Pleiotropy in complex traits

Sophie Hackinger

Wellcome Sanger Institute

University of Cambridge

Jesus College

This dissertation is submitted for the degree of Doctor of Philosophy

September 2018

Abstract

Thesis title: Pleiotropy in complex traits
Name: Sophie Hackinger

Genome-wide association studies (GWAS) have uncovered thousands of complex trait loci, many of which are associated with multiple phenotypes. The dedicated study of these pleiotropic effects is becoming increasingly common due to the availability of sample collections with high-dimensional phenotype data, such as the UK Biobank, and can yield important insights into the aetiology underlying complex disorders.

In my PhD, I performed multi-trait analyses of medically relevant complex phenotypes to identify shared genetic factors.

My first project involved a genome-wide overlap analysis of osteoarthritis (OA) and bone-mineral density (BMD), using summary statistics from two large-scale GWAS. OA and BMD are known to be inversely correlated, yet the genetics underlying this link remain poorly understood. I found robust evidence for association with OA at the *SMAD3* locus, which is known to play a role in bone remodeling and cartilage maintenance.

My second project aimed to elucidate the increased prevalence of type 2 diabetes (T2D) in schizophrenia (SCZ) patients. I used GWAS summary statistics of SCZ and T2D from the PGC and DIAGRAM consortia, respectively, to perform polygenic risk score analyses in three patient groups (SCZ only, T2D only, comorbid SCZ and T2D) and population-based controls. I find that the comorbid patient group have a higher genetic risk for both T2D and SCZ compared to controls, supporting the hypothesis that the epidemiologic link between these disorders is at least in part due to genetic factors.

In my third project, I leveraged the correlation structure of over 274 protein biomarkers and 57 quantitative traits to perform multivariate GWAS on correlated trait clusters in a Greek isolated population. This approach uncovered several novel cis-associations not identified in single-trait GWAS, and highlights the power advantage of multivariate analysis.

An important consideration for future studies will be the interpretation and follow-up of cross-phenotype associations, and the translation of these insights into clinical use.

Declaration

I hereby declare that the work described in this thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration, except where otherwise stated.

It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text

This thesis does not exceed 60,000 words (excluding Abstract, Table of Contents, Figure and Tables and References) as prescribed Degree Committee for the Faculty of Biology at the University of Cambridge.

Sophie Hackinger
September 2018

"Let's not try to figure out everything at once."

– Matt Berninger

Acknowledgements

First and foremost, I owe an immeasurably large “Thank you” to my supervisor, Prof Ele Zeggini, for nearly four years of support, guidance and encouragement.

I also owe my gratitude to my thesis committee members Carl Anderson and Angela Wood, whose feedback helped to shape the projects described here.

Since this is the only section of a thesis that generally has a realistic chance of attracting a readership > 4, I will try to keep the rest of it light and perhaps even moderately entertaining:

A big thank you to past members of Team144, with special shout-outs to Kostas (<3), for being “the normal guy”; Bram, for many meme-rable moments; and Arthur, whose French I will continue to excuse and who always “know what do”. Of course, this list would not be complete without wanna-be-full-time-Team144-member and excellent Mariachi player, Fernando, who sits behind the wall.

Casting the net of gratitude past N333, I also have to thank Loukas, for verbal sparring; Dan (de Lion), for many punderful conversations; Alex, for keeping the spice flowing; Joanna, for being a very reliable provider of procrastination material, as well as loyal consumer of my baking experiments; and José, for the daily embodiment of the “Morgan building spirit”.

Now moving beyond sunny Hinxton (in random order):

Thank you to Bianca and Nita, my oldest friends, who are always in my corner and – somewhat miraculously – still have not gotten bored of me. Nita also deserves a special thank you for being the link to my non-genetic family – Agron, Aferdita, Rron and Vesa. I would not be the person I am today without them (and how many people can say they are lucky enough to have two mums and two dads?). I am of course immensely grateful to my genetic parents for unconditional support and for supplying me with: my DNA (mum & dad), curiosity and

an appetite for adventure (mum), culinary curiosity and an adventurous appetite (dad), as well as the super-power of irony (also dad). A big thank you to Hannah, for being the best partner in crime and 50% of the Dubious Duo; Sun-Gou, who lives on in the memory of past and present Anderson team members (as well as across the pond, in Boston), for being the signal that drowns out the noise (Tejas); and my late cat Sina, for providing almost 18 years of furry stress-relief.

A special paragraph devoid of irony and jokes needs to be dedicated to the people who laid the intellectual foundation this thesis is built on: my former high school teachers. In particular, Peter Pail, whose eight years of English classes not only equipped me with the linguistic skills to undertake this doctorate, but also challenged and helped develop my capacity for critical thinking. I am honoured to have had him as my mentor.

Last but not least, I would like to thank Scott and Brian Devendorf, Aaron and Bryce Dessner, and Matt Berninger, for providing the soundtrack to this PhD, and to my life in general.

Publications

From this thesis:

Hackinger S, Prins B, Mamakou V, Zengini E, Marouli E, et al. Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia. *Transl Psych.* 2018 23;8(1):252

Hackinger S, Zeggini E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol.* 2017; 7(11).

Hackinger S, Trajanoska K, Styrkarsdottir U, Zengini E, Steinberg J, Ritchie GRS, et al. Evaluation of shared genetic aetiology between osteoarthritis and bone mineral density identifies SMAD3 as a novel osteoarthritis risk locus. *Hum Mol Genet.* 2017; 26(19):3850-3858.

Arising elsewhere:

Mahajan A, Taliun D, Thurner M, Robertson NR, Torres JM, [...], **Hackinger S**, Hattersley AT, Herder C, Ikram MA, Ingelsson M. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat Gene*t. 2018;50(11):1505-1513

Mahajan A, Wessel J, Willems SM, Zhao W, Robertson NR, Chu AY, […], **Hackinger S**, Hai Y, Han S, Tybjærg-Hansen A, et al. Refining the accuracy of validated target identification through coding variant fine-mapping in type 2 diabetes. *Nat Genet* 2018; 50(4):559-571.

Mamakou V, **Hackinger S**, Zengini E, Tsompanaki E, Marouli E, Serafetinidis I, et al. Combination therapy as a potential risk factor for the development of type 2 diabetes in patients with schizophrenia: the GOMAP study. *BMC Psychiatry* 2018; 18(1):249.

Zengini E, Hatzikotoulas K, Tachmazidou I, Steinberg J, Hartwig FP, Southam L, **Hackinger S,** et al. Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis. *Nature Genetics* 2018:1.

Casalone E, Tachmazidou I, Zengini E, Hatzikotoulas K, **Hackinger S**, Suveges D, et al. A novel variant in GLIS3 is associated with osteoarthritis. *Ann Rheum Dis* 2018; 77(4):620-623.

Marouli E, Kanoni S, Mamakou V, **Hackinger S**, Southam L, Prins B, et al. Evaluating the glucose raising effect of established loci via a genetic risk score. *PLoS One* 2017; 12(11):e0186669.

Hackinger S, Kraaijenbrink T, Xue Y, Mezzavilla M, Asan, van Driem G, et al. Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas. *Hum Genet* 2016; 135(4):393-402.

Table of Contents

List of Figures

List of Tables

Abbreviations

BMD	Bone mineral density
bp	base pair
CCA	Canonical correlation analysis
CI	Confidence interval
EA(F)	Effect allele (frequency)
eQTL	Expression quantitative trait locus
GO	Gene Ontology
GRCh	Genome Reference Consortium (human)
GWAS	Genome-wide association study
HWE	Hardy-Weinberg equilibirum
ICD-10	International Statistical Classification of Diseases and Related Health Problems, 10th Revision
Indel	Insertion/deletion
kb	kilobases
LD	Linkage disequilibrium
LMM	Linear mixed model
MAF	Minor allele frequency
Mb	megabases
MDS	Multi-dimensional scaling
MR	Mendelian randomisation
NEA	Non-effect allele
OA	Osteoarthritis
OR	Odds ratio
PC(A)	Principal component (analysis)
pQTL	Protein quantitative trait locus
PRS	Polygenic risk scores
QC	quality control
QTL	Quantitative trait locus
SCZ	Schizophrenia
SE	Standard error
SNP	Single nucleotide polymorphism
T2D	Type 2 diabetes
WES	Whole exome sequencing
WGS	Whole genome sequencing

Chapter 1 – Introduction

Since their inception in the early 2000s[1-3], genome-wide association studies (GWAS) have become the tool of choice for complex trait analysis. In the classical GWAS approach, the association of genetic variants across the entire genome with a single phenotype of interest is tested in a group of individuals. However, genes exert their function not as stand-alone units, but within complex networks of biomolecules that are often redundantly regulated. Likewise, phenotypes are interconnected by shared genetic and environmental factors. In recognition of this, and due to the increased availability of large-scale datasets, recent years have seen a shift towards the joint analysis of related phenotypes. Studies aiming to identify cross-phenotype associations can shed light onto the aetiology underlying epidemiologically related traits, or even uncover hitherto unknown links between traits that have seemingly very little in common. Additionally, they could provide important insights into genes and pathways specific to certain disease subtypes, possibly leading to more accurate disease classifications.

1.1 Advances and challenges in human genetics

1.1.1 The genetic basis of human disease

The field of human genetics aims to elucidate how genetic variation affects differences in phenotypes. Understanding the basis of heritable traits requires three main pieces of information: the number of genetic variants affecting the trait; the magnitude of their effects; and their frequency at a population level[4]. Together, these factors constitute the genetic architecture of a trait[4].

Originally, human diseases were categorised either as 'monogenic' (or 'Mendelian'), meaning that a single gene or mutation explains almost the entire variation in phenotype, with little to no environmental contribution[5]; or as 'polygenic', meaning that many (hundreds or thousands) genetic variants each contribute a fraction of the total genetic risk, together with environmental factors[6]. The mutations leading to Mendelian diseases typically lie within protein-coding regions and are therefore less common, as they will have been subject to purifying selection. Extending this line of reasoning to complex traits led to the 'common disease-common variant' hypothesis[6]: since complex disorders often have a late age of onset (i.e. post-adolescence), they have a comparatively small impact on reproductive fitness; consequently, variants with small risk-increasing effects should not be negatively selected against and over time will have risen in frequency at the population level.

The dichotomised view of mono- versus polygenic has shifted in past years as scientists began to take stock of the vast amount of data generated by genetic association studies: for many traits, common variants only explained a small amount of the total estimated heritability, even though the datasets used obtain these estimates were well-powered[4]. Furthermore, sequencing studies have identified rare variants of large effect[7-10] that contribute to complex traits in addition to common variants of small effect[11-13]. It is now believed that most human traits lie on a spectrum ranging from mono- to oligo- to polygenic, with both common and rare variants affecting phenotypic variation.

1.1.2 From linkage to GWAS

Before whole-genome genotyping of large sample sizes became feasible, linkage studies based on family data were the mainstay of human genetics research. Several study designs exist for the analysis of family data, including parent-offspring trios, extended pedigrees and affected sibling pairs[14]. While analysis methods differ depending on the design chosen, the basic premise is to test for statistically significant co-segregation ('linkage') of a trait of interest with genetic markers. Having become an established method for genetic mapping in model organisms, linkage studies were first used in humans in the 1980s and led to the successful identification of mutations responsible for Huntingdon's disease[15] and cystic fibrosis[16]. Despite these successes, the method soon proved inadequate for the mapping of loci for common diseases, for which the risk to unaffected relatives is lower than in Mendelian disorders[17]. In order to overcome this obstacle, complex disease research moved towards the use of case-control data, in which allele frequencies are compared between affected and unaffected individuals. Although the collection of unrelated individuals is less cumbersome than that of family data, initial efforts were hampered by the availability of genetic data: the systematic genotyping of whole genomes was not yet possible even in modest sample sizes, and studies therefore focused on candidate genes suspected to be involved in disease pathogenesis based on the function of their gene products[17]. Unfortunately, this approach was largely unsuccessful and most published findings were irreproducible[18].

Around the turn of the millennium massively parallel genotyping using microarray chips became feasible[17, 19, 20]. Due to the correlation of alleles at nearby variants (linkage disequilibrium), a map of 500,000 variants across the genome is sufficient to capture over 90% of the genetic variation in non-African populations[17]. This made it possible to conduct genome-wide association studies of complex traits, which abolished the need to have an *a priori* hypothesis about which genes or regions might harbour risk variants.

Early GWAS arrays were aimed primarily at common variants with minor allele frequencies (MAF) > 5%. However, it soon became clear that this end of the frequency spectrum only explained a fraction of the heritability of most traits and diseases studied[21]. Whole-genome (WGS) and whole-exome sequencing (WES) provide a more complete picture of an individual's allelic landscape, but were prohibitively expensive to be carried out routinely for large sample sizes[22, 23]. A workaround to this problem was the development of genotype imputation algorithms, which could predict genotypes not directly typed based on a reference panel of sequenced samples[24-26]. Within one decade GWAS had successfully identified variants contributing to numerous disease traits, including autoimmune, psychiatric and metabolic disorders, as well as quantitative traits such as anthropometric measurements or blood metabolite levels[11, 12, 27-33].

More recently, sequencing-based GWAS have shed light onto rare variant contributions to common complex traits[7, 8, 34, 35].

1.2 Pleiotropy in the GWAS era

In 2011, a systematic evaluation of associations reported in the NIHGR GWAS catalogue found that 4.6% of variants were associated with more than one trait[36]. This number is likely to have grown, as GWAS signals have been continuously added to the database.

Many cross-phenotype effects are not surprising. For example, variants in the *DSP* gene are associated with chronic obstructive pulmonary disease, as well as pulmonary fibrosis and lung function traits[37]. Others are perhaps less intuitive and can shed light into hitherto unknown connections between traits: variants in the *ASTN2* gene have been shown to affect both risk to osteoarthritis[38] and migraine[39, 40]. These seemingly unrelated diseases might share pathways involved in pain perception.

Until a few years ago, the focus of many consortia was to combine datasets of one phenotype for large-scale GWAS and meta-analyses[11, 30, 41]. For many traits, results from these studies are now publicly available, providing an excellent resource for cross-phenotype analyses using summary statistics. With the growing appreciation of pleiotropic effects in the scientific community, cross-disorder analyses of several related traits have been carried out to disentangle shared and disease-specific genetic determinants[42-45].

The establishment of genome-wide genotyped biobanks[46] and cohorts with in-depth phenotype information[47] has also made it possible to perform multi-trait analyses on the same sample set[27, 48], for example through phenome-wide association studies (PheWAS), where the association of each genetic variant with all phenotypes in a dataset is tested[49-52]. One challenge of the PheWAS approach is the high multiple testing burden that grows as the number of traits and variants tested increases[53]. Although this can be partly circumvented by performing targeted PheWAS at a selected number of variants hypothesized to exert pleiotropic effects[53, 54], other challenges such as consistent phenotyping and selection of appropriate covariates remain[49].

1.3 Types of pleiotropy

The term “pleiotropy” was coined over 100 years ago by German scientist Ludwig Plate to describe the phenomenon of a hereditary unit affecting more than one trait of an organism[55]. Since then, pleiotropy has been a topic of extensive research and debate. Before human genetics began to gain traction as a research field, pleiotropy was mainly studied in model organisms and, on a more theoretical level, in evolutionary biology[55, 56].Over the course of the past decades there have been several proposals on how to classify different types of pleiotropy[53, 55, 57, 58].

With regards to GWAS, it is important to note that cross-phenotype associations can arise due to several reasons, not all of which are biologically meaningful[57, 58]. Solovieff and colleagues[58] described three broad categories of pleiotropy in the context of complex traits:

In the case of biological pleiotropy, causal variants of different traits fall into the same gene or regulatory unit (e.g. transcription factor binding sites)[58]. In GWAS this could manifest itself in the form of two different variants in the same region tagging the same or two separate causal variants or as one variant tagging the causal one (Figure 1.1a-b). In practice, fine-mapping and molecular studies are required to confidently distinguish between these different scenarios[58].

