQTLs.

Min Population Count	Probes	Genes
8	564	543
6	641	616
1	718	691

Table 8.4: Counts of *cis* eQTLs detected in fixed effect meta-analysis at GWS.

8.3.2 ADME

This section presents *cis* eQTLs detected within the ADME gene list, which includes 298 genes. See methods section 2.15.2 for more information on this dataset. The following information is presented:

- **Dataset:** Counts of probes and genes passing the average intensity inclusion threshold within the GSK dataset.
- **GWS:** Count of probes and genes at GWS, within the GSK dataset. (p-value ≤ $5x10^{-8}$)

Table 8.5 presents counts of *cis* eQTL signals detected within the ADME dataset.

Population	Dataset		GWS	
	Probes	Genes	Probes	Genes
LWK	40	38	3	3
MKK	35	34	4	3
YRI	41	38	4	3
CHB	35	35	2	2
JPT	35	35	2	2
CEU	40	38	5	3
GIH	41	41	4	4
MEX	43	39	1	1

Table 8.5: Summary *cis* eQTL signals detected in the ADME dataset.

A total of 21 probes were detected that had SNPs reported in all eight populations; these are summarized in table 8.7. The column name List specifies which ADME gene list the gene belongs in, the column name Category specified the ADME category that the gene belongs in. See section 8.1.2 for full description of ADME gene list and category.

Min SNP Count	Probes	Genes
8	21	17
6	21	17
1	24	20

Table 8.6: Fixed effect meta-analysis results detected in the ADME dataset.

Probe	Gene	HGNC Symbol	List	Category	SNP	Beta	SE	z-score	p-value	Q p-value
ILMN_16478_1710170	ENSG00000155465	*SLC7A7*	Extended	Transporter	rs12884337	-0.55	0.033	-16.79	2.93×10^{-63}	5.45×10^{-4}
ILMN_15545_5670059	ENSG00000157379	*DHRS1*	Extended	Phase I	rs10134537	-0.56	0.037	-14.87	5.19×10^{-50}	0.6405
ILMN_22479_2900379	ENSG00000171234	*UGT2B7*	Core	Phase II	rs1454247	-1.11	0.076	-14.68	9.29×10^{-49}	8.13×10^{-5}
ILMN_138375_7650093	ENSG00000171234	*UGT2B7*	Core	Phase II	rs1454247	-1.13	0.077	-14.60	2.85×10^{-48}	2.46×10^{-5}
ILMN_2021_2350243	ENSG00000197888	*UGT2B17*	Core	Phase II	rs1454247	-1.25	0.087	-14.42	4.12×10^{-47}	3.57×10^{-5}
ILMN_5225_7050768	ENSG00000213759	*UGT2B11*	Extended	Phase II	rs1454247	-0.94	0.066	-14.36	1.00×10^{-46}	7.03×10^{-5}
ILMN_2021_5670180	ENSG00000197888	*UGT2B17*	Core	Phase II	rs1454247	-1.06	0.078	-13.70	1.11×10^{-42}	0.0014
ILMN_15293_3190328	ENSG00000072210	*ALDH3A2*	Extended	Phase I	rs12602112	0.26	0.023	11.41	3.81×10^{-30}	0.0067
ILMN_15891_10280	ENSG00000134184	*GSTM1*	Core	Phase II	rs11807	1.26	0.133	9.52	1.70×10^{-21}	0.2741
ILMN_15891_6480091	ENSG00000134184	*GSTM1*	Core	Phase II	rs12745189	-0.55	0.066	-8.36	6.55×10^{-17}	1.32×10^{-4}
ILMN_9522_3830538	ENSG00000116157	*GPX7*	Extended	Phase I	rs835341	0.35	0.042	8.35	6.75×10^{-17}	0.1911
ILMN_137799_10333	ENSG00000164904	*ALDH7A1*	Extended	Phase I	rs6870785	0.44	0.055	8.03	9.61×10^{-16}	0.0181
ILMN_19760_360577	ENSG00000112096	*SOD2*	Extended	Modifier	rs732498	-0.16	0.021	-7.64	2.18×10^{-14}	0.1632
ILMN_22401_730255	ENSG00000063854	*HAGH*	Extended	Phase I	rs2492883	0.13	0.017	7.54	4.83×10^{-14}	0.0034
ILMN_8534_940474	ENSG00000143198	*MGST3*	Extended	Phase II	rs4147596	-0.16	0.024	-6.49	8.58×10^{-11}	0.8413
ILMN_14614_6760255	ENSG00000138061	*CYP1B1*	Extended	Phase I	rs336031	0.43	0.067	6.35	2.22×10^{-10}	0.1551
ILMN_13962_6580397	ENSG00000121691	*CAT*	Extended	Modifier	rs769214	-0.11	0.019	-5.76	8.52×10^{-9}	0.8904
ILMN_10152_1940398	ENSG00000157326	*DHRS4*	Extended	Phase I	rs8022613	-0.17	0.030	-5.75	8.78×10^{-9}	0.0012
ILMN_37870_3930112	ENSG00000157326	*DHRS4*	Extended	Phase I	rs8022613	-0.17	0.031	-5.52	3.45×10^{-8}	0.0038
ILMN_137100_5270672	ENSG00000167468	*GPX4*	Extended	Phase I	rs2024092	0.10	0.019	5.47	4.57×10^{-8}	0.0130
ILMN_18916_4850209	ENSG00000187630	*DHRS4L2*	Extended	Phase I	rs8022613	-0.17	0.031	-5.46	4.75×10^{-8}	2.47×10^{-4}

Table 8.7: Table of probes detected at GWS in the ADME gene list with no missing variants.

8.4 ADME Enrichment Analysis

This section presents the results of an enrichment analysis of the *cis* eQTL probes and corresponding genes detected in the ADME dataset. In order to do this a Fisher exact test was carried out using the GSK enrichment analysis tool, Multiverse. See Methods section 2.15.1 for more information regarding Multiverse Fisher exact test.

Enrichment analysis was performed on the results of the fixed effect meta-analysis. Briefly the number of ADME genes with *cis* eQTLs within an enrichment category are compared to the number of ADME genes without *cis* eQTLs using a Fisher Exact test.

The results are presented with the following information:

- **Enrichment Term:** The enrichment term.
- **Description:** Description of the enrichment term.
- **Origin:** The origin of the enrichment term (MeSH, GeneGO or Reactome). More information on enrichment terms origins in methods section 2.15.1.
- **p-value:** Fisher exact p-value.

Table 8.8 presents the results of the enrichment analysis using the fixed effect meta-analysis results for ADME gene list.

Enrichment Term	Description	Origin	p-value
Placental Insufficiency	Insufficient blood flow to placenta during pregnancy	MeSH	3.21×10^{-4}
Dermatitis, Occupational	Also called eczema, inflammation of the skin.	MeSH	4.06×10^{-4}
Disease Progression	The worsening of a disease over time. Often used for chronic and incurable diseases.	MeSH	9.92×10^{-4}
Hemoglobinopathies	Genetic defect that results in abnormal structure of globin chains in the hemoglobin molecule.	MeSH	1.22×10^{-3}
Vitamin B6 Deficiency	A nutritional condition produced by a deficiency of Vitamin B6 in the diet.	MeSH	1.22×10^{-3}
Semen		MeSH	1.23×10^{-3}
Aberrant Crypt Foci	Clusters of abnormal tube like glands in the lining of the colon and rectum.	MeSH	4.91×10^{-3}
Oocysts	A cyst containing a zygote formed by parasital protozoan such as the malaria parasite,	MeSH	4.91×10^{-3}
Receptor Aggregation	The chemically stimulated aggregation of cell surface receptors.	MeSH	4.91×10^{-3}
Urea Cycle Disorders, Inborn	Rare congenital metabolism disorder of the urea cycle.	MeSH	4.91×10^{-3}

Table 8.8: Enrichment terms from the ADME gene set, fixed effect meta-analysis.

8.5 ADME Heterogeneity

This section presents examples of ADME genes with peak *cis* eQTL SNPs (eSNPs) that have significant effect size heterogeneity. These are important as they potentially show differences in drug target metabolism between ethnic groups.

In total, *cis* eQTLs for 21 probes (17 genes) of the 296 ADME genes were detected at GWS with SNPs reported in all eight populations. (Table 8.9 shows the probes and genes detected). Out of these *cis* eQTL signals significant heterogeneity for detected at 7 probes (6 genes) using the Cochran's Q statistic at p-value ≤ $1x10^{-3}$. At Cochran's Q statistic p-value ≤ 0.05, 14 probes (11 genes) were detected.

Table 8.9 shows the ADME probes and corresponding genes for which heterogeneous *cis* eQTLs were detected. Both the probes ILMN_138375_7650093 and ILMN_22479_2900379 correspond to the gene *UGT2B7*. Four of the probes have the same peak SNP: rs1454247, which is located within an intron of the gene *TMPRSS11E*. See Appendix A.8

The following sections review the ADME probes in detail; the probe ILMN_22479_2900379 is excluded as it encodes the same gene as ILMN_138375_7650093 to avoid repetition.

Probe	Gene	HGNC	SNP	Beta	SE	p-value	Cochran's Q p-value
ILMN_138375_7650093	ENSG00000171234	*UGT2B7*	rs1454247	-1.13	0.077	$2.85x10^{-48}$	$2.46x10^{-5}$
ILMN_22479_2900379 (*)	ENSG00000171234	*UGT2B7*	rs1454247	-1.11	0.076	$9.29x10^{-49}$	$8.13x10^{-5}$
ILMN_2021_2350243	ENSG00000197888	*UGT2B17*	rs1454247	-1.25	0.087	$4.12x10^{-47}$	$3.57x10^{-5}$
ILMN_5225_7050768	ENSG00000213759	*UGT2B11*	rs1454247	-0.94	0.066	$1.00x10^{-46}$	$7.03x10^{-5}$
ILMN_16478_1710170	ENSG00000155465	*SLC7A7*	rs12884337	-0.55	0.033	$2.93x10^{-63}$	$5.45x10^{-4}$
ILMN_15891_6480091	ENSG00000134184	*GSTM1*	rs12745189	-0.55	0.066	$6.55x10^{-17}$	$1.32x10^{-4}$
ILMN_18916_4850209	ENSG00000187630	*DHRS4L2*	rs8022613	-0.17	0.031	$4.75x10^{-8}$	$2.47x10^{-4}$

Table 8.9: Table of heterogeneous genes from the ADME dataset at Cochran's Q p-value ≤ $1x10^{-3}$. ILMN_22479_2900379 is not included in analysis due to being same gene as ILMN_138375_7650093 to avoid repetition.

8.5.1 ILMN_138375_7650093 (ENSG00000171234, UGT2B7)

The probe ILMN_138375_7650093 corresponds to the gene *UGT2B7*, this gene is in the core ADME list and is a phase II metabolism enzyme. The peak signal SNP is rs1454247, a phase III HapMap SNP. The SNP is significant at GWS (z-score = -14.61, p-value = 2.85×10^{-48}) and has significant effect size heterogeneity (Cochran's Q = 33.17, p-value = 2.46×10^{-5}).

Gene name and location

The probe's corresponding gene has the Ensembl identifier ENSG00000171234 and the HGNC symbol *UGT2B7*. Its full name is UDP Glucuronosyltransferase 2 Family, Polypeptide B7, the gene's start-site is chr4:69917081. The SNP rs1454247 is located at chr4:69341579. The SNP is within an intron of the gene *TMPRSS11E* which has the full name Transmembrane Protease serine 11E.

UGT2B7 encodes a protein which is a member of the UDP-glycosyltransferase (UGT) family. UGTs catalyze the transfer of carbohydrate moieties from UDP donor molecules to an acceptor molecule. UGTs have a major role in the elimination of xenobiotics and endogenous compounds. *UGT2B7* is a phase II metabolism enzyme and has unique specificity for 3,4-catechol estrogens and estriol suggesting it may have a role in regulating the level of these estrogen metabolism. Diseases associated with *UGT2B7* include hepatocellular adenoma (a benign liver tumour). *UGT2B7* is the gene reported for a GWAS for obesity related traits (not at GWS: rs4356975, 2.00×10^{-7})

TMPRSS11E is a serine protease which possesses both gelatinolytic (breaks down gelatine) and caseinolytic activities (breaks down caseine)

Figure 8.2 shows a forest plot for association and fixed effect meta-analysis results for probe ILMN_138375_7650093 with SNP rs1454247. Visually, the African and East Asian ancestry group's effect sizes form distinct clusters whilst the Eurasian-Hispanic ancestry group's do not. All eight populations are significantly different from zero at $p \leq 0.05$.

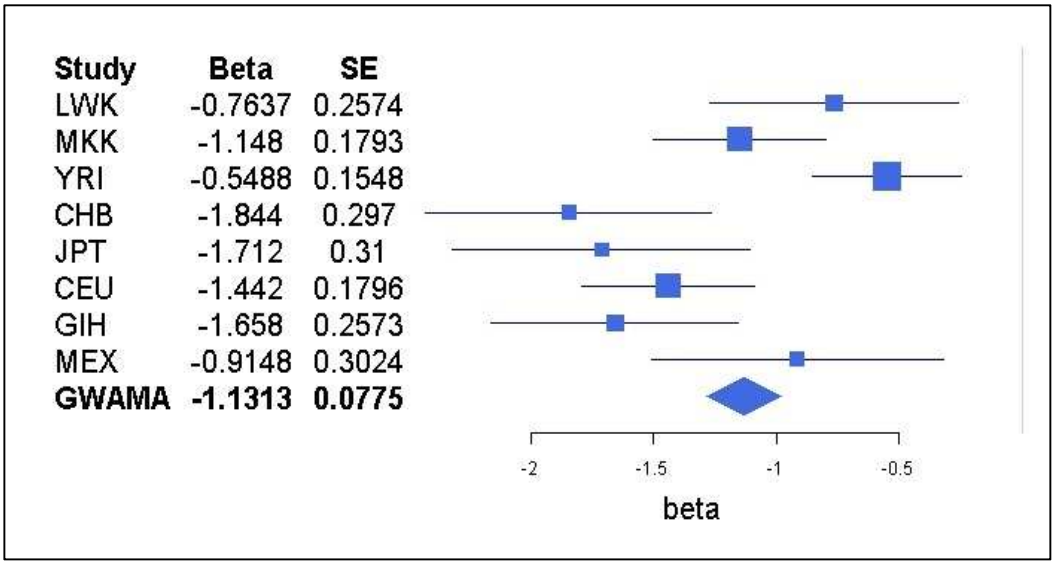


Figure 8.2: Forest plot of association and fixed effect meta-analyses for probe ILMN_138375_7650093 with SNP rs1454247.

Table 8.10 shows p-values, allele frequencies and peak signals for the probe ILMN_138375_7650093 at SNP rs1454247. Allele frequencies range from 0.21 to 0.56. The populations MKK, CHB and GIH are significant at GWS ($p \leq 5x10^{-8}$). One population (JPT) shares the same peak SNP (rs1454247) with that of the fixed-effect meta-analysis.

Population	Minor Allele (C / T)	Allele Frequency (Allele C)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	C	0.44	-0.7637	0.2574	4.00×10^{-3}	N
MKK	C	0.47	-1.148	0.1793	2.48×10^{-9}	N
YRI	C	0.36	-0.5488	0.1548	6.00×10^{-4}	N
CHB	C	0.24	-1.844	0.297	2.60×10^{-8}	N
JPT	C	0.21	-1.712	0.31	4.25×10^{-7}	Y
CEU	T	0.54	-1.442	0.1796	1.61×10^{-12}	N
GIH	C	0.43	-1.658	0.2573	1.15×10^{-8}	N
MEX	T	0.56	-0.9148	0.3024	4.4×10^{-3}	N
FIXED	--	--	-1.1313	0.0775	2.85×10^{-48}	Y

Table 8.10: Table of association analysis and fixed effect meta-analysis results for phase III HapMap SNP rs1454247 and probe ILMN_138375_7650093. Peak SNP indicates whether the association analysis peak SNP matches the fixed effect meta-analysis. The peak length specifies the number of peak SNP in each analysis.

Figures on the following pages show signal plots for association analysis (figures 8.3 and 8.4) and fixed effect meta-analysis results (figure 8.5). If fixed effect meta-analysis results with non-reported SNPs are included rs1454247 is not the peak SNP, instead rs35293564 (p-value=1.17×10^{-52}) is peak when SNPs for six or more populations are reported. The SNP rs35293564 is the peak signal in three populations: MKK, CHB and MEX and is missing in one population, JPT. The SNP rs35293564 (chr4:69356931) has a Cochran's Q p-value of 7.19×10^{-6} having greater effect size heterogeneity than rs1454247 (Q p-value = 2.46×10^{-5}). Like rs1454247, rs35293564 is also located within an intron of the gene *TMPRSS11E*. There is evidence of a signal in all eight populations at p-value ≤ 0.05; the weakest signal is in LWK (p-value=4.00×10^{-3}). LD blocks are narrow in all eight populations, peak signals comprise of 1 SNP in all eight populations. Both the CEU and GIH populations have a peak signal at the SNP rs28879970. This SNP is not part of the LD block, but forms an isolated peak.

The results indicate that the heterogeneity may be due to causal variant not being present in the Phase III HapMap genetic data, and the effect sizes of peak SNPs differ due to differing LD between the peak SNP and the causal variant. This assertion cannot be tested using the HapMap data, but that 1000G imputation would allow further investigation of this hypothesis.

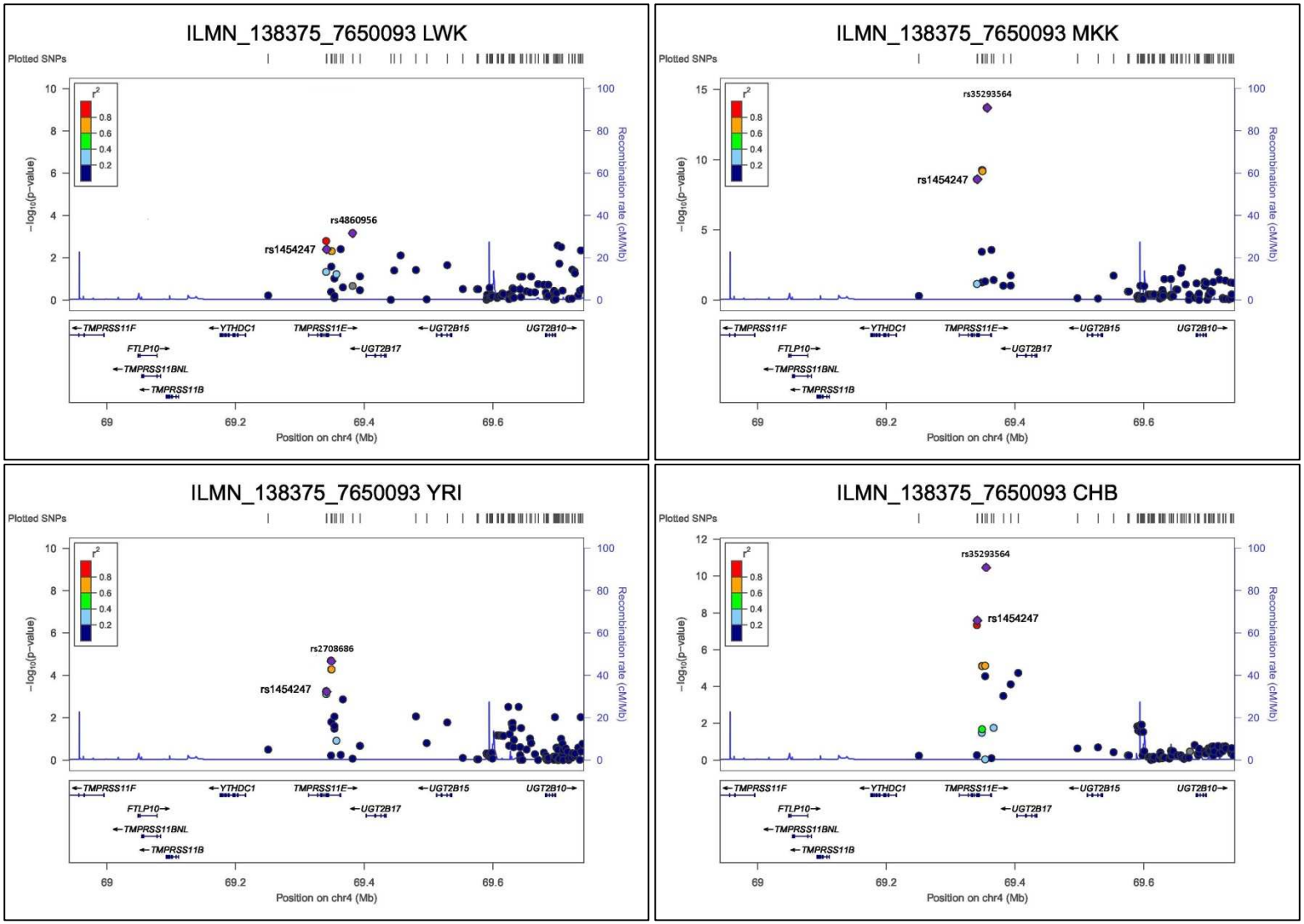


Figure 8.3: Signal plots for peak phase III HapMap SNPs for probe ILMN_138375_7650093

