# *Multiple trait genome-wide association studies: Applications and methods*

Marissa Erin LeBlanc

Dissertation presented for the degree of Philosophiae Doctor (PhD)


Department of Clinical Molecular Biology, Institute of Clinical Medicine
and
Oslo Centre of Biostatistics and Epidemiology

UNIVERSITY of OSLO

Oslo, February 2016

*Series of dissertations submitted to the*
*Faculty of Medicine, University of Oslo*

ISBN 978-82-8333-241-4



Cover: Hanne Baadsgaard Utigard
Printed in Norway: 07 Media AS – www.07.no

**Acknowledgements**

A degree in (bio)statistics has been a dream of mine since completing my MSc in Genetics in 2003. I would like to express my deepest gratitude to my statistical supervisor Dr. Bettina Kulle Andreassen and co-supervisor Professor Arnoldo Frigessi for taking a chance and hiring me, a non-traditional candidate for a PhD in Biostatistics. I would like to thank my clinical co-supervisor Dr. Ole A. Andreassen for providing the opportunity to collaborate on projects involving interesting applications of statistical genomics. To all three of my supervisors, thank you for your support, patience and for interesting discussions. I look forward to continuing to collaborate with all of you in the future.

I would like to thank Norway and the University of Oslo for providing an environment where one can simultaneously earn a doctorate degree, have a young family and earn a reasonable salary while doing so.

I would like to thank my co-authors and my colleagues at the Oslo Centre of Biostatistics and Epidemiology, both at the University of Oslo and at Oslo University Hospital, for contributing to a positive, stimulating and motivating work environment. Christian Page and Dr. Verena Zuber deserve particular thanks. I hope to continue working directly and indirectly with all of you for many years to come.

To my family, thank you for your patience. This has been a long time coming. Rasmus, Alma, Leona and Esben, this is for all of us.

**List of papers**

*Paper 1*
LeBlanc, M., Kulle, B., Sundet, K., Agartz, I., Melle, I., Djurovic, S., Frigessi, A. and Andreassen, O.A., 2012. Genome-wide study identifies *PTPRO* and *WDR72* and *FOXQ1-SUMO1P1* interaction associated with neurocognitive function. *Journal of Psychiatric Research*, *46*(2), pp.271-278.

*Paper 2*
LeBlanc, M., Zuber, V., Andreassen, B.K., Witoelar, A., Zeng, L., Bettella, F., Wang, Y., McEvoy, L.K., Thompson, W.K., Schork, A.J., Reppe, S., Barrett-Connor, E., Ligthart, S., Dehghan, A., Gautvik, K.M., Nelson, C.P., Schunkert, H., Samani, N.J., CARDIoGRAM Consortium, Ridker, P.M, Chasman, D.I., Aukrust, P., Djurovic, S., Frigessi, A., Desikan, R.S., Dale, A.M and Andreassen, O.A., 2016. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes with Several Cardiovascular Risk Factors. *Circulation Research*, 118(1), 83-94.

*Paper 3*
LeBlanc, M.*, Zuber V.*, Thompson W.K., Andreassen O.A., Frigessi A. and Andreassen., B.K, 2016. A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. (Submitted to *Plos Genetics*)
**contributed equally*

**Table of contents**

Followed by *Papers 1-3* with Supplements

## 1 Introduction

This thesis falls into the realm of *statistical genomics*. Statistical genomics is special in that it involves the integration of theory from two fields: statistics and genetics, in particular quantitative and population genetics. Much like statistics, genetics has an extensive theoretical foundation in the form of mathematical models that show how different evolutionary pressures, namely selection, mutation, migration and random genetic drift, affect gene frequencies and genetic variation. Statistical genomics builds on this knowledge and provides the statistical tools needed to make inference from genomic data, which is universally complex, high dimensional and fraught with multiple testing issues.

The central focus of this thesis is the genome-wide association study (GWAS) and associated applications and statistical methods. In brief, the goal of GWAS is to identify *polymorphic* loci, specific positions of variation in human DNA, that are *associated* with a given disease or trait. As will become apparent, this is not a straightforward task, and is filled with both genetic and statistical issues. These include, but are not limited to, dealing with the non-independence of alleles along a chromosome (linkage disequilibrium; LD), the frequency distribution of risk alleles, the statistical modeling of the relationship between disease and genotype, statistical correction for testing up to millions of genetic variants often in a hypothesis-free context, and particularly in this thesis, dealing with multiple correlated traits.

This thesis is divided in to five chapters. The first chapter provides an introduction including a primer to human genetics and historical description of GWAS together with the advent of the genomic era. The second chapter describes the materials and methods used in this thesis, including the GWAS analytical pipeline and new methodology in the so-called post-GWAS era. The second chapter also gives

details of the false discovery rate methodology and the phenotypes used in this thesis. The third chapter states the specific aims of this thesis and the fourth chapter gives a brief summary of each PhD paper. Finally, in the fifth chapter, the papers are discussed, and concluding remarks are made.

## 1.1 A primer to human genetics

### *1.1.1 Organization of the genetic material*

The cells of every life form contain a special molecule called *dioxyribonucleic acid* (DNA), the so-called "blueprint of life", that influences how each organism develops, functions and passes traits to the next generation. Humans are diploid and have approximately $5x10^{13}$ cells, each of which has a nucleus where the DNA is organized on 23 pairs of chromosomes. Within an individual, each cell contains the identical DNA sequence of nucleic acids, of which there are four types: adenine (A), guanine (G), cytosine (C) and thymine (T). The Human Genome Project, completed in 2003, decoded the human genome and provided the first map of these ATCGs along the 23 chromosomes. It is possible to provide such a "reference genome" because about 97% of genome is fixed (Auton et al., 2015). The remainder of the genome shows variation between individuals, and potentially contributes to the similarity between relatives. Similarity between relatives implies *heritable variation*, *i.e.* the fraction of phenotypic variability between individuals that can be attributed to genetic variation. Notably heritability is possible to estimate from phenotype data alone and does not require any genotyping. Complex traits that have high heritability are exactly those that are targeted by GWAS.

### *1.1.2 Transmission of genetic material from parent to offspring*

The genetic material is passed on from parent to offspring via a process called *meiosis* that occurs exclusively in the sex cells and leads to the formation of haploid gametes, containing only one set of chromosomes; these are the sperm in males and

eggs in females. Crucial to this process is *recombination*, where novel non-parental combinations of genetic variants are formed along the chromosomes. This is an important contributor to variation in human populations, and is a critical factor to consider when applying statistical methods to genomic data.

If recombination did not exist, all *loci* along a chromosome would be linked, causing strong statistical dependencies for all loci on a given chromosome. But since recombination does occur, loci that are far apart on a chromosome are inherited independently from each other. LD, *i.e.* the non-random association between genetic variants (*alleles)* on a chromosome, tends to be strong for physically close loci and tends to get weaker and weaker as a function of distance. LD decays over time and the pattern and extent of LD seen in human populations has been shaped by population history and events such as bottlenecks (sudden reductions in population size) and periods of rapid growth (Reich et al., 2001). When only genotyping a subset of variants, exploiting patterns of LD is an essential element for a good genome-wide genotyping strategy.

## 1.2 A brief historical description of GWAS

### *1.2.1 Advent of GWAS*

Long before the Human Genome Project was started, geneticists understood that genetic variation was key to understanding heritable complex traits and disease. Developing a strategy to identify specific trait loci, however, was and is a complex issue. Critical to this is the *allelic spectrum of disease, i.e*. the frequency distribution of risk alleles. This topic was heavily debated around the turn of the millennia in anticipation of the genomic era (Pritchard and Cox, 2002, Reich and Lander, 2001, Weiss and Clark, 2002). How many loci contribute to a common trait? Should the risk alleles be common or rare? Several lines of reasoning lead to the conclusion that with a few known exceptions, complex traits are highly polygenic, that is have hundreds or

thousands of contributing risk loci (Weiss and Clark, 2002, Pritchard and Cox, 2002, Reich and Lander, 2001). The polygenic nature of complex traits is widely accepted as fact. More controversial, particularly in the pre-genomic era, was the allelic spectrum of disease. The debate centered on whether risk variants (alleles) for common disease would be common or rare. Proponents of the so-called common disease/common variant hypothesis (CD/CV) argued that late-onset common diseases should be largely caused by common variants of modest effect size. Common variants are appealing to work with because they are relatively easy and cheap to identify, can be detected in smaller samples (power) and are generally *older* and geographically dispersed. But, neighbouring *older* variants have undergone more recomination, breaking down LD compared to more rare, local or more recent variants. The disadvantage here is that with weak LD, more variants need to be typed in order to have reasonable coverage of the common variation in the human genome. The strongest evidence for CD/CV came from the simulation studies of Reich and Lander (2001) who showed the CD/CV was *plausible*. However Weiss and Clark (2002) provided very strong, theoretical evidence showing that most risk variants would likely not be common. Despite the lack of solid evidence for the CD/CV hypothesis, and despite compelling evidence against this model of common disease, there was a strong push to go forward with a strategy to *map* the common variation in the human genome. This strategy was likely pursued because, at that time, the technology did not exist for large-scale sequencing studies (essential to map rare variants), and the genetics community largely believed that common variants were the best hope for genetic mapping of disease.

### *1.2.2 The Hapmap Project*

The HapMap project (http://hapmap.ncbi.nlm.nih.gov) was initiated late in 2002

in order to map all of the common single nucleotide variants in the human genome, known as single nucleotide polymorphism (SNPs), and its first phase was completed in 2005 (Gibbs et al., 2003, International HapMap Consortium, 2005). The HapMap project focused solely on common SNPs, in this case defined as those where the least frequent allele (called the *minor allele*) occurs in at least 1% of the population. Using the publically-available Hapmap data, and its corresponding LD structure, so-called *tag SNPs* can be identified. Tag SNPs are SNPs that are very well correlated with all the other SNPs in a defined region. Such a strategy employed when designing the first commercially-available GWAS genotyping panels, such as those offered by Affymetrix and Illumina.

*1.2.3 What is a GWAS?*

A GWAS is a type of study that is typically conducted in hundreds or thousands of *unrelated* subjects. Unrelated subjects can be treated as independent observations, whereas with related subjects, the genetic correlations due to relatedness need to be taken into account. The samples typically come from retrospective cohort studies or cases-control studies. For each subject, a trait or several traits (outcomes, also called phenotypes in genetics) and covariates of interest are measured, and hundreds of thousands to a few millions common genetic variants are genotyped. For case-control studies, independently for each of these SNPs, it is then investigated if the allelic or genotype frequencies are significantly altered between the case and the control groups. This is typically done using logistic regression and the effect size is reported as an *odds ratio*. Similarly for quantitative traits, linear regression is used independently for each SNP to see if a given SNP is associated with the outcome. The effect size in this case is reported as the regression coefficient for the SNP term. Clearly with so many statistical tests being performed, a formal statistical correction

for multiple testing is a mandatory step in GWAS, and this subject will be explored in detail in Sections 2.1.3 and 2.4.

Although there are several options for modeling SNP genotype, the most common approach is to use an *additive model*, where SNP genotypes are ordered and then modeled on a continuous scale. For instance a SNP with minor allele "A" and major allele "G" would have 3 possible genotypes: "AA", "AG" and "GG". With the additive model, we assume that genotype contributes in an additive manner to the phenotype, which implies that the heterozygous genotype "AG" lies exactly in between the two homozygous genotypes. As such we can translate the genotype categories "AA", "AG" and "GG" to a continuous scale: 0, 1, 2. Nearly all of the published GWAS studies to date use this approach.

*1.2.4 The GWAS era*

After the completion of the 1st Hapmap phase in 2005, it was theoretically possible to create a genome-wide panel of tag SNPs. Simultaneously, advances in genotyping technology meant that it was suddenly efficient both in terms of cost and time to carry out genome-wide genotyping, thanks to commercially-available GWAS chips from Illumina and Affymetrix. The first large-scale, well-designed GWAS for complex disease was published in Nature in 2007 and performed by the Wellcome Trust Case Control Consortium (WTCCC; Messerli et al., 2007). This study seeded a massive publication boom, where the number of GWAS studies has increased at an increasing rate. This is clearly evident in Figure 1, from in the Catalog of Published Genome-Wide Association Studies, showing the number of GWAS publications per calendar year from 2005-2013 (Welter et al., 2014). As of 2013, the catalog contained 1751 curated publications of 11 912 SNPs associations at $p < 10^{-5}$.

Clearly the GWAS approach for discovery of disease-associated genetic

variants has been widely adapted and a large number of trait-associated SNPs have been found. But has GWAS really been a success? It turns out that this is not a simple question to answer. Very few common variants of moderate to major effect have been found via GWAS. GWAS trait-associated common variants have very small effect sizes (Figure 2; Manolio et al., 2009). In fact, the effect sizes of common variants on common disease are universally so small that tens of thousands or even over one hundred thousand subjects are required to conduct a reasonably powered GWAS. Such sample sizes are impossible to achieve in individual studies and necessitate the formation of international consortia. These consortia perform meta-analysis of essentially all globally available samples for a given trait, with genotyping data available and that meet the inclusion criteria (*e.g.* often limited to one ethnicity). Examples include the Psychiatric Genetics Consortium (https://www.med.unc.edu/pgc) and the CardiogramplusC4D Consortium (www.cardiogramplusc4d.org/) whose data are used in this thesis. Even when including every virtually sample available globally, GWAS are still notoriously underpowered to identify common variants with extremely small effect sizes (odds ratio<1.1) independent of the choice of genotyping platform (Spencer et al., 2009).

In 2009, Manolio et al. coined the term "missing heritability", referring to the fact that the genetic variants identified by GWAS explain very little of the heritability for most complex traits and common diseases. The so-called missing heritability is likely due to a wide variety of factors including, but not limited to epigenetics, disease-causing rare variants, gene-gene interactions, gene-environment interactions and the underpowered nature of typical GWAS analysis (Eichler et al., 2010). One analytical approach to uncovering part of the missing heritability is to apply more sophisticated statistical methods to existing GWAS data, especially if the methods

involve the addition of biological knowledge. But before exploring these more advance methods, it is important to establish an understanding of the conventional GWAS statistical analysis pipeline.