Mediated pleiotropy refers to the case where a variant directly affects one trait, which in turn affects another (Figure 1.1c). GWAS will still pick up an association of the variant with the second trait, but this association will disappear when conditioned on the first. Causal inference can be achieved through Mendelian randomisation studies, which have been widely used in genetic epidemiology[27, 58-60]. An example is the association of the *FTO* gene with osteoarthritis (OA)[38], which was shown to exert its effect on OA through body mass index (BMI)[61].

Finally, cross-phenotype associations can also arise due to spurious pleiotropy. At the planning stage of a study, design artefacts may lead to inaccurate results. For example, ascertainment bias or misclassification of cases can both inflate genetic overlap estimates. At the analysis stage, causal variants in different genes may be tagged by the same GWAS variant (Figure 1.1d). A classic example of this is the human leukocyte antigen (HLA) region on chromosome 6. Due to its high gene density and extensive linkage disequilibrium (LD), GWAS signals within the HLA region are difficult to fine-map. While the HLA locus has been associated with a range of diseases[30, 38, 62-66], most prominently immune-mediated ones, it remains unclear to what extent these disorders share the same causal risk variants or genes.

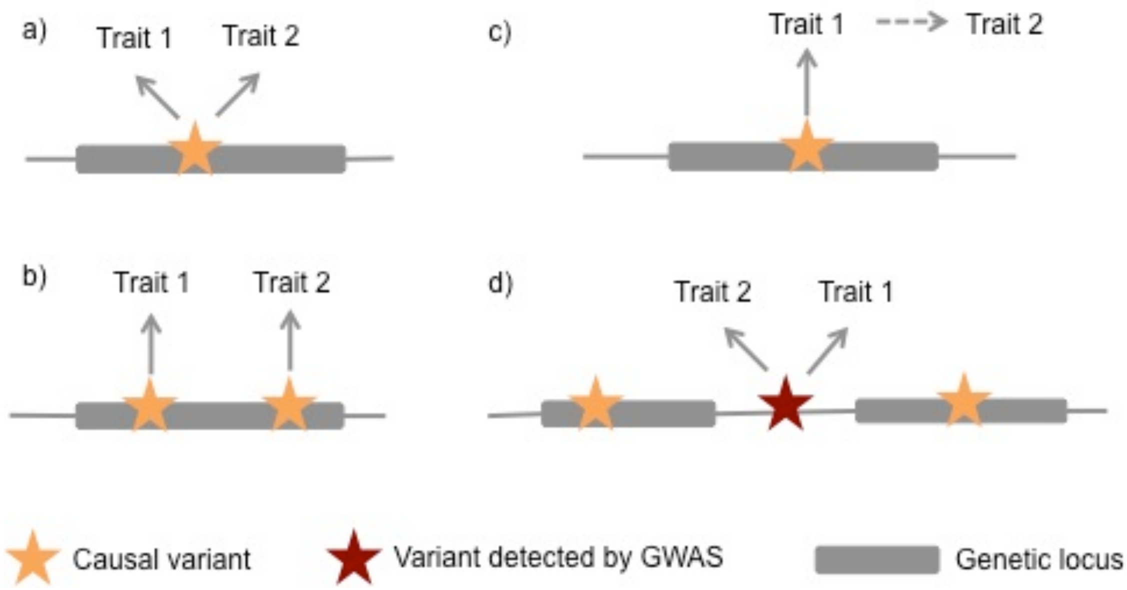

Figure 1.1. *Schematic representation of different scenarios for cross-phenotype associations. Such effects might arise due to biological pleiotropy, whereby causal variants for two traits colocalise in the same locus (a,b), due to mediated pleiotropy, whereby a variant exerts an effect on one trait through another one (c) , or due to spurious pleiotropy, whereby causal variants for two traits fall into distinct loci but are in LD with a variant associated with both traits (d). Adapted from Solovieff et al.*[58]

1.4 Analytical approaches

1.4.1 Overview of methods

Multi-trait analysis methods can be broadly classified into three categories according to the level at which they assess genetic overlap: genome-wide, regional and single variant. Genome-wide methods are currently only available for pairwise trait comparisons and can be used as an initial assessment of the global genetic overlap between two traits. The latter two approaches aim to detect cross-phenotype effects at distinct genomic regions and at single variants, respectively.

Region-based methods bin variants into groups based on pre-defined criteria, such as LD-blocks or gene boundaries, and then test for cross-phenotype effects within each group. An advantage of such approaches is that they alleviate the multiple testing penalty incurred by single-point analyses; furthermore, they can increase power by combining information across biologically meaningful units.

Since variant-level methods test each variant separately, they provide the highest resolution. On the other hand, they are less powerful in situations where each trait is associated with a different variant in the same functional unit, and might fail to identify these cross-phenotype effects unless all relevant variants are in at least moderate LD.

The above analysis approaches can be further sub-divided into univariate and multivariate, based on their underlying statistical framework. Univariate methods combine summary statistics of single-trait GWAS to search for cross-phenotype effects. This means that analyses can be carried out with each trait measured on a distinct set of individuals. Multivariate methods, on the other hand, jointly model all traits, which requires that all individuals included in the study have phenotype information for all traits analysed. The statistical difference between uni- and multivariate methods is best illustrated by the example of linear regression analysis: for univariate regression, the response variable (i.e. the phenotype) will be a vector, with one data point for each individual in the study; for multivariate regression, the response variable will be a matrix, where each row represents an individual and each column represents one phenotype. Although there are exceptions, these categories are often analogous to distinguishing between methods requiring only summary data and individual-level information, respectively.

Several comparisons of different multi-trait methods have been conducted to date, testing power and type I error rates, as well as computational performance under different scenarios[67-71]. Since each report includes a different combination of methods and settings (e.g. MAF, sample size, genetic effect size and trait number/correlation), it is difficult to pinpoint an overall winner. Nevertheless, some key insights have emerged from this body of work: Generally, multi-trait methods, both uni- or multivariate, are more powerful than testing each trait separately, as the latter approach incurs a multiple testing penalty dependent on the effective number of traits[67, 68, 72, 73]. When inter-trait correlations are high, the effective number of traits will be close to one, resulting in a low multiple testing penalty; however, many multi-trait methods only perform a single test of association and additionally can explicitly model trait correlations.

Multivariate methods outperform univariate ones in most simulation scenarios, whereas different types of multivariate methods perform similarly well in most simulation scenarios[67, 68, 71]. The gain in power seems to be highest if trait correlations are high and genetic effects on associated traits are in opposite direction[67]. However, it should be noted that this loss of power for univariate methods is in large part due to the heterogeneity in genetic effects for different traits, which methods based on effect estimates do not handle well. Even methods designed to account for heterogeneous effects, such as the S_{het} statistic applied in Chapter 2, loses power compared to multivariate methods in such scenarios[74]. Less data is available for categorical phenotypes. Porter and O'Reilly simulated two case-control datasets and found power curves appear similar to those of quantitative traits, with multivariate methods performing best[74]. MAF does not seem to influence power when only considering common variants ($0.05 \leq MAF \geq 0.5$)[67, 72].

There are some scenarios where the use of summary statistics-based univariate methods can be advantageous. For example, when the tested genetic variant is associated with all or most tested traits, with similar effect sizes, and traits are strongly positively correlated, methods using summary data, such as S_{het}/S_{hom} or TATES, outperform individual-level data methods[74]. In practice, however, it is difficult to know whether this scenario is the case in advance. Perhaps a more relevant consideration is study design. Datasets with multiple traits measured on a large number of people might not be available; since power is affected by sample size, it might be preferable to combine summary statistics from large GWAS on distinct traits[71]. Even when individual-level data for multiple phenotypes is available at a large enough sample size, missingness might pose another problem, since multivariate methods rely on complete data (see section 1.4.6.1). Furthermore, most multivariate methods assume normally distributed phenotypes, and are therefore not appropriate for trait combinations with mixed distributions. Reverse regression models circumvent this problem, but lose power when the number of traits included is very high (>40).

It is advisable to perform both multivariate and univariate association tests in a complementary way[75]. This will not only enable the detection of additional signals, but also aid the interpretation of a multivariate association (i.e. which trait(s) is/are driving the signal). Since only a handful of currently available methods explicitly test for cross-phenotype effects, considering univariate association statistics also guards against reporting false-positive multi-trait associations.

When combining summary statistics across multiple traits in a univariate fashion, an important consideration is the power of individual studies. As for regional or genome-wide methods, single-point methods will fail to detect cross-phenotype associations if the input datasets are underpowered. Another important aspect is the ancestry of input study samples, especially for methods requiring the specification of reference panels[76], for which combining studies from different populations might lead to spurious results.

1.4.2 Study design considerations

There are some practical considerations when selecting an appropriate method for multi-trait analysis:

Firstly, the type of data available will determine which statistical approach is applicable. Due to limitations of data sharing policies it might not be possible to obtain individual-level genotype data for all traits analysed.

Secondly, the type and number of traits to include needs to be considered: some approaches require all traits to be continuous, while others also allow for dichotomous traits or a combination of both. Several methods, such as colocalisation tests[77, 78] or genetic correlation analyses[79], can currently only accommodate two traits at a time, while others lose power with an increasing number of traits[80].

Finally, if each trait is measured on a different set of individuals, sample overlap between datasets will need to be accounted for. This has been implemented in several methods[78, 79, 81]. Ideally, the exact number of overlapping individuals will need to be accounted for. However, this is often not possible when using data from publicly available GWAS. One way to estimate the extent of overlap is to calculate the Pearson's correlation of the Z scores of all independent, non-associated variants from two studies[78], although other methods have also been proposed[82-84].

1.4.3 Genome-wide methods

Polygenic risk scores (PRS; or genetic risk scores) were initially used in genetic epidemiology to test how well a set of variables could predict, or distinguish between, case-control status in a study sample[85-88]. In the context of GWAS, the risk variables comprise variants known to be associated with a given trait. Odds ratios (ORs) for these variants from a "base" GWAS are then used to construct scores for each individual in an independent "target" dataset. Using logistic (binary trait) or linear (continuous trait) regression to relate phenotype and score, the proportion of phenotypic variance explained in the target data by the base risk variants can be directly estimated.

This framework can also be applied to two different traits[42, 51, 89, 90]. Purcell and colleagues[89] showed that risk scores for bipolar disorder are significantly associated with schizophrenia, and that the variance in phenotype captured by the score could be increased by relaxing the p-value threshold for variant inclusion (rather than using only genome-wide significant variants). One reason for this could be that many variants with a true effect on the phenotype did not reach genome-wide significance in the base study. This is especially likely in the case of highly polygenic traits, for which only a fraction of the heritability can be explained by currently known risk variants.

Genetic correlation (r_g) captures the extent to which genetic factors influence the covariance of two traits. While there are multivariate methods for genetic correlation analysis, such as GCTA[43, 91], BOLT-REML[92] and mvLMM[93], a univariate method based on LD score regression (LDSC) has gained popularity in recent years[79, 94]. LDSC only requires summary statistics, can handle any combination of traits and is not confounded by sample overlap. However, it requires the specification of a reference panel for LD estimation, which should be chosen with care when analysing two GWAS performed in populations of different ancestries. The LD Hub database, which acts as both a central aggregation of public summary statistics and an online interface for LDSC, enables systematic comparisons between a range of traits[95]. As the authors of LDSC point out, it is important to distinguish genetic correlation from pleiotropy[79]. A near-zero estimate of genetic correlation between two traits does not necessarily mean that they share no common risk loci. For example, there could be no directionality to their genetic relationship, i.e. at some shared loci the risk allele is the same for both traits, while at others the risk allele for one trait is protective of the other. An example of the latter scenario is the rs7501939 variant in *TCF2*, for which the C allele confers increased risk for prostate cancer and decreased risk for type 2 diabetes[9]. Similar to PRS, if either or both of the input datasets are underpowered this could also lead to a falsely low estimate of r_g. Conversely, in the case of disease traits, genetic correlation could be inflated due to ascertainment bias or misclassification of cases[58].

1.4.4 Regional methods

In 2013 Giambartolomei and colleagues developed a Bayesian colocalisation model to identify genomic regions of colocalising expression quantitative trait loci (eQTL) and GWAS signals[77]. This method was then extended to account for sample overlap, and implemented in a software package (gwas-pw) to enable simplicity of use for the pairwise comparison of GWAS summary statistics[78]. The model integrates the effects of all variants in a pre-defined region, such as approximately independent LD blocks[96]. It generates posterior probabilities for each of five hypotheses, the two most relevant being that in a given region the traits share one causal variant, and that they each have a separate causal variant. An advantage of this approach over many variant-level methods is that it evaluates the evidence for *both* traits being associated with a given regions, thus making it possible to distinguish from the scenario of one trait alone driving an observed signal[65].

Multivariate methods for locus-based analysis include extensions to canonical correlation analysis (CCA)[76, 97, 98], functional linear models[99], non-parametric tests[100] and multivariate mixed models[101].

1.4.4.1 *Rare variant methods*

The substantial drop in sequencing costs over the past decade together with the establishment of better reference panels for imputation have made association studies of low frequency and rare variants feasible[102, 103]. Methods for rare variant studies usually group several variants together and perform an association test with this composite genotype. They are generally more powerful than testing individual rare variants[104], and have been the tool of choice for single-trait studies[105]. Two of the most popular rare variant test methods are kernel-based tests, such as SKAT[106], and collapsing tests[107].

While some of the multi-trait methods are applicable to both common and low frequency markers[99-101], approaches have also been specifically designed for rare variants. These methods all rely on individual-level data with phenotypes measured in the same set of individuals. Wu and Pankow extended univariate SKAT for the application to multiple continuous traits[108]. Another method, MAAUSS, also builds on the SKAT algorithm, including a variance-covariance matrix that allows for the joint modeling of multiple phenotypes[109]. Multiple binary or a mixture of binary and continuous traits can be analysed by MAAUSS through integration of the generalised estimating equation framework. In adaptive weighting reverse regression (AWRR)[110], the genotypes in a set of variants are first combined, weighted by the strength of association and direction of effect of each variant; the resulting variable is then regressed on multiple traits and a score test used to assess significance. This reverse regression approach is similar to other methods discussed here and can incorporate large numbers of traits of any kind. Similarly, multi-phenotype analysis of rare variants (MARV) uses reverse regression combined with a burden-based method to combined rare variants in a region[111, 112]. In short, rare variants in a genomic region are combined into a single variable denoting the proportion of minor alleles an individual carries in that region. In addition to the full model where all phenotypes are included in the analysis, MARV also allows for a models selection procedure where all possible phenotype combinations are analysed. One downside to MARV is that, like other burden tests, it suffers a loss of power when the effects of rare variants in a region are in different direction, or if only a very small number of variants in a region are associated.

1.4.5 Single-point univariate methods

With the increasing availability of summary data from large-scale GWAS, an important question has been how to harness these data to perform pleiotropy analyses. Perhaps the simplest way to search for cross-phenotype effects is to decide on a p-value threshold and declare all variants that fall below this threshold for a group of traits as cross-phenotype associations[58]. However, this approach can be underpowered, as even with large sample sizes truly associated variants with sub-threshold p-values will be missed. Consequently, a number of methods to statistically combine summary data for multiple traits have been developed, many of which are based on meta-analytic approaches[80] [113-116].