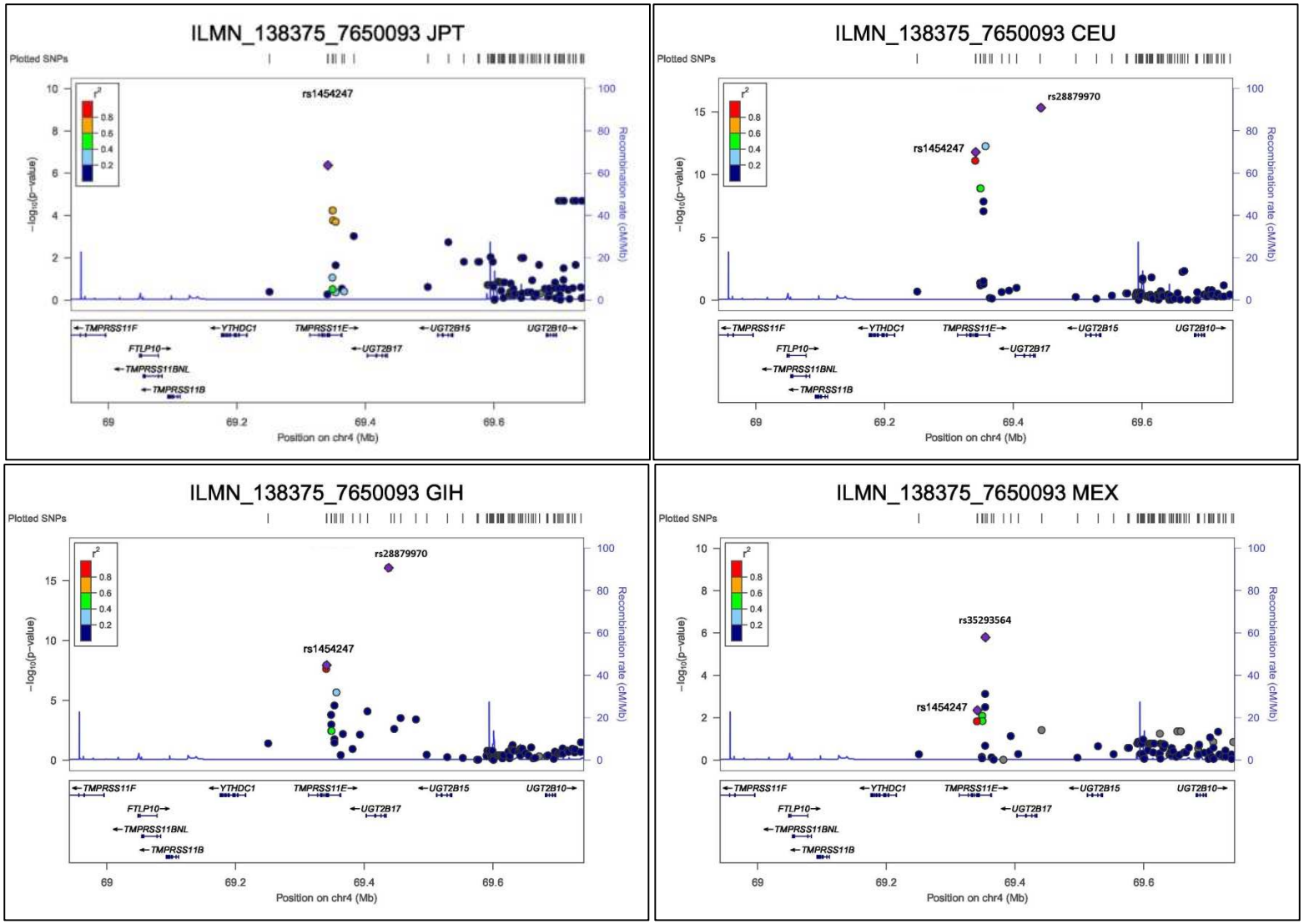


Figure 8.4: Signal plots for peak phase III HapMap SNPs for probe ILMN_138375_7650093

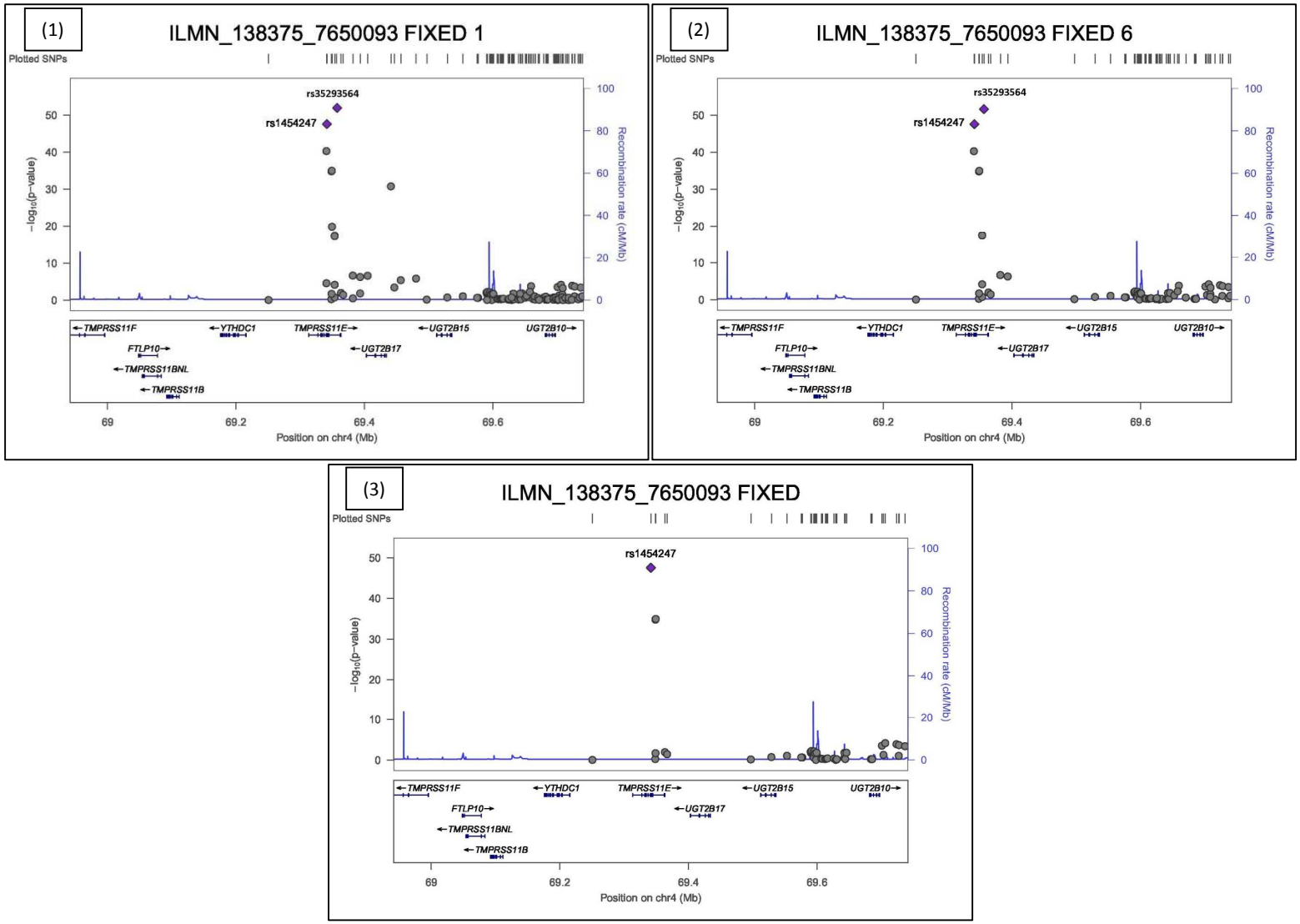


Figure 8.5: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_138375_7650093 (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.5.2 ILMN_2021_2350243 (ENSG00000197888, UGT2B17)

This gene is within the core ADME list and is a phase II metabolism enzyme. It has significant heterogeneity (Cochran's Q = 32.30, p-value = 3.57×10^{-5}). The peak SNP is rs1454247, a phase III HapMap SNP (Z-score = -14.42, p-value = 4.12×10^{-47}).

Gene name and location

The probe's gene has the Ensembl identifier ENSG00000197888 and the HGNC symbol *UGT2B17*. Its full name is UDP Glucuronosyltransferase 2 Family, Polypeptide B17, the gene's start-site is chr4:69434245. The SNP rs1454247 is located at chr4:69341579 (NCBI B37). The SNP is intronic within the gene *TMPRSS11E*.

UGT2B17 is a phase II metabolism enzyme which catalyses the transfer of glucoronic acid from uridine diphosphoglucuronic acid to a diverse array of substrates including steroid hormones and lipid-soluble drugs. This process known as glucoronidation is an intermediate step in the metabolism of steroids.

Figure 8.6 presents a forest plot for the probe ILMN_2021_2350243 with the SNP rs1454247. Visually there is some clustering of effect sizes in the African ancestry group, as in the previous example. In this case clustering of East Asian populations is less pronounced. All effect sizes are significantly different from zero at $p \leq 0.05$.

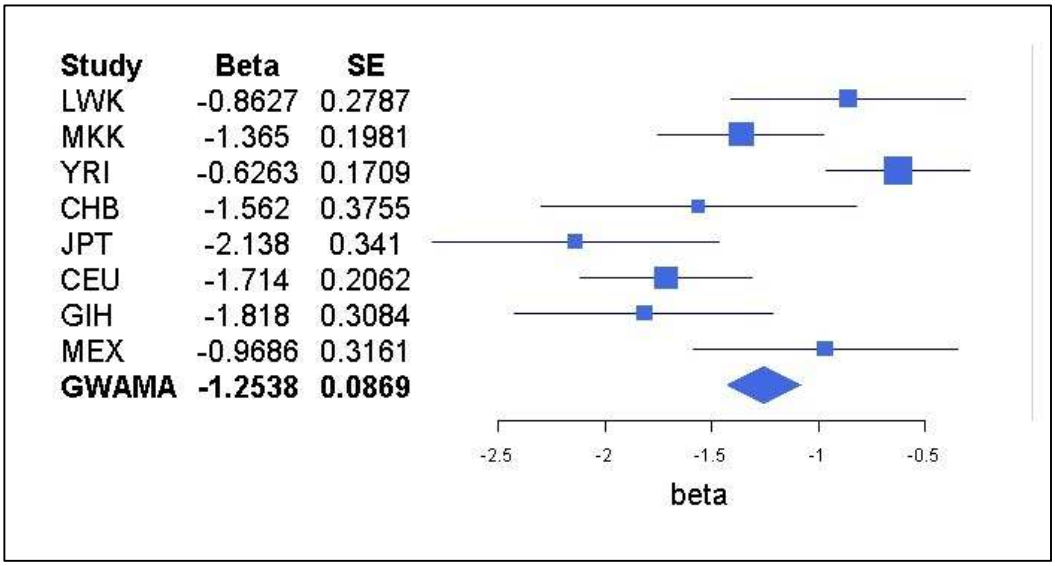


Figure 8.6: Forest plot of association and fixed effect meta-analyses for probe ILMN_2021_2350243 with SNP rs1454247.

Table 8.11 shows an overview of p-values, allele frequencies and peak SNPs for the eight populations at SNP rs1454247. The following observations can be made from this table: The populations MKK, JPT and CEU have significant effect sizes at GWS ($p \leq 5x10^{-8}$). Allele frequencies range from 0.21 to 0.56.

Population	Minor Allele (C / T)	Allele Frequency (Allele C)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	C	0.44	-0.8627	0.2787	$2.70x10^{-3}$	N
MKK	C	0.47	-1.365	0.1981	$2.07x10^{-10}$	N
YRI	C	0.36	-0.6263	0.1709	$4.00x10^{-4}$	N
CHB	C	0.24	-1.562	0.3755	$8.30x10^{-5}$	N
JPT	C	0.21	-2.138	0.341	$1.87x10^{-8}$	Y
CEU	T	0.54	-1.714	0.2062	$3.81x10^{-13}$	N
GIH	C	0.43	-1.818	0.3161	$1.12x10^{-7}$	N
MEX	T	0.56	-0.9686	0.3161	$4.00x10^{-3}$	N
FIXED	--	--	-1.2538	0.0869	$4.12x10^{-47}$	Y

Table 8.11: Table of association analysis and fixed effect meta-analysis results for phase III HapMap SNP rs1454247 and probe ILMN_2021_2350243. Peak SNP indicates whether the association analysis peak SNP matches the fixed effect meta-analysis. The peak length specifies the number of peak SNP in each analysis.

Figures on the following pages show signal plots for association analysis (figures 8.7 and 8.8) and fixed effect meta-analysis results (figure 8.9). The following observations can be made. First, the results are very similar to that of ILMN_138375_7650093, *UGT2B7* (section 8.5.1). When not reported variants are taken into account in the fixed effect meta-analysis, there is a more significant peak SNP at rs35293564 p-value = $3.53x10^{-51}$ Cochran's Q p-value = $2.77x10^{-6}$. This is the peak SNP in three populations: (MKK, CHB, MEX). The SNP is not reported in one population (JPT). Again, both CEU and GIH populations have peak signals at SNP rs28879970, which is not part of the main LD block, but forms an isolated peak SNP. The heterogeneity may be due to differing LD between causal variant and tag SNPs.

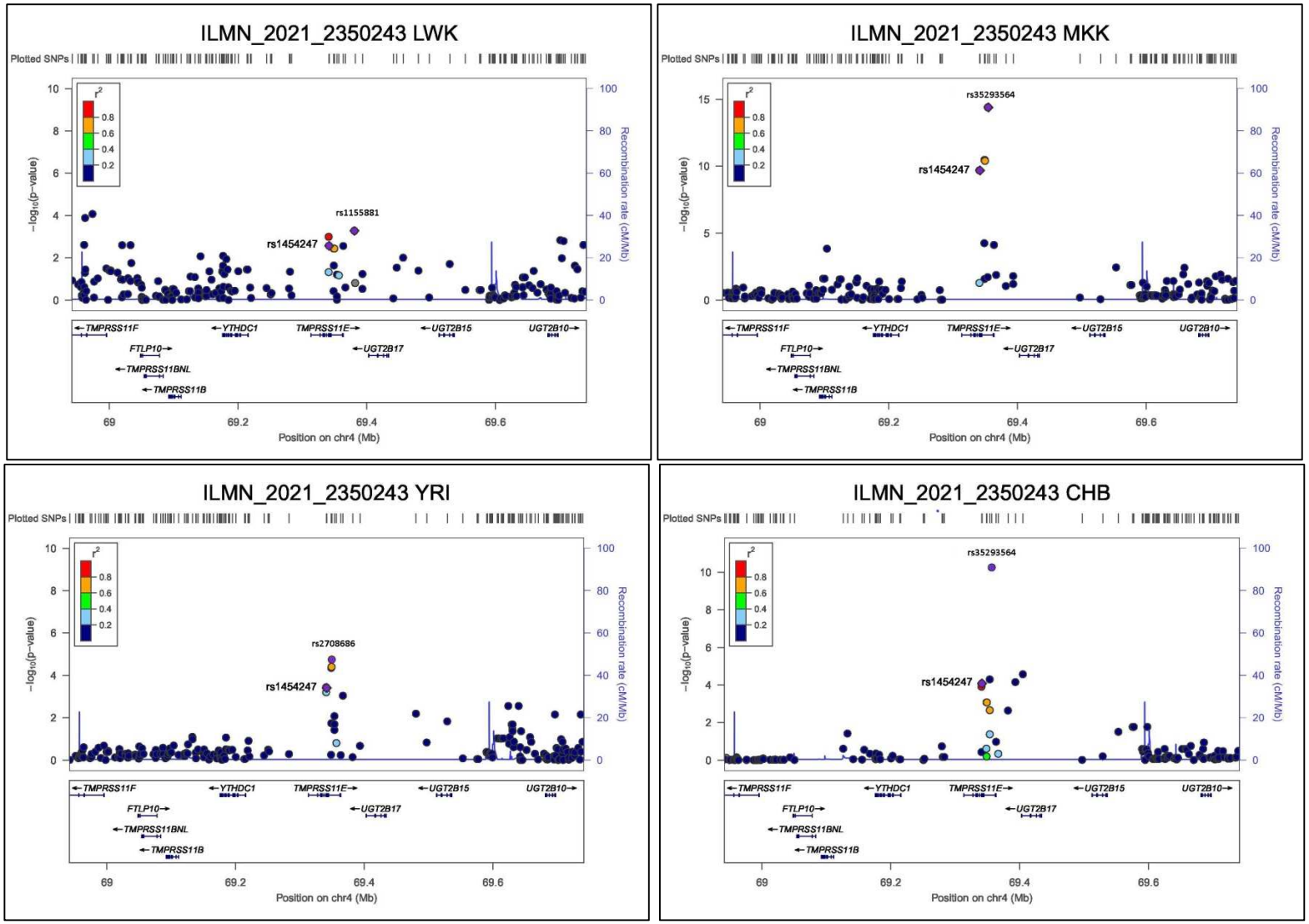


Figure 8.7: Signal plots for peak phase III HapMap SNPs for probe ILMN_2021_2350243

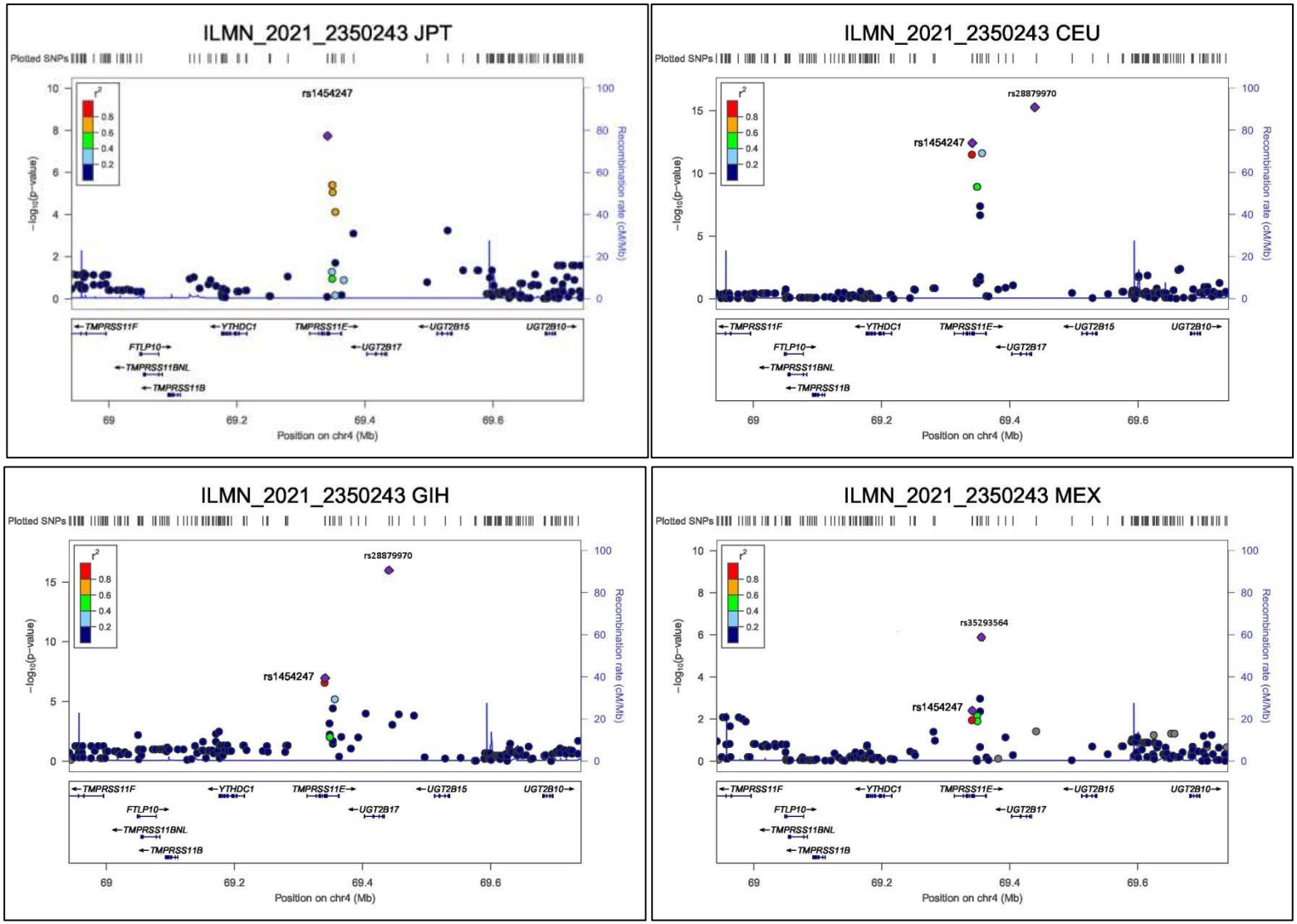


Figure 8.8: Signal plots for peak phase III HapMap SNPs for probe ILMN_2021_2350243

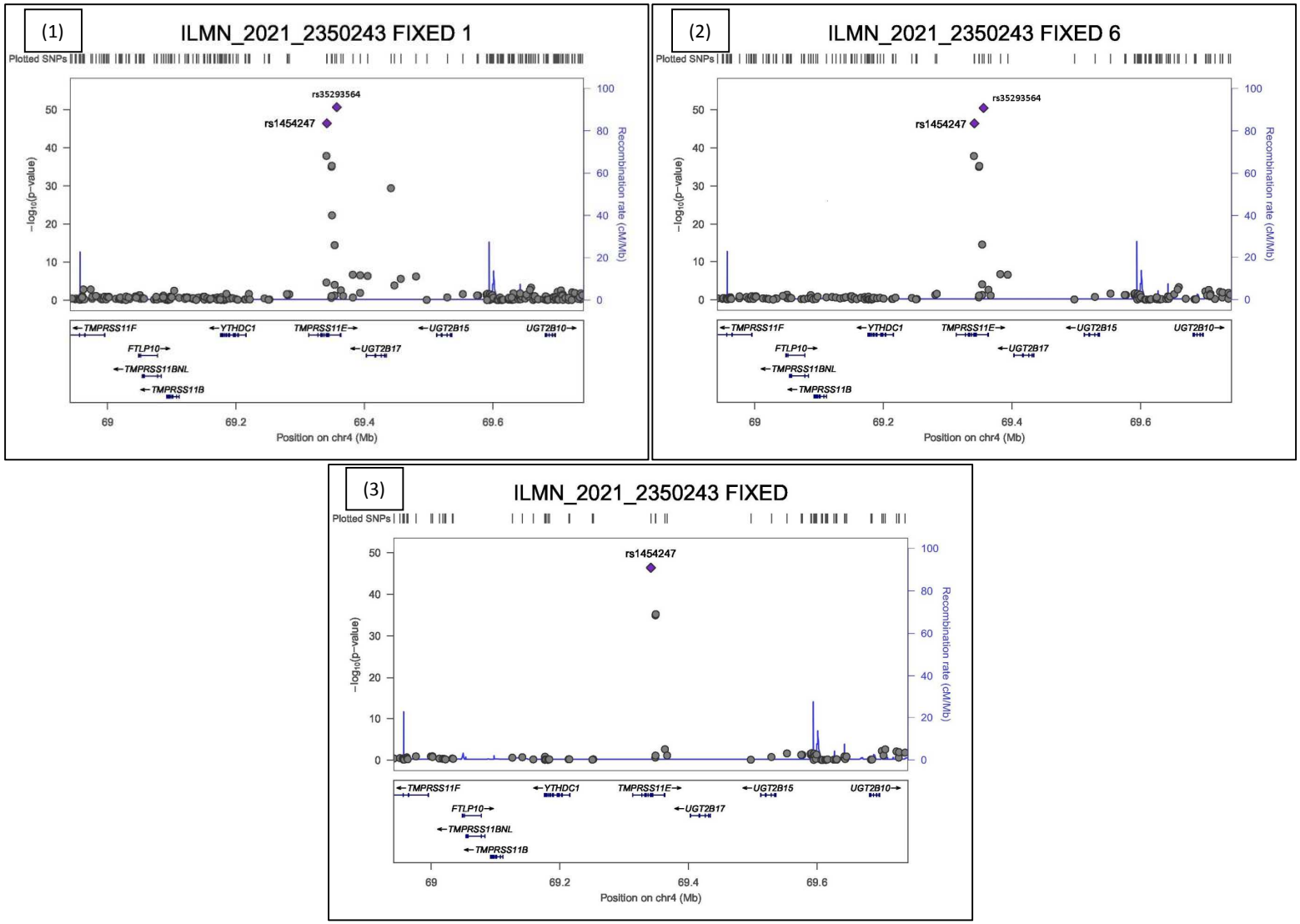


Figure 8.9: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_2021_2350243. (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.5.3 ILMN_5225_7050768 (ENSG00000213759, UGT2B11)

This probe is within the ADME extended gene list and is a Phase II metabolism enzyme. It has significant heterogeneity (Cochran's Q = 30.71, p-value = 7.03×10^{-5}). The peak SNP is rs1454247, a phase III HapMap SNP (Z-score = -14.36, p-value = 1.00×10^{-46}).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000213759 and the HGNC symbol *UGT2B11*. Its full name is UDP Glucuronosyltransferase 2 Family, Polypeptide B11, the gene's start-site is chr4: 70080449. The phase III HapMap SNP rs1454247 is located at chr4:69341579. The SNP is intronic within the gene *TMPRSS11E*.