## 2 Methods

### 2.1 GWAS – Analytical pipeline

*2.1.1 Review*

Let us first recall that the goal of GWAS is to detect loci associated with variation in a trait of interest, usually in a sample of independent subjects. Let us also recall that, because of the statistical dependencies between loci (*i.e.* LD), a properly chosen panel of ~1,000,000 SNPs is sufficient to tag most of the common genetic variation in the (European) human genome. As was introduced in Section 1.2.3, assuming the additive genetic model allows us to treat the three genotype categories at a bi-allelic SNP as a continuous variable with coding 0 (minor allele homozygote), 1 (heterozygote) and 2 (major allele homozygote).

We will first consider a simple GWAS design, with one retrospective sample, one phenotype and a one-SNP-at-a-time regression analysis. In GWAS, the phenotype is either categorical (usually binary case/control) or continuous (called "quantitative" in genomics literature). An underlying assumption for a successful GWAS is that the chosen genotyping platform has reasonable genomic coverage for the population from which the samples have been drawn.

*2.1.2 Association testing in GWAS*

Quantitative phenotypes are analyzed using a linear regression approach, with one SNP and clinical covariates are predictor variables. For simplicity, let us consider a regression model without covariates, but in practice covariates can easily be added. For $n$ samples with outcome $y_j, j = 1,\ldots,n$, and additively-modelled genotype $x_{jg}$ in

individual $j$ for SNP $g$, the linear regression model is $y_j = \alpha_g + \beta_g x_{jg} + \varepsilon_{jg}$, where $\varepsilon_{jg}$ is normally distributed with mean 0 and describes the error term of the relationship between outcome and genotype. The null hypothesis is for each SNP $g$ that the coefficient of the SNP term, $\beta_g$, is equal to zero and the test statistic is a Wald test, $\widehat{\beta_g} \Big/ se(\widehat{\beta_g})$, where *se* is the standard error. Here the estimated coefficient of the SNP term is reported as the effect size. Binary phenotypes are usually analyzed using a logistic regression approach, again with one SNP and clinical covariates as predictor variables. Covariates are again dropped for simplicity. Here, we code the outcome $Y_j$ as 1 cases and 0 for controls, and the logistic regression model is $\log\left(\Pr(Y_j = 1) \Big/ \Pr(Y_j = 0)\right) = \alpha_g + \beta_g x_{jg}$. The null hypothesis for each SNP $g$ is that the probability of being a case or a control is not associated with genotype. The effect size of a given SNP is reported as an odds ratio.

In the regression analysis step, it is possible to correct for genetic within sample differences (*population stratification*) by including the first few principle components derived from the genome-wide SNP panel, which can be interpreted as a sort of "origin score". This correction is desirable because it is protective against spurious associations in the case of both different phenotypic distributions and different allelic frequency distributions in the different subpopulations that may exist in the dataset.

After one regression model is built for each SNP on the genotyping panel, the strength of association between each SNP and the phenotype can be summarized by an effect size, associated confidence interval and a *p*-value. Given that on the order of one million SNPs are included in a GWAS, correction for multiple testing is a critical step in any GWAS analysis. There are several options here, as well as clear conventions established in the GWAS literature.

*2.1.3 Corrections for multiple testing in GWAS*

A *p*-value, which is the probability of seeing a test statistic equal to or greater than the observed test statistic if the null hypothesis is true, is generated for each statistical test. Statistical tests are generally called significant (*i.e.* the null hypothesis is rejected) if the *p*-value falls below a predefined $\alpha$, most often set to 0.05 and known as the *type I error rate*. This probability is for a single statistical test but in GWAS on the order of $10^6$ tests are conducted. If we were to declare SNPs as significantly associated with phenotype based on their *p*-values and a cut-off of 0.05, the cumulative type I error rate over all statistical tests is much greater than 0.05. As such, formal corrections for multiple testing are necessary, in order to maintain an *overall* type I error rate of 0.05 for the *entire* GWAS.

The simplest approach to correcting for multiple testing is the Bonferroni correction. The Bonferroni correction adjusts the alpha value from $\alpha = 0.05$ to $\alpha = (0.05/m)$ where *m* is the number of statistical tests conducted, *i.e.* the number of SNPs in the GWAS.

A related approach is to use the Bonferroni correction for genome-wide significance. The Bonferroni corrected significance threshold for a million tests is $0.05/1{,}000{,}000 = 5\text{x}10^{-8}$, and this cut off very commonly used as the "gold standard" for declaring an association significant in GWAS, regardless of the number of SNPs on the genotyping panel. This is because, for the European population it is estimated that there are approximately one million independent common SNPs in the genome, once the dependencies due to LD are taken into account (Clarke et al., 2011). Another estimate is 7.2x10-8 but $p<5\text{x}10^{-8}$ is the most common choice in the literature (Dudbridge and Gusnanto, 2008, Pe'er et al., 2008).

The Bonferroni correction is appropriate when a *single* false positive in a set of tests would be a problem, otherwise is a very conservative approach and potentially

leads to a large number of false negatives. An alternative, less conservative approach for correcting for multiple testing involves controlling for the expected *proportion* of false discoveries amongst the rejected null hypothesis instead. In GWAS, this is the proportion of trait-associated SNPs that are actually false positives. The first statistical procedure for controlling the *false discovery rate* (FDR) was proposed by Benjamini and Hochberg (1995). In brief, the *p*-values are ordered from smallest to largest, and assigned a corresponding rank *i*. For instance, for the smallest *p*-value, $i = 1$. Compare each individual *p*-value to its Benjamini-Hochberg critical value, $(i/m)*q$, where i is the rank, $m$ is the total number of tests, and $q$ is the false discovery rate you choose. The largest *p*-value that is less than $(i/m)*q$ is significant, and all of the *p*-values smaller than it are also significant.

Importantly, the FDR and Bonferroni corrections do not re-order the SNPs compared to their raw *p*-value rankings; they simply suggest different cut-off points as to what is declared as statistically significant. In Section 2.4 we will further explore procedures for correcting for multiple testing, including some methods that re-order the SNPs compared to their raw *p*-value rankings. But for the time being, let us continue to describe the typical GWAS pipeline.

*2.1.4 Validation of GWAS "discoveries"*

The gold standard for validation of a GWAS association is the replication of the association in an independent sample. Here the burden of multiple testing is less severe and the correction only needs to be made for the number of SNPs in the "associated" SNP set carried forward to the validation step, often on the order of 50 – 100 SNPs.

*2.1.5 Meta-analysis in GWAS*

The description of the GWAS analytical pipeline above assumes that the individuals are from one sample. In practice, nearly all GWAS studies of major

impact are conducted by consortia who collect as many studies as possible to be combined in a meta-analysis. Here issues such as phenotype definition, inclusion criteria, population stratification and genotyping platform become critically important. Imputation of missing genotype data is usually required. Genotype imputation exploits known LD patterns and haplotype frequencies in a reference population (*e.g.* from HapMap or the 1000 Genomes project) to estimate genotypes for SNPs not directly genotyped in the study. Meticulous routines for data storage, security, privacy and access are required. Detailed discussion of these issues is beyond the scope of this thesis but it is important to keep these potentially complicating factors in mind.

Assuming all of the issues above have been dealt with in a reasonable manner, conducting a GWAS meta-analysis for a given phenotype is straightforward. Each contributing study provides regression-derived effect size and associated standard error and the sample size for each SNP. Importantly, each study must specify the reference allele at each SNP; otherwise the effect direction cannot be aligned between studies. Subsequently, meta-analysis, such as inverse variance meta-analysis is conducted. The meta-analysis effect size estimate and associated *p*-value are then reported for each SNP and correction for multiple testing is performed. Usually the GWAS consortium will exclude some of its contribution studies from the meta-analysis and reserve them for a second phase of analysis (*i.e.* validation of the associated SNPs).

### *2.1.6 Multiple related phenotypes in GWAS*

It is common that several, related phenotypes are investigated by GWAS. This can be carried out in *one* study using the same sample set (*e.g.* the Global Lipids Consortium used the same sample to investigate several outcomes including

triglycerides, high-density lipoprotein, low-density lipoprotein and total cholesterol (Teslovich et al., 2010). Related phenotypes can also be investigated in (approximately) independent samples by separate consortia and published in separate publications (*e.g.* blood pressure (Ehret et al., 2011) and triglycerides (Teslovich et al., 2010)). When related phenotypes are investigated in the same sample, it is often because there is not one obvious primary phenotype, and it cost-effective to look at as many heritable phenotypes as possible using the same dataset. Other motivations for investigating related phenotypes (in one sample or in independent samples) include that the genetic basis so-called *endophenotypes* (stable phenotypes with a clear genetic connection) should be easier to identify than broader clinical definitions of disease or other quantitative traits (such as body mass index). Some related complex phenotypes, like type 2 diabetes and coronary artery disease, clearly merit their own consortium-level investigation.

Until recently, it was not common to integrate cross-phenotype results in any formal way. However, informal investigations of overlapping "discoveries", usually at the gene level, were often made. An example of this is the Venn diagram summarizing the findings of the Global Lipids Consortium, which gives a visual display of the overlapping gene sets for the four investigated lipids phenotypes (Figure 3). It is perhaps expected that related lipids phenotypes will also have overlapping gene sets, given their strong phenotypic correlation.

Statistics has well-developed methodology for dealing with multivariate data but these methods are rarely applied to GWAS data in order to deal with multiple, related phenotypes. The reasons for this are not entirely clear, but likely just have to do with conventions in the field of genomics. For a summary of multivariate methods for GWAS see Galesloot et al. (2014).

### *2.1.7 Relevant phenotypes for this thesis*

#### *2.1.7.1 Neurocognitive function*

Neurocognitive function broadly refers to multiple inter-correlated cognitive domains including attention, psychomotor speed, learning and memory, intelligence and executive functioning. In *Paper 1*, we investigate twenty-four neurocognitive tests falling into these five clinical domains via GWAS.

Heritability estimates for different aspects of neurocognitive function range from approximated 50 to 80% (e.g. Lee et al., 2010). Despite its high heritability, neurocognitive function is a particularly challenging phenotype to investigate via GWAS. Reasons for this include: the multivariate nature of neurocognition, the lack of a clear primary phenotype and a lack of consistent phenotype definitions across studies (due to different test batteries for neurocognition). Additionally, there is no consensus on how to deal with important covariates and confounders such as age, education and underlying diseases. Options here include using these as inclusion/exclusion criteria or including them as covariates in the statistical model. All in all, these particular challenges encountered for GWAS of neurocognitive function result in highly underpowered studies with limited or no options for replication.

Presently, nine loci have been associated with the key words "general cognitive ability" or "intelligence" or "cognitive test" or "neurocognive function" in the GWAS catalogue (http://www.ebi.ac.uk/gwas/home) at a *p*-value implying genome-wide significance ($p$-value < $5x10^{-8}$). The results are summarized in Table 1. We include the results from *Paper 1* in this list since it was published already in 2012.

**Table 1.** Single nucleotide polymorphisms associated with neurocognition at *p*-value < $5x10^{-8}$. Chr, chromosome.

| rs# | Gene | Chr | Reference |
|---|---|---|---|
| rs10457441 | intergenic | 6 | (Davies et al., 2015) |
| rs17522122 | *AKAP6* | 14 | (Davies et al., 2015) |
| rs10119 | *TOMM40* | 19 | (Davies et al., 2015) |
| rs2300290 | *PTPRO* | 12 | (LeBlanc et al., 2012, *i.e Paper 1*) |
| rs719714 | *WDR72* | 15 | (LeBlanc et al., 2012, *i.e. Paper 1*) |
| rs6043979 | *KIF16B* | 20 | (Loo et al., 2012) |
| rs3758171 | *PAX5* | 9 | (Loo et al., 2012) |
| rs3815908 | *ELSPBP1* | 19 | (Loo et al., 2012) |
| rs17518584 | *CADM2* | 3 | (Ibrahim-Verbaas et al., 2015) |

*2.1.7.2 Coronary artery disease*

In *Paper 2*, coronary artery disease (CAD) is investigated via GWAS. CAD is a leading cause of death worldwide. CAD happens when the arteries that supply blood to the heart acquire a build up of cholesterol and plaque causing them to be hardened and narrowed. Less blood is able to flow through the arteries causing less oxygen to get to the heart, leading to heart attack and often to permanent heart damage or even death. CAD also leads to heart failure and irregular beating of the heart. The heritability of CAD is approximately 40-50% (Peden and Farrall, 2011).

Several related consortia have investigated the genetics of CAD via GWAS leading to the identification of 46 CAD-associated loci achieving both *p*-value<$5x10^{-8}$ and validation in an independent dataset (CARDIoGRAMplusC4D Consortium et al., 2013; Table 2). These 46 loci were for the most part identified via consortium-based efforts including that of the CARDIoGRAMplusC4D Consortium whose summary statistic data is used in *Paper 2*.