In meta-analysis p-values or effect sizes are combined across multiple studies of the same trait[117]. For the latter, effects are typically either assumed to be consistent across studies (fixed effects meta-analysis) or allowed to vary (random effects meta-analysis). However, a genetic variant might have the opposite effect on two traits. While this can be circumvented by applying a directionality-agnostic p-value based meta-analysis, there are some limitations, such as the inability to obtain an overall effect estimate[117]. Therefore, these standard approaches are best-suited to groups of traits/disorders assumed to have similar underlying biological mechanisms[42]. The meta-analysis framework has been adapted to accommodate this and other issues that arise when combining several different traits:

Cotsapas and colleagues developed a cross-phenotype meta-analysis (CPMA) method that tests for the presence of two or more trait associations at a variant[113]. This has the advantage of protecting against the scenario of one trait driving the association. CPMA only requires p-values as input and is thus robust to heterogeneous effect directions. Since CPMA compares the distribution of p-values for all traits at a variant to the null hypothesis of uniformity, it is well suited for moderate to large numbers of phenotypes, but less so for pairs of traits.

In a generalisation of fixed-effects meta-analysis, all possible subsets of traits are evaluated to identify the one with the maximum absolute Z-statistic at a variant[80]. The approach, termed ASSET, takes effect estimates as input and can also be used to identify disease subtypes within case-control data. Extensions were also proposed to account for sample overlap and effect heterogeneity between traits[80]. Using ASSET investigators have identified three loci associated with five autoimmune disorders, as well as risk loci associated with different cancers[118].

Zhu and colleagues developed two meta-analysis test statistics to detect cross-phenotype associations assuming homogeneous and heterogeneous effects across studies, respectively[114]. The tests are implemented in the R package CPASSOC, and work with both univariate (i.e. one trait per cohort) and multivariate summary statistics (i.e. several traits measured in each cohort). CPASSOC requires the specification of an inter-cohort correlation matrix. Since the true phenotypic correlation is unknown in the absence of raw data, this can be derived from summary statistics and – similarly to approaches outlined above – accounts for overlapping samples. Applying CPASSOC to anthropometric trait summary data from the GIANT consortium identified one novel genome-wide significant locus within the *TOX* gene missed by conventional meta-analysis[119].

1.4.6 Single-point multivariate methods

As the availability of large-scale genetic datasets with multiple phenotype measurements increases, the focus of method development for multi-trait analyses has shifted towards multivariate methods that use individual-level data rather than summary statistics[120, 121]. These approaches are generally more powerful than combining test statistics from univariate GWAS, as the inter-trait covariance can be modelled directly from the data[67, 121].

One efficient way to deal with multivariate phenotypes is to first apply a dimension reduction technique that collapses the individual trait values, and then perform an association between genotype and this new set of variables. Principal component and canonical correlation analysis (PCA and CCA, respectively) are examples of such techniques; the former derives linear combinations of the phenotypes that explain the greatest possible covariance between them[72, 121-124], whereas the latter derives linear combinations of the traits that explain the greatest amount of covariance between a genetic locus and the traits[125] [97, 98].

Linear mixed models (LMMs) are an extension of standard regression analysis incorporating both fixed and random effects and have gained popularity in GWAS due to their ability to handle relatedness amongst individuals[126, 127]. Multivariate LMMs can be used for association testing with multiple phenotypes. They model association between a genetic marker and the traits as the fixed effect, and the inter-trait covariance as the random effect[121]. While multivariate mixed models are generally more powerful than standard univariate association tests, they perform less well when the traits under consideration are only weakly correlated[128]. Korte and colleagues first applied multivariate LMMs to pairwise quantitative trait measurements in a human cohort[128]. Several other methods based on multivariate LMMs exist[75, 93, 128, 129], including a multivariate extension to the GEMMA algorithm[75].

Bayesian statistics allow for a model comparison between several alternative hypotheses, making them an attractive tool for pleiotropy analysis[77, 78, 130-133]. A model-selection framework proposed by Stephens returns Bayes factors for each possible partitioning of phenotypes into one of three categories: unassociated, directly associated, or indirectly associated with a genetic marker[131]. At markers where the evidence against the global null is strong, the individual Bayes factors can be used to determine which traits are likely to drive the association. The framework is implemented in the software mvBIMBAM and has been used to identify variants associated with low- and intermediate density lipoprotein subfractions[134].

Another way to allow for the inclusion of traits with mixed distribution is to reverse the regression of phenotype on genotype routinely employed in GWAS. MultiPhen performs ordinal regression of the genotype (number of minor alleles at a marker) on multiple phenotypes and tests for association using a likelihood ratio test[74]. An advantage over other multivariate methods is the MultiPhen maintains appropriate type I error rates when applied to non-quantitative traits. MultiPhen has similar power to detect associations to other multivariate methods, such as mvBIMBAM and CCA, with negative phenotypic correlations leading to increased power[67, 71]. SCOPA is another method relying on reversed regression, with the added advantage of being able to model dosage data from imputed variants[135]. It additionally applies a model selection procedure to discern which traits underlie an association signal. A framework for meta-analysis of SCOPA-derived (META-SCOPA) summary statistics is implemented. One consideration for reverse regression methods is that any adjustments to phenotypes (e.g. age, sex, population structure) must be performed prior to the association analysis[135].

1.4.6.1 *Handling of missing data*

One potential obstacle of multivariate methods is the handling of incomplete data. Individuals for whom one or more of the analysed traits are missing will be excluded from the analysis, which can lead to substantial sample loss and to biased results, depending on the missingness patterns and number of traits included. Considering the reason why certain trait values might be missing is important in deciding on appropriate analysis approaches. Data points can either be missing completely at random (MCAR), missing at random (MAR) or missing not at random (MNAR)[136-138]. In the case of MCAR, the reason why a trait value is missing is unrelated to both observable and unobservable variables. MAR means that the reason a value is missing can be entirely explained by an observed variable. For example, if young people are less likely to fill in survey questions related to ethnicity, then self-reported ethnicity missingness will be random once age is adjusted for. Conversely, in MNAR situations the value of the missing data point is related to the reason it is missing. For example, people suffering from severe depression might not respond to surveys on mental health. In practice it is not possible to distinguish with certainty between MAR and MNAR scenarios, as this would require the researcher to know the true values of missing data points. Retaining only complete cases when there is MNAR will lead to spurious analysis results. If each trait is only measured once per individual included in the study, complete case analyses will not lead to bias if the MAR (or MCAR) assumption holds[136]. However, this might still result in substantial sample loss. Alternatively, missing values can be recapitulated using single or multiple imputation methods[136, 137, 139, 140]. As the name suggests, single imputation obtains a single estimate for each missing value based on the imputation model. While fast and relatively easy to implement, a downside of this approach is that it may result in biased results due to not accounting for the uncertainty of the imputed values[136, 138]. Multiple imputation, on the other hand, repeats the imputation procedure multiple times, which can guard against bias, but it relies on the MAR assumption, violation of which might impact the results[136, 140, 141]. For a more detailed discussion of phenotype imputation procedures see section 4.4.2.

1.4.7 Inferring causality

Determining whether the correlation between two traits is due to a causal link (i.e. trait 1 is a causal risk factor for trait 2), or due to confounding factors such as environmental exposures, can be achieved through Mendelian randomisation (MR). Notably, while most methods outlined in this chapter aim to detect biological pleiotropy and are confounded by mediated pleiotropy, the opposite is true for MR. In MR one or several genetic markers – so-called instrumental variables (IVs) – are used to infer whether or not trait 1 (the exposure) causally influences trait 2 (the outcome) [60, 142-145] (Figure 1.2). An early example of MR is a study published in 2005 which concluded that, contrary to prior belief, C-reactive protein levels were not causal for metabolic syndrome[60].

In order to be a valid IV, three key assumptions about the genetic marker must be met: first, the marker is associated with trait 1; second, the marker is not associated with any confounding variables, such as environmental factors that might affect trait 2 independently of trait 1; and third, the marker is not associated with trait 2 when conditioning on trait 1. The first two assumptions are usually easy to fulfill in a GWAS context. The first assumption also implies that the function of the gene or marker used as an IV is known *a priori*. Consequently, MR is not a method to detect new genotype-phenotype associations[144]. Some consideration should be given to assumption two, which can be violated in the case of population stratification[144, 146].

Arguably the biggest uncertainty is the third assumption, which will not hold if the variant(s) used independently affect both trait 1 and trait 2, i.e. if there is horizontal pleiotropy. The risk of this can be mitigated in several ways, all of which rely on the inclusion of more than one genetic marker in the MR analysis. One is to design the MR study so that the exposure of interest is a protein biomarker[147]. Proteins have the advantage that they are more proximal to the genetic effects acting on them, compared to metabolites or other circulating biomarkers. By restraining the selection of IVs to variants acting in *cis* to the gene encoding the protein, it should be possible – in theory – to minimise the chance of horizontal pleiotropy. Bioinformatics approaches can also be used to obtain functional annotation of variants and support their validity as IVs[147]. In addition, the included variants should not be in LD with nearby variants affecting the expression of other proteins, as this might lead to confounding if those proteins also affect the outcome[147].

There are several approaches to MR analysis, depending on the type of data available as well as the underlying assumptions about the genetic markers and traits. If both traits were measured on the same samples, MR can be performed via two-sided least squares analysis, where trait 1 is first regressed onto the IVs, and trait 2 is then regressed on the predicted values of trait 1 from the first regression; the effect size derived from the second regression is the MR estimate. A downside to this approach is that availability of samples with multiple trait measurements is still limited compared to the sample sizes achieved by GWAS consortia focusing on individual traits. As a result, a number of approaches have been developed to perform MR in a setting where each trait is measured on distinct samples (so called two-sample MR). In its most simplistic form, two-sample MR can be performed by obtaining a Wald estimator from the ratio of the effect of the variant on trait 1 over its effect on trait 2[148]. For multi-instrument MR, the selected variants can be combined into a weighted (based on variant-exposure effect sizes) or unweighted score which is then tested for association with the outcome of interest[149, 150].

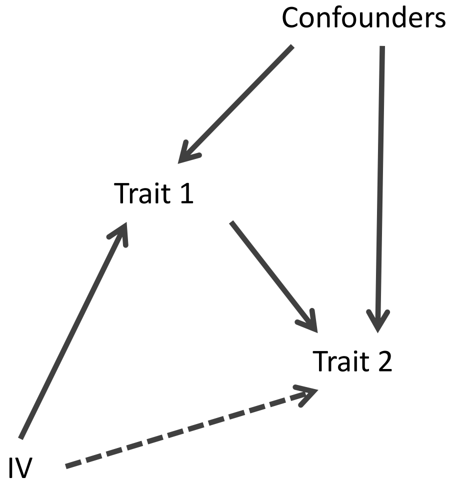

Figure 1.2. *Directed acyclic graph of the Mendelian randomisation model. IV=instrumental variable*

Several analytical approaches have been proposed to both detect and account for pleiotropy in MR settings[151-155]. If all of the variants satisfy the IV assumptions there should be no heterogeneity between their individual MR estimates[144, 151, 156]. In other words, in the case of no pleiotropy MR estimates of each variant will only vary by chance. The Cochran Q statistic and the related I^2 index can be used to test for heterogeneity between individual IV estimates[151]. If individual-level data are available and both traits have been measured on the same sample, the Sargan test can be used to assess evidence against the null of all MR estimates being the same[146]. For two-sample MR, an adaptation of inverse-variance weighted meta-analysis can be used to combine Wald estimators across several variants[157], either in a fixed or random effects model. The former assumes that the variants used are not pleiotropic, whereas the latter assumes that on average the pleiotropic effects of the variants cancel each other out. In Egger regression, variant-outcome effect sizes are regressed on the variant-trait 1 effect sizes with an unconstrained intercept[153]. An intercept term significantly different from zero is indicative of pleiotropy. Bowden and colleagues proposed a summary data-based step-wise analysis framework which applies all three of the above methods to differentiate between the scenarios of no pleiotropy, pleiotropy without heterogeneity and pleiotropy with heterogeneity[152, 158]. By applying this framework to summary data from two GWAS the authors showed that the observed association between plasma urate levels and cardiovascular disease was likely due to pleiotropy rather than a causal link, as evident from heterogeneity in the MR estimates from the 31 variants analysed.

While MR analyses are valuable tools to investigate causal relationships between complex phenotypes without the need to collect longitudinal data, it is not a replacement for experimental follow-up and characterisation of identified associations. Even when care is taken in the study design and several scenarios of pleiotropy tested, the possibility of confounding remains. In two-sample MR, where exposure and outcome are measured on different samples, the potential of unmeasured environmental variables to influence the results cannot be completely excluded. Furthermore, non-linear relationships between traits will lead to inaccurate MR estimates[159]. This is also true for genetic effects that do not follow an additive model, i.e. dominant or recessive effects. If the relationship between exposure and outcome is sex- and/or age-specific, ignoring these variables in the analysis will lead to inaccurate MR estimates. However, data from age- and sex-stratified GWAS is not readily available for many phenotypes. Lastly, if the exposure is a trait comprised of more than one sub-phenotype, it is possible that the effect on the outcome is driven by one of those rather than the composite exposure trait[160].

1.5 Conclusion

Investigating pleiotropy in human traits not only holds the potential to uncover additional associations, but could also help to redefine disease classifications. This is of particular interest in disorders for which the aetiopathology is unclear, and for which current diagnostic tools might be inadequate. For example, psychiatric conditions are highly comorbid and until recently[30, 161] have been mostly refractory to GWAS[162]. Comparisons of different psychiatric disorders have shown that the genetic overlap among them is extensive[42, 163], and that certain pairs of diseases are genetically more similar than others[163]. Together these findings suggest that shared biological mechanisms cross diagnostic boundaries and might aid the development of more accurate disease classification systems.

As personalised medicine becomes more established, pleiotropic effects will need to be taken into account for genetic risk prediction and counselling, especially where variants have opposite effects on disorders. For example, variants in the interleukin-10 and -27 genes increase the risk of type 1 diabetes, but have protective effects for Crohn's disease. A more comprehensive understanding of pleiotropy could also aid drug repurposing efforts.

1.6 Aims and overview of this thesis

The overarching aim of my doctoral work was to search for shared genetic determinants of medically relevant complex traits, with an emphasis on musculoskeletal and cardiometabolic phenotypes reflecting the main focus of our research group. The phenotypes investigated here were chosen based on their established epidemiologic link (Chapters 2 and 3) or their correlation with each other (Chapter 4). To explore the potential of different multi-trait approaches, I chose phenotype groups that would allow for summary statistics-based approaches (Chapter 2), individual-level data approaches with both phenotypes measured on all samples (Chapter 3), and individual-level methods aimed at exploring data with high-dimensional quantitative trait measurements (Chapter 4). I employed both uni- and multivariate methods to test for evidence of genetic overlap at a genome-wide, regional and variant level. By looking at different traits and disorders in a joint framework, I hoped to gain a better understanding of their genetic architecture.

In Chapter 2 of this thesis, I describe an overlap analysis of osteoarthritis (OA) and bone mineral density (BMD) at a genome-wide scale. OA has until recently been mostly refractory to GWAS, and genetic mechanisms influencing disease subtype are not well-understood. I therefore sought to leverage data from two published GWAS on OA and BMD, respectively, to identify common risk factors between those two traits. Increased BMD has been associated with a higher risk of OA. This epidemiologic link has been established through both prospective and ascertained studies, yet the underlying biological reason for this association is still not clear. So far, only a few studies have looked at the shared genetics of OA and BMD, and none have used genome-wide data. By searching for evidence of genetic overlap between summary statistics of OA and BMD GWAS, I hoped to identify common loci with potential biological relevance. Furthermore, I planned to leverage the BMD GWAS data to prioritise variants for follow-up and replication in independent OA datasets.