From the previous sections, it can be observed that the SNP rs1454247 is the peak SNP in four of the ADME probes with significant heterogeneity. This SNP lies within the intron of the gene *TMPRSS11E*. The imputed 1000 Genomes dataset introduced in chapter 5 was used to fine map the peak SNP for these probes. It was found that in all four probes, when using 1000 Genomes imputed data, the peak SNP was rs139440909, located at chr4:69365052. The 1000 Genomes SNP is no longer within the *TMPRSS11E,* but is intergenic and maps 1730 bps away from the gene. Apart from this, no other annotation was found for this SNP.

UGT2B11 is a phase II metabolism enzyme and is of major importance in the conjugation and subsequent elimination of potentially toxic xenobiotics and endogeneous compounds. The gene is associated with the disease Gilbert Syndrome (a genetic liver disorder).

Figure 8.10 shows the forest plot for ILMN_5225_7050768 with SNP rs1454247. As can be seen all populations effect sizes significantly differ from zero at $p \leq 0.05$. Visually the African and East Asian ancestry groups form effect size clusters.

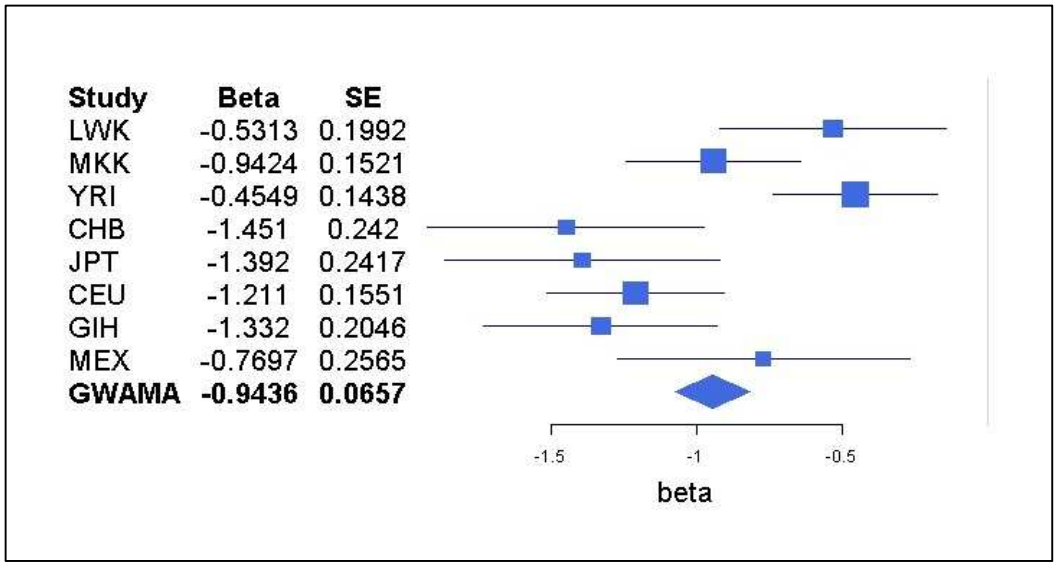


Figure 8.10: Forest plot of association and fixed effect meta-analyses for probe ILMN_5225_7050768 with SNP rs1454247.

Table 8.12 shows p-values, allele frequencies and peak signals for probe ILMN_5225_7050768 with SNP rs1454247. The following observations can be made: Three populations (MKK, CEU and GIH) have effect sizes significant at GWS ($p \leq 5x10^{-8}$). For this probe the peak SNP (rs1454247) is shared with the fixed effect meta-analysis in three populations (MKK, CHB and JPT). Allele frequencies range from 0.21 to 0.56

Population	Minor Allele (C / T)	Allele Frequency (Allele C)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	C	0.44	-0.5313	0.1992	$9.30x10^{-3}$	N
MKK	C	0.47	-0.9424	0.1521	$6.93x10^{-9}$	Y
YRI	C	0.36	-0.4549	0.1438	$2.00x10^{-3}$	N
CHB	C	0.24	-1.451	0.242	$6.34x10^{-8}$	Y
JPT	C	0.21	-1.392	0.2417	$1.60x10^{-7}$	Y
CEU	T	0.54	-1.211	0.1551	$4.84x10^{-12}$	N
GIH	C	0.43	-1.332	0.2046	$8.72x10^{-9}$	N
MEX	T	0.56	-0.7697	0.2565	$4.70x10^{-3}$	N
FIXED	--	--	-0.9436	0.0657	$1.00x10^{-46}$	Y

Table 8.12: Table of allele frequencies for phase III HapMap SNP rs1454247 and probe ILMN_5225_7050768.

Figures on the following pages show signal plots for association analysis (figures 8.11 and 8.12) and fixed effect meta-analysis results (figure 8.13). The following observations can be made. First, the SNP rs1454247 is still the peak SNP when non-reported variants are included in the fixed effect meta-analysis. rs1454247 is the peak SNP in three populations MKK, CHB and JPT. Second, in the populations CEU and GIH, there is an isolated peak SNP (rs28879970), not in LD with rs1454247. Heterogeneity may be due to differing LD with causal variant.

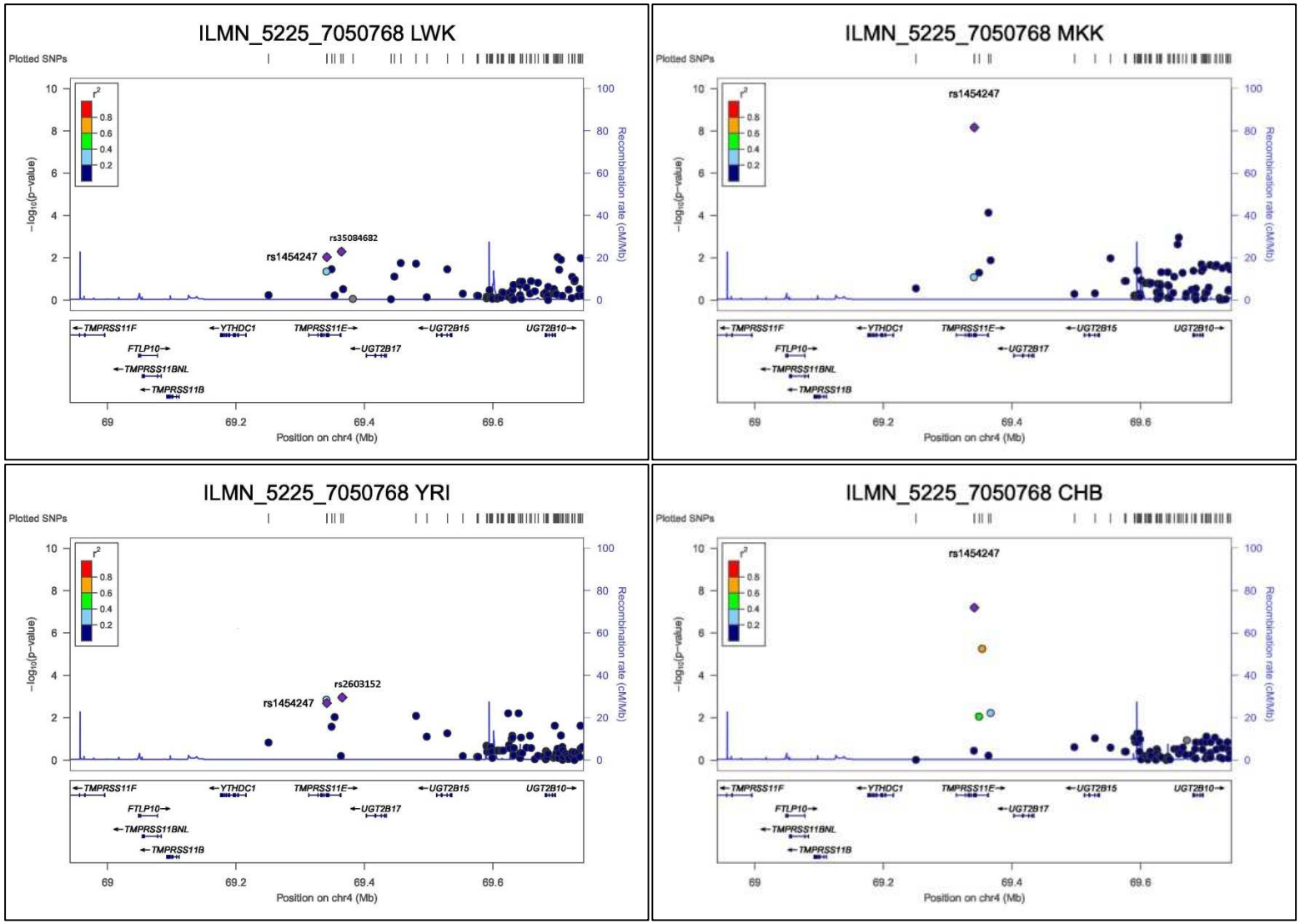


Figure 8.11: Signal plots for peak phase III HapMap SNPs for probe ILMN_5225_7050768

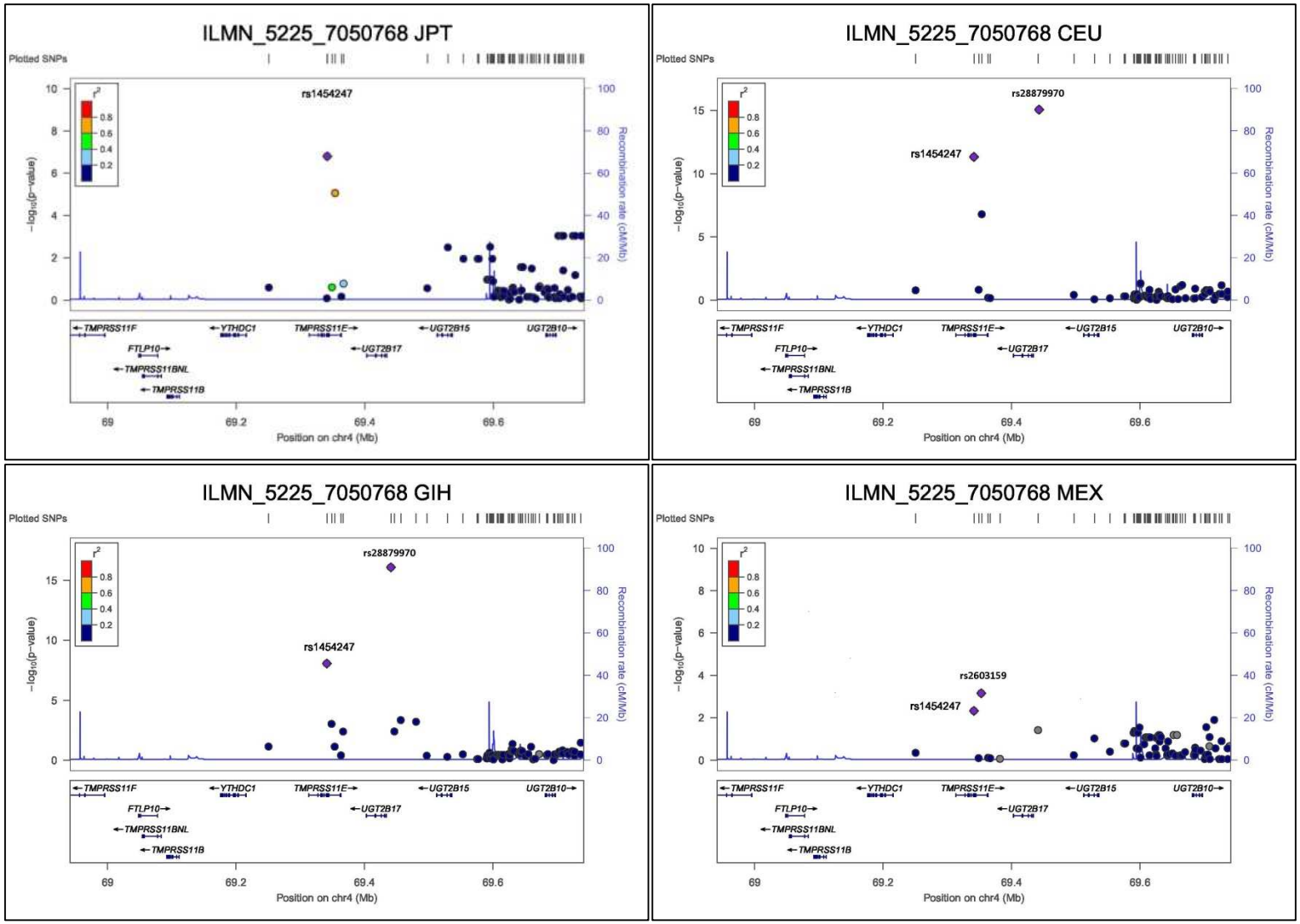


Figure 8.12: Signal plots for peak phase III HapMap SNPs for probe ILMN_5225_7050768

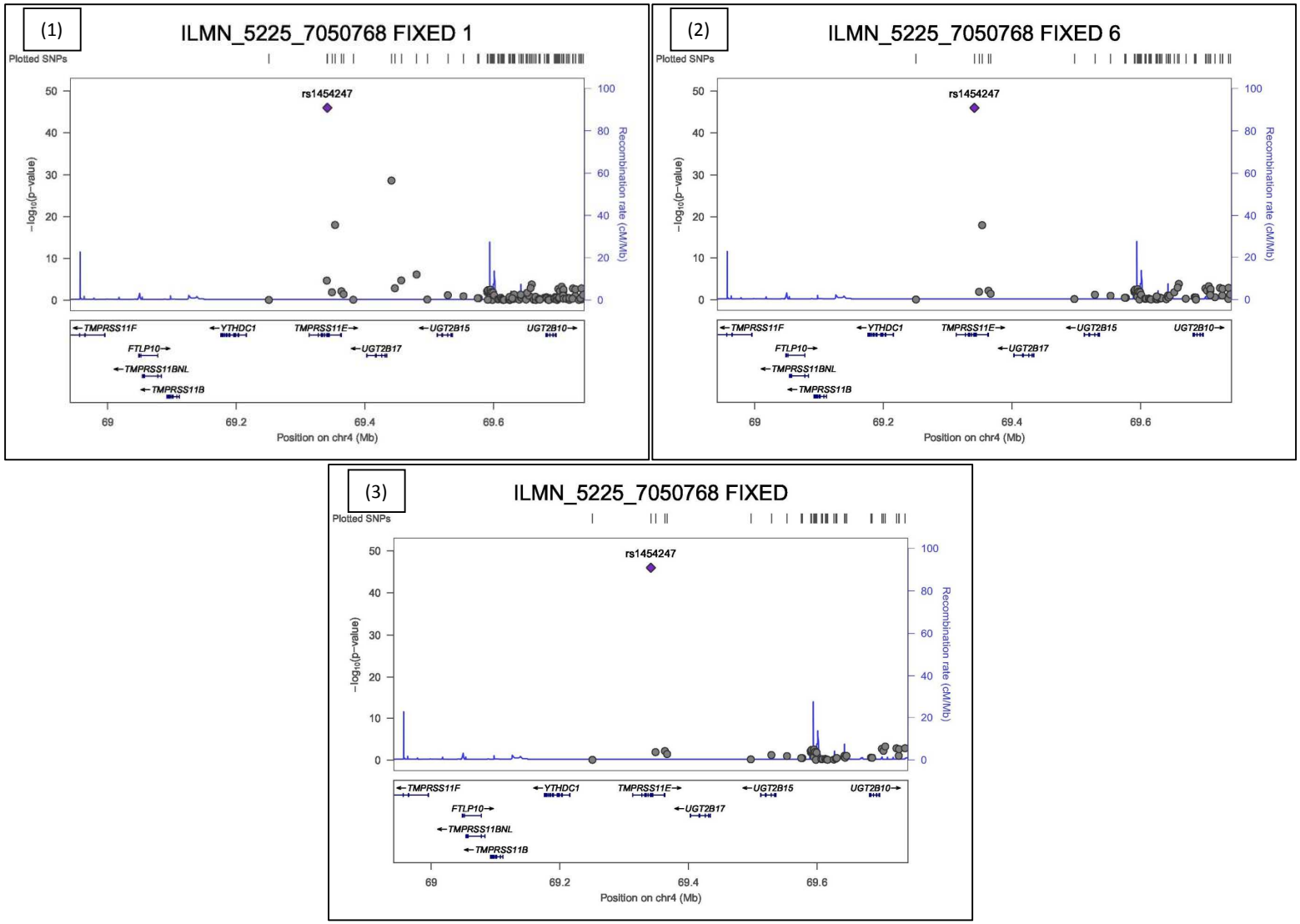


Figure 8.13: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_5225_7050768. (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.5.4 ILMN_16478_1710170 (ENSG00000155465, SLC7A7)

This probe is within the ADME extended gene list with class transporter and has significant heterogeneity (Cochran's Q = 25.81, p-value = 5.45×10^{-4}). The peak SNP is rs12884337, a phase III HapMap SNP (Z-score = -16.8, p-value = 2.93×10^{-63}), this is also the most significant *cis* eQTL in the ADME list.

Gene name and location

The probe's gene has the Ensembl ID ENSG00000155465 and the HGNC symbol *SLC7A7*. Its full name is Solute Carrier Family 7 (Cationic Amino Acid Transporter, y+ System), Member 7, the gene's start-site is chr14: 23299029. The SNP rs12884337 is located at chr14:23273114. The SNP is intronic, within the gene *SLC7A7*.

This protein is a transporter and forms the light subunit of a cationic amino acid transporter. This transporter is found in epithelial cell membranes where it transfers cationic and large neutral amino acids from the cell to the extra cellular space. This gene is associated with the disease lysinuric protein intolerance (LPI) a metabolic disorder affecting amino acid transport.

Figure 8.14 shows the forest plot for probe ILMN_16478_1710170 with SNP rs12884337. Visually clustering of effect sizes is observed within all ancestry groups. All signals are significantly different from zero at $p \leq 0.05$ with the exception of YRI.

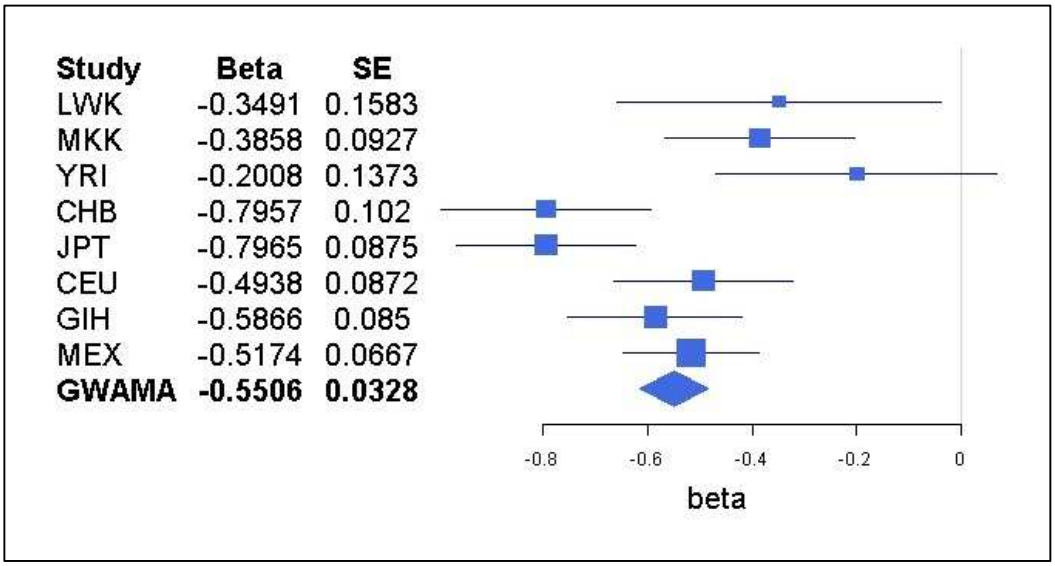


Figure 8.14: Forest plot of association and fixed effect meta-analyses for probe ILMN_16478_1710170 with SNP rs12884337.

Table 8.13 shows p-values, allele frequencies and peak SNP data for ILMN_16478_1710170 with rs12884337. The following observations can be made: Four populations (CHB, JPT, GIH, MEX) are significant at GWS ($p \leq 5x10^{-8}$). Five of the populations share the same peak SNP as the fixed effect meta-analysis, all populations which have a different peak SNP are within the African ancestry group. Allele frequencies range from 0.21 to 0.61.

Population	Minor Allele (C / T)	Allele Frequency (Allele C)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	C	0.21	-0.3491	0.1583	0.0303	N
MKK	C	0.36	-0.3858	0.0927	$5.61x10^{-5}$	N
YRI	C	0.26	-0.2008	0.1373	0.1465	N
CHB	C	0.32	-0.7957	0.102	$2.61x10^{-11}$	Y
JPT	C	0.43	-0.7965	0.0875	$6.77x10^{-14}$	Y
CEU	T	0.56	-0.4938	0.0872	$1.34x10^{-7}$	Y
GIH	T	0.59	-0.5866	0.085	$1.69x10^{-9}$	Y
MEX	T	0.61	-0.5174	0.0667	$2.41x10^{-9}$	Y
FIXED	--	--	-0.5506	0.0328	$2.93x10^{-63}$	Y

Table 8.13: Table of allele frequencies for phase III HapMap SNP rs12884337 and probe ILMN_16478_1710170.

Figures on the following pages show signal plots for association analysis (figures 8.15 and 8.16) and fixed effect meta-analysis results (figure 8.17). The following observations can be made: The SNP rs12884337 remains the peak SNP when non-reported variants are included in the fixed effect meta-analysis. All eight populations have a peak signal. The African populations have peak signals which are not in LD with rs12884337. Heterogeneity is being driven by difference in effect size in the African population, however signals are observed in the African populations.