**Table 2.** Single nucleotide polymorphisms associated with coronary artery disease at *p*-value < $5x10^{-8}$. Loci are reported at least one of the following publications: (CARDIoGRAMplusC4D Consortium et al., 2013, Schunkert et al., 2011, Samani et al., 2007, Clarke et al., 2009, Kathiresan et al., 2009, Soranzo et al., 2009, Wang et al., 2011, IBC 50K CAD Consortium, 2011).

| **rs#** | **Chr** | **Gene** |
|---|---|---|
| rs4845625 | 1 | *IL6R* |
| rs515135 | 2 | *APOB* |
| rs2252641 | 2 | *ZEB2-AC074093.1* |
| rs1561198 | 2 | *VAMP5-VAMP8-GGCX* |
| rs7692387 | 4 | *GUCY1A3* |
| rs273909 | 5 | *SLC22A4-SLC22A5* |
| rs10947789 | 6 | *KCNK5* |
| rs4252120 | 6 | *PLG* |
| rs264 | 8 | *LPL* |
| rs9319428 | 13 | *FLT1* |
| rs17514846 | 15 | *FURIN-FES* |
| rs2954029 | 8 | *TRIB1* |
| rs6544713 | 2 | *ABCG5-ABCG8* |
| rs1878406 | 4 | *EDNRA* |
| rs2023938 | 7 | *HDAC9* |
| rs602633 | 1 | *SORT1b* |
| rs11206510 | 1 | *PCSK9* |
| rs6725887 | 2 | *WDR12* |
| rs9818870 | 3 | *MRAS* |
| rs12190287 | 6 | *TCF21* |
| rs3798220 | 6 | *SLC22A3-LPAL2-LPA* |
| rs11556924 | 7 | *ZC3HC1* |
| rs1333049 | 9 | *CDKN2BAS1* |
| rs579459 | 9 | *ABO* |
| rs12413409 | 10 | *CYP17A1-CNNM2-NT5C2* |
| rs2505083 | 10 | *KIAA1462* |
| rs974819 | 11 | *PDGFD* |
| rs3184504 | 12 | *SH2B3* |
| rs4773144 | 13 | *COL4A1-COL4A2* |
| rs2895811 | 14 | *HHIPL1* |
| rs12936587 | 17 | *RAI1-PEMT-RASD1* |
| rs1122608 | 19 | *LDLR* |
| rs9982601 | 21 | *Gene desert (KCNE2)* |
| rs17114036 | 1 | *PPAP2B* |
| rs17609940 | 6 | *ANKS1A* |
| rs12526453 | 6 | *PHACTR1* |
| rs501120 | 10 | *CXCL12* |
| rs1412444 | 10 | *LIPA* |
| rs46522 | 17 | *UBE2Z* |

| | | |
|---|---|---|
| rs216172 | 17 | *SMG6* |
| rs2075650* | 19 | *ApoE-ApoC1* |
| rs445925* | 19 | *ApoE-ApoC1* |
| rs17464857 | 1 | *MIA3* |
| rs12539895 | 7 | *7q22* |
| rs9326246 | 11 | *ZNF259-APOA5-APOA1* |
| rs7173743 | 15 | *ADAMTS7* |

*not in high LD

**2.2 Methodology in the post-GWAS era**

*2.2.1 Are further discoveries possible with existing GWAS data?*

We have already described the "missing heritability" problem in GWAS. Although it is tempting to abandon common variants altogether and look for the missing heritability elsewhere, several lines of evidence suggest that there are still discoveries to be made in existing GWAS data. By looking at the quantile-quantile plot from almost any given large GWAS, it is clear that the *p*-value distribution has many more small *p*-values than expected by chance, and that the typical Bonferroni threshold used to declare statistical significance results in a large number of false negatives. A quantile-quantile plot for a typical Consortium-based GWAS is shown in Figure 4. By convention, these plots are displayed on the $-\log_{10}$ scale. Clearly, observed *p*-value distribution departs from the null distribution before the typical GWAS significance threshold of *p*-value $< 5 \times 10^{-8}$. This is strong empirical evidence that there are many false negatives in GWAS when a standard analytical pipeline is used. In many ways this is not surprising since the typical GWAS analysis is highly conservative, underpowered and done in a hypothesis-free manner (*i.e.* SNPs are treated as exchangeable). The question arises: Can we do better? Is there anything in statistics or biology that can help us to get more out of the existing data? The short answer is yes, that by using more advanced statistical methods, particularly those that incorporate additional biological knowledge, it is possible to make new discoveries in the GWAS data we already have available.

So what type of biological knowledge can be used to aid in the hunt for disease-causing genes? Prior to the GWAS era (pre-2007), it was common to focus the hunt for disease-variants in tens or hundreds of known protein coding genes. This so-called "candidate gene approach", where genes are selected according to *a priori* knowledge of the gene's biological function, has not been a useful way to identify *genetic variation* associated with disease or traits. By focusing on biologically-relevant genes, the burden of multiple testing is reduced. Even with less stringent significance thresholds, the candidate gene approach was not been very successful at identifying trait-associated loci. It turns out that, even if we know which genes are important for disease etiology, this is not equivalent to knowing *where* important genetic variation lies. Again, this is not entirely surprising since population genetics tells us that the more important a gene is for survival and function, the less natural variation we expect to see in the gene. Other reasons for the failure of the candidate gene approach may be that we know altogether too little about which genes may play a role in disease, too little about regulatory or other non-coding genetic variants, or maybe we simply know too little about important common variation in the genome in general.

In the last few years, our knowledge about the structure of the genome and particularly regulatory elements has exploded. Our understanding of the human genome has moved long past the "central dogma of molecular biology" that says that a gene is a piece of DNA that codes for a piece of messenger RNA (mRNA) that in turn codes for a protein. Although the central dogma is a good description of bacterial genomics, it is not an adequate description of how the human genome works. We now know that a large part of the important genetic variation lies outside of the tiny bits of the genome that code for proteins, and instead are involved in gene regulation. The ENCODE (Encyclopedia of DNA Elements) Consortium aims to identify all

functional elements in the human genome and maintains a comprehensive webpage and database (https://www.encodeproject.org). A better understanding of the regulatory elements of the genome has largely been driven by new technology and clever application of the new technology. Paired with *bioinformatic tools* (*i.e.* methods and software tools for understanding biological data), so-called *"-omics"* studies have led to several major genome-level insights about how gene regulation works.

The second-wave analysis of GWAS, characterized by improved use of bioinformatics, statistics and genetical knowledge is still in its infancy but has lead to the development of exciting and promising approaches for the discovery of disease-associated genetic variants. A detailed review of all of these new –omics technologies and related methodology is beyond the scope of this thesis, so we will instead focus on examples of how particular new insights into how human genetics works have been incorporated into GWAS analysis.

### *2.2.2 Example 1: Expression quantitative trait loci (eQTL) and GWAS*

Increasing evidence suggests that single nucleotide polymorphisms (SNPs) associated with complex traits are more likely to be expression quantitative trait loci (eQTLs) than would be expected by chance alone. Beginning around 2007, researchers (Stranger et al., 2007) began innovative genome-level studies in humans to find genetic variants (usually SNPs) that associate with variation in gene expression (usually at the mRNA level), termed eQTL experiments (see early review in Gilad et al., 2008). Here the goal is to identify SNPs that exhibit genotype-dependent gene expression (mRNA), with focus usually being on nearby protein-coding genes. The focus on nearby protein-coding genes is in part to reduce the burden of multiple testing, because we know that genetic variants can also influence

gene expression of distant genes such as genes on other chromosomes. Nearby eQTLs are called cis-eQTLs and distant eQTLs are called trans-eQTLs. Since gene expression is tissue dependent, eQTLs are specific to a given tissue type, for instance adipose tissue or blood. The basic idea for any cis-eQTL analysis involves first defining "nearby" (*e.g.* limit association analysis to genes +/-1000 kb from a given SNP), then calculating an association statistic between the given SNP and the mRNA expression data one at a time for all "nearby" genes. This results in one *p*-value for each SNP-nearby gene pair.

It has been shown that eQTLs are enriched for SNPs associated with complex diseases and traits using GWAS (Cookson et al., 2009, Nicolae et al., 2010). As such, eQTLs are one type of biological information that can be used to re-prioritize GWAS findings. For example, Westra et al. (2013) incorporate eQTL information into GWAS using *p*-value weighting methods. In brief, the GWAS *p*-values are reweighted by weights based on the eQTL *p*-value for each SNP. This is just one example of how eQTL can be incorporated into a GWAS analysis, in this case at the summary statistic level.

*2.2.3 Example 2: Genome annotation and GWAS*

Genome annotation can also be used to improve gene discovery in existing GWAS data. SNPs can be annotated to different genomic regions such as regulatory elements, coding genic elements, introns, and intergenic regions. The annotation is based not only on the exact physical location of a given SNP but also on LD with the SNP and the different genomic elements. Schork et al. (2013) show that certain genomic elements are enriched for small *p*-values in GWAS, indicating that genomic annotation is useful for breaking the exchangeability assumption of the standard GWAS pipeline. This assumption is broken when SNPs come from pre-determinable

categories or clusters, within which they can be dependent and share distributions of effects. Using genome annotation as informative prior information means that *a posteriori* SNPs are no longer exchangeable and are no longer identically-distributed. The suggestion of Shork et al. is to incorporate the annotation information in a conditional false discovery rate setting (see Section 2.4.3.1 for more on the conditional false discovery rate).

### *2.2.4 Example 3: Multiple traits, pleiotropy and GWAS*

The idea that one gene can influence more than one phenotype is well established in genetics (*pleiotropy*). The phenotypes may be obviously connected (*e.g.* high-density lipoprotein and low-density lipoprotein) or less obviously connected (*e.g.* the sickle cell anemia gene leads to *both* changes in red blood cell morphology *and* to improved resistance to malaria). When pleiotropy is also highly polygenic (*i.e.* there are many genes effecting both phenotypes), it should be detectable at the genome-level. Andreassen et al. (2013) use stratified quantile-quantile plots (Figure 5) to visualize this. When stratifying the GWAS *p*-values for a first trait based on their significance in a second related trait, there is more and more leftward deflection on the plot. This indicates that on the genomic scale, the *p*-values in the second trait are informative about significance in the first trait, indicating polygenic pleiotropy. This implies that the *p*-values from trait 2 can be useful prior information to incorporate into the analysis to discover SNPs associated with trait 1. Using conditional false discovery rate (Section 2.4.3.1), the exchangeability assumption implicit in standard GWAS is broken, and the analysis favors those SNPs that are associated with both trait 1 and 2. The methods of Andreassen et al. (2013) are used in *Paper 2* and the related polygenic-pleiotropy informed methods of Zablocki et al. (2014) are used in *Paper 3*.

**2.3 Sample overlap in cross-trait analysis of GWAS**

Analysis of GWAS data in the post-GWAS era often requires the integration of GWAS data for related traits, usually at the summary statistic level. There are several potential advantages when working with cross-trait GWAS, including increasing the power and sophistication of the statistical methodology, and the possibility to ask more sophisticated biological questions. Since summary statistics do not contain any sensitive information, it is now common practice for GWAS consortia to release their summary statistics for public download from their homepages. Summary statistics are efficient to work with compared to genotype-phenotype data, and when a sufficient statistic is used, they contain all information needed for further inference.

When the GWAS sample for a first trait overlaps with the GWAS sample for a second trait, the test statistics for a given SNP will be spuriously correlated, even when genotype is independent from both phenotypes. Lin and Sullivan (2009) were the first to address the methodological challenge of integrating GWAS with overlapping subjects. Using the correlation between the maximum likelihood estimates for the regression coefficients for a given SNP *g*, correlation due to overlap for two case control-studies is:

$$cor(\hat{z}_1, \hat{z}_2) \approx \frac{1}{\sqrt{n_1}\sqrt{n_2}}\left(n_{c0}\sqrt{\exp{(\propto_1+\propto_2)}} + \frac{n_{c1}}{\sqrt{\exp(\propto_1+\propto_2)}}\right) \qquad [2.3.1]$$

where $\exp(\alpha_1 + \alpha_2) \approx n_{11}n_{21}/n_{10}n_{20}$, and $n_1$ is the sample size of study 1, and $n_2$ the sample size of study 2, and where we denote the number of cases in study 1 and 2 as $n_{11}$ and $n_{21}$ respectively, similarly $n_{10}$ and $n_{20}$ for the number of controls in study 1 and 2 respectively, and denote the overlap in controls by $n_{c0}$ and in cases by $n_{c1}$. To calculate this, one needs only the summary statistics and the numbers of overlapping and non-overlapping subjects, which in practice can often be determined from the original GWAS publications. In *Paper 3*, we use approach of Lin and Sullivan to

provide analogous formulas for the correlation due to overlap for all possible pairings of GWAS studies.

In some situations, it is not possible to determine the actual number of overlapping subjects. In this case the GWAS summary statistics for the two traits can be used to estimate the correlation. If all SNPs in both GWAS were null (*i.e.* truly independent from both phenotypes), and if the samples were non-overlapping, the correlation of the summary statistics would be approximately 0. If all SNPs in both GWAS were null but samples overlapped, the correlation of summary statistics would give an estimate of the spurious correlation due to overlap. But in reality, GWAS summary statistics contain both null and non-null SNPs (*i.e.* those with a true genetic effect on one or both phenotypes). Thus, correlation of the summary statistics would include both the effect of the overlapping subjects and the effect of the non-null SNPs, which may be truly correlated (*i.e.* pleiotropy). To date there are two proposed methods for estimating the correlation due to sample overlap from GWAS summary statistics.

Province and Borecki (2013) propose using the tetrachoric correlation of a binary transformation of summary statistics to estimate the correlation due to overlap. In their proposed method they categorize the GWAS summary statistics for each study ($z$-scores) as $z<0$ and $z>0$. They then calculate the tetrachoric correlation of the resulting categorized vectors. They argue and show with simulation studies that this protects against the influence of the non-null SNPs in estimating the correlation due to sample overlap.

Zhu et al. (2015) also derive a formula for estimating the correlation due to sample overlap based on summary statistics. First, prune the GWAS summary statistics down to an independent set of SNPs (based on known LD structure in the

data). Second, calculate:

$$corr(T_1,T_2)=\{\Sigma_g(T_{g1}-\mu_1)(T_{g2}-\mu_2)\}/sqrt\{\Sigma_g(T_{g1}-\mu_1)^2(T_{g2}-\mu_2)^2\} \quad [2.3.2]$$

where $T_1,T_2$ are the test statistics for the SNPs for traits 1 and 2 in their corresponding cohorts, and $\mu_1$ and $\mu_2$ are their corresponding means. Their method assumes that all correlation in the test statistics can be either attributed to overlapping or related samples in the two studies. Therefore this method for estimating correlation due to sample overlap will not work well if there is also polygenic pleiotropy.

## 2.4 A primer to false discovery rate methodology

### *2.4.1 Benjamini-Hochberg false discovery rate*

As discussed in Section 2.1.3, correction for multiple testing is an essential part of GWAS analysis. Although the historical "gold standard" correction is a Bonferroni-based cut-off of $5x10^{-8}$, this is indisputably an overly conservative approach (for example see empirical evidence of this in the typical GWAS quantile-quantile in Figure 4). A viable and more liberal alternative is instead controlling the false discovery rate (FDR). As introduced in Section 2.1.3, the FDR was first introduced in the landmark paper 'Controlling the false discovery rate: a new and powerful approach to multiple comparisons' by Benjamini and Hochberg (1995). Interestingly, further development of FDR-based methodology has been largely inspired by problems arising in genomics research, where studies involving gene expression microarray experiments were the first application to present with multiple testing challenges on such an enormous scale (Benjamini, 2010).