In Chapter 3, I aimed to elucidate the genetic contribution to schizophrenia (SCZ) and type 2 diabetes (T2D) comorbidity. SCZ patients are at an elevated risk of developing T2D compared to the general population. While antipsychotic medications are known to cause metabolic side effects, impaired glucose homeostasis was also found in drug-naïve SCZ patients. To assess the extent to which this association can be explained by genetics, I used summary statistics from published GWAs on SCZ and T2D, respectively, in conjunction with individual-level data from a cohort comprising patients with either T2D, SCZ or both disorders.

In Chapter 4, I outline an analysis framework for multi-trait GWAS in a sample collection with high-dimensional quantitative phenotype information. Including multiple correlated traits in an association model can increase power to detect associations. However, as datasets with hundreds of trait measurement become more common, selecting meaningful trait groups is not always straight forward. For this project, I used a Greek isolated population cohort with whole-genome sequencing data (average of 22x) and over 300 quantitative traits to perform phenotype imputation, clustering and association analysis using a multivariate linear mixed model.

Finally, in Chapter 5, I summarise key lessons and insights gained throughout my PhD and discuss the current landscape and future outlook of pleiotropy research.

Chapter 2 – A genome-wide evaluation of the shared aetiology between osteoarthritis and bone mineral density

2.1 Introduction

2.1.1 Pathobiology of osteoarthritis

Osteoarthritis (OA) is a degenerative disease of the synovial joint affecting over 40% of people over 70 years of age[164]. Synovial (or diarthrodial) joints connect the end of two bones through a joint capsule filled with synovial fluid, which provides lubrication. They are the most common type of joint in mammals and allow for a variety of movements. Depending on body site, the joint capsule may also comprise meniscal discs composed of fibrocartilage (such as in the knee). The bone ends (epiphyses) connected by the joint are lined with articular cartilage that acts as a shock absorbent and diffuses friction.

Hallmarks of OA include cartilage degradation, joint-space narrowing, formation of osteophytes (bony protrusions) within the joint and subchondral bone remodeling[165, 166]. While the pathologic processes taking place in the osteoarthritic joint are well understood on a macroscopic level, their timing and causal mechanisms are not clear. Consequently, there are no preventive treatments or early detection methods (e.g. biomarkers), and clinical diagnosis relies on the presence of radiographic features[167]. Since there is no curative therapy, the main treatment strategy consists of pain management and, in severe cases, joint replacement surgery (arthroplasty)[165]. As a result of its high prevalence and lack of effective therapeutic options, OA poses a high economic health burden, further motivating efforts to better understand risk factors and biological processes involved in OA onset and progression.

2.1.2 Genetics of osteoarthritis

OA is a complex disorder with the 27 currently known risk loci accounting for approximately 26.3% of disease heritability[168, 169]. Until recently, OA had been mostly refractory to GWAS. The first large-scale OA GWAS was conducted by the arcOGEN consortium in a two-stage design, culminating in a total discovery sample of 7,410 cases and 11,009 controls[38, 170]. These sample numbers were further increased by collaborative efforts such as the deCODE project and treatOA[171], as well as more recently in a GWAS using the first release of the UK Biobank resource, with a discovery stage dataset of over 10,000 cases and up to 50,000 matched controls[172]. OA is a heterogeneous disorder, with heritability varying depending on the affected joint. Of the 27 published risk loci to date, 6 and 10 are associated with knee OA only and hip OA only, respectively, while 11 are associated with both hip and knee OA[172-174]. This further highlights how phenotypic variation is reflected by genetics, and demonstrates the need for strict phenotype definitions.

2.1.3 Bone mineral density

Bone mineral density refers to the mineral content in bone tissue and serves as a clinical indicator of fracture risk and, consequently, osteoporosis. The most common measurement method is dual X-ray absorptiometry (DXA), although other methods, such as quantitative computer tomography or quantitative ultrasound, exist.

BMD is determined by a set of interdependent processes collectively termed bone remodeling. Bone remodeling includes both bone formation (osteogenesis; mediated primarily by osteoblasts) and bone breakdown (resorption; mediated primarily by osteoclasts).

The heritability of BMD varies depending on body site, with estimates ranging from 50 to 85%[175]. A GWAS carried out by the Genetic Factors for Osteoporosis (GEFOS) consortium 2012 found 56 loci associated with BMD[13]. Three years later, a rare variant of large effect was identified by GEFOS combining whole-genome sequencing and GWAS imputation[8]. The largest genetic study on BMD to date was performed using heel bone estimates in almost 150,000 individuals of the UK Biobank. This effort almost tripled the number of known BMD loci and also provided extensive *in-silico* functional follow-up of novel associations[176].

2.1.4 Shared mechanisms of osteoarthritis and bone mineral density

The link between bone mineral density (BMD) and OA was first reported in 1972 by Foss and Byers, who observed higher BMD in femoral heads excised during OA-related hip replacement surgery[177]. Since then, a number of cross-sectional and longitudinal studies have found higher femoral neck (FN) and lumbar spine (LS) BMD, as well as total body BMD to be associated with incident OA at the hip, knee and other joint sites[178-181].

Findings with regards to the relationship between BMD and OA progression are less clear[182]. Elevated bone turnover – usually a marker for decreased BMD – was reported in patients with progressive knee OA compared to patients with stable OA[183]. Decreased baseline femoral neck BMD (FNBMD) has also been associated with knee OA progression[184, 185]. Conversely, data from the Rotterdam Study showed a non-significant trend of higher odds or knee OA progression with increased lumbar spine BMD (LSBMD)[186], while another study found no link between knee OA progression and total body- or FNBMD[178].

Several biological mechanisms are implicated in both OA and BMD, such as bone remodeling, mesenchymal stem cell differentiation and inflammation[13, 38, 165, 187]. *RUNX2*, a key transcription factor regulating endochondral ossification and osteoblast differentiation[188, 189], has been associated with both OA and BMD based on its proximity to genome-wide significant variants[13, 38]. The other locus with known GWAS hits for both traits is *KLHL42* (or *KLHDC5*), although its biological relevance remains unclear[13, 38].

In addition, Yerges-Armstrong et al. have previously shown nominal association of 4 BMD-linked single nucleotide polymorphisms (SNPs) with knee OA[190]. However, despite the long-established epidemiologic link and shared biology, the genetic overlap of OA and BMD has not yet been assessed on a genome-wide level. Here, I present results from the first genome-wide analysis establishing shared genetic aetiology between OA and BMD.

2.1.5 Chapter overview

In this chapter I describe the first systematic overlap analysis of OA and BMD on a genome wide scale, using summary statistics from the GEFOS consortium for lumbar spine (n=31,800) and femoral neck (n=32,961) BMD, and from the arcOGEN consortium for three OA phenotypes (hip, n_{cases}=3,498; knee, n_{cases}=3,266; hip and/or knee, n_{cases}=7,410; $n_{controls}$=11,009).

First, I assess genome-wide correlation using pairwise LD score regression. Second, I employ a Bayesian colocalisation method as well as an overlap analysis based on incremental p-value thresholds. The former aims to pinpoint specific regions across the genome that have a high probability of harbouring pleiotropic signals; the latter tests for an excess of shared variants at different significance cut-offs between two datasets. This can be used both to estimate genetic overlap and to follow-up individual variants shared at more stringent p-value thresholds.

Third, I aggregated the genome-wide summary statistics of each dataset into gene- and pathway-level associations. Using a false discovery rate of 5% I then searched for genes and pathways that were significant for at least one OA and one BMD phenotype.

Fourth, I took forward 143 variants identified through the colocalisation and p-value based overlap analyses for replication in two large-scale GWAS of hip and/or knee OA in the UK Biobank and the deCODE cohort. I subsequently meta-analysed those variants across both replication cohorts and the arcOGEN combined dataset.

2.1.6 Publication note and contributions

All analyses outlined in this chapter were carried out by me, with the exception of the functional follow-up of *SMAD3* and the genome-wide correlation analysis between arcOGEN and the ALSPAC/Generation R study. The look-up of *SMAD3* expression in cartilage was performed by Julia Steinberg. The genetic correlation analysis in ALSPAC/Generation R was carried out by Katerina Trajanoska using the same parameters as described here under the section “2.2.4. Genome-wide genetic correlation analysis”.

The work described in this chapter has been peer-reviewed and published in *Human Molecular Genetics*[168].

2.2 Materials and Methods

2.2.1 Datasets

The analyses outlined in this chapter were conducted using summary association statistics from the arcOGEN[38] and GEFOS consortia[13]. The arcOGEN data comprised three OA phenotypes: knee OA, hip OA, and knee and/or hip OA (combined OA). A detailed description of the contributing studies and phenotype definitions can be found in Ref. 6. OA case status was determined radiographically as a Kellgren-Lawrence grade score ≥ 2. Most cases included in arcOGEN had progressed to a severe disease endpoint, as evident from the fact that 80% had undergone total joint replacement surgery. Samples were genotyped on the Illumina Human 610-Quad BeadChips (Illumina, San Diego, CA, USA) and variant QC was performed for cases and controls separately: SNPs were excluded if they had MAF≥5% and call rate<95%, or MAF<5% and call rate<99%, and an exact HWE p<0.0001. Population stratification was assessed by PCA, and the first ten principal components were included in the analyses. Genotype imputation was carried out using IMPUTEv2[191] with the HapMap III reference panel (all populations)[38]. Case-control association analyses were carried out using SNPTESTv2[133], and additional GWAS stratified by sex, joint replacement surgery and joint site were also performed[38]. In this chapter, I used summary data from the non-stratified combined OA GWAS, as well as from the joint site stratified analyses (hip only and knee only).

The BMD data consisted of meta-analysis summary statistics for FN and LSBMD[13]. The 17 individual studies contributing data to the GEFOS discovery stage comprised samples from North America, Europe, Australia and East Asia. BMD was measured by dual-energy X-ray absorptiometry. Genotyping using chip arrays was performed by each participating study and genotypes were filtered for MAF≥1% for all studies, as well as HWE p-value and call rate at varying thresholds (see Ref [13], Supplementary Table 18D for a list of genotyping platforms and QC measures applied). Genotype imputation was carried out with BIM-BAM[192], IMPUTE[191] or MACH[193] using HapMap Phase 2 release 22 reference data (CEU or Han Chinese in Beijing and Japanese in Tokyo as appropriate. Genome-wide association analyses for FN-BMD and LS-BMD were conducted by each participating study separately, using an additive model and sex-specific standardised residuals adjusted for and age-, weight- and principal components. Results from individual studies were meta-analysed under a fixed-effects model, retaining only variants that were present in more than three studies[13].

In addition, summary data for skull and total body BMD of 9,142 samples from the Avon Longitudinal Study of Parents and their Children (ALSPAC) and the Generation R[194] study were used to calculate genetic correlation with arcOGEN.

For replication, I used summary statistics from two OA GWAS: the UK Biobank[46] and the deCODE study. Hospital episode statistics were used to define case status for OA in the UKBB sample. Inclusion and exclusion criteria were based on the International Statistical Classification of Diseases and Related Health Problems (ICD). Cases were defined as having OA (hip and/or knee) ICD-9 or ICD-10 codes only, and no inflammatory arthritis syndromes or other musculoskeletal disorders. Age-matched controls were selected on the condition that they did not have a hospital diagnosed (ICD-9 or ICD-10) or self-reported musculoskeletal disorders or symptoms.

For the deCODE dataset The information on hip, knee and vertebral osteoarthritis was obtained from Landspitali University Hospital electronic health records, Akureyri Hospital electronic health records and from a national Icelandic hip or knee arthroplasty registry[195]. Samples with secondary osteoarthritis (e.g. Perthes disease, hip dysplasia), post-trauma osteoarthritis (e.g. anterior cruciate ligament rupture) and those also diagnosed with rheumatoid arthritis were excluded from these lists. Only those diagnosed with osteoarthritis after the age of 40 were included. Hand osteoarthritis patients were drawn from a database of over 9,000 hand osteoarthritis patients that was initiated in 1972[196]. The study was approved by the Data Protection Authority of Iceland and the National Bioethics Committee of Iceland. Informed consent was obtained from all participants.

2.2.2 Reduced arcOGEN GWAS

For the p-value based overlap analysis as well as the gene and pathway analysis I excluded samples from London-based cohorts (TwinsUK and Chingford Study) from the arcOGEN datasets to avoid overlap with GEFOS samples. The full arcOGEN dataset was used for all other analyses described, since these methods are either not biased by sample overlap[79] or can correct for it via the correlation of summary statistics[78]. After exclusion of 714 samples, I carried out genome-wide association analyses on the arcOGEN dataset for each of the three phenotype groups using the "--method score" option in SNPTEST v2.5[133].

	arcOGEN			**arcOGEN excl. London**			**GEFOS**		**deCODE**	**UK Biobank**
	combined	**hip**	**knee**	**combined**	**hip**	**knee**	**LSBMD**	**FNBMD**	**combined**	**combined**
Cases	7,410	3,498	3,266	6,694	3,032	3,088	-	-	9,429	6,586
Controls	11,009	11,009	11,009	10,968	10,968	10,968	-	-	199,421	26,384
Total	18,419	14,507	14,275	17,662	14,000	14,056	31,800	32,961	208,850	32,970

Table 2.1. *Sample numbers for datasets analysed in this chapter.*

2.2.3 Estimating sample overlap

The GEFOS summary statistics used here are the result of a large-scale meta-analysis consisting of 17 cohorts. Since I did not have access to individual-level genotype data for the participating studies, it was not possible to perform identity checks with samples included in arcOGEN.

Several methods[78, 82, 197] have been proposed to estimate the extent of sample overlap, which can arise due to duplicated samples across studies or due to relatedness, based on only summary data. I use two correlation estimates to quantify the extent of sample overlap between each GEFOS and arcOGEN dataset: Pearson's and tetrachoric correlation of Z-scores[82]. The advantage of using tetrachoric correlation over Pearson's correlation lies in the fact that the former truncates all Z-scores into two bins (0 or 1), depending on whether they are positive or negative. This effectively attenuates the effect of significant associations, which might otherwise contribute to an inflated correlation estimate. To calculate tetrachoric correlation between arcOGEN and GEFOS, I transformed the Z-scores of the intersection of SNPs to a binomial distribution as described above and constructed a 2x2 table of the resulting counts. I then computed tetrachoric correlation using the "psych" package in R[198].

In the second approach I computed Pearson's correlation of summary statistics using only independent, non-associated variants[119, 197]:

$$corr(T_1, T_2) = \frac{\sum_i (T_{i1} - \mu_1)(T_{i2} - \mu_2)}{\sqrt{\sum_i (T_{i1} - \mu_1)^2 (T_{i2} - \mu_2)^2}}$$

Where T_1 and T_2 correspond to the test statistics (Z-scores) for each SNP *i* for study 1 and 2, and μ_1 and μ_2 correspond to their means. For each pairwise combination of the OA and BMD traits I took the intersection of SNPs and kept only those that were not associated with either trait (-1.96 > Z-score < 1.96). I then LD-pruned this set of SNPs in PLINK[199], using a window size of 250kb shifted by 200 variants at each iteration and an r^2 threshold of 0.2; the unimputed genotypes from the full arcOGEN data were used to calculated LD. An estimate of sample overlap was obtained by calculating Pearson's correlation of the Z-scores of all independent SNPs. Both methods gave low, non-significant correlation estimates, indicating that the effect of sample overlap is minimal (Table 2.2).