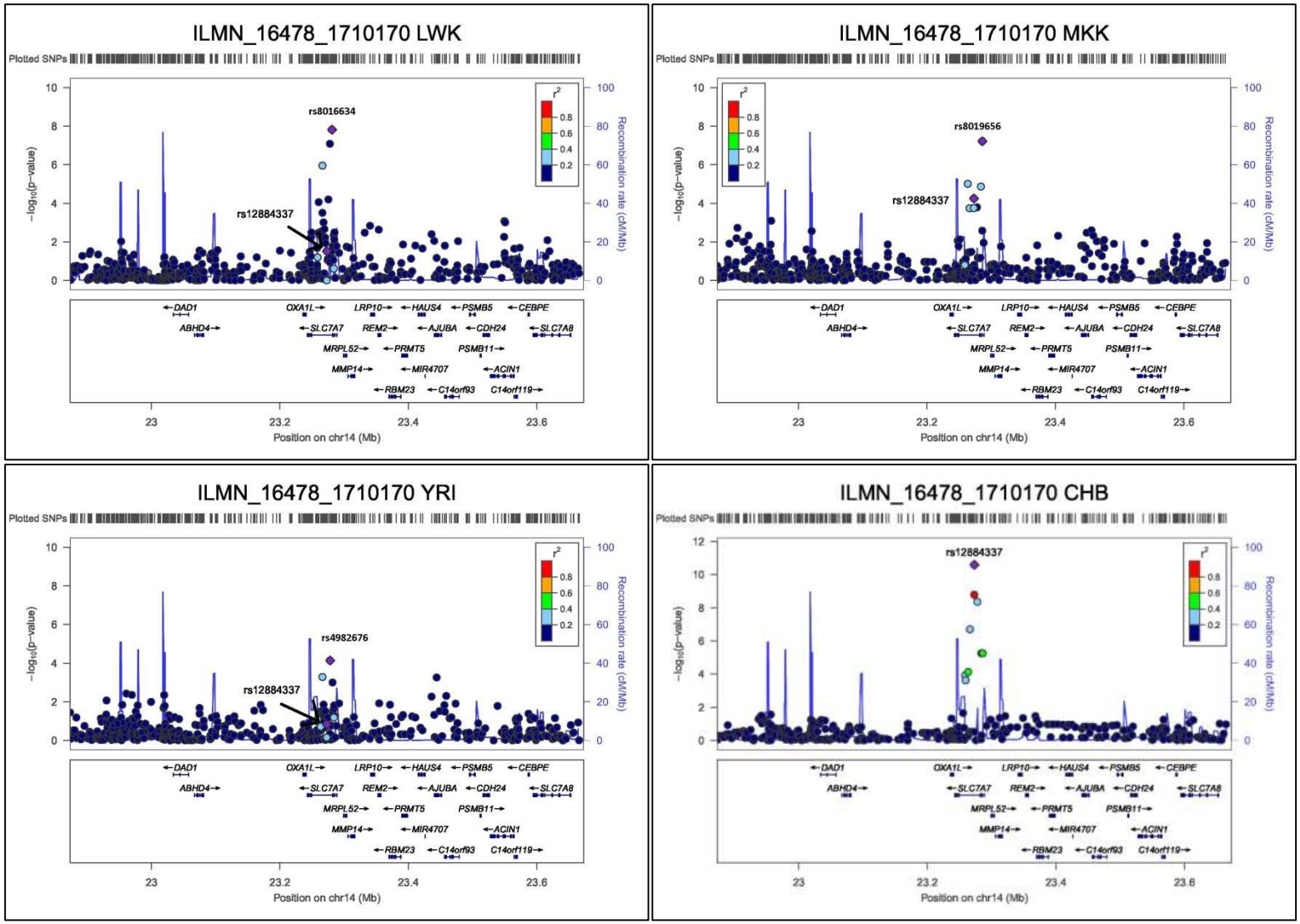


Figure 8.15: Signal plots for peak phase III HapMap SNPs for probe ILMN_16478_1710170

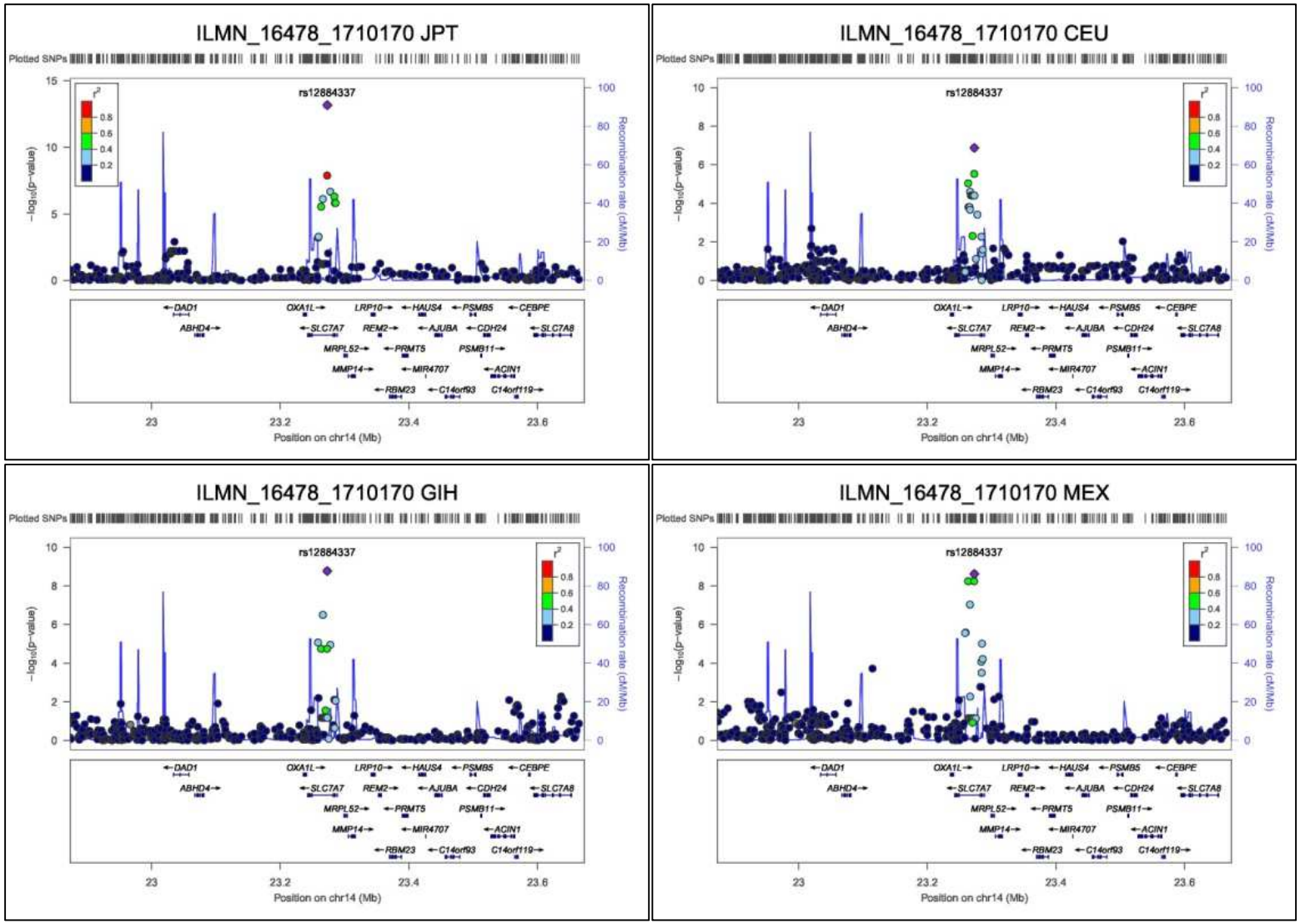


Figure 8.16: Signal plots for peak phase III HapMap SNPs for probe ILMN_16478_1710170

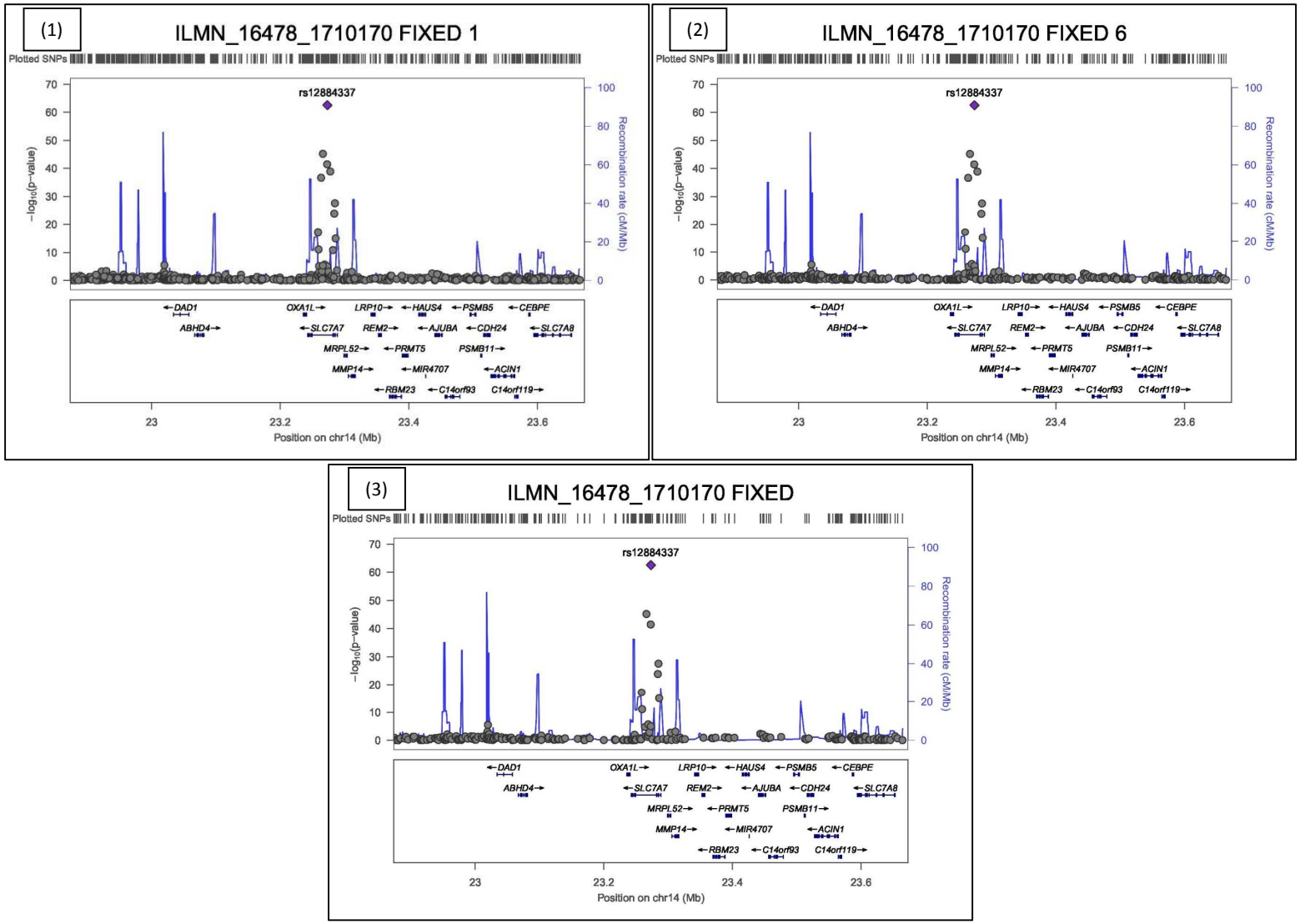


Figure 8.17: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_16478_1710170. (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.5.5 ILMN_15891_6480091 (ENSG00000134184, GSTM1)

This probe is within the ADME Core gene list with class phase II metabolism enzyme and has significant heterogeneity (Cochran's Q = 29.22, p-value = 1.32×10^{-4}). The peak SNP is rs12745189, a phase III HapMap SNP (Z-score = -8.36, p-value = 6.55×10^{-17}).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000134184 and the HGNC symbol *GSTM1*. Its full name is Glutathione S-Transferase Mu 1 the gene's start-site is chr1:110230436. The SNP rs12745189 is located at chr1:110186563.

The gene is within the ADME Core list and is a Phase II metabolism enzyme. This gene encodes a glutathione S-transferase that belongs to the mu class. The enzyme functions in the detoxification of electrophilic compounds, including carcinogens, therapeutic drugs, environmental toxins and products of oxidative stress, by conjugation with glutathione. Diseases associated with *GSTM1* include lung cancer (gstm1-related) and asbestosis. GSTM1 has been reported as the reported gene is a GWAS for bladder cancer at GWS (p-value = 5.00×10^{-31})

Figure 8.18 shows a forest plot of probe ILMN_15891_6480091 with SNP rs12745189. LWK, YRI and CEU effect sizes do not differ significantly from zero at $p \leq 0.05$. Visually, the East Asian ancestry groups cluster by effect size.

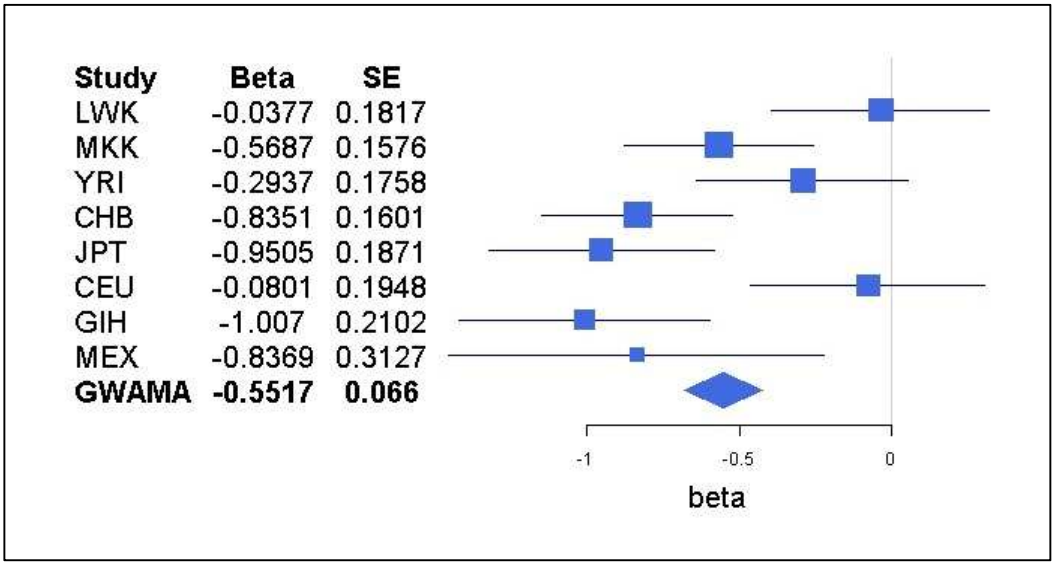


Figure 8.18: Forest plot of association and fixed effect meta-analyses for probe ILMN_15891_6480091 with SNP rs12745189.

Table 8.14 shows p-values, allele frequencies and peak SNPs for ILMN_15891_6480091 with SNP rs12745189. From the table the following observations can be made: None of the eight populations are significant at GWS ($p \leq 5x10^{-8}$). None of the peak SNPs matches the fixed effect peak SNP. Allele frequencies range from 0.41 to 0.70.

Population	Minor Allele (T / C)	Allele Frequency (Allele T)	Beta (Allele T)	SE	p-value	Peak SNP
LWK	T	0.44	-0.0377	0.1817	0.8361	N
MKK	T	0.47	-0.5687	0.1576	$4.37x10^{-4}$	N
YRI	T	0.37	-0.2937	0.1758	0.0977	N
CHB	C	0.59	-0.8351	0.1601	$1.54x10^{-6}$	N
JPT	C	0.68	-0.9505	0.1871	$2.52x10^{-6}$	N
CEU	T	0.41	-0.0801	0.1948	0.6818	N
GIH	T	0.46	-1.007	0.2102	$8.65x10^{-6}$	N
MEX	C	0.70	-0.8369	0.3127	0.0109	N
FIXED	--	--	-0.5517	0.0660	$6.55x10^{-17}$	Y

Table 8.14: Table of allele frequencies for phase III HapMap SNP rs12745189 and probe ILMN_15891_6480091.

Figures on the following pages show signal plots for association analysis (figures 8.19 and 8.20) and fixed effect meta-analysis results (figure 8.21). The following observations can be made: When non-reported variants are included in the fixed effect meta-analysis rs12745189 is no longer the peak SNP. With six or more SNPs present: rs1056806 is the peak SNP (p-value=7.84×10^{-18}), this SNP is missing in MKK and CEU. When one or more SNPs are present: rs3754446 is the peak SNP (p-value=1.12×10^{-25}), this SNP is missing in LWK, MKK and YRI. There is evidence of a signal in all eight populations. There is also a strong signal in YRI, even though rs12745189 is not significant at 0.05. CHB and JPT both have signal, but rs12745189 is not the peak SNP.

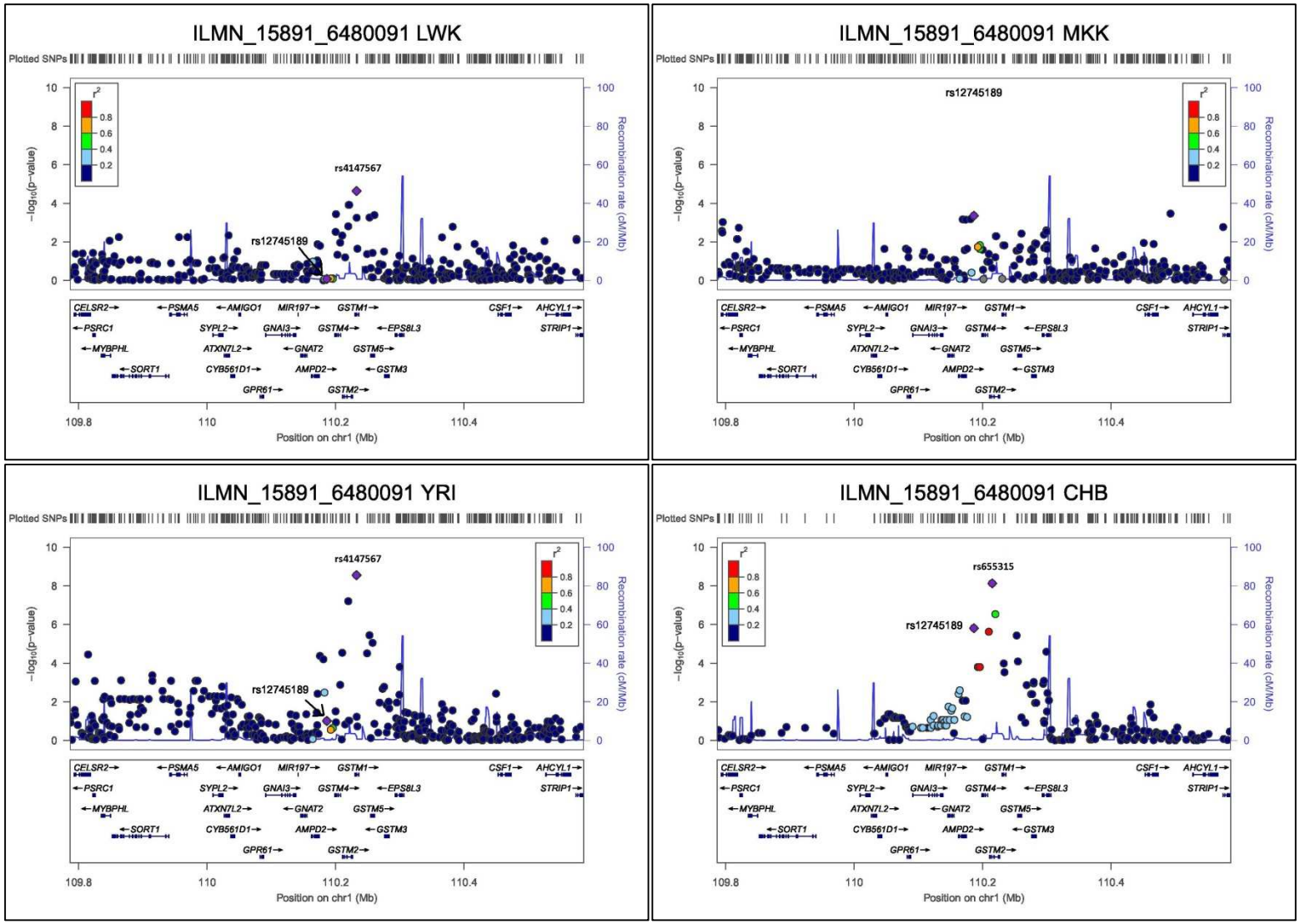


Figure 8.19: Signal plots for peak phase III HapMap SNPs for probe ILMN_15891_6480091

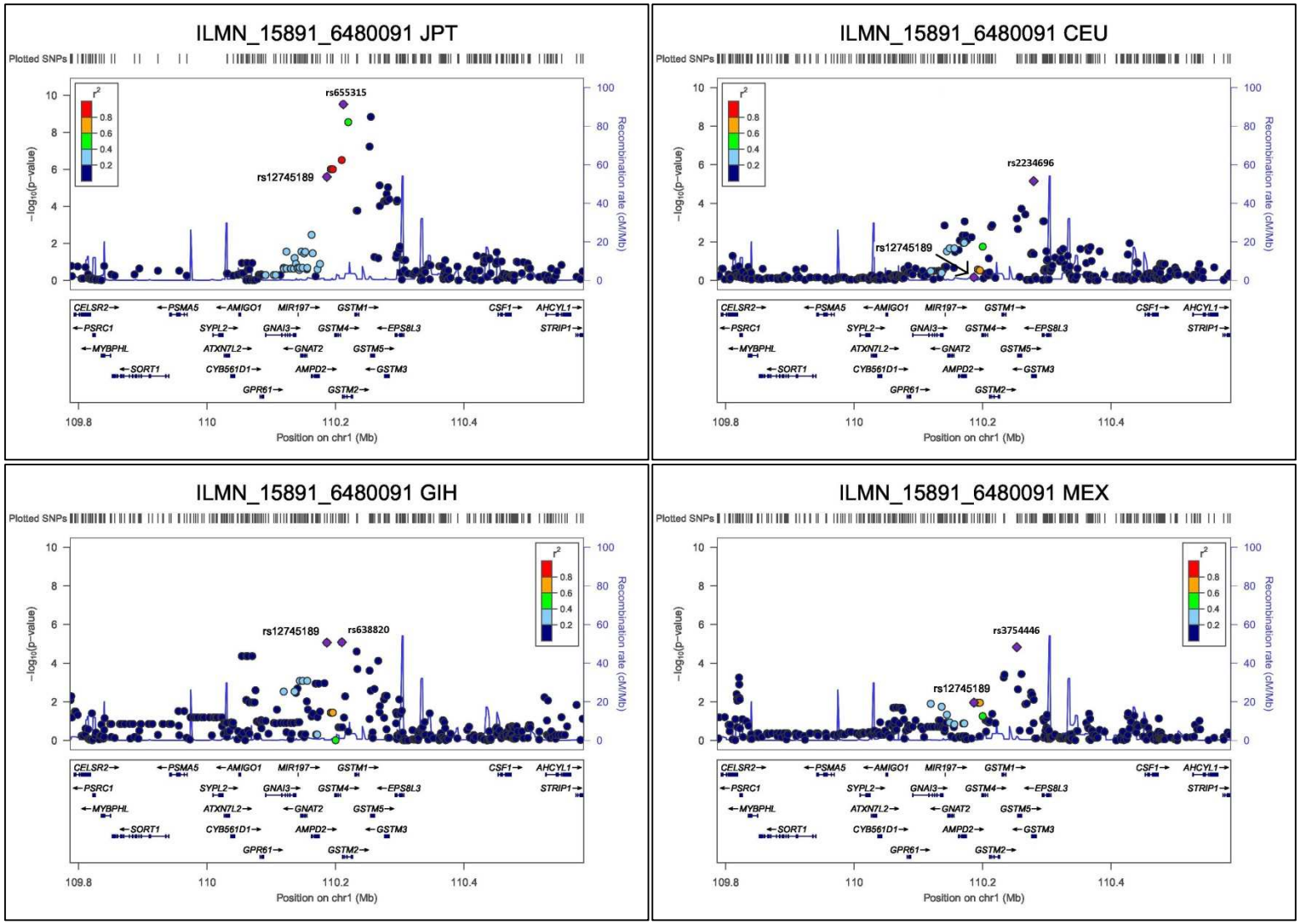


Figure 8.20: Signal plots for peak phase III HapMap SNPs for probe ILMN_15891_6480091