To control the number of false discoveries, *i.e.* the expected ratio, *E(V/R)*, of the number of false positives *V* among all significant tests *R*, Benjamini and Hochberg introduced a step-up procedure that is guaranteed to control *E(V/R)* at a level less than *q*, the desired FDR control. We revisit this procedure, first introduced in Section

2.1.3, with the addition of more formal notation. First, order the $m$ $p$-values from smallest to largest, $p_{(1)} \leq ... \leq p_{(m)}$ and assign a corresponding rank $i$ to each $p$-value. Compare each individual $p$-value to its Benjamini-Hochberg critical value, $(i/m)*q$. Define $k = max(i : p(i) \leq (i/m)*q)$ and all hypotheses belonging to $p_{(1)},...,p_{(k)}$ are rejected. Thus the largest $p$-value that is less than $(i/m)*q$ is significant, and all of the $p$-values smaller than it are also significant as well.

*2.4.2 The Bayesian approach to the false discovery rate*

The FDR has subsequently been approached from a Bayesian perspective (see Storey, 2002, Efron et al., 2001, Efron and Tibshirani, 2002, Efron, 2008). Fundamental to the Bayesian approach is the two-group model, where each of the $m$ tests is either null or non-null with prior probability $\pi_0$ or $\pi_1 = 1 - \pi_0$ respectively. The $p$-value, $p_{1g}$, or more generally the test statistic for SNP $g$, $z_{1g}$, has a different distribution based on whether it is null or non-null. In the following we drop the $g$ subscript for simplicity. Let $F_0(z_1)$ and $F_1(z_1)$ denote the cumulative distribution functions of the density functions $f_0(z_1)$ and $f_1(z_1)$, for the null and non-null densities functions respectively. As such, $z_1$ follows a two-group mixture model with cumulative distribution function:

$$F(z_1) = \pi_0 F_0(z_1) + \pi_1 F_1(z_1) \qquad [2.4.2.1]$$

and density function

$$f(z_1) = \pi_0 f_0(z_1) + \pi_1 f_1(z_1). \qquad [2.4.2.2]$$

From here we can use Bayes theorem and define the tail area-based FDR (*Fdr*) as

$$Fdr(z_1) = Pr(null \mid Z \geq z_1) = \pi_0 F_0(z_1) / F(z_1) \qquad [2.4.2.3]$$

and the local FDR (*fdr*) as

$$fdr(z_1) = Pr(null \mid Z = z_1) = \pi_0 f_0(z_1) / f(z_1). \qquad [2.4.2.4]$$

*Fdr* is very much like a corrected $p$-value and connects very closely to the Benjamini

and Hochberg FDR (Efron, 2008). In order to estimate *Fdr* or *fdr*, one proceeds by fitting the mixture model in either *Equation 1.6.1* or *1.6.2* to the observed data. This can be done using either a theoretical null model (*e.g.* standard normal distribution for *z*-scores or *uniform(0,1)* distribution for *p*-values), or an empirical null model (*e.g.* specifying a distribution type but estimating the parameters from the data). Additionally, an estimate of $f(z_1)$ or $F(z_1)$ is required, as is an estimate of $\pi_0$ (which in GWAS can reasonable and conservatively be set to 1). $F(z_1)$ can be estimated by the empirical cumulative distribution funtion $m_p/m$, where $m_p$ is the number of tests with a *z*-score greater than or equal to $z_1$ and $m$ is the total number of tests. GWAS data is particularly well-suited to *Fdr* or *fdr* estimation since $m$ is very large (on the order of $10^6$) and $\pi_0$ well approximated by 1 (that is, only a few dozen to a few hundred common variants out of ~one million are expected to be non-null).

*2.4.3 Bivariate extensions of the false discovery rate*

The FDR methods described above implicitly assume the exchangeability of SNPs. Breaking this assumption and incorporating prior information on each SNP will improve power, as long as this prior information is truly a useful covariate. The prior information could be different kinds of annotation (see Sections 2.2.2 to 2.2.4 for examples), but in this thesis we focus on pleiotropy. The basic idea is that in the presence of polygenic pleiotropy, the GWAS summary statistic of a second trait ($z_2$) can be informative for FDR modeling for the first trait ($z_1$). *Papers 2* and *3* in this thesis use FDR methodology involving bivariate extensions to *Equations 2.4.2.3* and *2.4.2.4*. *Paper 2* uses the conditional FDR (*condFdr*) and the related conjunctional FDR (*conjFdr*), extensions of the *Fdr*, using estimating procedures described in Andreassen et al. (2013). *Paper 3* uses the covariate-modulated FDR (*cmfdr*), an extension of the *fdr*, proposed by Ferkingstad et al. (2008) using estimating

procedures first described in Zablocki et al. (2014).

Conceptually, a full mixture model for two traits is a four-group mixture model, given by the following density function:

$$f(z_1, z_2) = \pi_0 f_0(z_1, z_2) + \pi_1 f_1(z_1, z_2) + \pi_2 f_2(z_1, z_2) + \pi_3 f_3(z_1, z_2) \quad [2.4.3.1]$$

and where $\pi_0$ is the proportion of SNPs for which both phenotypes are null, $\pi_1$ is the proportion of SNPs where both phenotype 1 and 2 are non-null (*i.e.* the pleiotropic SNPs), $\pi_2$ is the proportion of SNPs where phenotype 1 is null and phenotype 2 is non-null, and $\pi_3$ is the proportion of SNPs where phenotype 2 is null and phenotype 1 is non-null. Likewise, $f_0(z_1, z_2)$ is the density function for the SNPs where both phenotypes are null, $f_1(z_1, z_2)$ is the density function for the SNPs where both phenotypes are non-null, $f_2(z_1, z_2)$ is the density function for the SNPs where phenotype 1 is null and phenotype 2 is non-null and $f_3(z_1, z_2)$ is the density function for the SNPs where phenotype 2 is null and phenotype 1 is non-null. This full specification of the four-group mixture model is a useful starting point since it classifies SNPs into four biologically-interpretable categories, which may be useful for future inference. In practice, a simplified mixture model is usually assumed for estimation procedures in bivariate extensions of the local false discovery rate. If we imagine that all non-null SNPs are non-null for *both* trait 1 and trait 2 (i.e., $\pi_2$ and $\pi_3$ are *0)*, the mixture model in *Equation* 2.4.3.1 simplifies to:

$$f(z_1, z_2) = \pi_0 f_0(z_1, z_2) + \pi_1 f_1(z_1, z_2). \quad [2.4.3.2]$$

*2.4.3.1 Conditional false discovery rate*

The *condFdr* is defined, using Bayes Theorem, as:

$$condFdr(z_1 \mid z_2) = Pr(\text{null for trait 1} \mid Z_1 \geq z_1 \text{ and } Z_2 \geq z_2)$$

$$= \pi_0 (z_2) F_0(z_1 \mid z_2) / F(z_1 \mid z_2) \quad [2.4.3.1.1]$$

or on the *p*-value scale,

$$condFdr(p_1 \mid p_2) = Pr(\text{null for trait 1} \mid P_1 \leq p_1 \text{ and } P_2 \leq p_2)$$

$$= \pi_0(p_2) F_0(p_1 \mid p_2) / F(p_1 \mid p_2) \qquad [2.4.3.1.2]$$

Under the null hypothesis $p_1$ and $p_2$ are independent so $F_0(p_1 \mid p_2) = F_0(p_1) = p_1$. This can be thought of as the expected quantile of $p_1$ under the null hypothesis. Therefore

$$condFdr(p_1 \mid p_2) = \pi_0(p_2) p_1 / F(p_1 \mid p_2). \qquad [2.4.3.1.3]$$

Conservatively, $\pi_0(p_2)$ is set to 1. The conditional cumulative distribution function, $F(p_1 \mid p_2)$, needs to be estimated from the data. This can be thought of as the observed quantile of $p_1$ conditioned on the *p*-value in the second trait being as small as or smaller than the observed *p*-value, $p_2$. The approach taken here is described in detail in Andreassen et al. (2013). In brief SNPs are binned into a "look-up table", with the *p*-value in the first trait in the rows and the *p*-value from the second-trait in the columns. From this table, the observed quantile of $p_1$ amongst the subset of SNPs for which the *p*-value for the second trait is as small as or smaller than $p_2$ is calculated.

*2.4.3.2 Covariate modulated local false discovery rate*

The local false discovery rate has also been extended to include information from a second variable. This extension was first proposed by Ferkingstad et al. (2008) and further developed by Zablocki et al. (2014). In *Paper 3*, we use the estimation procedures of Zablocki et al. (2014).

The *cmfdr* is defined, using Bayes Theorem, as:

$$cmfdr(z_1 \mid z_2) = Pr(\text{null for trait 1} \mid Z_1 = z_1 \text{ and } Z_2 = z_2)$$

$$= \pi_0(z_2) f_0(z_1) / f(z_1 \mid z_2)$$

$$= \pi_0(z_2) f_0(z_1) / \{ \pi_0(z_2) f_0(z_1) + \pi_1(z_2) f_1(z_1 \mid z_2) \}. \qquad [2.4.3.2.1]$$

Here it is required to estimate the proportion of SNPs that are null for trait 1 given that $Z_2 = z_2$, the parameters for the null density function for $z_1$, which is assumed

independent from $z_2$ and the non-null density function for $z_1$ given that $Z_2=z_2$. A fully Bayesian estimation procedure is followed, where $f_0(z_1)$ follows a folded normal distribution with mean 0 and $f_1(z_1|z_2)$ follows a gamma distribution. Here the shape parameter is modeled as dependent on $z_2$ and but the rate parameter is assumed independent from $z_2$. The proportion of non-null SNPs for trait 1 is dependent on $z_2$, and is modeled using a logistic regression procedure. The implementation of this procedure in R is available from the authors at: https://sites.google.com/site/covmodfdr/.

## 3 Aims

This thesis aims to apply and improve analyses of GWAS data, specifically using a standard GWAS pipeline for the genotype-phenotype data from TOP study for multiple neurocognitive traits (*Paper 1*), and using pleiotropy-informed false discovery rate methodology for summary statistic data from the CARDIoGRAMplusC4D Consortium and related cardio-metabolic traits for CAD, in order to find trait-associated genetic variants (*Paper 2*). We aimed to propose a method to adjust for sample overlap in cross-trait analysis of GWAS data when only summary statistics are available (*Paper 3*).

## 4 Summary of papers in this thesis

### 4.1 *Paper 1*

LeBlanc, M., Kulle, B., Sundet, K., Agartz, I., Melle, I., Djurovic, S., Frigessi, A. and Andreassen, O.A., 2012. Genome-wide study identifies PTPRO and WDR72 and FOXQ1-SUMO1P1 interaction associated with neurocognitive function. *Journal of psychiatric research*, *46*(2), pp.271-278.

The aim of this paper was to find SNPs/genes associated with neurocognitive function using the standard GWAS approach. Samples were from the Thematically Organized Psychosis (TOP) Study conducted at Oslo University Hospital. The sample

included healthy individuals (n = 377) and patients with schizophrenia spectrum disorders (n = 204) and bipolar disorders (n = 177) having genotype (Affymetrix Genome-Wide Human SNP Array 6.0) and neurocognitive data available. Twenty-four neurocognitive tests falling into five clinical domains (Attention, Executive Functioning, Psychomotor Speed, Learning and Memory, Intelligence) were explored as outcome variables using a standard GWAS approach. Two independent associations achieve genome-wide significance based on Bonferroni correction and these were annotated to the *PTPRO* and *WDR72* genes. Additionally, we looked for interaction in the subset of SNPs with $p$-value $< 3.6 \times 10^{-7}$, corresponding to an overall $\alpha$ of 0.2, and found a significant *FOXQ1-SUMO1P1* interaction. The findings should be replicated in independent samples, but indicate a role of *PTPRO* in Learning and Memory, *WDR72* with Executive Functioning, and an interaction between *FOXQ1* and *SUMO1P1* for Psychomotor Speed.

**4.2 *Paper 2***

LeBlanc, M., Zuber, V., Andreassen, B.K., Witoelar, A., Zeng, L., Bettella, F., Wang, Y., McEvoy, L.K., Thompson, W.K., Schork, A.J., Reppe, S., Barrett-Connor, E., Ligthart, S., Dehghan, A., Gautvik, K.M., Nelson, C.P., Schunkert, H., Samani, N.J., CARDIoGRAM Consortium, Ridker, P.M, Chasman, D.I., Aukrust, P., Djurovic, S., Frigessi, A., Desikan, R.S., Dale, A.M and Andreassen, O.A., 2016. Identifying Novel Gene Variants in Coronary Artery Disease and Shared Genes with Several Cardiovascular Risk Factors. *Circulation Research*, 118(1):83-94.

The main aim of this paper was to find SNPs/genes associated with coronary artery disease (CAD) using a post-GWAS era approach. Here we used the summary statistics from a large-scale genomic study conducted by CARDIoGRAMplusC4D Consortium together with GWAS summary statistics from eight related cardiovascular risk factors to improve gene discovery for CAD. The eight risk factors were: type 1 diabetes, type 2 diabetes, high-density lipoprotein, low-density lipoprotein, triglycerides, C-reactive protein, body mass index and systolic blood

pressure. Using the conditional FDR pairwise with CAD as the primary trait and each risk factor as the secondary trait, we found a significant polygenic pleiotropy enrichment for each pair. We identified 67 novel loci associated with CAD (overall conditional FDR < 0.01). Further, we identified 53 loci with significant effects in both CAD and at least one risk factor. The observed polygenic overlap between CAD and cardio-metabolic risk factors indicates an etiological relationship that warrants further investigation. The new genetic loci identified implicate novel genetic mechanisms related to CAD. A favorable editorial was published in the same issue of *Circulation Research* highlighting the importance of the results to the cardiac genetic community (Quertermous and Ingelsson, 2016).