Datasets	**Pearson's (95% CI)**	**Tetrachoric (95% CI)**
allOA and LSBMD	-0.0001 (-0.0061-0.0058)	0.0033 (-0.002-0.005)
allOA and FNBMD	0.0045 (-0.0014-0.0105)	0.0013 (-0.003-0.002)
hipOA and LSBMD	0.0041 (-0.0019-0.0102)	0.0018 (-0.002-0.003)
hipOA and FNBMD	0.0036 (-0.0023-0.0097)	0.0034 (-0.002-0.005)
kneeOA and LSBMD	-0.0014 (-0.0070-0.0041)	0.0025 (-0.005-0.004)
kneeOA and FNBMD	0.0017 (-0.0038-0.0073)	0.0002 (-0.0005-0.0006)

Table 2.2. *Sample overlap between each pairwise OA and BMD dataset as estimated by Pearson's and tetrachoric correlation. 95% confidence intervals (CI) are given in parentheses.*

2.2.4 Genome-wide genetic correlation

I performed LD score regression analysis[79] on each pairwise combination between the arcOGEN and GEFOS datasets, using pre-computed LD scores based on the European (EUR) sample of the 1000 Genomes Project[94]. In addition, LD score regression was also performed by our collaborators (see "2.1.6. Publication note and contributions") between all three arcOGEN datasets and a paediatric BMD sample.

LD score regression relies on the assumption that variants in strong LD with a causal variant will have a higher association statistic than variants in low LD. In a single-study scenario, this fact can be harnessed to assess whether genome-wide inflation of test statistics is due to true polygenicity or to confounding factors such as cryptic relatedness; in the latter case inflation will not correlate with LD between variants. To assess the relative contribution of polygenicity and confounding in a GWAS, one can regress the association statistics on LD scores, which are given by:

$$l_j = \sum\nolimits_k r_{jk}^2$$

Where r_{jk}^2 is the LD between the index variant j and another variant k[94]. Thus, LD score can be interpreted as the extent of genetic variation captured by the index variant *j*. Extending this to a two-study scenario gives the following regression framework:

$$E\left[z_{1j}z_2\middle|l_j\right] = \frac{\sqrt{N_1N_2}\rho_g}{M}l_j + \frac{\rho N_S}{\sqrt{N_1N_2}}$$

where z_{ij} is the Z-score for study i and variant j, N_i is the number of samples in study i, ρ_g denotes the genetic covariance between the studies, M is the number of variants with MAF≥5% present in the reference panel used for LD score calculation, N_S is the number of samples overlapping between the two studies and ρ is the trait correlation. Genetic correlation can then be calculated by dividing ρ_g by the SNP heritabilities of both studies. Sample size for binary traits is defined as total sample size (cases and controls) (see Ref [79], Supplementary Note section 1.4, page 6).

2.2.5 Assessment of shared association signals

For each pairwise combination between the two BMD and three OA phenotypes I assessed the extent of shared association signals at different p-value cutoffs, following the approach described by Elliott and colleagues[200]. I filtered both datasets to a common set of SNPs on which p-value-informed linkage disequilibrium pruning was performed. To this end SNPs were sorted based on their association with OA, as this was our primary trait of interest and we therefore aimed to maximise retention of associated variants. Starting with the top SNP (i.e. the lowest p-value), any SNP in LD with that index SNP (r^2>0.05) was removed. A more stringent LD threshold was used here than for the estimation of sample overlap, in order to minimise the potential for inflating the test statistics. The next most-strongly associated SNP was then considered; if this SNP had already been excluded based on a previous iteration, it was skipped. This process was repeated until an independent list of SNPs was generated.

I assessed the extent of shared association signals between OA and BMD by constructing 2x2 contingency tables of the number of overlapping variants above and below ten different p-value thresholds (P_t: 0.5, 0.1, 0.05, 0.04, 0.03, 0.02, 0.01, 0.005, 0.001, 5 x 10^{-5}). To test for significance of overlap, a chi-squared test was performed at each P_t.

Empirical overlap p-values were obtained by repeating the chi-squared test after randomly permuting the GEFOS p-values. This was done 1,000,000 times to obtain a null distribution of overlap p-values against which the original overlap p-value could be compared.

2.2.6 Colocalisation analysis

I employed a Bayesian colocalisation method to search for genomic regions harbouring cross-phenotype associations between OA and BMD[78]. This is an extension of a method previously developed by Giambartolomei et al[77]., with the added option to correct for sample overlap. The model uses Z-scores and standard errors from two association studies to generate posterior probabilities for each of five hypotheses:

H_0: the region contains no variants associated with trait 1 or trait 2

H_1: the region contains one variant associated with trait 1

H_2: the region contains one variant associated with trait 2

H_3: the region contains one variant associated with both trait 1 and trait 2

H_4: the region contains one variant associated with trait 1 and a second variant associated with trait 2

Splitting the genome into uniform segments without accounting for LD structure can result in the double-counting of signals if segment boundaries happen to fall within an associated region. I used LD-blocks pre-computed using the LDetect algorithm[96] and the European sample of the 1000 Genomes Phase 1 data[201].

2.2.7 Gene and pathway analysis

Gene- and pathway analyses were performed on each OA and BMD dataset using MAGMA[202]. First, SNPs are assigned to genes, which are tested for their association with the phenotype. Results from this step are then combined into pathway-based association statistics.

Database	**URL**
BioCarta	http://cgap.nci.nih.gov/Pathways/BioCarta_Pathways
KEGG	http://www.genome.jp/kegg
Matrisome	http://matrisomeproject.mit.edu
Pathway Interaction Database	http://pid.nci.nih.gov
Reactome	http://www.reactome.org
SigmaAldrich	http://www.sigmaaldrich.com/life-science.html
Signaling Gateway	http://www.signaling-gateway.org
Signal Transduction KE	http://stke.sciencemag.org
SuperArray	http://www.superarray.com

Table 2.3. *Pathway databases included in the Molecular Signatures Database Canonical Pathways collection.*

For the gene analysis, I grouped variants into genes using SNP locations from dbSNP version 135 and NCBI 37.3 gene definitions. I performed this step twice, once annotating SNPs to a gene only if they fell within the gene's transcription start and stop site, and once including SNPs that fell within a 20 kilobase window of the gene.

I ran two separate pathway analyses, one using the Molecular Signatures Database canonical pathways collection[203], comprising 1,329 manually curated gene-sets from nine databases (Table 2.3), and one using 6,166 gene-sets from the Gene Ontology pathway database[204]. Significance was defined using a 5% FDR equivalent to a q-value of 0.05 for both the gene and pathway analyses[205].

2.2.8 Cross-phenotype meta-analysis

I used a multi-trait meta-analysis approach to search for novel associations in each pairwise combination of arcOGEN and GEFOS datasets[114]. The method, CPASSOC, requires only summary data and generates two test statistics:

The first, S_{hom}, assumes homogeneous effects across studies and is equivalent to performing an inverse variance weighted meta-analysis if no sample overlap between the studies exists. The second, S_{het}, is more powerful if effects are heterogeneous between studies. Both statistics require the specification of a correlation matrix of dimensions KxK, where K is the number of studies or traits included. I used tetrachoric correlation to construct this matrix as described in section "2.2.3. Estimating sample overlap".

To investigate whether any of the genome-wide significant signals were novel, I extracted a list of independent top variants ($r^2<0.1$ with any SNP within 500 kb) for both S_{het} and S_{hom} in each analysis. I then looked up their p-value in GEFOS and arcOGEN to see whether the signal could be explained entirely by either of the cohorts. Variants that did not fall within a genome-wide significant OA or BMD locus ($r^2>0.2$ or within 500 kb of genome-wide significant SNPs) were followed-up using the GWAS catalogue resource (https://www.ebi.ac.uk/gwas, date accessed: 23/03/2017). I performed an *in-silico* lookup in the UK Biobank hip and/or knee OA data of top SNPs with $p < 5x10^{-8}$ in any of the CPASSOC analyses that did not fall into known OA or BMD loci.

2.2.9 Replication and meta-analysis for OA

I took forward a total of 143 SNPs for *in silico* replication. This set comprises the two most strongly associated variants (one for each trait) in each region from the Bayesian colocalisation test, as well as all variants overlapping at P_t=0.005 in the SNP-based overlap analyses. I used the METAL[206] software package to perform inverse variance weighted meta-analysis of these SNPs in using summary statistics from the arcOGEN combined OA dataset (including London samples), the UK Biobank[46] and the deCODE[207] study. I first performed a meta-analysis across the replication datasets (UK Biobank and deCODE) and then across all three datasets (UK Biobank, deCODE and arcOGEN) (Appendix A).

2.2.10 Functional follow-up of *SMAD3*

Details of sample description and processing can be found elsewhere[208]. Briefly, articular cartilage was obtained from 12 patients undergoing total joint replacement for knee OA, and 9 patients for hip OA. Cartilage was graded using the OARSI cartilage classification system[209, 210].

2.3 Results

2.3.1 Genome-wide genetic correlation

I used linkage disequilibrium (LD) score regression to estimate the genome-wide genetic correlation between OA and BMD. There was a significant correlation between combined OA and LSBMD (r^2=0.18; p=0.022), as well as combined OA and peadiatric total body BMD (r^2=0.22; p=0.019, Figure 2.1).

2.3.2 Extent of shared association signals

I found evidence for significant overlap of association signals at different p-value thresholds (P_t) between all three OA categories and LSBMD (permutation adjusted p-value (p_{perm})<0.05) (Table 2.4).

Analysis of the combined OA and LSBMD data resulted in significant overlap p-values at P_t=0.001 and 0.005, as well as at less stringent P_t. Four SNPs overlap at P_t=5x10^{-4} (rs17158899, rs4536164, rs11826287 and rs630765), one of which (rs11826287) is genome-wide significantly associated with FNBMD (p=3.61x10^{-14}) and maps to an intron in *LRP5*. The highest overlap was observed between hip OA and LSBMD, with six SNPs overlapping at P_t=5x10^{-4} (p_{perm}=5.7x10^{-5}). Two of these SNPs are genome-wide significantly associated with BMD in GEFOS (rs1524928, p=5.29x10^{-9} and rs716255, p=2.07x10^{-11}). A significant overlap was also observed for P_t of 0.001, 0.005 and 0.01 in the hip OA-LSBMD comparison. Compared to the hip OA and LSBMD analysis, overlap p-values for the knee OA and LSBMD comparison were at least one order of magnitude smaller.

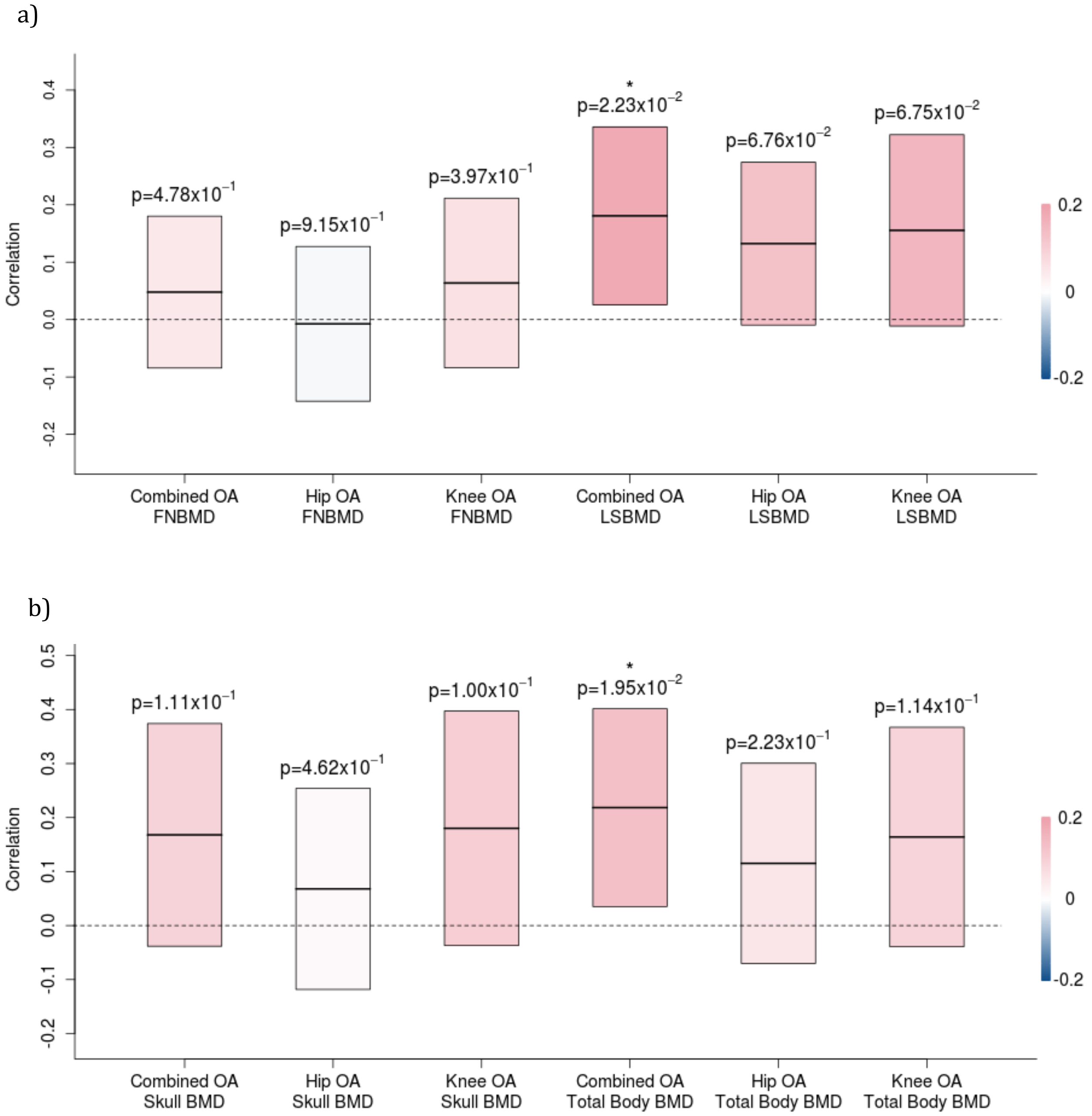

Figure 2.1. *Genetic correlation between osteoarthritis (OA) and bone mineral density (BMD) as estimated by LD score regression. Correlations were calculated between each pairwise comparison of phenotypes in arcOGEN and GEFOS (a), and arcOGEN and a paediatric BMD cohort (b). Rectangles show the correlation estimate (middle horizontal line) and standard errors (upper and lower bounds) of each comparison. Rectangles are coloured according to the strength of correlation. Significant correlation estimates are marked by an asterisk. LSBMD=lumbar spine BMD; FNBMD=femoral neck BMD*

	knee OA				hip OA				combined OA			
	LSBMD		FNBMD		LSBMD		FNBMD		LSBMD		FNBMD	
Total number of SNPs	75,125		75,270		74,999		75,147		75,015		75,161	
Pt	**SNPs**	**P_perm**	**SNPs**	**P_perm**	**SNPs**	**P_perm**	**SNPs**	**P_perm**	**SNPs**	**P_perm**	**SNPs**	**P_perm**
0.5	28,910	7.95x10^{-1}	27,534	5.84x10^{-2}	28,563	1.64x10^{-1}	27065	7.33x10^{-1}	28,940	8.53x10^{-2}	27,599	2.74x10^{-1}
0.1	2,620	3.67x10^{-2}	2,131	2.33x10^{-1}	2,513	1.46x10^{-2}	2057	1.74x10^{-1}	2,633	7.92x10^{-2}	2,119	5.41x10^{-1}
0.05	939	8.98x10^{-4}	700	3.10x10^{-2}	868	1.42x10^{-2}	647	2.47x10^{-1}	931	1.25x10^{-2}	698	2.35x10^{-1}
0.04	680	4.74x10^{-4}	474	1.27x10^{-1}	641	6.48x10^{-4}	434	4.99x10^{-1}	682	2.41x10^{-3}	486	2.33x10^{-1}
0.03	437	2.87x10^{-3}	291	3.27x10^{-1}	416	1.02x10^{-3}	272	4.20x10^{-1}	451	1.20x10^{-3}	314	6.57x10^{-2}
0.02	244	1.59x10^{-3}	152	3.03x10^{-1}	239	6.10x10^{-5}	150	6.31x10^{-2}	251	1.95x10^{-3}	160	1.07x10^{-1}
0.01	88	3.01x10^{-3}	53	2.18x10^{-1}	87	2.97x10^{-4}	52	4.94x10^{-2}	87	2.79x10^{-2}	60	5.84x10^{-2}
0.005	29	1.97x10^{-1}	16	6.94x10^{-1}	38	3.04x10^{-4}	26	6.39x10^{-4}	37	6.89x10^{-3}	24	2.43x10^{-2}
0.001	6	1.12x10^{-2}	4	3.39x10^{-2}	7	1.68x10^{-3}	1	1	10	3.90x10^{-5}	3	4.52x10^{-2}
5x10^{-4}	4	4.21x10^{-3}	2	7.12x10^{-2}	6	5.70x10^{-5}	1	5.64x10^{-2}	4	6.81x10^{-3}	0	1

Table 2.4. *SNP-based overlap analysis of OA and BMD. For each comparison the total number of SNPs present in both datasets after LD-pruning pruning is given, as well as the number of SNPs falling below each p-value threshold (Pt). P_perm=empirical overlap p-value obtained through permutation analysis*

Four SNPs (rs7104420, rs9466056, rs881803 and rs4536164) overlapped at P_t=5x10^{-4} in this analysis (p_{perm} =4.21x10^{-3}). Two of these, rs4536164 and rs9466056, fall within known BMD risk loci[13].