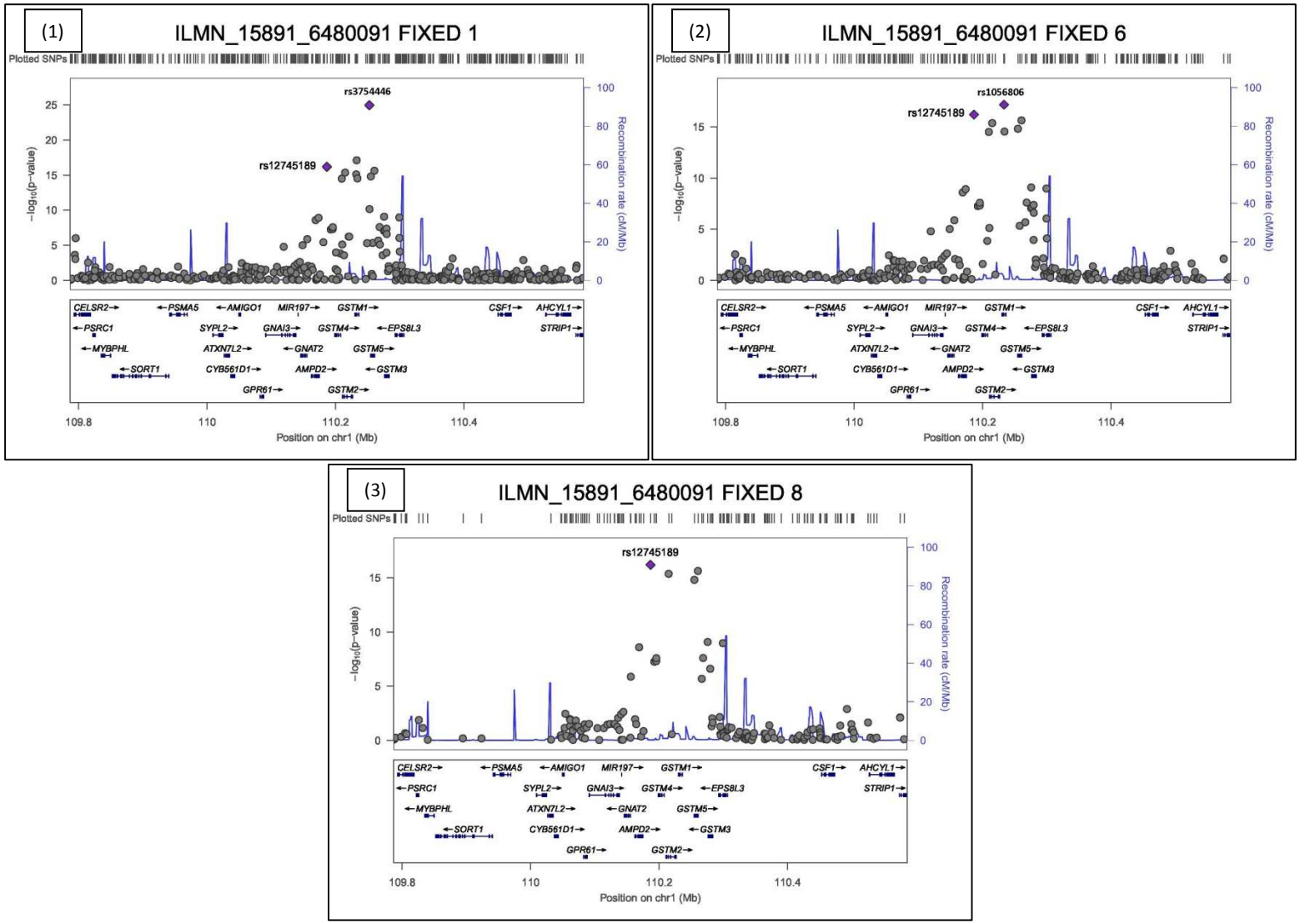


Figure 8.21: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_16478_1710170. (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.5.6 ILMN_18916_4850209 (ENSG00000187630, DHRS4L2)

This probe is within the ADME Extended gene list with class Phase I metabolism enzyme and has significant heterogeneity (Cochran's Q = 27.71, p-value = 2.47×10^{-4}). The peak SNP is rs8022613, a phase III HapMap SNP (Z-score = -5.46, p-value = 4.75×10^{-8}).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000187630 and the HGNC symbol *DHRS4L2*. Its full name is Dehydrogenase /Reductase (SDR Family) Member 4 Like 2 the gene's start-site is chr14:24439148. The phase III HapMap SNP rs8022613 is located at chr14:24406918.

The gene is within the ADME Extended gene list with class Phase I metabolism enzyme. It is a member of the short chain dehydrogenase reductase family. It may be a NADPH dependent retinol oxidoreductase (retinol is an animal form of vitamin A).

Figure 8.22 shows a forest plot for ILMN_18916_4850290 with SNP rs8022613. Visually there is some evidence of clustering within the African ancestry group. Five of the populations do not significantly differ from zero $p \leq 0.05$.

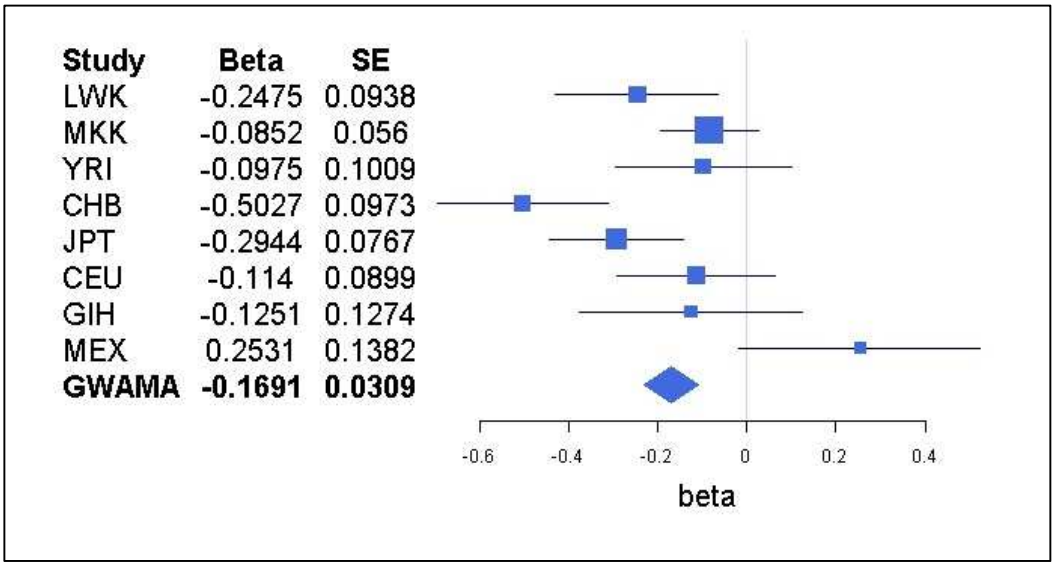


Figure 8.22: Forest plot of association and fixed effect meta-analyses for probe ILMN_18916_4850209 with SNP rs8022613.

Table 8.15 shows p-values, allele frequencies and peak signals for ILMN_18916_4850209 at SNP rs8022613. From the table the following observations can be made: Allele frequencies range from 0.15 – 0.95. None of the populations are significant at GWS ($p \leq 5x10^{-8}$). CHB has the same peak SNP as the fixed effect meta-analysis.

Population	Minor Allele (T / C)	Allele Frequency (Allele T)	Beta (Allele C)	SE	p-value	Peak SNP	Peak Length (SNP count)
LWK	T	0.22	-0.2475	0.0938	$9.99x10^{-3}$	N	1
MKK	T	0.33	-0.0852	0.056	0.1305	N	2
YRI	T	0.15	-0.0975	0.1009	0.3360	N	1
CHB	C	0.87	-0.5027	0.0973	$1.84x10^{-6}$	Y	1
JPT	C	0.89	-0.2944	0.0767	$2.53x10^{-4}$	N	1
CEU	C	0.89	-0.114	0.0899	0.2075	N	1
GIH	C	0.89	-0.1251	0.1274	0.3293	N	4
MEX	C	0.95	0.2531	0.1382	0.0748	N	1
FIXED	--	--	-0.1691	0.009	$4.75x10^{-8}$	Y	1

Table 8.15: Table of allele frequencies for phase III HapMap SNP rs8022613 and probe ILMN_18961_4850209.

Figures on the following pages show signal plots for association analysis (figures 8.23 and 8.24) and fixed effect meta-analysis results (figure 8.25). The following observations can be made: When non-reported variants are included in the fixed effect meta-analysis rs8022613 is no longer the peak SNP. Six or more SNPs are present the peak signal is rs2146432 ($p=9.49x10^{-10}$) this SNP is missing in one population: YRI. When one or more SNPs are present the peak signal is rs9805889 ($p=2.32x10^{-14}$) this SNP is missing in three populations: CHB, JPT, GIH, MEX. There is evidence of a signal in six of the populations, but GIH and MEX do not appear to have any signal at all. rs8022613 is the peak signal in the CHB population. The heterogeneity due to non-reported variants and no signal detected in GIH and MEX.

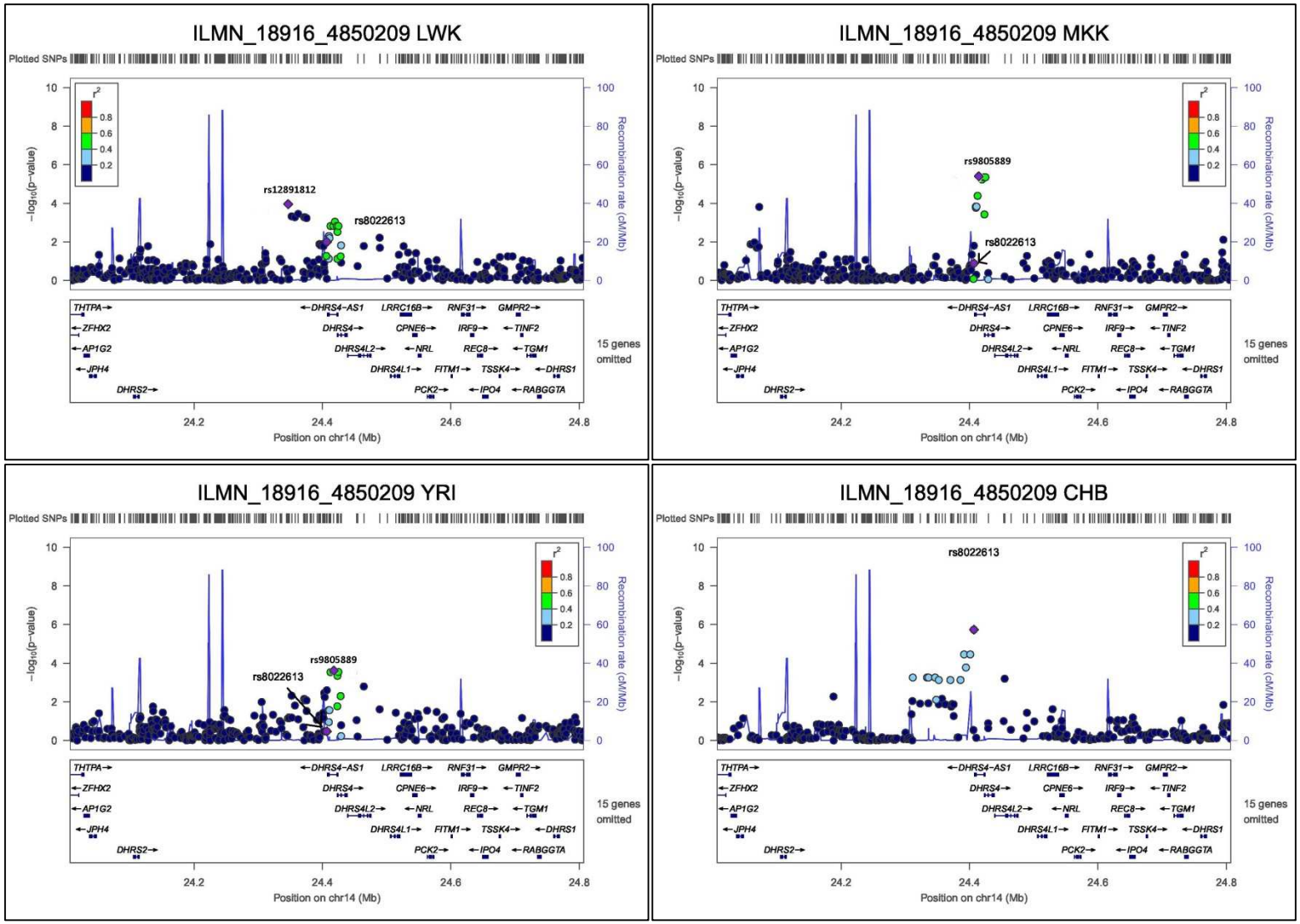


Figure 8.23: Signal plots for peak phase III HapMap SNPs for probe ILMN_18916_4850209

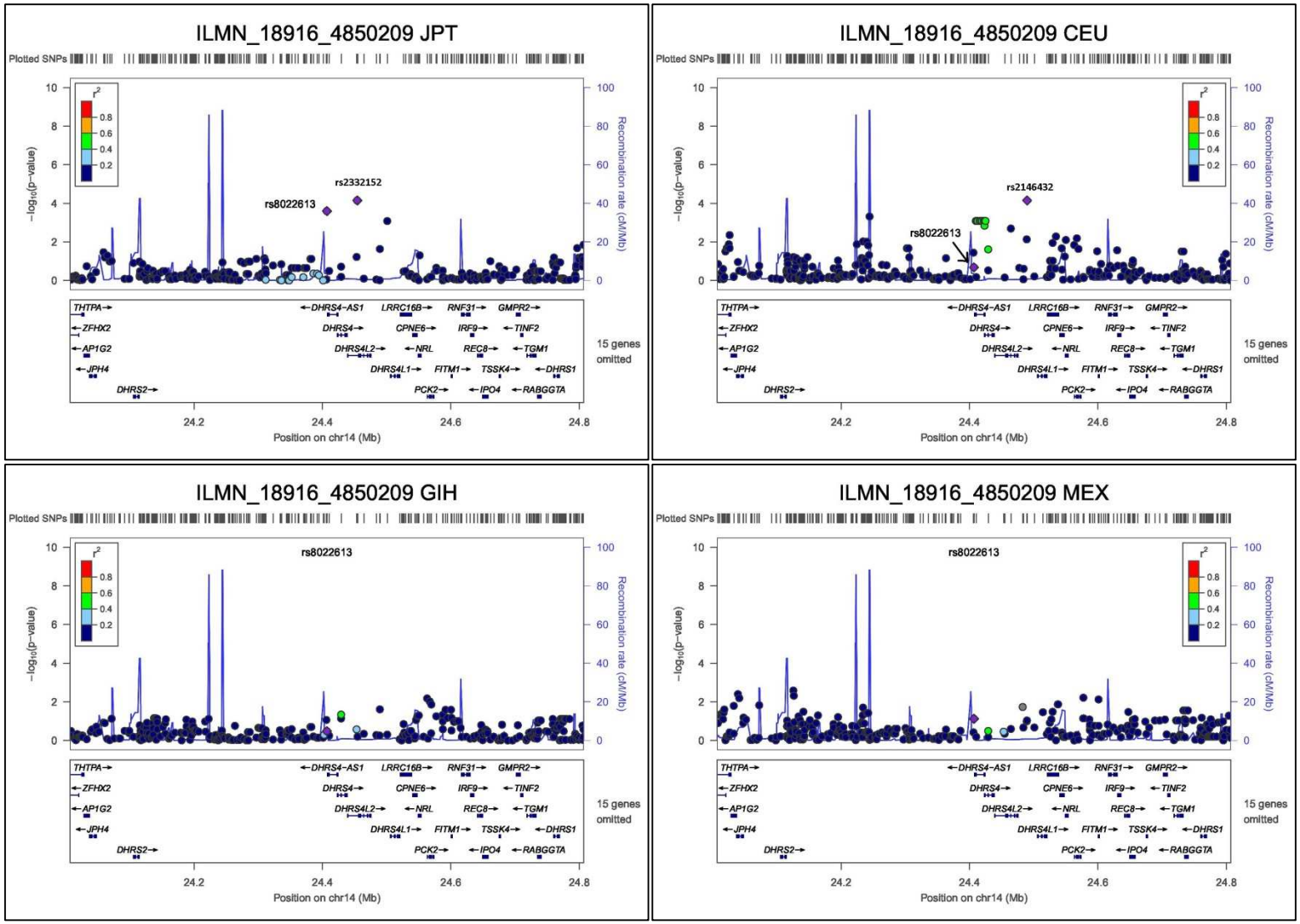


Figure 8.24: Signal plots for peak phase III HapMap SNPs for probe ILMN_18916_4850209

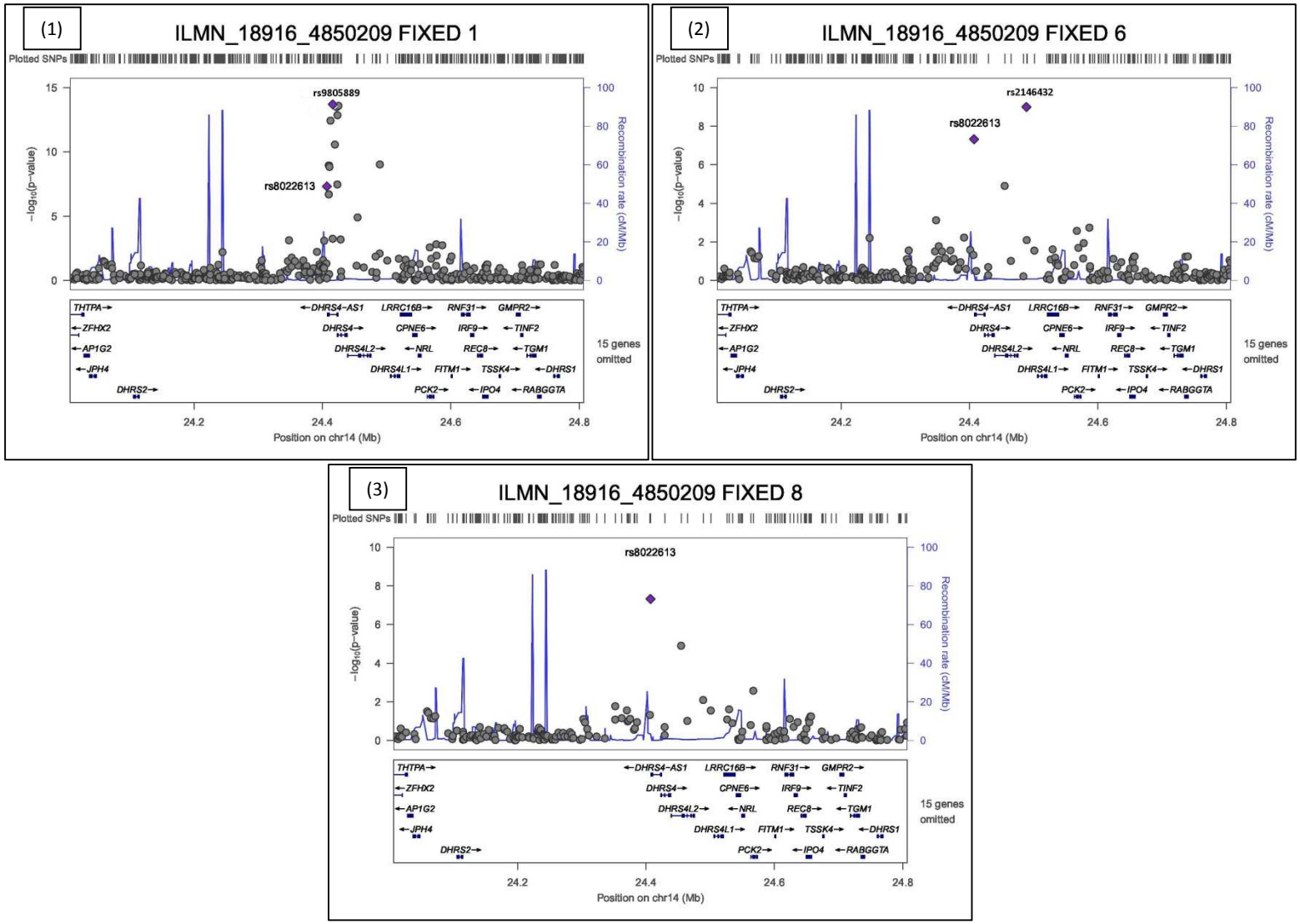


Figure 8.25: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_18916_4850209. (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.6 ADME Functional variants

ADME functional variants are markers within ADME genes that are determined to be important. This section presents results of an analysis which integrates ADME functional variants with peak *cis* eQTLs. Functional variants were retrieved from the ADME Safari website (https://www.ebi.ac.uk/chembl/admesarfari).

One variant was found, the probe ILMN_15545_5670059 for the gene DHRS1. The overlapping variant is rs10134537 (chr14:24760764; Amino Acid position: 241 (Isoleucin (I) to Threonine (T))).