**4.3 *Paper 3***

LeBlanc, M.*, Zuber V.*, Thompson W.K., Andreassen O.A., Frigessi A. and Andreassen, B.K., 2016. A correction for sample overlap in genome-wide association studies in a polygenic pleiotropy-informed framework. (Submitted to *Plos Genetics*)
**Contributed equally*

The main aim of this paper was to propose a method for correcting for sample overlap when integrating GWAS data across several traits, at the summary statistic level. When two GWAS contain overlapping subjects, their summary statistics for a given SNP are correlated even under the null hypothesis of no genetic effects. The proposed correction is based on the correlation between the maximum likelihood (ML) estimates for the regression coefficients from the first and second GWAS, for a given SNP under the null hypothesis. We derive the correlation for any pairwise combination of quantitative or case-control GWAS, and then use this correlation in a linear transformation to de-correlate the GWAS summary statistics. Using the covariate-modulated false discovery rate and simulated GWAS for two traits and with sample overlap, we show that without correction for sample overlap, the false discovery proportion greatly exceeds that of simulated independent GWAS.

When applying the proposed correction to the simulated data with overlap, proper control of the false discovery rate is restored. The proposed correction is then applied to genotype-phenotype data from the Psychiatric Genetics Consortium. We generate summary statistics for GWAS for schizophrenia and bipolar disorder using first independent control sets and then overlapping control sets. We show that before applying the proposed correction, using the covariate-modulated false discovery rate leads to more "discoveries" that are most likely false positives, and this gets worse as the extent of sample overlap increases. After applying the proposed correction, the number of discoveries in the overlapping GWAS is comparable to the number of discoveries in the independent GWAS, implying successful control of the false discovery rate.

## 5. Discussion

### 5.1 Thesis overview

This thesis focuses on specific applications and methods pertaining to genome-wide association studies. In *Paper 1*, we worked with genotype-phenotype data from the TOP study and conduct a basic genome-wide association analysis for several neurocognitive traits. The main aim of the paper was to identify SNPs associated with neurocognition. The analysis plan mainly followed a standard GWAS analytical pipeline, but did not include a replication step. The work presented in *Paper 1* was conducted during the "GWAS era". The work presented in *Paper 2* and *Paper 3* was conducted in the "post-GWAS era". In *Paper 2*, we used summary statistics from GWAS for coronary artery disease and related cardio-metabolic traits, all sourced from international consortia-based studies, to improve discovery of CAD associated common genetic variants and to quantify the extent of polygenic pleiotropy for CAD and each related trait. Here we used the conditional false discovery rate to integrate summary statistics for pairs of traits, with CAD conditioned on each of the secondary, related traits. We performed an internal validation step and attempted an external validation step. In conducting the analysis for *Paper 2*, we encountered overlapping samples for CAD and several of the secondary traits. Having access only to the GWAS summary statistics and not the underlying genotype-phenotype data meant that we could not remove the sample overlap by splitting the samples. Therefore, in *Paper 3*, we propose two-step method for correcting for sample overlap when working with summary statistics for two GWAS having non-distinct sample sets. The first step of our proposal is a de-correlation step and the second step is user-defined. In *Paper 3* the covariate-modulated false discovery rate was used in the second step. We also applied the proposed correction in *Paper 2* using the proposed

de-correlation step followed by the conditional false discovery rate analysis.

This discussion is organized as follows. First we discuss the main contributions, strengths and weaknesses, and suggestion possible future work and provide a conclusion for each paper. We then conclude with a general discussion of how these papers together fit into the general themes raised in this thesis.

## 5.2 *Paper 1*

### *5.2.1 Paper 1 – main contributions*

In *Paper 1*, we aimed to identify neurocognitive-trait associated SNPs. At the time that this analysis was conducted (mainly in 2009/2010), there was not a comparable published GWAS of neurocognitive traits, although GWAS for certain neurocognitive traits we beginning to emerge (Bates et al., 2009, Butcher et al., 2008, Papassotiropoulos et al., 2006). The main contribution of *Paper 1* is the identification of two novel loci associated with neurocognition in the TOP sample. A SNP annotated to *PTPRO* was associated with a Learning and Memory trait, and a SNP annotated to *WDR72* was associated with an Executive Functioning trait. A more speculative finding was the identified interaction between *FOXQ1* and *SUMO1P1* for a Psychomotor Speed trait.

### *5.2.2 Paper 1 – strengths and weaknesses*

The main weaknesses in *Paper 1* are its small sample size (lack of power) and its lack of a replication sample. It can also be argued that the 24 different neurocognitive traits could have been analyzed in a more cohesive, elegant and powerful way. However, these weaknesses are particularly challenging to address for neurocognitive traits. The challenges here include that the neurocognitive test battery is never the same from one study of neurocognition to the next, making multi-study data alignment and meta-analysis difficult to conduct. Within one study, statistical methods for multivariate data may be difficult to apply because of missing data. In the

TOP study, the 24 neurocognitive tests were rarely conducted on all subjects and data imputation would thus be required for multivariate methods. In analyses not published or presented here, we did the required imputation and followed this by principal components analysis (PCA). Here we found that the neurocognitive traits were largely independent from each other and did not strongly cluster into the clinically-defined neurocognitive domains. This presented an obstacle for multivariate analysis, which is of course most powerful when the traits are correlated. In earlier studies of neurocognition, the first principle component itself is used at the outcome. We did not want to do this, however, as it essentially removes the possibility for future replication efforts. The TOP study is in all likelihood unique in its test battery and no other study would be able to replicate the exact same principle component.

When we corrected for multiple testing, we only did a within-trait correction. In other words, we did not account for multiple testing across traits. Surprisingly, this is the common approach in GWAS, but strictly speaking the topic should at least be addressed. However, our study was hypothesis-generating, and we did not want employ an overly conservative approach and increase the false negatives.

As a secondary result, we identified a SNP-SNP interaction. A strength of this analysis is that we only carried forward the top SNPs from the GWAS analysis, reducing the multiple testing burden, but simultaneously this may also be a weakness, since association in GWAS may not be a good criteria for pre-screening SNPs likely to interact with each other.

*5.2.3 Paper 1 – future work*

Since *Paper 1* was published, there has been one major GWAS of neurocognition (Ibrahim-Verbaas et al., 2015) that specifically tried to replicate our *WDR72* association but the finding did not replicate. The replication effort did not use

exactly the same neurocognitive test and was likely underpowered, but the result did not even nominally replicate. To improve discovery of neurocognitive-related common genetic variants, in addition to the obvious need for larger sample sizes, it may be useful to explore neurocognitive GWAS data using post-GWAS era methodology such as those we used in *Paper 2* and *Paper 3*. Conditional quantile-quantile plots, followed by further analysis when polygenic pleiotropy is seen could be helpful. Incorporating brain-eQTLs into the analysis may also help in identifying neurocognitive-trait associated SNPs.

*5.2.4 Paper 1 – conclusion*

Our GWAS of neurocognition in a sample of 700 individuals identified two neurocognitive-trait associated SNPs that have yet to be replicated in an independent sample. It may be possible to revisit the dataset in the future with more advanced statistical methods from the post-GWAS era, but this will never "overcome" the small sample size, which is 10x to 100x smaller than the most successful GWAS for other common human traits.

### 5.3 *Paper 2*

*5.3.1 Paper 2- main contributions*

In *Paper 2*, we aimed to identify coronary artery disease associated SNPs and to illustrate the extent of polygenic overlap between CAD and related cardio-metabolic traits. The main contributions of *Paper 2* are the identification of 67 novel CAD associated loci, 53 loci associated with both CAD and at least one related trait, and the identification of significant polygenic overlap between CAD and 8 related traits. This provides strong evidence the known phenotypic correlations between CAD and the related traits are accompanied by genetic correlations.

*5.3.2 Paper 2 – strengths and weaknesses*

A strength of *Paper 2* is that it uses the best-available data available.

Specifically, it uses summary statistics from consortia-based GWAS for CAD and related cardio-metabolic traits. Without exception, the underlying samples are the largest available and the studies have been published in high-impact journals. The original studies for CAD and each related trait was based on a standard GWAS pipeline, and then replication in an independent dataset. A weakness of our analysis is lack of a stringent independent replication step, but has several other strengths compared to the standard GWAS pipeline. We provide an internal validation, called the replication rate, which is essentially a cross-validation procedure using sub-studies instead of individuals as the unit in the derivation and validation sets. We also nominally replicate some of the findings in an independent prospective cohort. By identifying a polygenic pleiotropic signal for CAD and each related trait, we establish a type of information that can be used to break the exchangeability assumption implicit in standard GWAS and improve power by incorporating this information into our analysis pipeline.

A limitation of *Paper 2* is that we again do not formally adjust for computing the *condFdr* for 8 pairs. We do use a relatively stringent threshold, following the method of Andreassen et al. (2014) but a more precise solution is given by Liley and Wallace (2015) for declaring the upper bound for the false discovery rate of all declared SNPs. We were already in the review process when this method was published and would have employed it had we been aware of it at the time of analysis.

Compared to earlier publications, we modified the *condFdr* estimation in two important ways. Since the "GWAS" summary statistics from the CardiogramplusC4D Consortium are based on the Metabochip (Voight et al., 2012), they are not actually from a GWAS in the strict sense. The Metabochip is a custom Illumina genotyping panel designed to test, ~200,000 SNPs of interest for metabolic

and atherosclerotic / cardiovascular disease traits. Content on the chip was selected on the basis of previously published GWAS results and of HapMap and 1000 Genomes Project SNP content. The Metabochip includes fine mapping around previous GWAS hits. This has important implications for all types of false discovery rate calculations using the Metabochip since the number of SNPs in LD on the chip is higher for non-null SNPs. To account for this, we estimated the *condFdr*, namely $F(p_1|p_2)$ using an LD-pruned set of SNPs. This was an important modification to avoid a substantial increase in the type 1 error rate in the *condFdr* and a loss of control of the false discovery rate.

We also quantified the extent of and adjusted for sample overlap using the methods developed in *Paper 3*. This was of critical importance and without this adjustment, the false discovery rate would have been greatly inflated. Liley and Wallace (2015) provide an alternative approach to calculating the *condFdr* when samples overlap for case-control studies but not for overlapping samples when one study has a quantitative outcome. Therefore, their method would not have been sufficient for our application, and their approach was published after we had completed the analysis for *Paper 2*.

The lack of a "perfect" replication sample is a definite weakness, but no perfect replication sample is available. We instead applied an internal validation analysis (replication rate) and attempted replication in a prospective dataset.

*5.3.3 Paper 2 – future work*

For the novel loci reported here, the next step will be to replicate them in an independent study, should such a dataset become available. Additional work to establish causality, and work in the laboratory to understand biological mechanisms connecting CAD and related cardio-metabolic traits should be conducted in the future.

*5.3.4 Paper 2 – conclusion*

Our paper makes a substantial contribution to the field of cardiac genetics and makes important polygenic connections between CAD and eight related cardio-metabolic traits. Additionally, we believe we wrote a paper that made the power and utility of advanced statistical methods clear to non-statisticians.

## 5.4 Paper 3

*5.4.1 Paper 3 – main contribution*

This manuscript proposes a correction for sample overlap for cross-trait analysis of GWAS data at the summary statistic level. Overlap in samples between two GWAS studies can introduce bias when combining and integrating the two studies and thus leads to spurious findings. Our correction is appropriate for both case-control studies and for quantitative outcomes. Additionally, our correction can be seen as a pre-processing step, which allows for any type of data integration downstream. Since consortia-based GWAS for related traits nearly always contain overlapping samples and data is released on the level of summary statistics, our proposed correction makes an important and useful contribution to the field. It allows for unbiased downstream integration of GWAS with overlap in samples and cross-trait analysis without the access to the original genotype data.

*5.4.2 Paper 3 – strengths and weaknesses*

A main strength of *Paper 3* is that is provides an easy to implement correction for sample overlap and strong evidence that this correction works using both simulation studies and genotype-phenotype data from the PGC. Our approach to correcting for sample overlap has the distinct advantage of not being tied to any particular downstream implementation of cross-trait analysis. The correction for sample overlap proposed by Liley and Wallace (2015) is tied to the *condFdr* and case-control studies. Their critical insight is that sample overlap renders $F_0(p_1)$

dependent on $p_2$, and they modify this part of the estimation procedure using the correlation due to overlap provided in the Lin and Sullivan (2009) paper.

*5.4.3 Paper 3 – future work*

The method of Liley and Wallace (2015) could be extended using the correlations due to sample overlap that we provide for quantitative trait studies and for the quantitative trait study-case control study combination. This may seem obvious, but it is not obvious for the many of the end users of these methods. The *cmfdr* method could also be extended in a similar way to the Liley and Wallace paper. Their insight that sample overlap renders $F_0(p_1)$ dependent on $p_2$ could be used to develop a new algorithm for estimating the *cmfdr*. Our paper again provides the correlations due to sample overlap for all possible study combinations necessary for such an estimation.

*5.4.4 Paper 3 – conclusion*

The proposed method provides a de-correlation step that can be incorporated into any cross-trait GWAS analytical pipeline, and can drive research forward by allowing for data integration of GWAS datasets containing overlapping subjects. Although this is not the first publication on overlapping subjects in GWAS, this is the first publication to provide the explicit formulas needed for calculating correlation due to sample overlap for *all* possible pairs of case-control and quantitative trait studies, when only summary statistics are available. Since our proposed method is easy to implement, and since the paper provides a detailed synthesis on the subject of sample overlap, we believe that researchers will better understand the bias induced by sample overlap and how to address the bias caused by sample overlap.

**5.5 Concluding Remarks**

This thesis aimed to apply and improve analyses for GWAS and to address the issue of sample overlap in cross-trait analysis of GWAS data at the summary statistic

level. Although at the surface, many of the statistical methods used here are standard and are based on well-established regression methods and correction for multiple testing, naïve application of these techniques without consideration of the special properties of genomic data will lead to incorrect conclusions. Heritability, linkage disequilibrium in the human genome, design of the genotyping chips and a basic understanding of the allelic spectrum of disease are all important consideration when conducting genomic analyses. A good overview of common practices in genomic studies is also important, otherwise critical issues like sample overlap for two datasets coming from two separate GWAS Consortia published in separate manuscripts could easily be missed, resulting in an extremely biased analysis. It seems that, going forward, the most successful approaches to genomic data, and certainly to GWAS data, will continue to combine biological knowledge in a clever way with improved and innovative statistical techniques. The key here is to use the biological knowledge in such a way that the exchangeability assumption implicit in a naïve GWAS pipeline is broken, but without being too biased by this prior biological knowledge and overly committed to which loci are important.