Overlap signal was much weaker for the OA and FNBMD comparisons, with only five P_t reaching statistical significance (P_t=0.001 for knee OA, P_t=0.005 and 0.01 for hip OA, and P_t=0.001 and P_t=0.005 for combined OA). The SNP overlapping at P_t=5x10^{-4} for hip OA and FNBMD (rs1524928) was also among the six SNPs identified in the hip OA and LSBMD analysis. Two SNPs overlapped at this P_t for knee OA, one being rs9466056 and the other rs1283614, which maps to an intron of the BMD locus *MEF2C*[211].

2.3.3 Evidence for colocalising regions

I employed a regional Bayesian colocalisation test that measures the posterior probabilities for each of four alternative hypotheses compared to one global null hypothesis (i.e. no associations for either trait in that region). I identified four independent genomic regions with a high posterior probability of harbouring one causal variant common to both traits analysed (posterior probability for hypothesis 3≥0.9) (Table 2.5).

	Analysis	SNPs	Chr	Start (bp)	Stop (bp)	Top SNP BMD	Top SNP OA	PP
Hypothesis 3	**allOA and LSBMD**	817	chr14	91297823	93129850	rs1286147; rs1286063	rs1286077	0.95
	hipOA and FNBMD	817	chr14	91297823	93129850	rs1286147	rs1286077	0.98
	hipOA and LSBMD	817	chr14	91297823	93129850	rs1286147; rs1286063	rs1286077	0.99
	hipOA and LSBMD	1242	chr10	78708452	80875213	rs7071206	rs716255	0.92
	hipOA and LSBMD	531	chr1	44974119	46897698	rs7554123	rs7545984	0.91
	kneeOA and FNBMD	1235	chr6	19208477	21677746	rs9466056	rs9466056	0.99
Hypothesis 4	**hipOA and LSBMD**	268	chr4	696848	1415698	rs3755955	rs3755920	0.97
	kneeOA and LSBMD	382	chr16	14464538	16152940	rs4985155	rs9935327	0.95
	kneeOA and LSBMD	1070	chr6	150255029	151910904	rs4869742	rs9384514	0.90

Table 2.5. *Regions with strong evidence of pleiotropy. For each region the number of SNPs, start and stop position in basepairs (bp) and most strongly associated SNPs for OA and BMD are given. Chromosome coordinates are in hg19. Hypothesis 3=one causal variant; hypothesis 4=two distinct causal variants; PP=posterior probability*

The region containing the *RPS6KA5* gene was identified by three comparisons (combined OA and LSBMD, hip OA and LSBMD, and hip OA and FNBMD). The most strongly associated SNPs in this region lie in introns of *RPS6KA5* and are genome-wide for increased BMD at both the lumbar spine and femoral neck (rs1286147 and rs1286063, $p<5x10^{-8}$) and nominally significant associations with increased risk of combined and hip OA (rs1286077, p<0.05); the three SNPs are in perfect LD (r^2=1.00 for each pairwise combination).

Two further regions were identified in the hip OA and LSBMD analysis. The first spans a known LSBMD locus upstream of the *KCNMA1* gene on chromosome 10 and contains a regulatory variant in a CTCF binding site that is nominally significant for hip OA (rs716255, p=0.001). The second lies on chromosome 1 and contains two nominally significant variants for LSBMD (rs7554123, p= $1.12x10^{-4}$) and hip OA (rs7545984, p=$1.29x10^{-4}$), respectively, which both fall within an intron of *RNF220*.

The region identified in the knee OA and FNBMD analysis contains one lead SNP for both traits, rs9466056, which is associated with high FNBMD (p=$1.8x10^{-8}$) and decreased risk of knee OA (p=$1.1x10^{-4}$), mapping to an intergenic region between *CDKAL1* and *SOX4*.

I also identified three regions (Table 2) with a high posterior probability of harbouring two distinct causal variants (PP for hypothesis 4≥0.9). All three of these contain a known BMD locus, with the top SNPs for LSBMD mapping to introns of *IDUA*, *CCDC170* and *PDXDC1*. The top SNPs for knee and hip OA are nominally associated (p<0.05) with these respective phenotypes in arcOGEN.

2.3.4 Gene and pathway analysis

Of the individual genes significantly associated (q<0.05) with at least one OA or BMD phenotype, *SUPTH3, COL11A1,* and *APCDD1* overlapped between OA and BMD (Table 2.6). All three include variants that were identified in the SNP-wise overlap analysis and taken forward for replication.

Gene	Combined OA	Hip OA	Knee OA	LSBMD	FNBMD
COL11A1	3.40E-01	1.04E-02	8.08E-01	2.31E-02	4.46E-04
SUPT3H	1.40E-01	3.74E-02	7.55E-01	7.27E-04	6.98E-01
APCDD1	3.72E-02	9.14E-01	3.58E-02	4.16E-02	3.62E-01

Table 2.6. *False discovery rate corrected p-values (q-values) for the three genes significantly associated with at least one osteoarthritis (OA) and one bone mineral density (BMD) phenotype. LSBMD=lumbar spine BMD; FNBMD=femoral neck BMD*

There were no pathways significantly associated with any OA phenotype in any of the analyses. One of the CP pathways was associated with FNBMD ("basal cell carcinoma", q=0.02) when allowing a 20 kilobase (kb) window around genes. Using GO annotations a total of 33 unique pathways were associated with either BMD phenotype using strict or lenient gene definitions (Table 2.8; Table 2.7), including several with direct biological relevance, such as "regulation of ossification" or "osteoblast development".

Pathway	Genes	Beta	SE	P	P_{BH}
Skeletal System Development	438	0.189	0.04	8.90E-06	2.74E-02
Positive Regulation Of Cartilage Development	28	0.657	0.16	2.14E-05	4.41E-02
Positive Regulation Of Chondrocyte Differentiation	19	0.836	0.19	6.04E-06	2.74E-02

Table 2.7. *GO pathways significantly associated with FNBMD when including a 20kb window around genes. P=raw p-values; P_{BH}=false discovery rate corrected p-values(q-values)*

Pathway	Genes	Beta	SE	P	P_{BH}
Formation Of Primary Germ Layer	107	0.30	0.08	5.58E-05	2.06E-02
Negative Regulation Of Fat Cell Differentiation	41	0.64	0.15	6.41E-06	6.39E-03
Branch Elongation Of An Epithelium	17	1.13	0.27	9.87E-06	6.39E-03
Mammary Gland Epithelium Development	51	0.51	0.12	1.33E-05	7.45E-03
Skeletal System Development	438	0.20	0.05	8.78E-06	6.39E-03
Embryo Development	861	0.13	0.03	2.12E-05	1.01E-02
Canonical Wnt Signaling Pathway	87	0.38	0.09	9.15E-06	6.39E-03
Somitogenesis	57	0.42	0.12	1.56E-04	3.85E-02
Phosphate Containing Compound Metabolic Process	1875	0.08	0.02	8.98E-05	2.52E-02
Gastrulation	147	0.24	0.07	1.75E-04	4.11E-02
Mammary Gland Development	112	0.30	0.08	8.33E-05	2.52E-02
Positive Regulation Of Peptidyl Threonine Phosphorylation	24	0.78	0.17	1.44E-06	2.96E-03
Embryonic Morphogenesis	524	0.14	0.04	2.00E-04	4.11E-02
Regulation Of Catenin Import Into Nucleus	26	0.58	0.16	1.91E-04	4.11E-02
Mammary Gland Alveolus Development	16	1.48	0.22	1.71E-11	5.27E-08
Regulation Of Peptidyl Threonine Phosphorylation	35	0.65	0.15	5.59E-06	6.39E-03
Mammary Gland Lobule Development	16	1.48	0.22	1.71E-11	5.27E-08
Osteoblast Development	18	0.99	0.24	1.58E-05	8.14E-03
Dorsal Ventral Axis Specification	20	0.74	0.18	2.49E-05	1.10E-02
Mammary Gland Epithelial Cell Proliferation	12	1.16	0.27	1.04E-05	6.39E-03
Muscle Cell Differentiation	230	0.22	0.06	7.21E-05	2.34E-02
Phosphorylation	1160	0.09	0.03	1.99E-04	4.11E-02
Somite Development	72	0.42	0.11	3.96E-05	1.63E-02
Embryonic Organ Development	391	0.21	0.05	4.84E-06	6.39E-03
Axis Elongation	26	0.69	0.18	5.69E-05	2.06E-02
Regulation Of Ossification	170	0.25	0.07	1.85E-04	4.11E-02
Regulation Of Stem Cell Differentiation	113	0.30	0.08	8.77E-05	2.52E-02
Beta Catenin Destruction Complex	14	0.79	0.22	1.22E-04	3.13E-02
Protein Complex Scaffold	66	0.44	0.12	6.37E-05	2.18E-02
Glutamate Receptor Binding	35	0.55	0.16	2.16E-04	4.30E-02
G Protein Coupled Receptor Binding	245	0.23	0.06	9.39E-05	2.52E-02

Table 2.8. *GO pathways significantly associated with LSBMD when including a 20kb window around genes. P=raw p-values; P_{BH}=false discovery rate corrected p-values(q-values)*

2.3.5 Cross-phenotype meta-analysis

To search for potential novel associations not identified by single-trait GWAS, I performed a cross-phenotype meta-analysis between each pairwise combination of OA and BMD datasets (Figure 2.2; Figure 2.3; Figure 2.4; Figure 2.5). Using the CPASSOC method[114], I computed two statistics, S_{hom} and S_{het}, which assume homogeneous and heterogeneous effects across studies, respectively. I identified 13 independent associations not previously reported for BMD or OA, which I followed up in the UK Biobank combined OA dataset (Table 2.9). One SNP, rs11164649, was nominally significant ($p<0.05$). This SNP lies in an intron of the *COL11A1* gene and is in strong LD ($r^2=0.92$) with a variant (rs1903787) identified in the SNP-wise overlap analysis which was taken forward for replication.

SNP	CHR	POS	EA	NEA	P	BETA	SE
rs7545984	1	45003893	C	T	7.78E-01	0.014	0.048
rs12060207	1	98335381	T	C	5.65E-01	0.018	0.031
rs11164649	1	103444679	G	T	3.83E-02	0.036	0.017
rs17578878	4	37900725	T	C	6.24E-01	-0.013	0.026
rs7735525	5	95981203	G	A	4.89E-01	-0.055	0.080
rs1748234	6	45140853	T	C	5.91E-01	-0.009	0.016
rs7853022	9	18930055	T	G	5.33E-01	0.014	0.022
rs996793	9	23676631	G	A	2.55E-01	-0.061	0.054
rs10491510	9	35040245	T	C	5.56E-01	0.041	0.070
rs11188469	10	97511901	T	G	7.42E-01	0.010	0.031
rs7908390	10	100147247	C	T	3.51E-01	0.267	0.279
rs17098135	14	61689992	C	A	9.38E-01	-0.005	0.064
rs2197166	18	10493908	A	G	7.41E-01	-0.006	0.018

Table 2.9. *In-silico look up in the UK Biobank combined OA data of top SNPs that reached genome-wide significance in any of the CPASSOC analyses for S_{het} or S_{hom}, and are not known osteoarthritis (OA) or bone mineral density (BMD) loci. For each SNP summary statistics in the form of effect estimates (BETA), standard errors (SE) and p-values (P). EA=effect allele; NEA=non-effect allele*

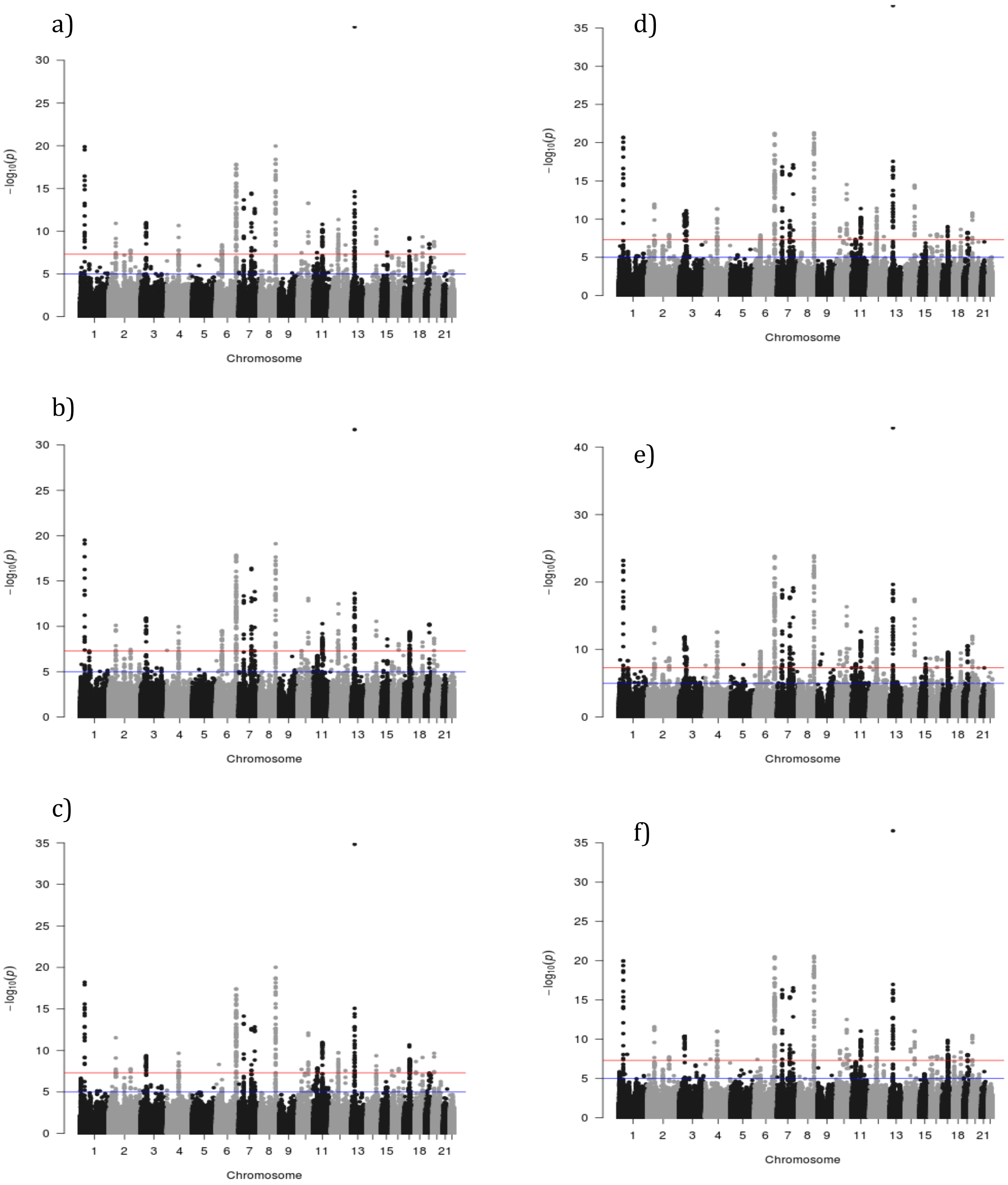

Figure 2.2. *Manhattan plots of multi-trait meta-analysis between osteoarthritis (OA) and lumbar spine bone mineral density (LSBMD). The CPASSOC method was used to calculate the Shom (a-c) and Shet (d-f) statistics for combined OA and LSBMD (a,d), hip OA and LSBMD (b,e) and knee OA and LSBMD (c,f).*