8.6.1 ILMN_15545_5670059 (ENSG00000157379, DHRS1)

This probe is for the gene DHRS1 which is in the ADME extended list and is a Phase I metabolism enzyme. The peak SNP for this probe is an ADME functional variant: rs10134537, a phase III HapMap SNP (z-score = -14.88, p-value = 5.19×10^{-50}), with no significant heterogeneity (Cochran's Q = 5.16, p-value = 0.641).

Gene name and location

The probe's gene has the Ensembl ID ENSG00000157379 and the HGNC symbol DHRS1. Its full name is Dehydrogenase/Reductase (SDR Family) Member 1, the gene's start-site is chr14: 24769039 (NCBI B37). The SNP is rs10134537 is located at chr14: 24760764 (NCBI B37). The SNP is exonic within gene DHRS1 (Amino Acid position: 241 (Isoleucin (I) to Threonine (T)). This gene is a Phase I metabolism enzyme which is a member of the short-chain dehydrogenase/ reductase (SDR) family. The encoded enzyme likely functions as an oxidoreductase.

Figure 8.26 shows a forest plot for the probe ILMN_15545_5670059. No evidence of heterogeneity and effect size clustering by ancestry group is observed. All eight populations differ from zero at $p \leq 0.05$.

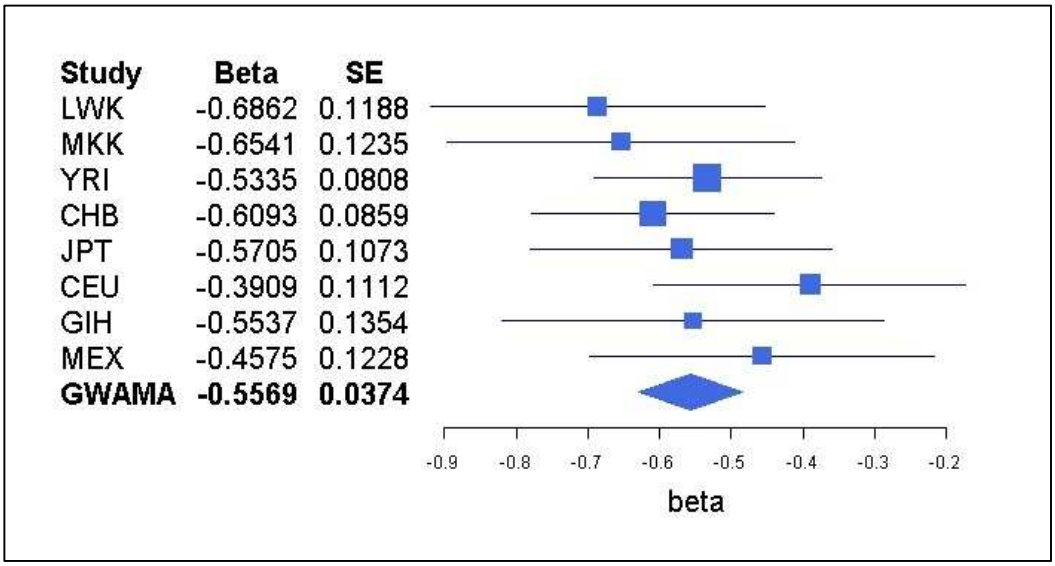


Figure 8.26: Forest plot of association and fixed effect meta-analyses for probe ILMN_15545_5670059 with SNP rs10134537.

Table 8.16 shows the p-values, allele frequencies and peak SNPs for ILMN_15545_5670059 at the SNP rs10134537. From the table the following observations can be made: All populations share the peak SNP with the fixed effect meta-analysis. Allele frequencies range from 0.05 to 0.13. Two populations YRI and CHB are significant at GWS ($p \leq 5x10^{-8}$). All eight populations peak SNP are also the fixed effect meta-analysis peak SNP.

Population	Minor Allele (A / G)	Allele Frequency (Allele A)	Beta (Allele A)	SE	p-value	Peak SNP
LWK	A	0.06	-0.6862	0.1188	$1.46x10^{-7}$	Y
MKK	A	0.06	-0.6541	0.1235	$4.80x10^{-7}$	Y
YRI	A	0.13	-0.5335	0.0808	$1.72x10^{-9}$	Y
CHB	A	0.06	-0.6093	0.0859	$5.88x10^{-10}$	Y
JPT	A	0.05	-0.5705	0.1073	$9.85x10^{-7}$	Y
CEU	A	0.06	-0.3909	0.1112	$7.00x10^{-4}$	Y
GIH	A	0.09	-0.5537	0.1354	$1.00x10^{-4}$	Y
MEX	A	0.11	-0.4575	0.1228	$6.00x10^{-4}$	Y
FIXED	--	--	-0.5569	0.0374	$5.19x10^{-50}$	Y

Table 8.16: Table of allele frequencies for phase III HapMap SNP rs10134537 and probe ILMN_15545_5670059.

Figures on the following pages show signal plots for association analysis (figures 8.27 and 8.28) and fixed effect meta-analysis results (figure 8.29). The following observations can be made the eSNP rs10134537 is the peak SNP in all the association analysis results. rs10134537 is still the peak SNP if non-reported variants are included in the fixed effect meta-analysis. African ancestry groups have a single peak with no LD block. Signals in Eurasian-Hispanic ancestry groups are weaker with no LD blocks. There is evidence of LD blocks in the East Asian ancestry groups.

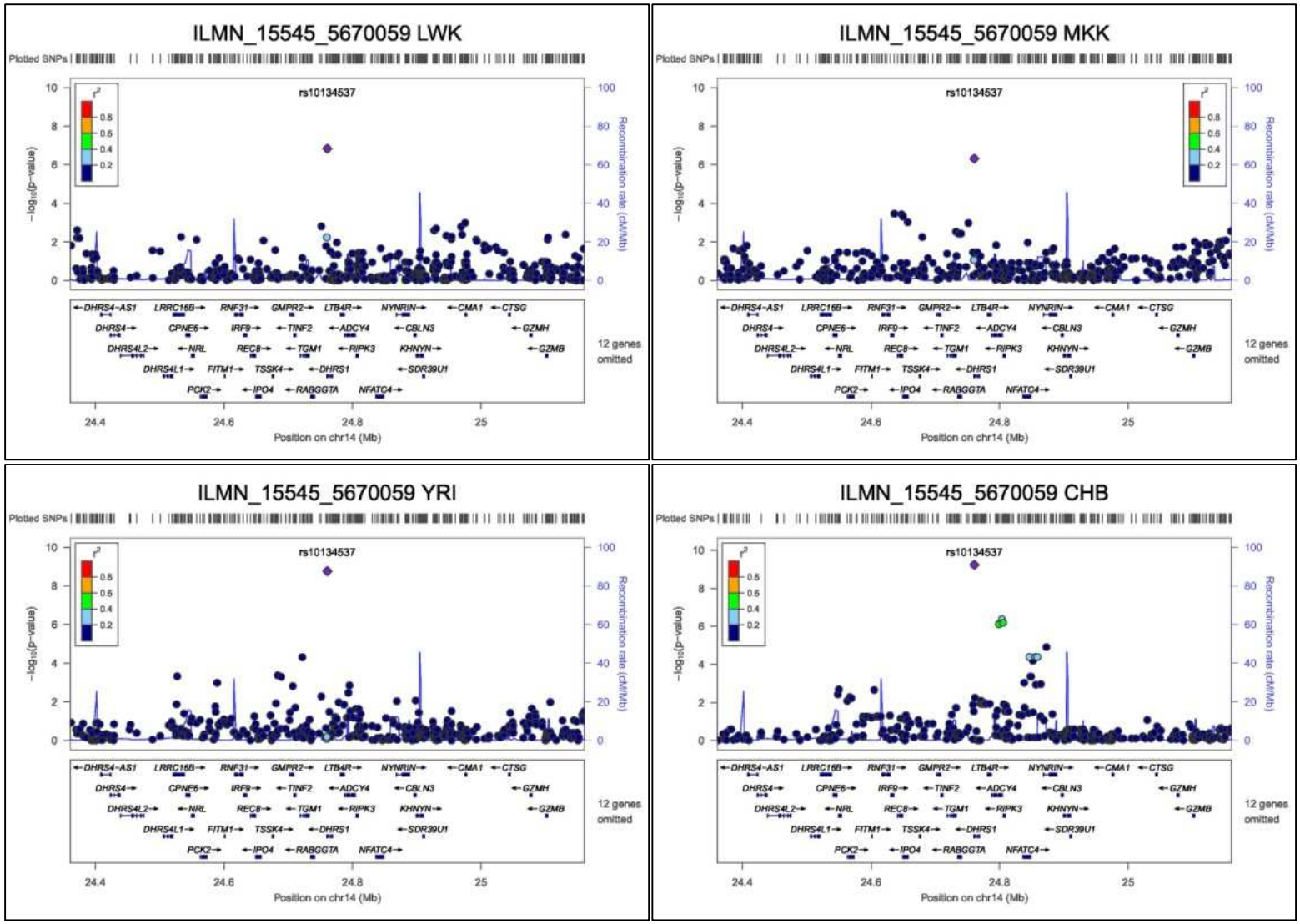


Figure 8.27: Signal plots for peak phase III HapMap SNPs for probe ILMN_15545_5670059

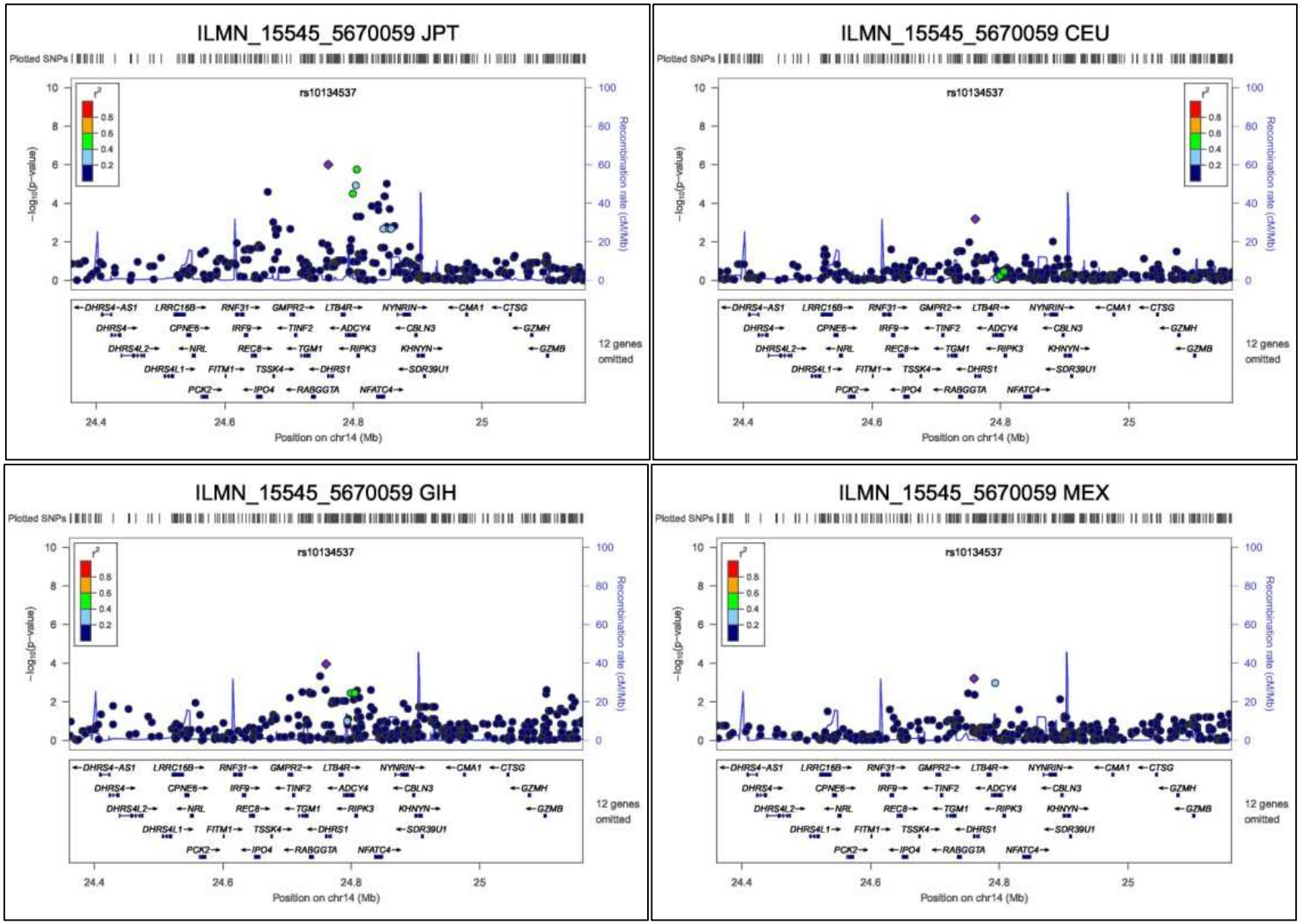


Figure 8.28: Signal plots for peak phase III HapMap SNPs for probe ILMN_15545_5670059

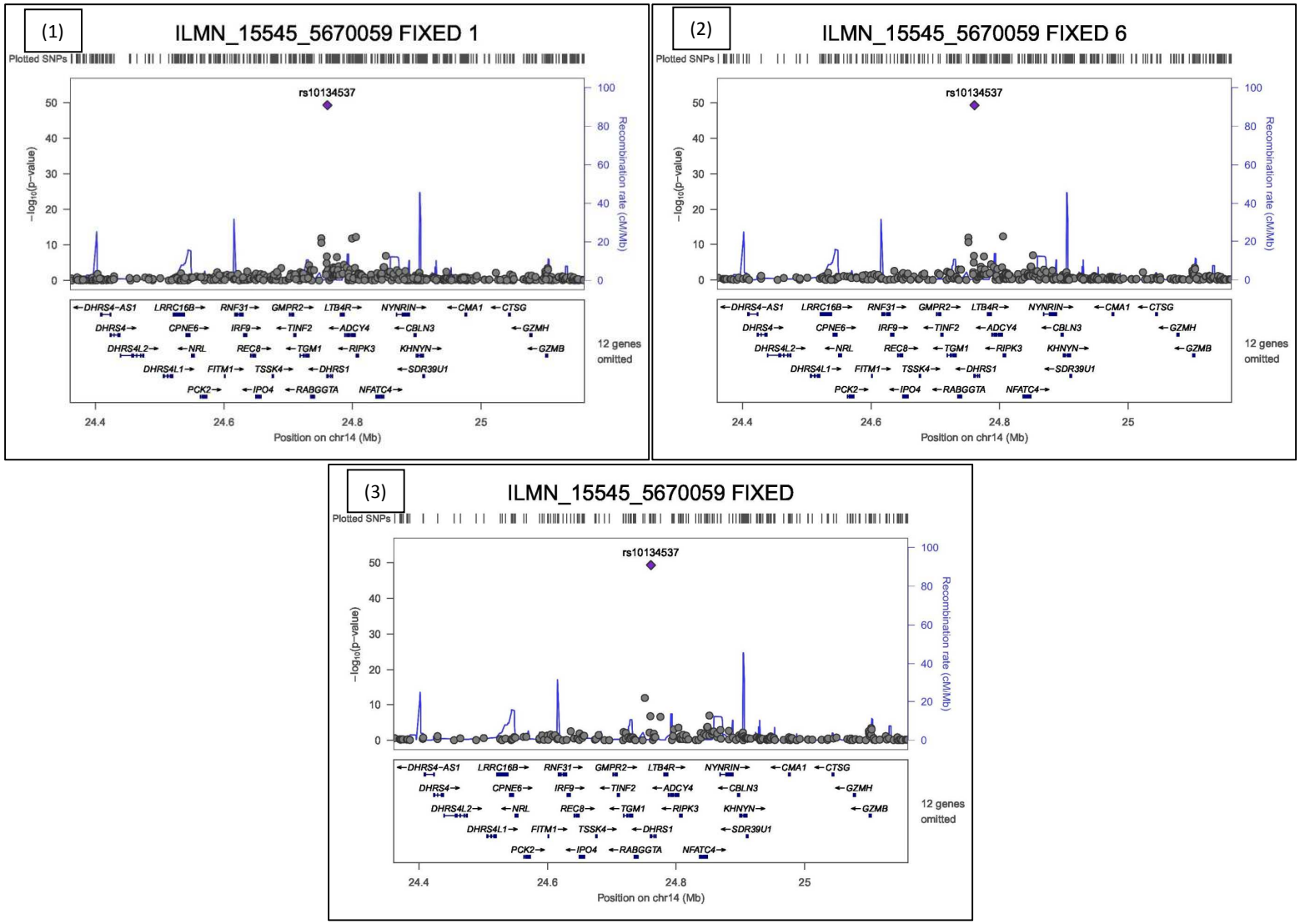


Figure 8.29: Signal plot of SNP $-\log_{10}$(p-values) from fixed effect meta-analysis for probe ILMN_15545_5670059. (1) Includes 1 or more missing variants, (2) Includes 2 or less missing variants, (3) no missing variants.

8.7 Summary

This chapter presented an analysis of the Phase III HapMap *cis* eQTLs (identified in chapter 4) using ADME gene lists. eQTL markers for ADME genes, are of particular importance as they can be used as a indicator of drug metabolism expression within each patient. The variants can be selected as an indicator of drug selection and dosage for personalized medicine.

Heterogeneity across ancestry groups / populations can be used to screen whether these eQTLs have effect sizes specific to one or more populations or whether are present at all (eSNP is monomorphic), these can be used to determine drug selection and dosage for entire ethnic groups, and can act as a starting point for investigating differences in metabolism across ethnic groups.

Heterogeneity has been analysed in a number of ADME genes, the main reason for this seems to be different LD between the tag SNP and the causal SNP, between populations. There are some examples where signals are not detected at all in some populations. Also four of the probes with significant heterogeneity have the same peak SNP: rs1454247, which is located within an intron of the gene *TMPRSS11E*, this suggest a link between these genes and the gene *TMPRSS11E*.

A *cis* eQTL (*DHRS1*) has been identified which has an ADME functional variant as the peak eSNP. This has a strong signal and is the peak signal in all eight populations. This suggests that this SNP could be the causal SNP, although this needs to be verified using 1000 Genomes Project imputed data.

CHAPTER 9 DISCUSSION

In my thesis, I have identified *cis* eQTLs in individuals from the Phase II and III HapMap datasets. Microarray expression and genotype data have been used to detect *cis* eQTLs. To do this summary statistics have been generated with association analysis and used in fixed effect and trans-ethnic Bayesian meta-analysis. The overall aims of this thesis are to: firstly improve power to detect *cis* eQTLs using meta-analysis of association analysis summary statistics; then to identify and characterize heterogeneity of summary statistics between populations; and thirdly to fine map *cis* eQTLs using LD differences between populations. In addition to these aims, an analysis has been performed to identify eSNP / dSNP pairs which are the tag SNPs for the same causal variant and identify *cis* eQTLs related to drug discovery and development, specifically those within the "absorption, distribution, metabolism, and excretion" (ADME) gene list.

Chapter 3 presented the results of an association analysis followed by fixed effect meta-analysis on the three populations (ASN, CHB and YRI) of the 210 Phase II HapMap individuals in order to identify *cis* eQTLs. The aims of this analysis were to firstly increase the power to detect eQTLs by combining the association analysis summary statistics using fixed effect meta-analysis, and secondly to determine the extent of heterogeneity detected in association analysis summary statistics between the populations. The analysis found evidence that combining association analysis summary statistics using fixed effect meta-analysis did lead to an increase in power to detect *cis* eQTLs. There was an increase in the number of *cis* eQTL probes detected with fixed effect meta-analysis at GWS (1024 SNPs reported in three populations), in comparison with those detected in association analysis at GWS (ASN: 483, CEU: 183 and YRI: 268). In addition to this, considerable heterogeneity in effect sizes was detected between the populations. A total of 106 out of 1024 probes reported had a Cochran's Q p-value < $1x10^{-3}$, more than is expected by chance. General trends of the heterogeneous probes were reported: 62 of 106 heterogeneous probes had one or more populations with no significant association signal, whilst only one of 106

probes had significant effect sizes in opposite directions. Correlation of effect sizes between populations showed that the ASN:CEU pair were the most in agreement, suggesting that much of the heterogeneity is due to differences in effect size of the YRI population, possibly due to smaller LD blocks.

In order to investigate the heterogeneity identified in the Phase II HapMap *cis* eQTLs, two examples were selected for further analysis. For the first example, *PDXDC2P*, the heterogeneity was due to a lack of any significant signal in the YRI population and possible differences in effect size between ASN and CEU populations due to differing LD with the causal SNP. Again in the second example, *USMG5* there was no significant signal in YRI and, in addition to this, the peak SNP from the fixed effect meta-analysis (when populations with one or more non-reported SNPs are included) failed QC in the YRI population. In both these examples, there is no signal in the YRI population, which could be due to the causal SNP, or a good tag SNP in the YRI population, not being present in the Phase II HapMap dataset.

Multi-ethnic meta-analysis can be used to fine map variants using differences in LD block size between populations. Two examples (*XKR9* and *SNHG17*) were selected, both of which demonstrated the use meta-analysis across populations for trans-ethnic fine mapping. Both examples had differing LD structure between populations, with the narrowest peak being in the YRI population. This chapter uncovered considerable amount of heterogeneity between populations, the examples selected show that heterogeneity is due to a lack of signal within the YRI population and differences in LD between the tag SNP and causal variant in the ASN and CEU populations. The analysis demonstrates the need for higher density genotype data to check whether lack of signal is due to causal variant (or a good tag SNP in all populations) not being present in the Phase II HapMap dataset.

The analysis in chapter 4 identified *cis* eQTLs using the 709 individuals in the Phase III HapMap dataset. The aims of the analysis were to increase power over the Phase II HapMap and improve

fine mapping resolution by including more individuals from more diverse populations. As in the Phase II HapMap analysis, evidence of an increase in power when combining association analysis results were combined using fixed effect meta-analysis was found. There was an increase in the number of *cis* eQTLs detected at GWS using fixed effect meta-analysis: A total of 1846 probes were detected in the meta-analysis with SNPs reported in all eight populations, compared with an average of 231 in population-specific association analysis. In addition to this, considerable heterogeneity was identified. Out of 1846 probes detected at GWS, 190 had significant heterogeneity at Cochran's Q p-value $\leq 1x10^{-3}$. In addition to this 140 / 190 had no association signal in at least one population and 2 / 190 had opposite effect sizes.