**References**


Andreassen, O. A., Thompson, W. K., Schork, A. J., Ripke, S., Mattingsdal, M., Kelso, J. R., Kendler, K. S., O'Donovan, M. C., Rujescu, D., Werge, T., Sklar, P., Psychiatric Genomics Consortium Schizophrenia Working Group, Psychiatric Genomics Consortium Bipolar Disorder Working Group, Roddey, J. C., Chi-Hua, C., McEvoy, L. K., Desikan, R. S., Djurovic, S. & Dale, A. M. 2013. Improved detection of common variants associated with schizophrenia and bipolar disorder using pleiotropy-informed conditional False Discovery Rate. *PLoS Genet,* 9**,** e1003455.

Andreassen, O. A., Zuber, V., Thompson, W. K., Schork, A. J., Bettella, F., Djurovic, S., Desikan, R. S., Mills, I. G. & Dale, A. M. 2014. Shared common variants in prostate cancer and blood lipids. *Int J Epidemiol*. 43, 1205-14.

Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A. & Abecasis, G. R. 2015. A global reference for human genetic variation. *Nature,* 526**,** 68-74.

Bates, T. C., Price, J. F., Harris, S. E., Marioni, R. E., Fowkes, F. G., Stewart, M. C., Murray, G. D., Whalley, L. J., Starr, J. M. & Deary, I. J. 2009. Association of KIBRA and memory. *Neurosci Lett,* 458**,** 140-3.

Benjamini, Y. 2010. Discovering the false discovery rate. *J R Stat Soc Series B Stat Methodol,* 72**,** 405-416.

Benjamini, Y. & Hochberg, Y. 1995. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *J R Stat Soc Series B Stat Methodol,* 289-300.

Butcher, L. M., Davis, O. S., Craig, I. W. & Plomin, R. 2008. Genome-wide quantitative trait locus association scan of general cognitive ability using pooled DNA and 500K single nucleotide polymorphism microarrays. *Genes Brain Behav,* 7**,** 435-46.

The CARDIoGRAMplusC4D Consortium, Deloukas, P., Kanoni, S., Willenborg, C., Farrall, M., Assimes, T. L., Thompson, J. R., Ingelsson, E., Saleheen, D., Erdmann, J., Goldstein, B. A., Stirrups, K., Konig, I. R., Cazier, J. B., Johansson, A., Hall, A. S., Lee, J. Y., Willer, C. J., Chambers, J. C., Esko, T., Folkersen, L., Goel, A., Grundberg, E., Havulinna, A. S., Ho, W. K., Hopewell, J. C., Eriksson, N., Kleber, M. E., Kristiansson, K., Lundmark, P., Lyytikainen, L. P., Rafelt, S., Shungin, D., Strawbridge, R. J., Thorleifsson, G., Tikkanen, E., Van Zuydam, N., Voight, B. F., Waite, L. L., Zhang, W., Ziegler, A., Absher, D., Altshuler, D., Balmforth, A. J., Barroso, I., Braund, P. S., Burgdorf, C., Claudi-Boehm, S., Cox, D., Dimitriou, M., Do, R., Consortium, D., Consortium, C., Doney, A. S., El Mokhtari, N., Eriksson, P., Fischer, K., Fontanillas, P., Franco-Cereceda, A., Gigante, B., Groop, L., Gustafsson, S., Hager, J., Hallmans, G., Han, B. G., Hunt, S. E., Kang, H. M., Illig, T., Kessler, T., Knowles, J. W., Kolovou, G., Kuusisto, J., Langenberg, C., Langford, C., Leander, K., Lokki, M. L., Lundmark, A., McCarthy, M. I., Meisinger, C., Melander, O., Mihailov, E., Maouche, S., Morris, A. D., Muller-Nurasyid, M., Mu, T. C., Nikus, K., Peden, J. F., Rayner, N. W., Rasheed, A., Rosinger, S., Rubin, D., Rumpf, M. P., Schafer, A., Sivananthan, M., Song, C., Stewart, A. F., Tan, S. T., Thorgeirsson, G., van der Schoot, C. E., Wagner, P. J., et al. 2013.

Large-scale association analysis identifies new risk loci for coronary artery disease. *Nat Genet,* 45**,** 25-33.
Clarke, G. M., Anderson, C. A., Pettersson, F. H., Cardon, L. R., Morris, A. P. & Zondervan, K. T. 2011. Basic statistical analysis in genetic case-control studies. *Nat Protoc,* 6**,** 121-33.
Clarke, R., Peden, J. F., Hopewell, J. C., Kyriakou, T., Goel, A., Heath, S. C., Parish, S., Barlera, S., Franzosi, M. G., Rust, S., Bennett, D., Silveira, A., Malarstig, A., Green, F. R., Lathrop, M., Gigante, B., Leander, K., de Faire, U., Seedorf, U., Hamsten, A., Collins, R., Watkins, H. & Farrall, M. 2009. Genetic variants associated with Lp(a) lipoprotein level and coronary disease. *N Engl J Med,* 361**,** 2518-28.
Cookson, W., Liang, L., Abecasis, G., Moffatt, M. & Lathrop, M. 2009. Mapping complex disease traits with global gene expression. *Nat Rev Genet,* 10**,** 184-94.
Davies, G., Armstrong, N., Bis, J. C., Bressler, J., Chouraki, V., Giddaluru, S., Hofer, E., Ibrahim-Verbaas, C. A., Kirin, M., Lahti, J., van der Lee, S. J., Le Hellard, S., Liu, T., Marioni, R. E., Oldmeadow, C., Postmus, I., Smith, A. V., Smith, J. A., Thalamuthu, A., Thomson, R., Vitart, V., Wang, J., Yu, L., Zgaga, L., Zhao, W., Boxall, R., Harris, S. E., Hill, W. D., Liewald, D. C., Luciano, M., Adams, H., Ames, D., Amin, N., Amouyel, P., Assareh, A. A., Au, R., Becker, J. T., Beiser, A., Berr, C., Bertram, L., Boerwinkle, E., Buckley, B. M., Campbell, H., Corley, J., De Jager, P. L., Dufouil, C., Eriksson, J. G., Espeseth, T., Faul, J. D., Ford, I., Gottesman, R. F., Griswold, M. E., Gudnason, V., Harris, T. B., Heiss, G., Hofman, A., Holliday, E. G., Huffman, J., Kardia, S. L., Kochan, N., Knopman, D. S., Kwok, J. B., Lambert, J. C., Lee, T., Li, G., Li, S. C., Loitfelder, M., Lopez, O. L., Lundervold, A. J., Lundqvist, A., Mather, K. A., Mirza, S. S., Nyberg, L., Oostra, B. A., Palotie, A., Papenberg, G., Pattie, A., Petrovic, K., Polasek, O., Psaty, B. M., Redmond, P., Reppermund, S., Rotter, J. I., Schmidt, H., Schuur, M., Schofield, P. W., Scott, R. J., Steen, V. M., Stott, D. J., van Swieten, J. C., Taylor, K. D., Trollor, J., Trompet, S., Uitterlinden, A. G., Weinstein, G., Widen, E., Windham, B. G., Jukema, J. W., Wright, A. F., Wright, M. J., et al. 2015. Genetic contributions to variation in general cognitive function: a meta-analysis of genome-wide association studies in the CHARGE consortium (N=53949). *Mol Psychiatry,* 20**,** 183-92.
Dudbridge, F. & Gusnanto, A. 2008. Estimation of significance thresholds for genomewide association scans. *Genet Epidemiol,* 32**,** 227-34.
Efron, B. 2008. Microarrays, empirical Bayes and the two-groups model. *Stat Sci,* 23**,** 1-22.
Efron, B., Storey, J. D. & Tibshirani, R. 2001. *Microarrays empirical bayes methods, and false discovery rates*, Department of Statistics, Stanford University.
Efron, B. & Tibshirani, R. 2002. Empirical Bayes methods and false discovery rates for microarrays. *Genet Epidemiol,* 23**,** 70-86.
Ehret, G. B., Munroe, P. B., Rice, K. M., Bochud, M., Johnson, A. D., Chasman, D. I., Smith, A. V., Tobin, M. D., Verwoert, G. C., Hwang, S. J., Pihur, V., Vollenweider, P., O'Reilly, P. F., Amin, N., Bragg-Gresham, J. L., Teumer, A., Glazer, N. L., Launer, L., Zhao, J. H., Aulchenko, Y., Heath, S., Sober, S., Parsa, A., Luan, J., Arora, P., Dehghan, A., Zhang, F., Lucas, G., Hicks, A. A., Jackson, A. U., Peden, J. F., Tanaka, T., Wild, S. H., Rudan, I., Igl, W., Milaneschi, Y., Parker, A. N., Fava, C., Chambers, J. C., Fox, E. R., Kumari, M., Go, M. J., van

der Harst, P., Kao, W. H., Sjogren, M., Vinay, D. G., Alexander, M., Tabara, Y., Shaw-Hawkins, S., Whincup, P. H., Liu, Y., Shi, G., Kuusisto, J., Tayo, B., Seielstad, M., Sim, X., Nguyen, K. D., Lehtimaki, T., Matullo, G., Wu, Y., Gaunt, T. R., Onland-Moret, N. C., Cooper, M. N., Platou, C. G., Org, E., Hardy, R., Dahgam, S., Palmen, J., Vitart, V., Braund, P. S., Kuznetsova, T., Uiterwaal, C. S., Adeyemo, A., Palmas, W., Campbell, H., Ludwig, B., Tomaszewski, M., Tzoulaki, I., Palmer, N. D., Aspelund, T., Garcia, M., Chang, Y. P., O'Connell, J. R., Steinle, N. I., Grobbee, D. E., Arking, D. E., Kardia, S. L., Morrison, A. C., Hernandez, D., Najjar, S., McArdle, W. L., Hadley, D., Brown, M. J., Connell, J. M., Hingorani, A. D., Day, I. N., Lawlor, D. A., Beilby, J. P., Lawrence, R. W., Clarke, R., et al. 2011. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. *Nature,* 478**,** 103-9.

Eichler, E. E., Flint, J., Gibson, G., Kong, A., Leal, S. M., Moore, J. H. & Nadeau, J. H. 2010. Missing heritability and strategies for finding the underlying causes of complex disease. *Nat Rev Genet,* 11**,** 446-50.

Ferkingstad, E., Frigessi, A., Rue, H. v., Thorleifsson, G. & Kong, A. 2008. Unsupervised empirical Bayesian multiple testing with external covariates. *Annal Appl Stat***,** 714-735.

Galesloot, T. E., van Steen, K., Kiemeney, L. A., Janss, L. L. & Vermeulen, S. H. 2014. A comparison of multivariate genome-wide association methods. *PLoS One,* 9**,** e95923.

Gibbs, R. A., Belmont, J. W., Hardenbol, P., Willis, T. D., Yu, F., Yang, H., Chang, L.-Y., Huang, W., Liu, B. & Shen, Y. 2003. The international HapMap project. *Nature,* 426**,** 789-796.

Gilad, Y., Rifkin, S. A. & Pritchard, J. K. 2008. Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends Genet,* 24**,** 408-15.

Global Lipids Genetics Consortium 2013. Discovery and refinement of loci associated with lipid levels. *Nat Genet,* 45**,** 1274-1283.

IBC 50K CAD Consortium 2011. Large-scale gene-centric analysis identifies novel variants for coronary artery disease. *PLoS Genet,* 7**,** e1002260.

Ibrahim-Verbaas, C. A., Bressler, J., Debette, S., Schuur, M., Smith, A. V., Bis, J. C., Davies, G., Trompet, S., Smith, J. A., Wolf, C., Chibnik, L. B., Liu, Y., Vitart, V., Kirin, M., Petrovic, K., Polasek, O., Zgaga, L., Fawns-Ritchie, C., Hoffmann, P., Karjalainen, J., Lahti, J., Llewellyn, D. J., Schmidt, C. O., Mather, K. A., Chouraki, V., Sun, Q., Resnick, S. M., Rose, L. M., Oldmeadow, C., Stewart, M., Smith, B. H., Gudnason, V., Yang, Q., Mirza, S. S., Jukema, J. W., deJager, P. L., Harris, T. B., Liewald, D. C., Amin, N., Coker, L. H., Stegle, O., Lopez, O. L., Schmidt, R., Teumer, A., Ford, I., Karbalai, N., Becker, J. T., Jonsdottir, M. K., Au, R., Fehrmann, R. S., Herms, S., Nalls, M., Zhao, W., Turner, S. T., Yaffe, K., Lohman, K., van Swieten, J. C., Kardia, S. L., Knopman, D. S., Meeks, W. M., Heiss, G., Holliday, E. G., Schofield, P. W., Tanaka, T., Stott, D. J., Wang, J., Ridker, P., Gow, A. J., Pattie, A., Starr, J. M., Hocking, L. J., Armstrong, N. J., McLachlan, S., Shulman, J. M., Pilling, L. C., Eiriksdottir, G., Scott, R. J., Kochan, N. A., Palotie, A., Hsieh, Y. C., Eriksson, J. G., Penman, A., Gottesman, R. F., Oostra, B. A., Yu, L., DeStefano, A. L., Beiser, A., Garcia, M., Rotter, J. I., Nothen, M. M., Hofman, A., Slagboom, P. E., Westendorp, R. G., Buckley, B. M., Wolf, P. A., Uitterlinden, A. G., Psaty, B. M., Grabe, H. J., Bandinelli, S., Chasman, D. I., et al. 2016. GWAS for executive function and