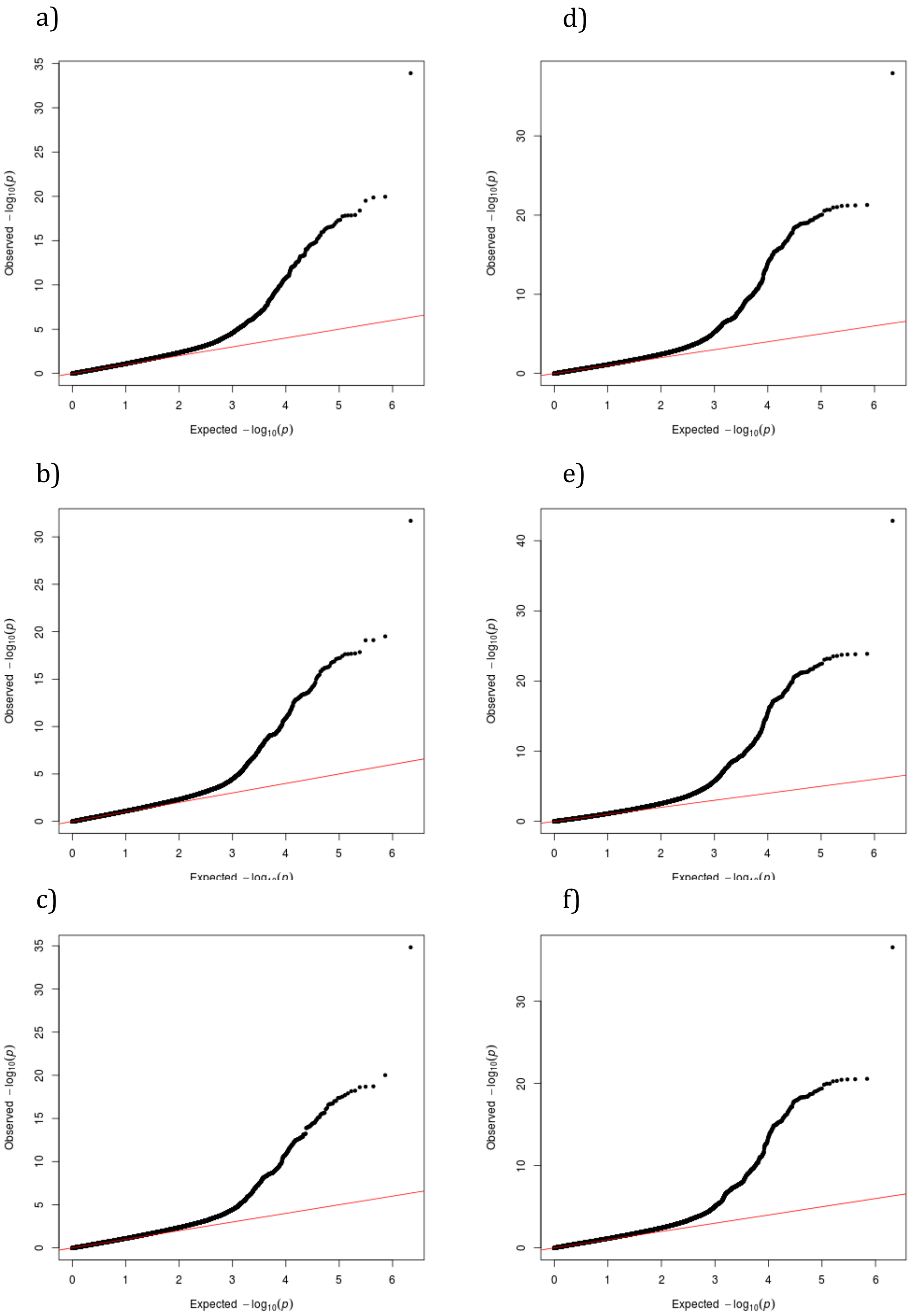

Figure 2.3. *Quantile-quantile plots of multi-trait meta-analysis between osteoarthritis (OA) and lumbar spine bone mineral density (LSBMD). The CPASSOC method was used to calculate the Shom (a-c) and Shet (d-f) statistics for combined OA and LSBMD(a,d), hip OA and LSBMD (b,e) and knee OA and LSBMD (c,f).*

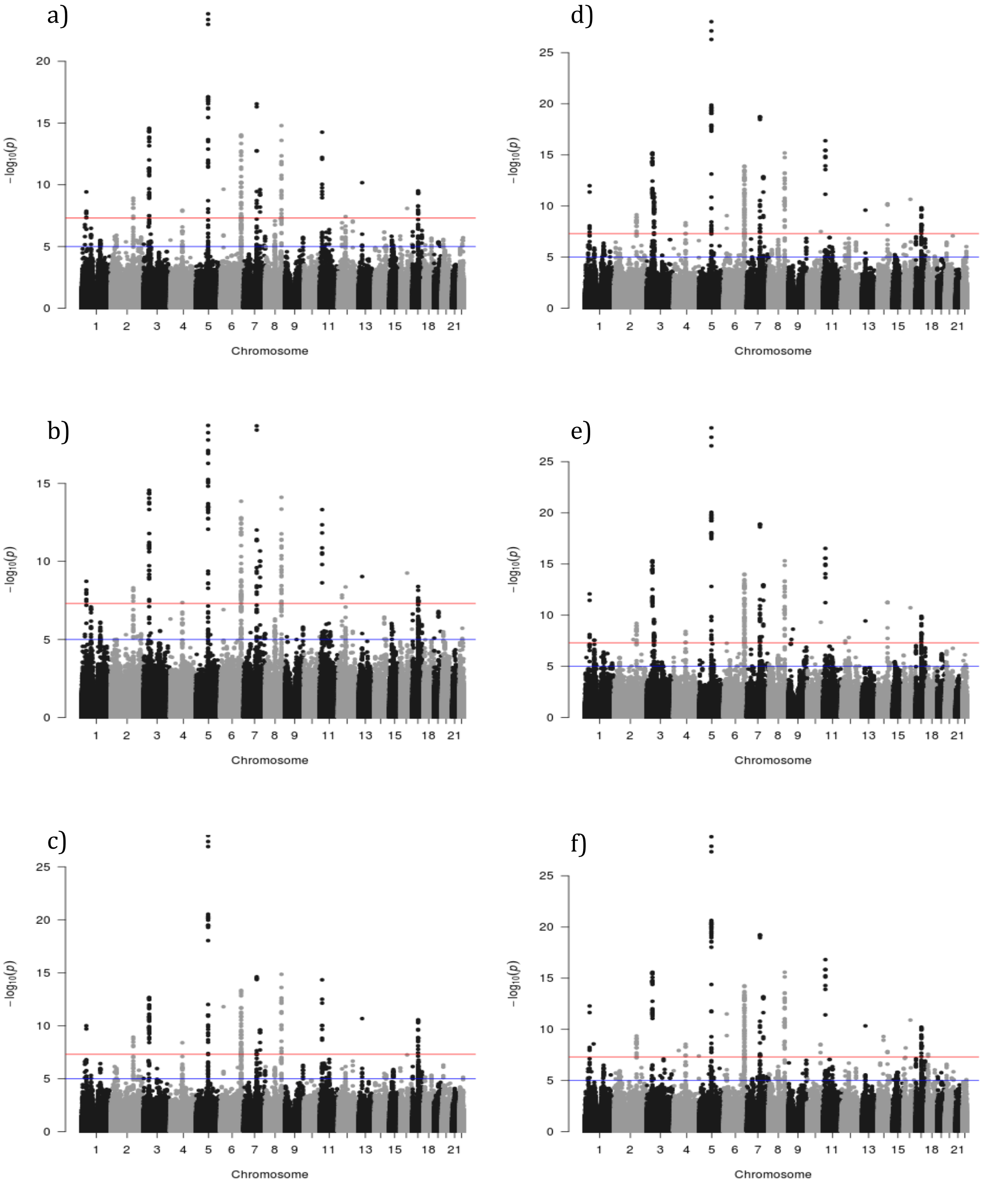

Figure 2.4. *Manhattan plots of multi-trait meta-analysis between osteoarthritis (OA) and femoral neck bone mineral density (FNBMD). The CPASSOC method was used to calculate the Shom (a-c) and Shet (d-f) statistics for combined OA and FNBMD(a,d), hip OA and FNBMD (b,e) and knee OA and FNBMD (c,f).*

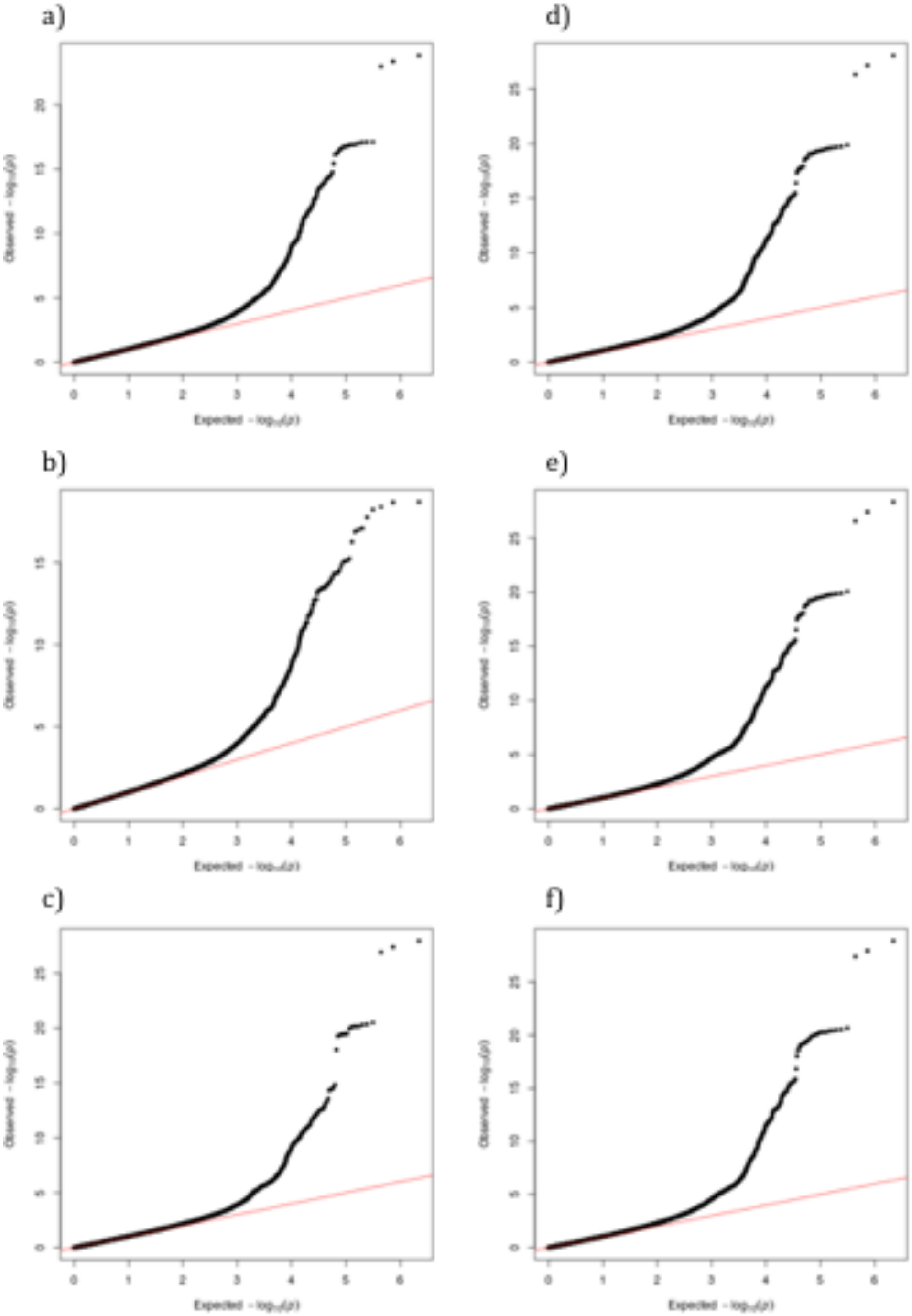

Figure 2.5. *Quantile-quantile plots of multi-trait meta-analysis between osteoarthritis (OA) and femoral neck bone mineral denisty (FNBMD). The CPASSOC method was used to calculate the Shom (a-c) and Shet (d-f) statistics for combined OA and FNBMD(a,d), hip OA and FNBMD (b,e) and knee OA and FNBMD (c,f).*

2.3.6 Replication and meta-analysis for OA

I took forward a total of 143 SNPs identified in the colocalisation and/or p-value based overlap analysis for replication in UK Biobank and deCODE (Appendix A). None of the SNPs taken forward are genome-wide significantly associated in arcOGEN.

I subsequently meta-analysed the above list of SNPs across UK Biobank and deCODE, and then across arcOGEN, UK Biobank and deCODE. I found a significant excess of independent SNPs with the same direction of effect among variants with $p_{meta}<0.05$ in the replication cohorts (binomial sign test $p= 2.62x10^{-06}$) and across all three cohorts (binomial sign test $p=7.75x10^{-11}$), as well as all independent SNPs included in the meta-analysis of the replication cohorts (binomial sign test p=0.002) and all three cohorts (binomial sign test p=0.03).

Variants within several genes linked to bone, cartilage and extracellular matrix biology, including *APCDD1, SUPTH3, COL11A1, NOTCH4, SEMA3A, LGR4, PTCH1* and *RPS6KA5*, were associated at $p_{meta}<0.05$ (Appendix A).

Two variants reached genome-wide significance in the meta-analysis across arcOGEN, deCODE and UK Biobank: rs12901071 (OR 1.08 [95% CI 1.05-1.11], $p_{meta}=3.12x10^{-10}$) and rs10518707 (OR 1.07, [95% CI 1.03-1.09], $p_{meta}=2.15x10^{-8}$). They were also nominally significant in the UK Biobank-deCODE meta-analysis (rs12901071: $p=2.46x10^{-07}$; rs10518707: $p=7.90x10^{-06}$). Both are intronic variants in the *SMAD3* gene ($r^2=0.645$) and were identified in the SNP-wise overlap analysis of combined OA vs. LSBMD and hip OA vs. LSBMD, respectively (Figure 2.6).