In order to characterize the heterogeneity in allelic effects between populations, three examples were selected for further analysis. The first example was the gene *CAV2* where heterogeneity was being caused by one or more SNPs failing QC or being monomorphic in the fixed effect meta-analysis peak SNP, and a lack of significant signal in the YRI population. The second example was the gene *FANCA* where heterogeneity was due to a lack of signals in the East Asian ancestry group (CHB, JPT) and, as before, one or more SNPs failing QC or being monomorphic at the fixed effect meta-analysis peak SNP. The final example, *USMG5,* was also reported in the Phase II HapMap dataset. Again, the reasons for heterogeneity are firstly a lack of signal in the YRI association analysis results (matching that in Phase II dataset), and SNPs failing QC or being monomorphic at the peak SNP in the analysis. The results of the *USMG5 cis* eQTL are very similar between the Phase II and Phase III HapMap results. Taking the examples investigated in this chapter, together, the heterogeneity in allelic effects between populations appears to be due to two reasons. The first is that one or more of the populations have no signal at the loci. These could be studied in more detail using higher density genotype data, such as 1000 Genomes Project reference panels, which are more likely to capture the causal variant in all populations. The second reason is due to one or more SNPs being monomorphic or failing QC at the peak SNP in the fixed effect meta-

analysis. This is an issue because peak signals with SNPs reported in all eight populations are being investigated, which means the peak SNP reported is not always the most significant.

Chapter 5 extends the analysis in chapter 4 by imputing the Phase III HapMap genotypes with the 1000 Genomes "all ancestries" March 2012 reference panel. The imputed genotypes generated are used to perform association analysis and fixed effect meta-analysis in order to identify *cis* eQTLs. This analysis has the following aims: firstly to identify the contribution of imputed genotypes to improve fine mapping of *cis* eQTLs; and secondly to investigate the effect of imputation on heterogeneity. The imputed genotypes had an intersection of 2,841,022 SNPs across the eight populations, in comparison the Phase III HapMap dataset without imputation had an intersection of 708,270 SNPs. Out of the 1811 *cis* eQTL probes detected at GWS, with SNPs reported in all eight populations, 1238 had an improvement in signal with the imputed SNPs. This indicates that 68% of the probes have an improvement in fine mapping when using imputed data. Out of these 1238 probes, 657 have a reduction in heterogeneity at the imputed SNP. This result was not entirely expected as we would think that the closer the imputed SNP is to the causal variant, the amount of heterogeneity would reduce. This could suggest that heterogeneity is not due to LD differences between populations at some *cis* eQTLs, but a result of causal variants that are specific to one population group, or which interact with factors that differ between populations. Annotation of the SNPs for the 1811 probes were investigated, approximately 50% were identified as within introns, with 25% in intergenic regions. Intronic regions suggest that factors such as transcript stability and RNA splicing could be causal. Intergenic regions suggest variants in *cis* elements, such as promoters and enhancers are causal.

Three examples from the previous chapter have been analysed using the imputed data. Very similar results were seen in these examples, but the imputed SNPs resulted in an increase of signal strength. An example was selected (*CORO2A*) where a large difference in z-score is observed between the imputed and non-imputed SNPs. Another example (*CHI3L2*) was selected

were the imputed SNP leads to a large reduction of heterogeneity when compared to the non-imputed SNP. The reduction is due to better tagging of association signals in the African populations after imputation.

In chapter 6, dSNPs from the NHGRI GWAS catalog were integrated with Phase III HapMap imputed *cis* eQTLs detected in chapter 5. This integration was performed to identify eQTL peak SNPs (eSNPs) which tagged or were the same causal variant of GWAS disease SNPs (dSNPs). A pipeline was developed to do this that uses reciprocal conditional analysis to determine whether the eSNP and dSNP pair tag the same causal variant. To do this, the results of the analysis are selected and prioritized using one of three criteria: criterion 1 selects SNP pairs which are the same SNP; criterion 2 select SNP pairs which eliminate the association signal of the other in conditional analysis; and criterion 3 selects pairs which partially remove the association signal of the other in conditional analysis, and which are also within immunological related phenotypes. For each of the criteria the following counts of dSNP – eSNP pairs were detected. For criterion 1, 12 eSNP dSNP pairs were detected out of these four had an immunological phenotype. For criterion 2, another 12 eQTLs were detected out of these six had an immunological phenotype. For criterion 3, 18 eSNP – dSNP pairs were identified. It was found that in many of these examples, the likely functional gene identified in the GWAS matches the inference from expression data in these analyses: *CARD9*, *CCDC88B*, *PRKCB*, *FCRL3*, *GPX4*, *STAT4* and *EOMES*. . However, the genes IL19 (Bechet's Disease) and ZPBP2 (cervical cancer, Primary Billiary Cirrhosis) contradicted GWAS candidate gene.

In chapter 7, the results of using a Bayesian trans-ethnic meta-analysis (MANTRA) on the imputed Phase III HapMap association summary statistics generated in chapter 5 were presented. The aim of this chapter was to determine improvements in fine mapping resolution by calculating credible sets with the results of the MANTRA analysis. Credible sets were calculated for all three ancestry groups together and separately, and provided an overview of the fine mapping resolution

differences between ancestry groups. Median credible sets were calculated for the 1811 probes detected in chapter 5 at GWS with SNPs reported in eight populations. Resolution was substantially better when all ancestry groups were used together (All – median: 7) compared to individual ancestry groups (African: 90.5, East Asian: 149, Eurasian-Hispanic: 129.5). As expected, African ancestry populations have the best resolution for fine-mapping.

To investigate the differences in resolution between the ancestry groups, several examples were selected for further investigation. Two examples were selected in which the credible set was smaller in all ancestry groups together than ancestry groups individually. The first of these *C12orf54* has an improvement in resolution due to a smaller LD block and stronger signal in the African ancestry group. The second, *GDE1,* demonstrates improvement due to a smaller LD block and stronger signal in the East Asian ancestry group. In addition, two examples were selected in which the credible set of all ancestry groups together was larger than the ancestry groups separately. The first example, *PAX8,* has a credible set of 10 with all three ancestry groups, but a credible set of 7 with the African ancestry group alone. The reason for this is that, although the African ancestry group has a narrower LD block than the East Asian and Eurasian-Hispanic, the African peak SNP is weaker than the other two ancestry groups, and so the signal is being swamped by the two other ancestry groups. The second example, *PEX6,* has a credible set of 3 for all ancestry groups, but credible sets of 1, 1 and 2 for the African, East Asian and Eurasian-Hispanic ancestry groups respectively. In this case there appears to be at least two independent signals at this locus, using all three ancestry groups detects these signals as peak, whilst the ancestry groups alone detect only a subset of these signals

The final chapter (chapter 8) presented an analysis of the Phase III HapMap *cis* eQTL dataset generated in chapter 4 to identify ADME genes with *cis* eQTLs. The analysis had the following aims: firstly to identify *cis* eQTLs which are also within the ADME gene list, then to perform enrichment analysis on this list of ADME *cis* eQTLs, and finally to review examples of these genes,

selected with significant heterogeneity or functional variants. In total out of 564 *cis* eQTL probes detected in this analysis at GWS, with reported SNPs in all eight populations 21 were probes for ADME genes. Out of these 21 probes, seven had significant heterogeneity at Cocrhan's Q p-value $\leq 1x10^{-3}$. GSK enrichment analysis using Fisher's exact test was performed on the 21 ADME probes identified. The top enrichment terms were Placental Insufficiency, Dermatitis, Occupational, Disease Progression, Hemoglobinopathies and Vitamin B6 Deficiency. Six examples of heterogeneity were selected for further analysis. The main cause of heterogeneity appears to be difference in LD between tag SNPs and causal SNPs between populations. However there are some examples where no signals are detected in one or more populations. It has also been found that four of the example probes which map to three genes (*UGT2B7, UGT2B17 and UGT2B11*) share the same peak SNP (rs1454247), which is located within an intron of the gene *TMPRSS11E*. In addition to examples of heterogeneity, a *cis* eQTL for the gene *DHRS1* has been identified which has an ADME functional variant as its peak eSNP. This has a strong signal of association, and is the peak signal in all eight populations. This suggests that this SNP could be the causal SNP, although this needs to be verified using 1000 Genomes Project imputed data.

There are a number of additional analyses that could be undertaken to extend the work in this thesis. For example, in chapter 7, 99% credible sets were determined to indicate resolution of signals. In many cases these credible sets consisted of more than one SNP. An analysis which could be carried out would be to look at overlap of annotation with credible set variants for each eQTL where the credible set is a tractable size (for example, less than ten variants). This would overcome the problem that there might be a number of SNPs in strong LD with the peak SNP for which association signals cannot be distinguished (the peak eQTL might only just be a fraction more significantly associated than others), and just focussing on annotation of this SNP may be misleading. Section 7.3.4 presented an example (*PEX6*) where multiple signals exist at the same locus, which may be shared across ancestry groups or specific to one population. This can lead to trans-ethnic credible sets that are wider than population-specific sets. In order to separate these

multiple signals, conditional analysis can first be performed to search for distinct eQTLs for each probe. Fine-mapping of each distinct signal could then be undertaken on the basis of conditioning on all other eQTLs for that probe. Also a sex-differentiated analysis could be performed to search for heterogeneity in effects between males and females. This analysis could try to determine whether sex-specific disease/trait association signals overlap with sex-differentiated eQTLs, aiding interpretation of GWAS signals.

Another future extension would be to use RNA-SEQ to generate expression data from the HapMap datasets. This would allow greater accuracy of gene expression intensity and facilitate the detection of novel transcripts, allele-specific expression and splicing QTLs. A further extension could be to look at the effect that gene-gene interactions have on eQTLs. Finally this eQTL data could be further integrated with other high throughput datasets, such as methylation QTLs, histone QTLs and DNase QTLs. In addition,some of the borderline significant results reported in this thesis could benefit from further evaluation of their significance empirically through simulation.

One of the limitations of this study is the use of LCLs as the cell type to detect eQTLs. Since LCLs exist in culture over a long period of time, they can be prone to confounding with batch effects. LCLs also have limited disease relevance, as the cells exist in culture over a long period of time, and have been transformed into an immortalized cell line. However, as LCLs have an immunological origin, they can be used to detect diseases with an immunological basis.

Another limitation of this study is that probes that contain SNPs were not removed prior to the analysis. Many of the top SNPs with very strong Mendelian significance have been found to contain SNPs. However, not all the probes contain SNPs and so some may be valid. The Mendelian probes detected in chapters 3, 4 and 5 were be contrasted with the weaker “complex” effect probes detected in the GWAS Integration analysis in chapter 6. It would be expected that

GWAS variants would map to eQTL SNPs with modest signal / effect size, as individual GWAS variants typically have modest effect size.

This dataset has many potential future uses, including integration with trans-ethnic disease fine mapping projects, where the location of an eQTL within a region can be indicative of the possible causal variant. This data could also be integrated with multiple QTLs, such as methylation, transcription binding site and protein QTLs and GWAS SNPs to produce a clearer picture of the function of eQTLs with respect to disease and the exact mechanism by which expression is altered. The data could also be added to existing eQTL databases, such as GTEX (GTEx Consortium 2013).

In summary, the results of these analyses may help in making the first steps in the interpretation of disease GWAS signals and help to fine-map causal variants by localising eQTLs. This might help explain a mechanism of regulation in the impact of the disease SNP on phenotype. These results could also help in explaining observed ethnic differences in disease association signals that might be acting through differential regulation of expression in different ancestry groups.

REFERENCES

1. 1000 Genomes Project Consortium *et al.* (2012) An integrated map of genetic variation from 1,092 human genomes. Nature. Nov 1;491(7422):56-65.

2. 1000 Genomes Project Consortium *et al*. (2010) A map of human genome variation from population-scale sequencing. Nature. Oct 28;467(7319):1061-73
3. Anderson *et al.* (2011) Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet. 2011 Mar;43(3):246-52.
4. arcOGEN Consortium *et al.* (2012) Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. Lancet. Sep 1;380(9844):815-23.
5. Australia and New Zealand Multiple Sclerosis Genetics Consortium (ANZgene). (2009) Genome-wide association study identifies new multiple sclerosis susceptibility loci on chromosomes 12 and 20. Nat Genet Jul;41(7):824-8.

6. Barrett *et al*. (2009) Genome-wide association study and meta-analysis find that over 40 loci affect risk of type 1 diabetes. Nat Genet. Jun;41(6):703-7.

7. Battle, A., et al. (2015). Impact of regulatory variation from RNA to protein. Science 347(6222): 664-667.

8. 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**, 289-300.

9. Berndt SI *et al.* (2013) Genome-wide association study identifies multiple risk loci for chronic lymphocytic leukemia. Nat Genet. Aug;45(8):868-76.
10. Bønnelykke K *et al*. (2013) Meta-analysis of genome-wide association studies identifies ten loci influencing allergic sensitization. Nat Genet. Aug;45(8):902-6.

11. Brem RB *et al*: (2002) Genetic dissection of transcriptional regulation in budding yeast. Science. Apr 26;296(5568):752-5.

12. Bystrykh, L.et al. (2005) Uncovering regulatory pathways that affect hematopoietic stem cell function using 'genetical genomics'. Nat. Genet.37, 225–232

13. Chesler, E.J.et al. (2005) Complex trait analysis of gene expression uncovers polygenic and pleiotropic networks that modulate nervous system function.Nat. Genet.37, 233–242

14. Chu *et al.* (2011) A genome-wide association study identifies two new risk loci for Graves' disease. Nat Genet. Aug 14;43(9):897-901.
15. Churchill *et al* (1994). "Empirical threshold values for quantitative trait mapping." Genetics 138(3): 963-971.
16. Clarke GM *et al* (2011) Basic statistical analysis in genetic case-control studies. Nat Protoc. 2011 Feb;6(2):121-33.

17. Dimas,A.S. *et al.* (2009) Common Regulatory Variation Impacts Gene Expression in a Cell Type Dependent Manner. *Science* **325**, 1246-1250.

18. ENCODE Project Consortium (2012) An integrated encyclopedia of DNA elements in the human genome. Nature. 6;489(7414):57-74.

19. Fairfax, B. P., et al. (2014). Innate Immune Activity Conditions the Effect of Regulatory Variants upon Monocyte Gene Expression. Science 343(6175): 1246949.
20. Fischer *et al.* (2012) A novel sarcoidosis risk locus for Europeans on chromosome 11q13.1. Am J Respir Crit Care Med. Nov 1;186(9):877-85.
21. Florez JC. *et al.* (2011) Does metformin work for everyone? A genome-wide association study for metformin response. Curr Diab Rep. 11(6):467-9.
22. Franke *et al.* (2010) Genome-wide meta-analysis increases to 71 the number of confirmed Crohn's disease susceptibility loci. Nat Genet. 42(12):1118-25.

23. GoDARTS and UKPDS Diabetes Pharmacogenetics Study Group *et al*. (2011) Common variants near *ATM* are associated with glycemic response to metformin in type 2 diabetes. Nat Genet. Feb;43(2):117-20.

24. Grundberg, E., et al. (2011). Global analysis of the impact of environmental perturbation on cis-regulation of gene expression. PLoS Genet 7(1): e1001279.

25. Grundberg E: *et al.* (2012) Multiple Tissue Human Expression Resource (MuTHER) Consortium. Mapping cis- and trans-regulatory effects across multiple tissues in twins. Nat Genet. 44(10):1084-9.

26. GTEx Consortium. (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet. 45(6):580-5.

27. Hindorff LA *et al*. A Catalog of Published Genome-Wide Association Studies. Available at:www.genome.gov/gwastudies. Accessed [05-03-2014].

28. Hinds DA *et al.* (2013) A genome-wide association meta-analysis of self-reported allergy identifies shared and allergy-specific susceptibility loci. Nat Genet. 45(8):907-11.

29. Howie BN *et al* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genetics* 5(6): e1000529

30. Hou S *et al.* (2012) Identification of a susceptibility locus in *STAT4* for Behçet's disease in Han Chinese in a genome-wide association study. Arthritis Rheum. Dec;64(12):4104-13.

31. Huang da, W., et al. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4(1): 44-57.

32. Hubner, N. et al. (2005) Integrated transcriptional profiling and linkage analysis for identification of genes underlying disease. Nat. Genet.37, 243–253

33. International HapMap Consortium *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature. Oct 18; 449(7164):851-61.

34. International HapMap 3 Consortium *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. Nature. Sep 2;467(7311):52-8.

35. International Multiple Sclerosis Genetics Consortium *et al*. (2011) Genetic risk and a primary role for cell-mediated immune mechanisms in multiple sclerosis. Nature. 10;476(7359):214-9.

36. Johnson, A. D. *et al*. (2008) SNAP: A web-based tool for identification and annotation of proxy SNPs using HapMap Bioinformatics, 24(24):2938-2939

37. Jostins *et* al. (2012) Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature. 1;491(7422):119-24.

38. Kirino Y *et al*. (2013) Genome-wide association analysis identifies new susceptibility loci for Behçet's disease and epistasis between HLA-B*51 and ERAP1. Nat Genet. 45(2):202-7.

39. Lappalainen T *et al.* (2013) Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 26;501(7468):506-11.

40. Lauc G *et al.* (2013) Loci associated with N-glycosylation of human immunoglobulin G show pleiotropy with autoimmune diseases and haematological cancers. PLoS Genet. ;9(1)

41. Liu *et al.* (2010)Genome-wide meta-analyses identify three loci associated with primary biliary cirrhosis. Nat Genet. 42(8):658-60.

42. Magi R *et al* (2010): GWAMA: software for genome-wide association meta-analysis. BMC Bioinformatics, 11:288

43. Marchini J. *et al* (2007) A new multipoint method for genome-wide association studies via imputation of genotypes. Nature Genetics 39 : 906-913

44. McGovern *et al*. (2010) Genome-wide association identifies multiple ulcerative colitis susceptibility loci. Nat Genet. 42(4):332-7.

45. Mi, H., et al. (2013). PANTHER in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Res 41(Database issue): D377-386.

46. Montgomery,S.B. *et al.* (2010) Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature* **464**, 773-777 .

47. Morris AP. (2011) Transethnic meta-analysis of genomewide association studies. Genet Epidemiol. 35(8):809-22

48. Morris AP *et al* (2012).Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet. 44(9):981-90.

49. Myers,A.J. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat Genet* **39**, 1494-1499.

50. Nakamura *et al*. (2012) Genome-wide association study identifies TNFSF15 and POU2AF1 as susceptibility loci for primary biliary cirrhosis in the Japanese population. Am J Hum Genet. 5;91(4):721-8.

51. Nica,A.C. *et al.* (2011) The Architecture of Gene Regulatory Variation across Multiple Human Tissues: The MuTHER Study. *PLoS Genet* **7**, e1002003 .

52. Patsopoulos *et al*. (2011) Genome-wide meta-analysis identifies novel multiple sclerosis susceptibility loci. Ann Neurol. 70(6):897-912.

53. Pickrell,J.K. *et al.* (2010) Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768-772 .

54. Price AL *et al*. (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet. 38(8):904-9

55. Purcell S *et al* (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. American Journal of Human Genetics, 81.

56. Remmers *et al*. (2010) Genome-wide association study identifies variants in the MHC class I, IL10, and IL23R-IL12RB2 regions associated with Behçet's disease. Nat Genet. 42(8):698-702.

57. Schadt, E.E.et al. (2003) Genetics of gene expression surveyed in maize, mouse and man.Nature 422, 297–302

58. Schadt E.E. *et al:* (2008) Mapping the Genetic Architecture of Gene Expression in Human Liver. *PLoS Biol* **6**, e107.

59. Shi *et al.* (2013) A genome-wide association study identifies two new cervical cancer susceptibility loci at 4q12 and 17q12. Nat Genet. 45(8):918-22.

60. Slager SL *et al*. (2012) Common variation at 6p21.31 (*BAK1*) influences the risk of chronic lymphocytic leukemia. Blood. 26;120(4):843-6.

61. Small, K. S., et al. (2011). Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. Nat Genet 43(6): 561-564.

62. Storey,J.D. & Tibshirani,R. (2003) Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences* **100**, 9440-9445 .

63. Stranger BE *et al:* (2012) Patterns of cis regulatory variation in diverse human populations. PLoS Genet.;8(4).

64. Stranger BE *et al*: (2007)Population genomics of human gene expression. Nat Genet. ;39(10):1217-24.

65. Veyrieras,J.B. *et al.* (2008) High-Resolution Mapping of Expression-QTLs Yields Insight into Human Gene Regulation. *PLoS Genet* **4**, e1000214 .

66. Wang K *et al*. (2010) ANNOVAR: Functional annotation of genetic variants from next-generation sequencing data Nucleic Acids Research, 38:e164.
67. Westra, H.-J., et al. (2013). Systematic identification of trans eQTLs as putative drivers of known disease associations. Nat Genet 45(10): 1238-1243.
68. Winkler TW *et al*. (2014) Quality control and conduct of genome-wide association meta-analyses. Nat Protoc. 9(5):1192-212.
69. Yvert G. *et al.* (2003) Trans-acting regulatory variation in Saccharomyces cerevisiae and the role of transcription factors. Nat Genet. 35(1):57-64.
70. Zeller,T. *et al.* (2010) Genetics and Beyond: The Transcriptome of Human Monocytes and Disease Susceptibility. *PLoS ONE* **5**, e10693.