processing speed suggests involvement of the CADM2 gene. *Mol Psychiatry* 21, 189-97.
International HapMap Consortiu,. 2005. A haplotype map of the human genome. *Nature,* 437**,** 1299-1320.
Kathiresan, S., Voight, B. F., Purcell, S., Musunuru, K., Ardissino, D., Mannucci, P. M., Anand, S., Engert, J. C., Samani, N. J., Schunkert, H., Erdmann, J., Reilly, M. P., Rader, D. J., Morgan, T., Spertus, J. A., Stoll, M., Girelli, D., McKeown, P. P., Patterson, C. C., Siscovick, D. S., O'Donnell, C. J., Elosua, R., Peltonen, L., Salomaa, V., Schwartz, S. M., Melander, O., Altshuler, D., Merlini, P. A., Berzuini, C., Bernardinelli, L., Peyvandi, F., Tubaro, M., Celli, P., Ferrario, M., Fetiveau, R., Marziliano, N., Casari, G., Galli, M., Ribichini, F., Rossi, M., Bernardi, F., Zonzin, P., Piazza, A., Yee, J., Friedlander, Y., Marrugat, J., Lucas, G., Subirana, I., Sala, J., Ramos, R., Meigs, J. B., Williams, G., Nathan, D. M., MacRae, C. A., Havulinna, A. S., Berglund, G., Hirschhorn, J. N., Asselta, R., Duga, S., Spreafico, M., Daly, M. J., Nemesh, J., Korn, J. M., McCarroll, S. A., Surti, A., Guiducci, C., Gianniny, L., Mirel, D., Parkin, M., Burtt, N., Gabriel, S. B., Thompson, J. R., Braund, P. S., Wright, B. J., Balmforth, A. J., Ball, S. G., Hall, A., Linsel-Nitschke, P., Lieb, W., Ziegler, A., Konig, I., Hengstenberg, C., Fischer, M., Stark, K., Grosshennig, A., Preuss, M., Wichmann, H. E., Schreiber, S., Ouwehand, W., Deloukas, P., Scholz, M., Cambien, F., Li, M., Chen, Z., Wilensky, R., Matthai, W., Qasim, A., Hakonarson, H. H., Devaney, J., Burnett, M. S., et al. 2009. Genome-wide association of early-onset myocardial infarction with single nucleotide polymorphisms and copy number variants. *Nat Genet,* 41**,** 334-41.
LeBlanc, M., Kulle, B., Sundet, K., Agartz, I., Melle, I., Djurovic, S., Frigessi, A. & Andreassen, O. A. 2012. Genome-wide study identifies PTPRO and WDR72 and FOXQ1-SUMO1P1 interaction associated with neurocognitive function. *J Psychiatr Res,* 46**,** 271-8 (***Paper 2***).
Lee, T., Henry, J. D., Trollor, J. N. & Sachdev, P. S. 2010. Genetic influences on cognitive functions in the elderly: a selective review of twin studies. *Brain Res Rev,* 64**,** 1-13.
Liley, J. & Wallace, C. 2015. A Pleiotropy-Informed Bayesian False Discovery Rate adapted to a Shared Control Design Finds New Disease Associations From GWAS Summary Statistics. *PLoS Genet,* 11**,** e1004926-e1004926.
Lin, D. Y. & Sullivan, P. F. 2009. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet,* 85**,** 862-72.
Loo, S. K., Shtir, C., Doyle, A. E., Mick, E., McGough, J. J., McCracken, J., Biederman, J., Smalley, S. L., Cantor, R. M., Faraone, S. V. & Nelson, S. F. 2012. Genome-wide association study of intelligence: additive effects of novel brain expressed genes. *J Am Acad Child Adolesc Psych,* 51**,** 432-440 e2.
Manolio, T. A., Collins, F. S., Cox, N. J., Goldstein, D. B., Hindorff, L. A., Hunter, D. J., McCarthy, M. I., Ramos, E. M., Cardon, L. R., Chakravarti, A., Cho, J. H., Guttmacher, A. E., Kong, A., Kruglyak, L., Mardis, E., Rotimi, C. N., Slatkin, M., Valle, D., Whittemore, A. S., Boehnke, M., Clark, A. G., Eichler, E. E., Gibson, G., Haines, J. L., Mackay, T. F., McCarroll, S. A. & Visscher, P. M. 2009. Finding the missing heritability of complex diseases. *Nature,* 461**,** 747-53.
Messerli, F. H., Williams, B. & Ritz, E. 2007. Essential hypertension. *Lancet,* 370**,** 591-603.

Nicolae, D. L., Gamazon, E., Zhang, W., Duan, S., Dolan, M. E. & Cox, N. J. 2010. Trait-associated SNPs are more likely to be eQTLs: annotation to enhance discovery from GWAS. *PLoS Genet,* 6**,** e1000888.

Papassotiropoulos, A., Stephan, D. A., Huentelman, M. J., Hoerndli, F. J., Craig, D. W., Pearson, J. V., Huynh, K. D., Brunner, F., Corneveaux, J., Osborne, D., Wollmer, M. A., Aerni, A., Coluccia, D., Hanggi, J., Mondadori, C. R., Buchmann, A., Reiman, E. M., Caselli, R. J., Henke, K. & de Quervain, D. J. 2006. Common Kibra alleles are associated with human memory performance. *Science,* 314**,** 475-8.

Pe'er, I., Yelensky, R., Altshuler, D. & Daly, M. J. 2008. Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol,* 32**,** 381-5.

Peden, J. F. & Farrall, M. 2011. Thirty-five common variants for coronary artery disease: the fruits of much collaborative labour. *Hum Mol Genet,* 20**,** R198-205.

Pritchard, J. K. & Cox, N. J. 2002. The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet,* 11**,** 2417-23.

Province, M. A. & Borecki, I. B. 2013. A correlated meta-analysis strategy for data mining "OMIC" scans. *Pac Symp Biocomput***,** 236-46.

Quertermous, T. & Ingelsson, E. 2016. Coronary Artery Disease and Its Risk Factors Leveraging Shared Genetics to Discover Novel Biology. *Circ Res,* 118**,** 14-16.

Reich, D. E., Cargill, M., Bolk, S., Ireland, J., Sabeti, P. C., Richter, D. J., Lavery, T., Kouyoumjian, R., Farhadian, S. F., Ward, R. & Lander, E. S. 2001. Linkage disequilibrium in the human genome. *Nature,* 411**,** 199-204.

Reich, D. E. & Lander, E. S. 2001. On the allelic spectrum of human disease. *Trends Genet,* 17**,** 502-10.

Samani, N. J., Erdmann, J., Hall, A. S., Hengstenberg, C., Mangino, M., Mayer, B., Dixon, R. J., Meitinger, T., Braund, P., Wichmann, H. E., Barrett, J. H., Konig, I. R., Stevens, S. E., Szymczak, S., Tregouet, D. A., Iles, M. M., Pahlke, F., Pollard, H., Lieb, W., Cambien, F., Fischer, M., Ouwehand, W., Blankenberg, S., Balmforth, A. J., Baessler, A., Ball, S. G., Strom, T. M., Braenne, I., Gieger, C., Deloukas, P., Tobin, M. D., Ziegler, A., Thompson, J. R. & Schunkert, H. 2007. Genomewide association analysis of coronary artery disease. *N Engl J Med,* 357**,** 443-53.

Schork, A. J., Thompson, W. K., Pham, P., Torkamani, A., Roddey, J. C., Sullivan, P. F., Kelsoe, J. R., O'Donovan, M. C., Furberg, H., Schork, N. J., Andreassen, O. A. & Dale, A. M. 2013. All SNPs are not created equal: genome-wide association studies reveal a consistent pattern of enrichment among functionally annotated SNPs. *PLoS Genet,* 9**,** e1003449.

Schunkert, H., Konig, I. R., Kathiresan, S., Reilly, M. P., Assimes, T. L., Holm, H., Preuss, M., Stewart, A. F., Barbalic, M., Gieger, C., Absher, D., Aherrahrou, Z., Allayee, H., Altshuler, D., Anand, S. S., Andersen, K., Anderson, J. L., Ardissino, D., Ball, S. G., Balmforth, A. J., Barnes, T. A., Becker, D. M., Becker, L. C., Berger, K., Bis, J. C., Boekholdt, S. M., Boerwinkle, E., Braund, P. S., Brown, M. J., Burnett, M. S., Buysschaert, I., Carlquist, J. F., Chen, L., Cichon, S., Codd, V., Davies, R. W., Dedoussis, G., Dehghan, A., Demissie, S., Devaney, J. M., Diemert, P., Do, R., Doering, A., Eifert, S., Mokhtari, N. E., Ellis, S. G., Elosua, R., Engert, J. C., Epstein, S. E., de Faire, U., Fischer, M.,

Folsom, A. R., Freyer, J., Gigante, B., Girelli, D., Gretarsdottir, S., Gudnason, V., Gulcher, J. R., Halperin, E., Hammond, N., Hazen, S. L., Hofman, A., Horne, B. D., Illig, T., Iribarren, C., Jones, G. T., Jukema, J. W., Kaiser, M. A., Kaplan, L. M., Kastelein, J. J., Khaw, K. T., Knowles, J. W., Kolovou, G., Kong, A., Laaksonen, R., Lambrechts, D., Leander, K., Lettre, G., Li, M., Lieb, W., Loley, C., Lotery, A. J., Mannucci, P. M., Maouche, S., Martinelli, N., McKeown, P. P., Meisinger, C., Meitinger, T., Melander, O., Merlini, P. A., Mooser, V., Morgan, T., Muhleisen, T. W., Muhlestein, J. B., Munzel, T., Musunuru, K., Nahrstaedt, J., Nelson, C. P., Nothen, M. M., Olivieri, O., et al. 2011. Large-scale association analysis identifies 13 new susceptibility loci for coronary artery disease. *Nat Genet,* 43**,** 333-8.

Soranzo, N., Spector, T. D., Mangino, M., Kuhnel, B., Rendon, A., Teumer, A., Willenborg, C., Wright, B., Chen, L., Li, M., Salo, P., Voight, B. F., Burns, P., Laskowski, R. A., Xue, Y., Menzel, S., Altshuler, D., Bradley, J. R., Bumpstead, S., Burnett, M. S., Devaney, J., Doring, A., Elosua, R., Epstein, S. E., Erber, W., Falchi, M., Garner, S. F., Ghori, M. J., Goodall, A. H., Gwilliam, R., Hakonarson, H. H., Hall, A. S., Hammond, N., Hengstenberg, C., Illig, T., Konig, I. R., Knouff, C. W., McPherson, R., Melander, O., Mooser, V., Nauck, M., Nieminen, M. S., O'Donnell, C. J., Peltonen, L., Potter, S. C., Prokisch, H., Rader, D. J., Rice, C. M., Roberts, R., Salomaa, V., Sambrook, J., Schreiber, S., Schunkert, H., Schwartz, S. M., Serbanovic-Canic, J., Sinisalo, J., Siscovick, D. S., Stark, K., Surakka, I., Stephens, J., Thompson, J. R., Volker, U., Volzke, H., Watkins, N. A., Wells, G. A., Wichmann, H. E., Van Heel, D. A., Tyler-Smith, C., Thein, S. L., Kathiresan, S., Perola, M., Reilly, M. P., Stewart, A. F., Erdmann, J., Samani, N. J., Meisinger, C., Greinacher, A., Deloukas, P., Ouwehand, W. H. & Gieger, C. 2009. A genome-wide meta-analysis identifies 22 loci associated with eight hematological parameters in the HaemGen consortium. *Nat Genet,* 41**,** 1182-90.

Spencer, C. C., Su Z Fau - Donnelly, P., Donnelly P Fau - Marchini, J. & Marchini, J. 2009. Designing genome-wide association studies: sample size, power, imputation, and the choice of genotyping chip. *PLoS Genet*, e1000477.

Storey, J. D. 2002. A direct approach to false discovery rates. *J R Stat Soc Series B Stat Methodol,* 64**,** 479-498.

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., Ingle, C. E., Dunning, M., Flicek, P., Koller, D., Montgomery, S., Tavare, S., Deloukas, P. & Dermitzakis, E. T. 2007. Population genomics of human gene expression. *Nat Genet,* 39**,** 1217-24.

Teslovich, T. M., Musunuru, K., Smith, A. V., Edmondson, A. C., Stylianou, I. M., Koseki, M., Pirruccello, J. P., Ripatti, S., Chasman, D. I., Willer, C. J., Johansen, C. T., Fouchier, S. W., Isaacs, A., Peloso, G. M., Barbalic, M., Ricketts, S. L., Bis, J. C., Aulchenko, Y. S., Thorleifsson, G., Feitosa, M. F., Chambers, J., Orho-Melander, M., Melander, O., Johnson, T., Li, X., Guo, X., Li, M., Shin Cho, Y., Jin Go, M., Jin Kim, Y., Lee, J. Y., Park, T., Kim, K., Sim, X., Twee-Hee Ong, R., Croteau-Chonka, D. C., Lange, L. A., Smith, J. D., Song, K., Hua Zhao, J., Yuan, X., Luan, J., Lamina, C., Ziegler, A., Zhang, W., Zee, R. Y., Wright, A. F., Witteman, J. C., Wilson, J. F., Willemsen, G., Wichmann, H. E., Whitfield, J. B., Waterworth, D. M., Wareham, N. J., Waeber, G., Vollenweider, P., Voight, B. F., Vitart, V., Uitterlinden, A. G., Uda, M., Tuomilehto, J., Thompson, J. R., Tanaka, T., Surakka, I., Stringham, H. M.,

Spector, T. D., Soranzo, N., Smit, J. H., Sinisalo, J., Silander, K., Sijbrands, E. J., Scuteri, A., Scott, J., Schlessinger, D., Sanna, S., Salomaa, V., Saharinen, J., Sabatti, C., Ruokonen, A., Rudan, I., Rose, L. M., Roberts, R., Rieder, M., Psaty, B. M., Pramstaller, P. P., Pichler, I., Perola, M., Penninx, B. W., Pedersen, N. L., Pattaro, C., Parker, A. N., Pare, G., Oostra, B. A., O'Donnell, C. J., Nieminen, M. S., Nickerson, D. A., Montgomery, G. W., Meitinger, T., McPherson, R., McCarthy, M. I., et al. 2010. Biological, clinical and population relevance of 95 loci for blood lipids. *Nature,* 466**,** 707-13.