APPENDIX

Probe	HGNC	SNP	Allele Aligned	Beta	SE	Minor Allele	Allele Frequencies
				(ASN, CEU, YRI)			
GI_21553312-S	*CHURC1*	rs4899152	A	1.23, 1.20, 1.03	0.048, 0.065, 0.046	A, A, A	0.14,0.19,0.43
GI_4504184-S	*GSTT1*	rs407257	C	1.64, -1.28, 1.08	0.049, 0.089, 0.193	C, G, C	0.35,0.61,0.28
GI_4507820-S	*UGT2B17*	rs3100f645	T	2.77, -1.59, -1.09	0.094, 0.163, 0.200	T, G, G	0.15,0.68,0.80
GI_4507822-S	*UGT2B11*	rs3100645	T	2.32, -1.36, -1.06	0.084, 0.135, 0.177	T, G, G	0.15,0.68,0.80
GI_21536345-A	*TCL6*	rs2296310	A	0.86, 0.53, 0.95	0.039, 0.056, 0.056	A, A, A	0.26,0.12,0.24
GI_39780596-S	*PNMAL1*	rs8107491	G	1.09, 0.79, 0.51	0.039, 0.145, 0.095	G, G, G	0.14,0.31,0.38
GI_31377831-S	*PKHD1L1*	rs2607624	C	0.75, 0.59, 0.57	0.030, 0.093, 0.047	C, C, C	0.33,0.15,0.38
GI_31377759-S	*ORMDL1*	rs6942	C	0.60, 0.61, -0.50	0.028, 0.035, 0.074	C, C, G	0.26,0.26,0.87
GI_38348363-S	*MXRA7*	rs2240773	A	1.14, 1.03, 0.96	0.047, 0.085, 0.148	A, A, A	0.14,0.09,0.22
GI_4507824-S	*UGT2B7*	rs3100645	T	2.14, -1.21, -0.99	0.083, 0.126, 0.169	T, G, G	0.15,0.68,0.80

Table A.1: Ten *cis* eQTLs with largest absolute z-score. Association summary statistics are presented See chapter 3 section 3.5.4

Probe	HGNC	SNP	Allele Aligned	Beta	SE	Minor Allele	Allele Frequencies
				(ASN, CEU, YRI)			
GI_31377759-S	*ORMDL1*	rs6942	C	0.60, 0.61, -0.50	0.028, 0.035, 0.074	C, C, G	0.26,0.26,0.87
GI_38348363-S	*MXRA7*	rs2240773	A	1.14, 1.03, 0.96	0.047, 0.085, 0.148	A, A, A	0.14,0.09,0.22
GI_18426974-S	*HLA-DQA1*	rs9272346	G	4.68, 4.44, 4.59	0.273, 0.303, 0.309	G, G, G	0.42,0.45,0.42
GI_28872737-I	*SEMA4G*	rs2863095	T	0.28, 0.28, 0.28	0.017, 0.021, 0.017	T, T, T	0.49,0.19,0.48
GI_28872735-A	*MRPL43*	rs2863095	T	0.50, 0.48, 0.45	0.031, 0.046, 0.027	T, T, T	0.49,0.19,0.48
GI_20070185-S	*PTER*	rs1055340	T	-0.79, -0.65, 0.80	0.045, 0.072, 0.055	T, T, A	0.44,0.46,0.79
GI_15082255-I	*LCMT1*	rs277886	C	0.34, -0.43, -0.37	0.017, 0.040, 0.038	C, T, T	0.46,0.80,0.52
GI_22265329-I	*NUDT2*	rs10972063	A	-0.69, -0.57, -0.47	0.031, 0.081, 0.117	A, A, A	0.43,0.12,0.10
GI_42659736-S	*XPRA1*	rs2304683	G	-0.24, 0.26, 0.29	0.016, 0.022, 0.037	G, A, A	0.46,0.68,0.92
GI_29735811-S	*XKR9*	rs6998786	G	0.19, -0.16, 0.17	0.015, 0.015, 0.015	G, A, G	0.11,0.53,0.15

Table A.2: Ten *cis* eQTLs with largest absolute z-score and no heterogeneity. Association summary statistics are presented See chapter 3 section 3.5.4

Probe	HGNC	SNP	Allele (Aligned)	Beta	Standard Error	Minor Allele	Allele Frequencies
				(ASN, CEU, YRI)			
GI_40217627-S	*PDXDC2P*	rs6499292	G	1.14, -0.87, 0.03	0.047, 0.058, 0.041	G, A, A	0.18,0.66,0.61
GI_14249375-S	*USMG5*	rs7831	C	1.34, 0.94, 0.07	0.053, 0.157, 0.081	C, C, C	0.31,0.39,0.24
GI_38176290-I	*CAV2*	rs12670840	T	0.38, 0.24, 0.01	0.028, 0.075, 0.019	T, T, T	0.21,0.28,0.17
GI_38176291-A	*CAV2*	rs12670840	T	0.45, 0.28, 0.01	0.031, 0.086, 0.023	T, T, T	0.21,0.28,0.17
GI_4507400-I (*)	*TFAM*	rs10826176	A	-0.56, 0.16, -0.01	0.047, 0.071, 0.044	A, A, G	0.26,0.22,0.57
GI_4507820-S	*UGT2B17*	rs3100645	T	2.77, -1.59, -1.09	0.094, 0.163, 0.200	T, G, G	0.15,0.68,0.80
GI_33859747-S	*LOC65996*	rs7249714	G	-0.31, -0.3 -0.04	0.023, 0.023, 0.030	G, G, G	0.25,0.39,0.34
GI_16306561-S	*RPL37A*	rs284565	G	-0.45, -0.04, -0.26	0.039, 0.028, 0.044	G, G, G	0.17,0.24,0.28
GI_5729867-S	*HERC2*	rs12593929	G	0.06, -0.04, 0.47	0.047, 0.117, 0.045	G, G, A	0.46,0.07,0.69
GI_4507822-S	*UGT2B11*	rs3100645	T	2.32, -1.36, -1.06	0.084, 0.134, 0.177	T, G, G	0.15,0.68,0.80

Table A.3: Ten *cis* eQTLs with most significant heterogeneity. Association summary statistics are presented See chapter 3 section 3.5.4

Probe	HGNC	SNP	Beta	SE	Alleles	Allele Frequencies
			(LWK,MKK,YRI,CHB,JPT,CEU,GIH,MEX)			
ILMN_24844_2100600	*IRF5*	rs6965542	(1.22,1.01,1.17, 1.39, 1.13, 1.03,1.15,1.12)	(0.075,0.071,0.068, 0.079,0.070, 0.084,0.086,0.084)	(C,C,C, T,T, C,C,C)	C:0.47,0.49,0.49, 0.51,0.57, 0.46,0.39,0.46
ILMN_22173_5690095	*ERAP2*	rs2548533	(-1.57,-1.39,-1.81, -1.88,-1.64, -1.66,-1.95,-1.82)	(0.106,0.119,0.110, 0.114,0.091, 0.123,0.133,0.190)	(G,G,A, A,A, A,A,A)	G:0.44,0.41,0.61, 0.61,0.52, 0.51,0.56,0.54
ILMN_27265_6330037 (*)	*CHURC1*	rs10141986	(1.29,1.35,1.20, 1.39,1.43, 1.43,1.44,1.42)	(0.073,0.080,0.090, 0.124,0.152, 0.107,0.108,0.140)	(G,G,G, G,G, G,G,G)	G:0.40,0.43,0.45, 0.16,0.11, 0.24,0.27,0.24
ILMN_2170_6860347	*WBSCR27*	rs13228435	(-0.79,-0.76,-0.71, -0.71,-0.68, -0.81,-0.64,-0.63)	(0.058,0.058,0.053, 0.044,0.051, 0.065,0.055,0.059)	(C,T,C, T,T, T,T,T)	C:0.45,0.66,0.47, 0.54,0.60, 0.79,0.59,0.57
ILMN_7731_3520685	*C17orf97*	rs11150882	(-0.75,-0.61,-0.79, -1.03,-1.06, -1.11,-0.92,-1.03)	(0.069,0.050,0.051, 0.083,0.066, 0.085,0.071,0.136)	(G,G,G, A,A, A,A,A)	G:0.34,0.36,0.41, 0.68,0.67, 0.79,0.65,0.73
ILMN_7731_2570703	*C17orf97*	rs11150882	(-1.10,-0.90,-1.12, -1.51,-1.57, -1.60,-1.40,-1.48)	(0.093,0.070,0.075, 0.123,0.095, 0.121,0.118,0.206)	(G,G,G, A,A, A,A,A)	G:0.34,0.36,0.41, 0.68,0.67, 0.79,0.65,0.73
ILMN_20550_7330093	*HLA-DRB1*	rs9271170	(3.91,3.92,3.89, 3.28,3.34, 3.52,3.59,4.04)	(0.292,0.390,0.309, 0.309,0.264, 0.213,0.270,0.497)	(T,T,T, T,T, T,T,T)	T:0.25,0.13,0.22, 0.18,0.25, 0.31,0.30,0.23
ILMN_15237_5860053	*IPO8*	rs11834524	(0.73,0.53,0.71, 0.66,0.58, 0.70,0.63,0.45)	(0.052,0.048,0.065, 0.042,0.042, 0.050,0.053,0.084)	(G,G,G, A,A, A,A,A)	G:0.22,0.27,0.19, 0.57,0.52, 0.52,0.61,0.83
ILMN_22465_2350368	*PTER*	rs7909832	(-0.66,-0.52,-0.50, -1.11,-1.12, -1.02,-1.08,-0.93)	(0.078,0.081,0.079, 0.072,0.063, 0.070,0.070,0.098)	(T,T,T, C,C, T,T,T)	T:0.31,0.43,0.33, 0.54,0.55, 0.47,0.43,0.30
ILMN_33941_1570382 (*)	-	rs9945924	(-1.09,-0.81,-1.15, -0.98,-0.89, -0.93,-0.92,-0.96)	(0.084,0.074,0.069, 0.082,0.084, 0.096,0.097,0.114)	(A,A,A, A,A, A,A,A)	A:0.38,0.35,0.45, 0.22,0.19, 0.23,0.22,0.34

Table A.4: Ten probes with largest absolute z-score from the Phase III HapMap fixed effect meta-analysis. (*) deprecated in NCBI Build 37. Association summary statistics are presented.

See chapter 4, section 4.6.3

Probe	HGNC	SNP	Beta	SE	Minor Allele	Allele Frequencies
			(LWK,MKK,YRI,CHB,JPT,CEU,GIH,MEX)			
ILMN_5108_2030195	*CAV2*	rs17138749	0.15,0.19,0.00, 0.79,1.04, 0.54,0.39,0.74	0.121,0.049,0.032, 0.083,0.080, 0.128,0.104,0.177	C,C,C, C,C, C,C,C	C:0.25,0.13,0.17, 0.16,0.23, 0.17,0.12,0.23
ILMN_5108_2260136	*CAV2*	rs17138749	0.05,0.11,0.01, 0.75,0.96, 0.51,0.35,0.69	0.123,0.047,0.034, 0.080,0.079, 0.114,0.100,0.170	C,C,C, C,C, C,C,C	C:0.25,0.13,0.17, 0.16,0.23, 0.17,0.12,0.23
ILMN_128814_5960180	TMED4	rs217378	0.11,0.10,0.14, 0.79,0.93, 0.09,0.19,0.11	0.082,0.071,0.085, 0.067,0.066, 0.053,0.086,0.104	A,A,A, G,G, A,G,A	A:0.40,0.37,0.34, 0.57,0.53, 0.45,0.55,0.45
ILMN_1950_6220333	MYRFL	rs4512898	0.02,0.11,0.01, 0.34,0.31, 0.09,0.14,0.09	0.015,0.021,0.015, 0.029,0.033, 0.027,0.024,0.023	A,A,A, A,A, A,A,A	A:0.22,0.27,0.23, 0.41,0.47, 0.06,0.20,0.22
ILMN_30273_4860327	UTS2	rs2493214	-0.14,-0.13,-0.15, -0.63,-0.69, -0.01,-0.02,-0.54	0.056,0.047,0.065, 0.079,0.081, 0.022,0.062,0.090	A,A,A, A,A, G,G,G	A:0.28,0.45,0.20, 0.36,0.45, 0.81,0.84,0.58
ILMN_4793_2490209	RP11-153K16.2	rs6435069	0.50,0.55,0.51, 0.01,-0.03, 0.10,-0.01,0.16	0.068,0.063,0.084, 0.044,0.046, 0.061,0.046,0.069	G,G,G, A,A, A,A,A	G:0.26,0.30,0.24, 0.82,0.87, 0.54,0.70,0.50
ILMN_28044_4830575	UTS2	rs2493214	-0.23,-0.42,-0.36, -1.40,-1.38, 0.03,-0.17,-1.49	0.173,0.124,0.219, 0.175,0.209, 0.056,0.157,0.254	A,A,A, A,A, G,G,G	A:0.28,0.45,0.20, 0.36,0.45, 0.81,0.84,0.58
ILMN_4387_7210035	FAM154B	rs1877242	0.37,0.25,0.32, -0.01,0.05, 0.03,0.03,0.15	0.063,0.034,0.046, 0.019,0.026, 0.030,0.028,0.088	C,C,C, C,C, C,C,C	C:0.28,0.28,0.19, 0.24,0.17, 0.30,0.40,0.38
ILMN_17555_5560494	ZP3	rs1754183	0.31,0.39,0.06, 0.65,0.01, 1.18,0.88,0.60	0.125,0.089,0.102, 0.118,0.111, 0.105,0.111,0.196	G,G,G, G,G, G,G,G	G:0.29,0.35,0.46, 0.29,0.12, 0.36,0.32,0.25
ILMN_13285_4210136	*FANCA*	rs2239360	0.36,0.29,0.24, 0.03,0.06, 0.40,0.35,0.43	0.035,0.030,0.050, 0.045,0.039, 0.037,0.050,0.043	C,C,C, C,C, T,C,C	C:0.35,0.38,0.21, 0.22,0.13, 0.64,0.37,0.42

Table A.5: Probes with most significant heterogeneity detected using the Cochran's Q statistic with association summary statistics. See chapter 4 section 4.6.5

Probe	HGNC	SNP	Allele Aligned	Beta	SE	Allele 1	Allele Frequencies
ILMN_24844_2100600	*IRF5*	rs10954213	A	-1.25,1.05,-1.17, -1.39,-1.13, -1.03,1.15,-1.12	0.073,0.069,0.068, 0.079,0.070, 0.084,0.087,0.084	A,G,A, A,A, A,G,A	0.52,0.49,0.49, 0.51,0.37, 0.54,0.57,0.55,
ILMN_22173_5690095	*ERAP2*	rs2549782	G	1.58,1.39,1.81, 1.91,1.64, 1.66,1.95,1.82	0.106,0.119,0.110, 0.113,0.092, 0.123,0.134,0.190	G,G,G, G,G, G,G,G	0.54,0.58,0.41, 0.39,0.47, 0.49,0.47,0.48,
ILMN_2170_6860347	*WBSCR27*	rs10225983	A	-0.79,0.76,-0.73, -0.71,-0.68, -0.81,0.64,-0.63	0.058,0.057,0.053, 0.045,0.051, 0.065,0.055,0.059	A,G,A, A,A, A,G,A	0.47,0.67,0.47, 0.55,0.56, 0.79,0.54,0.52,
ILMN_7731_3520685	*C17orf97*	rs7502594	C	0.75,-0.62,0.78, 1.03,1.06, 1.09,-0.92,1.03	0.063,0.053,0.053, 0.083,0.066, 0.081,0.071,0.136	C,T,C, C,C, C,T,C	0.63,0.60,0.58, 0.31,0.34, 0.21,0.33,0.22,
ILMN_7731_2570703	*C17orf97*	rs7502594	C	1.10,-0.91,1.11, 1.51,1.57, 1.57,-1.40,1.48	0.086,0.075,0.076, 0.123,0.095, 0.115,0.118,0.206	C,T,C, C,C, C,T,C	0.63,0.60,0.58, 0.31,0.34, 0.21,0.33,0.22,
ILMN_15237_5860053	*IPO8*	rs28661513	C	-0.73,0.53,-0.71, -0.66,-0.58, -0.70,0.63,-0.45	0.052,0.048,0.065, 0.042,0.042, 0.050,0.053,0.084	C,T,C, C,C, C,T,C	0.75,0.72,0.80, 0.43,0.46, 0.48,0.37,0.15,
ILMN_20550_7330093	*HLA-DRB1*	rs9271073	G	3.88,3.91,3.93, 3.28,3.34, 3.53,3.59,3.78	0.296,0.390,0.328, 0.310,0.264, 0.213,0.270,0.432	G,G,G, G,G, G,G,G	0.24,0.11,0.21, 0.16,0.27, 0.31,0.31,0.22,
ILMN_25790_1110273	*SLFN5*	rs11080327	A	0.75,-0.83,0.64, 0.67,0.83, 0.82,-1.00,0.75	0.054,0.058,0.061, 0.062,0.059, 0.071,0.085,0.083	A,G,A, A,A, A,G,A	0.60,0.59,0.67, 0.63,0.43, 0.48,0.47,0.33,
ILMN_3178_4390692	*HLA-DRB5*	rs115466555	A	-4.61,3.86,-4.22, -4.25,-4.08, -4.83,4.56,-4.72	0.396,0.475,0.381, 0.445,0.352, 0.333,0.333,0.723	A,C,A, A,A, A,C,A	0.82,0.95,0.84, 0.87,0.78, 0.80,0.73,0.89,
ILMN_18858_160242	*PKHD1L1*	rs1783166	A	-0.48,-0.39,-0.55, -0.76,-0.71, -0.55,-0.52,-0.51	0.055,0.059,0.057, 0.049,0.045, 0.057,0.065,0.067	A,A,A, A,A, A,A,A	0.54,0.53,0.46, 0.65,0.65, 0.61,0.39,0.35,

Table A.6: Ten probes from fixed effect meta-analysis with largest absolute Z score in 1000 Genomes. with association summary statistics. See chapter 5 section 5.5.4

Probe	HGNC	SNP	Allele Aligned	Beta	SE	Allele 1	Allele Frequencies
ILMN_5108_2030195	*CAV2*	rs13235183	G	-0.24,-0.26,-0.03, -0.79,-1.01, -1.31,-0.82,-1.09	0.139,0.063,0.053, 0.083,0.089, 0.146,0.121,0.209	G,G,G, G,G, G,G,G	0.84,0.94,0.93, 0.85,0.79, 0.92,0.93,0.91,
ILMN_128814_5960180	*TMED4*	rs217378	A	0.11,0.10,0.14, 0.79,0.93, 0.09,0.19,0.11	0.082,0.072,0.086, 0.067,0.066, 0.053,0.086,0.104	A,A,A, A,A, A,A,A	0.39,0.38,0.33, 0.60,0.50, 0.44,0.54,0.48,
ILMN_5108_2260136	*CAV2*	rs13235183	G	-0.22,-0.19,-0.04, -0.75,-0.94, -1.17,-0.78,-1.03	0.140,0.060,0.055, 0.080,0.087, 0.131,0.114,0.200	G,G,G, G,G, G,G,G	0.84,0.94,0.93, 0.85,0.79, 0.92,0.93,0.91,
ILMN_4387_7210035	*FAM154B*	rs11635460	A	0.37,-0.25,0.32, -0.01,0.06, 0.04,-0.03,0.16	0.063,0.034,0.046, 0.019,0.026, 0.030,0.028,0.086	A,G,A, A,A, A,G,A	0.25,0.28,0.20, 0.23,0.20, 0.30,0.38,0.37,
ILMN_13285_4210136	*FANCA*	rs8051231	C	-0.31,0.31,-0.25, -0.03,-0.06, -0.40,0.38,-0.43	0.036,0.028,0.044, 0.042,0.037, 0.037,0.046,0.043	C,G,C, C,C, C,G,C	0.58,0.66,0.71, 0.75,0.88, 0.35,0.67,0.57,
ILMN_2228_6250228	*STOX1*	rs4746792	C	-0.01,-0.01,-0.04, -0.11,-0.14, -0.18,-0.14,-0.21	0.011,0.014,0.012, 0.025,0.028, 0.025,0.025,0.035	C,C,C, C,C, C,C,C	0.61,0.81,0.64, 0.91,0.88, 0.83,0.78,0.74,
ILMN_27741_6650402	*SERF2*	rs2467402	A	0.14,-0.01,0.09, 0.22,0.19, 0.11,-0.09,0.19	0.018,0.014,0.019, 0.022,0.036, 0.043,0.027,0.024	A,C,A, A,A, A,C,A	0.44,0.45,0.39, 0.31,0.26, 0.06,0.16,0.23,
ILMN_2021_2350243	*UGT2B17*	rs139440909	G	1.79,1.92,0.57, 3.04,3.21, 1.95,2.61,1.73	0.223,0.193,0.178, 0.297,0.334, 0.232,0.413,0.281	G,G,G, G,G, G,G,G	0.55,0.44,0.52, 0.16,0.12, 0.47,0.31,0.29,
ILMN_22479_2900379	*UGT2B7*	rs139440909	G	1.60,1.64,0.49, 2.58,2.70, 1.67,2.35,1.66	0.196,0.169,0.158, 0.250,0.281, 0.193,0.336,0.267	G,G,G, G,G, G,G,G	0.55,0.44,0.52, 0.16,0.12, 0.47,0.31,0.29,
ILMN_21130_5090184	*HEBP2*	rs6919872	A	1.93,1.65,1.86, 0.82,1.02, 0.86,0.78,1.20	0.125,0.114,0.130, 0.160,0.141, 0.172,0.141,0.201	A,A,A, A,A, A,A,A	0.51,0.57,0.47, 0.91,0.86, 0.94,0.91,0.85,

Table A.7: Ten probes with greatest heterogeneity (largest Cochran's Q-statistic) for imputed SNPs with association summary statistics. See chapter 5, section 5.5.6

Probe	HGNC	SNP	Beta	SE	Allele 1	Allele Frequencies
ILMN_138375_7650093	*UGT2B7*	rs1454247	(0.76,1.15,0.55, 1.84,1.71, 1.44,1.66,0.91)	(0.257,0.179,0.155, 0.297,0.310, 0.180,0.257,0.302)	(C,C,C, C,C, T,C,T)	C:0.44,0.47,0.36, 0.24,0.21, 0.54,0.43,0.56,
ILMN_22479_2900379	*UGT2B7*	rs1454247	(0.75,1.10,0.57, 1.79,1.57, 1.41,1.64,0.98)	(0.248,0.176,0.151, 0.286,0.304, 0.176,0.252,0.297)	(C,C,C, C,C, T,C,T)	C:0.44,0.47,0.36, 0.24,0.21, 0.54,0.43,0.56
ILMN_2021_2350243	*UGT2B17*	rs1454247	(0.86,1.37,0.63, 1.56,2.14, 1.71,1.82,0.97)	(0.279,0.198,0.171, 0.375,0.341, 0.206,0.308,0.316)	(C,C,C, C,C, T,C,T)	C:0.44,0.47,0.36, 0.24,0.21, 0.54,0.43,0.56
ILMN_5225_7050768	*UGT2B11*	rs1454247	(0.53,0.94,0.45, 1.45,1.39, 1.21,1.33,0.77)	(0.199,0.152,0.144, 0.242,0.242, 0.155,0.205,0.256)	(C,C,C, C,C, T,C,T)	C:0.44,0.47,0.36, 0.24,0.21, 0.54,0.43,0.56
ILMN_16478_1710170	*SLC7A7*	rs12884337	(0.35,0.39,0.20, 0.80,0.80, 0.49,0.59,0.52)	(0.158,0.093,0.137, 0.102,0.087, 0.087,0.085,0.067)	(C,C,C, C,C, T,T,T)	C:0.21,0.36,0.26, 0.31,0.43, 0.56,0.59,0.61,
ILMN_15891_6480091	*GSTM1*	rs12745189	(-0.04,-0.57,-0.29, -0.84,-0.95, -0.08,-1.01,-0.84)	(0.182,0.158,0.176, 0.160,0.187, 0.195,0.210,0.313)	(T,T,T, C,C, T,T,C)	T:0.44,0.47,0.37, 0.59,0.68, 0.41,0.46,0.70,
ILMN_18916_4850209	*DHRS4L2*	rs8022613	(0.25,0.09,0.10, 0.50,0.29, 0.11,0.13,-0.25)	(0.094,0.056,0.101, 0.097,0.077, 0.090,0.127,0.138)	(T,T,T, C,C, C,C,C)	T:0.22,0.33,0.15, 0.87,0.89, 0.89,0.89,0.95,

Table A.8: Table of ADME heterogeneous probes, with association summary statistics see chapter 8, section 8.5