Voight, B. F., Kang, H. M., Ding, J., Palmer, C. D., Sidore, C., Chines, P. S., Burtt, N. P., Fuchsberger, C., Li, Y., Erdmann, J., Frayling, T. M., Heid, I. M., Jackson, A. U., Johnson, T., Kilpelainen, T. O., Lindgren, C. M., Morris, A. P., Prokopenko, I., Randall, J. C., Saxena, R., Soranzo, N., Speliotes, E. K., Teslovich, T. M., Wheeler, E., Maguire, J., Parkin, M., Potter, S., Rayner, N. W., Robertson, N., Stirrups, K., Winckler, W., Sanna, S., Mulas, A., Nagaraja, R., Cucca, F., Barroso, I., Deloukas, P., Loos, R. J., Kathiresan, S., Munroe, P. B., Newton-Cheh, C., Pfeufer, A., Samani, N. J., Schunkert, H., Hirschhorn, J. N., Altshuler, D., McCarthy, M. I., Abecasis, G. R. & Boehnke, M. 2012. The metabochip, a custom genotyping array for genetic studies of metabolic, cardiovascular, and anthropometric traits. *PLoS Genet,* 8**,** e1002793.

Wang, F., Xu, C. Q., He, Q., Cai, J. P., Li, X. C., Wang, D., Xiong, X., Liao, Y. H., Zeng, Q. T., Yang, Y. Z., Cheng, X., Li, C., Yang, R., Wang, C. C., Wu, G., Lu, Q. L., Bai, Y., Huang, Y. F., Yin, D., Yang, Q., Wang, X. J., Dai, D. P., Zhang, R. F., Wan, J., Ren, J. H., Li, S. S., Zhao, Y. Y., Fu, F. F., Huang, Y., Li, Q. X., Shi, S. W., Lin, N., Pan, Z. W., Li, Y., Yu, B., Wu, Y. X., Ke, Y. H., Lei, J., Wang, N., Luo, C. Y., Ji, L. Y., Gao, L. J., Li, L., Liu, H., Huang, E. W., Cui, J., Jia, N., Ren, X., Li, H., Ke, T., Zhang, X. Q., Liu, J. Y., Liu, M. G., Xia, H., Yang, B., Shi, L. S., Xia, Y. L., Tu, X. & Wang, Q. K. 2011. Genome-wide association identifies a susceptibility locus for coronary artery disease in the Chinese Han population. *Nat Genet,* 43**,** 345-9.

Weiss, K. M. & Clark, A. G. 2002. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet,* 18**,** 19-24.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L. & Parkinson, H. 2014. The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res,* 42**,** D1001-6.

Westra, H. J., Peters, M. J., Esko, T., Yaghootkar, H., Schurmann, C., Kettunen, J., Christiansen, M. W., Fairfax, B. P., Schramm, K., Powell, J. E., Zhernakova, A., Zhernakova, D. V., Veldink, J. H., Van den Berg, L. H., Karjalainen, J., Withoff, S., Uitterlinden, A. G., Hofman, A., Rivadeneira, F., t Hoen, P. A., Reinmaa, E., Fischer, K., Nelis, M., Milani, L., Melzer, D., Ferrucci, L., Singleton, A. B., Hernandez, D. G., Nalls, M. A., Homuth, G., Nauck, M., Radke, D., Volker, U., Perola, M., Salomaa, V., Brody, J., Suchy-Dicey, A., Gharib, S. A., Enquobahrie, D. A., Lumley, T., Montgomery, G. W., Makino, S., Prokisch, H., Herder, C., Roden, M., Grallert, H., Meitinger, T., Strauch, K., Li, Y., Jansen, R. C., Visscher, P. M., Knight, J. C., Psaty, B. M., Ripatti, S., Teumer, A., Frayling, T. M., Metspalu, A., van Meurs, J. B. & Franke, L. 2013. Systematic identification of trans eQTLs as putative drivers of known disease associations. *Nat Genet,* 45**,** 1238-43.

Zablocki, R. W., Schork, A. J., Levine, R. A., Andreassen, O. A., Dale, A. M. & Thompson, W. K. 2014. Covariate-modulated local false discovery rate for genome-wide association studies. *Bioinformatics,* 30**,** 2098-104.

Zhu, X., Feng, T., Tayo, B. O., Liang, J., Young, J. H., Franceschini, N., Smith, J. A., Yanek, L. R., Sun, Y. V., Edwards, T. L., Chen, W., Nalls, M., Fox, E., Sale, M., Bottinger, E., Rotimi, C., Liu, Y., McKnight, B., Liu, K., Arnett, D. K., Chakravati, A., Cooper, R. S. & Redline, S. 2015. Meta-analysis of correlated traits via summary statistics from GWASs with an application in hypertension. *Am J Hum Genet,* 96**,** 21-36.

**Figures**

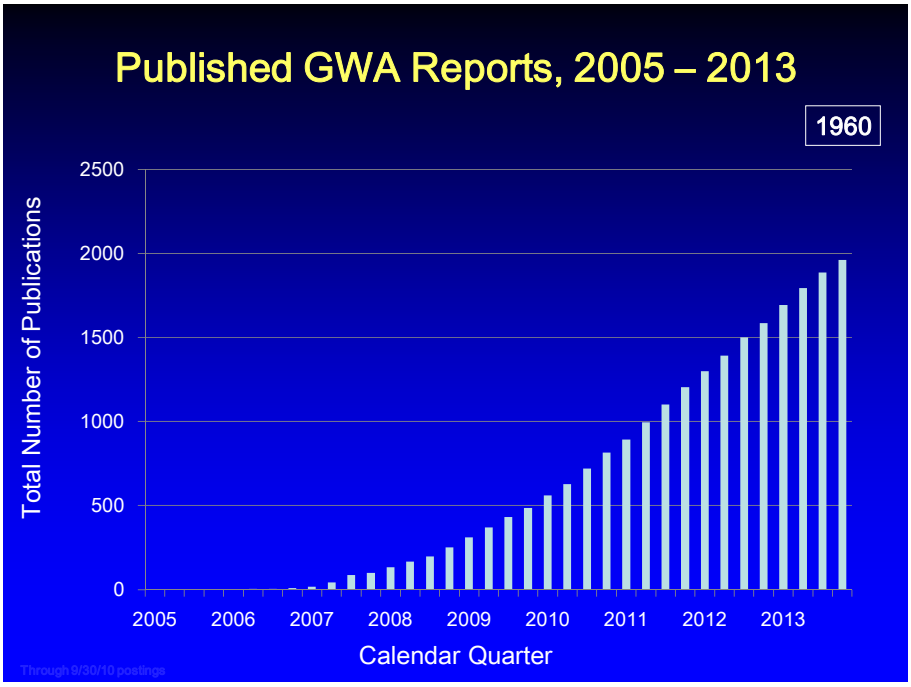


**Figure 1.** Published GWAS, 2005-2013. Figure from the Catalog of Published Genome-Wide Association Studies. Figure is downloaded from https://www.genome.gov/26525384 and the Catalog is currently maintained at http://www.ebi.ac.uk/gwas/ (Welter et al., 2014).

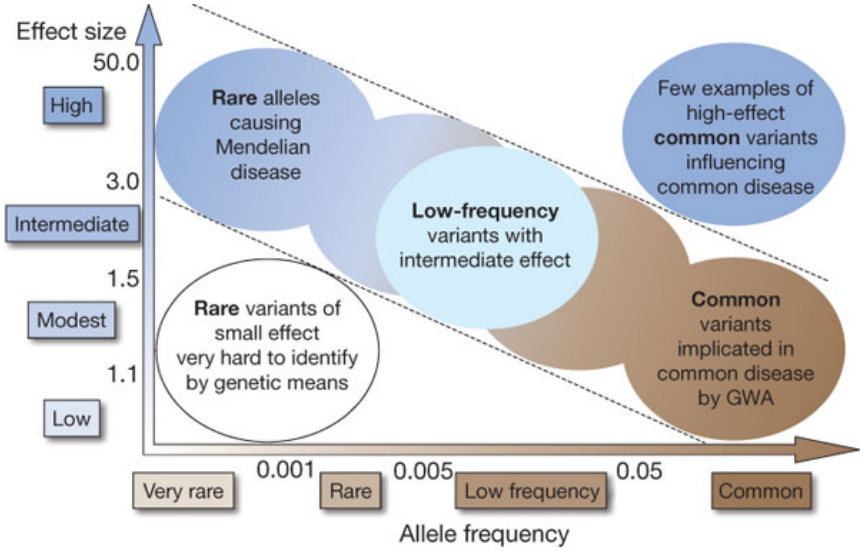


**Figure 2.** Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Figure from Manolio et al. (2009).

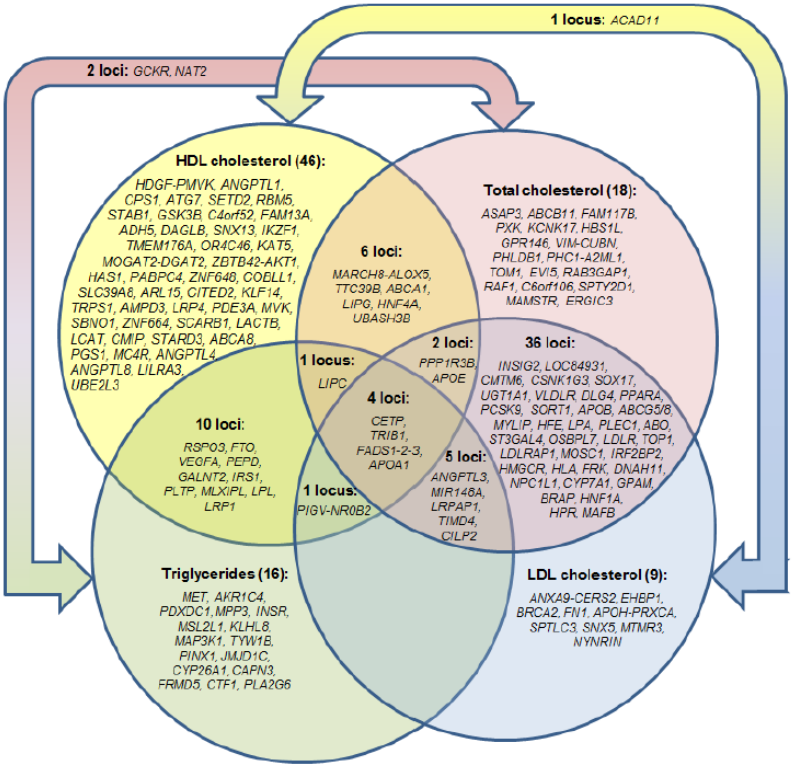


**Figure 3.** Venn diagram summarizing the genes annotated to genome-wide significant SNPs for four different outcomes in the Global Lipids Consortium GWAS. Figure from the Global Lipids Genetics Consortium (2013).

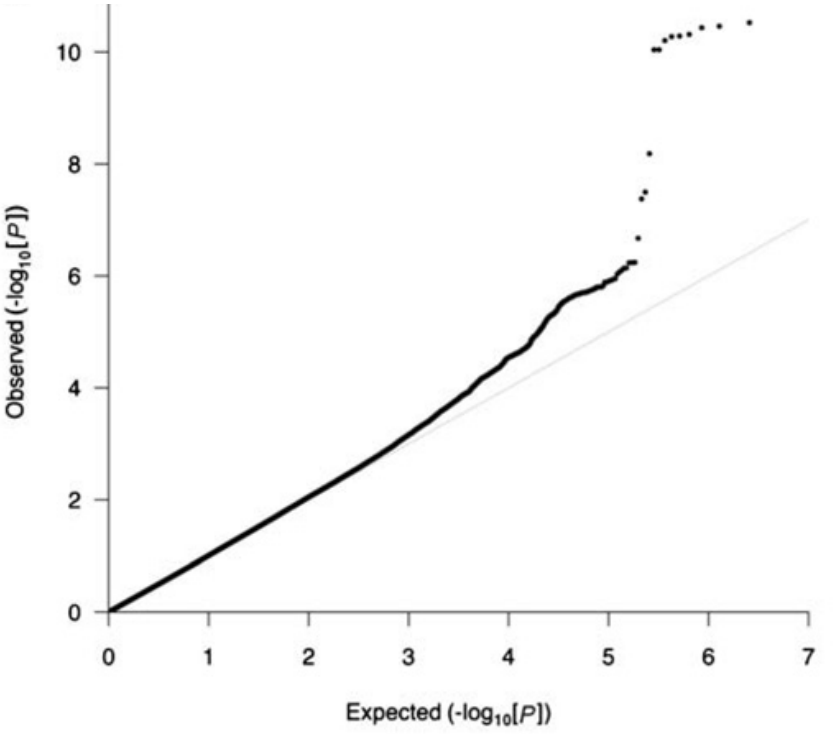


**Figure 4.** A typical quantile-quantile plot for a Consortium-based GWAS.

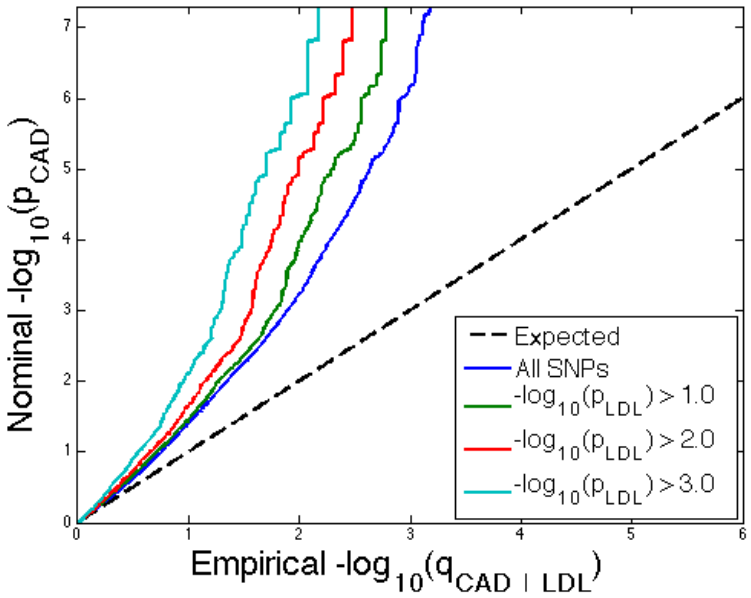


**Figure 5.** Stratified quantile-quantile plot. Conditional quantile-quantile plot of nominal versus empirical -$\log_{10}$ *p*-values in Coronary Artery Disease (CAD) as a function of significance of association with low density lipoprotein cholesterol (LDL). Dotted line indicates the null hypothesis.