Both new genome-wide significant SNPs for OA were imputed in the arcOGEN data (imputation info score>0.95) and are nominally associated with combined OA (Appendix A). They are also nominally associated with increased LSBMD in GEFOS (rs12901071, $p=1.58x10^{-3}$ and rs10518707, $p=3.47x10^{-5}$), but not FNBMD (rs12901071, $p=3.72x10^{-1}$ and rs10518707, $p=2.46x10^{-1}$). *SMAD3* is associated (p<0.05) with LS and FNBMD, hip and combined OA in the gene analysis (Table 2.10), although this association only holds for LSBMD when using false discovery rate (FDR) correction ($q=6.92x10^{-6}$).

In the colocalisation analysis, the region in which both top SNPs reside (chr15:67,095,629-69,017,421) has a posterior probability of containing a single pleiotropic variant associated with hip OA and LSBMD (hypothesis 3) of 0.88.

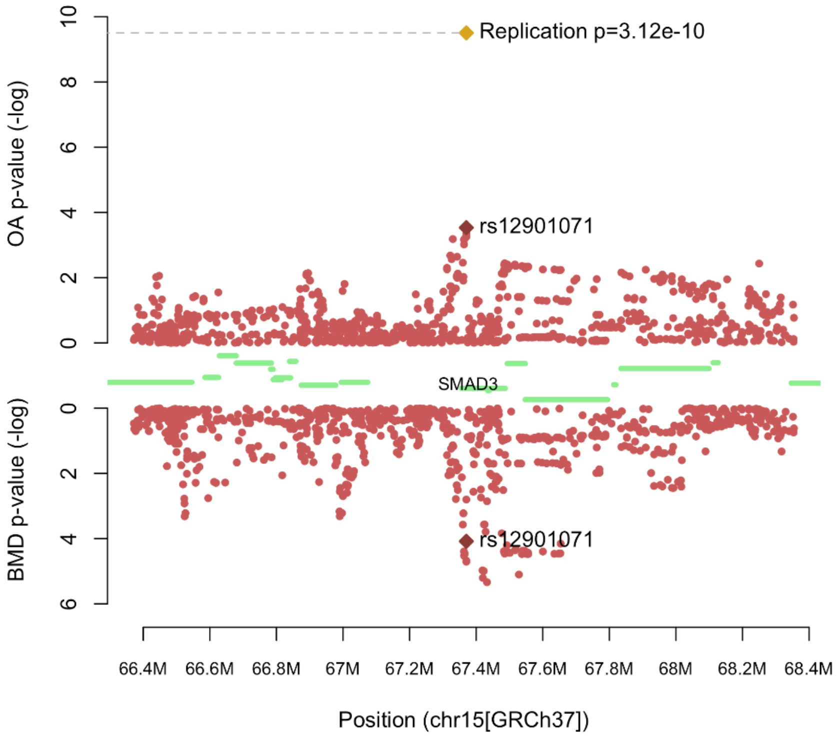

Figure 2.6. *Regional association plot of SMAD3. The -log(p-values) of SNPs in the arcOGEN combined osteoarthritis (OA) data (top) and GEFOS lumbar spine bone mineral density (LSBMD) data (bottom) are plotted against their chromosomal position. The meta-analysis p-value of rs12901071 is plotted as a golden diamond. Protein coding genes are represented by green bars.*

Trait	**Variants**	**P**	**P_{BH}**
Combined OA	100	4.96E-03	3.45E-01
Hip OA	99	9.18E-04	2.17E-01
Knee OA	100	4.43E-01	9.21E-01
FNBMD	202	3.79E-03	1.98E-01
LSBMD	202	1.11E-08	6.92E-06

Table 2.10. *Results for* SMAD3 *in the gene analysis. P=raw p-values; P_{BH}=false discovery rate corrected p-values(q-values)*

2.3.7 Functional follow-up of *SMAD3*

Using RNA sequencing data, I confirmed the expression of *SMAD3* in low-grade degenerate articular cartilage of 12 knee and 9 hip OA patients undergoing total joint replacement[208] (Figure 2.7). *SMAD3* is among the 30% most expressed genes in the knee articular cartilage samples, and among the 15% most expressed genes in the hip articular cartilage samples.

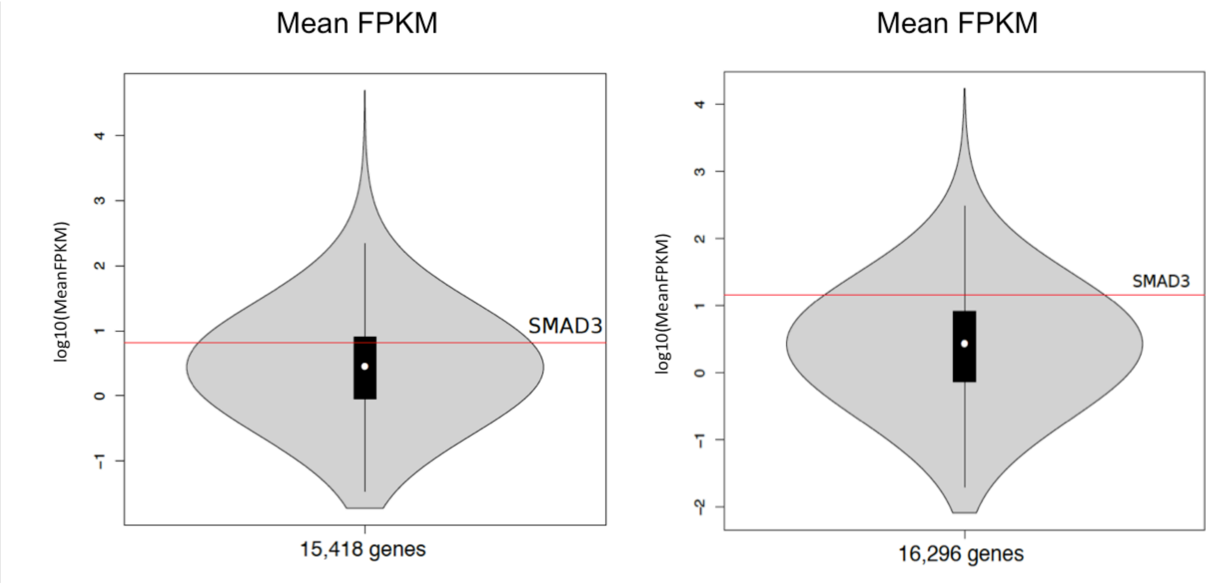

Figure 2.7. *Mean expression of 15,418 genes in low-grade degenerate articular cartilage of the knee (left) and 16,296 genes in intact articular cartilage of the the hip (right). Mean* SMAD3 *expression is shown by a red line. FPKM: fragments per kilobase per million mapped reads. Boxplots represent the median (white dot), interquartile range (IQR; black box) and the lowest and highest value still within 1.5 IQR of the lower and upper quartile, respectively (whiskers).*

2.4 Discussion

The analysis of shared genetic aetiology across epidemiologically linked traits can enhance power to identify disease variants and shed light into the biological mechanisms underpinning these associations. I conducted the first genome-wide overlap analysis of BMD and OA using summary statistics from two large-scale GWAS of these traits, respectively. It should be note that test statistics from individual analyses were not corrected for multiple testing. Bonferroni correction would have been overly conservative, as the the OA and BMD datasets were comprised of highly correlated phenotypes, respectively. A more appropriate approach for multiple testing correction would have been calculating the effective number of phenotypes among the five OA and BMD groups; however, this was not possible as individual-level data was not available for BMD.

2.4.1 Differential overlap of FN- and LSBMD with OA

There was a stronger overlap between OA and LSBMD than for OA and FNBMD, both in the SNP-based and genetic correlation analyses. The fact that only the correlation between combined OA and LSBMD was significant could be due to the bigger sample size in this OA dataset compared to the hip or knee OA data. While the FN- and LSBMD datasets are very similar in size, the knee OA and hip OA datasets each contain approximately half the number of cases compared to the combined OA dataset. This difference in power might at least partly explain why the genetic correlation estimates for joint-specific OA and LSBMD did not achieve statistical significance.

Epidemiological data from the Chingford study have shown increased baseline BMD to be associated with incident radiographic knee OA, with the mean increase in LSBMD being approximately twice as high as the increase in FNBMD[179, 184]. Incident knee OA was also linked to higher baseline LSBMD, but not FNBMD, in the Baltimore Longitudinal Study of Ageing[212]. The reasons for this differential association of FN- and LSBMD with OA remain unclear. One possible explanation could be the comorbidity of knee and spinal OA, characterised by spinal osteophytes, which could lead to increased LSBMD measurements. However, in one study, adjustment for the presence of osteophytes at the lumbar spine did not change the strength of association between OA and LSBMD[179]. Damage to the spine accumulates over time and can lead to changes such as breakdown of the invertebral discs, scoliosis and osteochondrosis, a process also referred to as degenerative disc disease (DDD). Although the association between DDD and LSBMD remains unclear[213-216], it is known that the presence of degenerative features can increase LSBMD measurements obtained via dual X-ray absorptiometry[217]. While this might have contributed to the observed association between LSBMD and OA, I found genetic correlations of a similar magnitude between OA and skull, as well as total body BMD measurements in a paediatric cohort[218]. As DDD and related features such as osteophytes are unlikely to be present in young individuals, these results suggest that the correlation between OA and LSBMD is not purely artefactual.

2.4.2 Genetics of hip vs knee OA

Both LD score regression and the SNP-based overlap analysis showed a greater degree of overlap between hip OA and both BMD measurements than between knee OA and BMD. Hip OA is estimated to have a higher heritability than knee OA[219, 220], with environmental risk factors such as physical activity and BMI more strongly associated with the latter[221]. A recent study of 9,000 twin pairs also found that genetics explained 73% of variation in hip arthroplasty due to OA were explained by genetics, compared to 45% in knee arthroplasty[219]. In other words, these findings suggest that progression to severe OA at the hip is more strongly influenced by genetics than at the knee. The same study also showed a stronger dependence of knee arthroplasty on BMI.

2.4.3 Variants and regions with potential pleiotropic effects

I identified 143 variants with evidence for potential pleiotropic effects on OA and BMD. Many of these reside in or near biologically relevant genes, two of which (*KLHL42*/*KLHDC5* and *SUPT3H/RUNX2*) are established loci for both traits[13, 38]. Variants in three loci (*SUPTH3*, *APCDD1,* and *COL11A1*) were also significantly associated with at least one OA and BMD phenotype in the gene analysis. *APCDD1* is an inhibitor of *WNT* signaling[222], which is implicated in both OA and BMD. *COL11A1* encodes collagen type 11, an important component of cartilage and bone, and has been associated with OA in a candidate gene meta-analysis[223]. Other examples include the *LGR4* gene, in which a rare variant in the Icelandic population has been associated with low BMD and osteoporotic fractures[224]; and *SEMA3A*, which affects bone remodeling in rats[225].

RNF220, which was identified in the hip OA-LSBMD colocalisation analysis, increases canonical Wnt signaling[226], a key pathway involved in bone remodeling and osteoarthritis[189, 227]. The protein product of *RNF220* de-ubiquitinates beta-catenin by forming a complex with a ubiquitin-specific peptidase[226].

2.4.4 *SMAD3* as a novel osteoarthritis risk locus

I identified novel genome-wide significant associations at two intronic SNPs in *SMAD3*, and confirm expression of this gene in primary chondrocytes from articular cartilage of OA patients undergoing total joint replacement surgery. Activated SMAD3 acts downstream of TGF-β, repressing osteoblast differentiation and the production of bone matrix[228, 229]. It also represses the cartilage-degrading enzyme matrix metalloproteinase 13 in chondrocytes[229]. Missense mutations in a conserved protein domain of *SMAD3* have been linked to aneurysm-osteoarthritis syndrome, a congenital disorder characterised by arterial aneurysms, heart abnormalities and early-onset OA[230].

Due to its role in bone and cartilage biology, *SMAD3* has been previously assessed in a candidate gene study of hip and knee OA[231]. Despite their small sample size (number of cases<400), the investigators found nominal associations ($p<0.05$) for both OA phenotypes in their discovery, which were further strengthened in a meta-analysis (hip OA $p=4x10^{-4}$; knee OA $p=7.5x10^{-6}$). Notably, their top signal (rs12901499) maps to the same locus as the lead SNP ($r^2=0.645$) in the meta-analysis presented here.

More recently, two studies have shown *SMAD3* expression to be correlated with the genotype at a 3'UTR SNP[232], and to be significantly higher in cartilage from OA patients compared to healthy controls[233]. The authors postulate that this could be a compensatory mechanism to counteract existing cartilage damage, or that *SMAD3* expression levels outside a narrow range have detrimental effects. The top SNP in the arcOGEN-deCODE-UKBB meta-analysis, rs12901071, is associated with increased *SMAD3* expression in skeletal muscle tissue ($p= 7.5x10^{-6}$) in GTEx[234].

2.4.5 Limitiations and future work

This work exemplifies the potential to uncover new disease risk loci by combining data of epidemiologically linked traits. Methods combining univariate summary statistics of different traits – such as the colocalisation analysis employed here[78] – often do not require a locus to be genome-wide significantly associated in any of the individual studies in order to detect a cross-phenotype association. Hence, they can increase power to identify associated variants or regions without the need to collect larger sample sizes[58]. A downside of such approaches is that they do inherently rely on the assumption that bot input datasets are well-powered to detect associations. Again, using the example of the colocalisation method, a failure to detect any regions with strong evidence for a shared causal variant could be due to the lack of shared signals or the fact that one or both datasets are underpowered. There is a stark difference in sample size and, consequently, statistical power between the arcOGEN and GEFOS GWAS datasets. Larger datasets where phenotype information for both OA and BMD is available in the same individuals will aid in further disentangling the extent of shared genetics between them. The full UK Biobank dataset, which was released in early 2018, comprises half a million people of European ancestry and includes both clinical and self-reported OA phenotypes. While only a small subset of participants currently have DXA BMD measurements, heel BMD measured by quantitative ultrasound is available for approximately 270,000 individuals, and has been used to successfully identify novel BMD loci in the first release UK Biobank samples[176].

There is currently no early detection protocol for osteoarthritis[235], and diagnosis consequently occurs only when the disease has become symptomatic, i.e. patients present with joint pain and discomfort. Lifestyle changes such as weight management and modified exercise are not as effective at a point where tissue degradation and localised bone remodeling has already taken place in the affected joint(s)[236]. Similarly, regenerating degraded cartilage in progressive OA is currently not feasible, and most putative drugs are aimed at halting degenerative processes[237]. Measurable biomarkers hold the promise of identifying individuals at high risk of developing OA and could facilitate early intervention before more severe damage occurs[238]. Coupled with genetic screening (and imaging data where appropriate), biomarkers could additionally help to stratify patients not only based on affected joint site, but molecular as well as physiological endophenotypes[239].

In this chapter, I identified genes with potential involvement in both OA and BMD. Their RNA and protein expression levels will need to be explored in OA cases and healthy controls, and should be cross-checked with the presence of osteophytes. The identification of pleiotropic loci could potentially help to determine whether patients are "bone formers" and at an increased risk of osteophytes, bone cysts and elevated bone turnover. Such molecular phenotyping could not only facilitate early detection, but also reveal pathways for potential pharmacological intervention. Currently the main line of medication for OA consists of analgesics to alleviate pain[165]. Several drugs aimed at cartilage or bone remodeling are being trialed for their use in OA treatment, but are either counter-indicated due to side effects or show limited efficacy[165, 236, 240]. A better understanding of the molecular processes underlying different hallmarks of OA will enable more refined drug targeting. Recent OA GWAS have also found OA-associated genes with effects on osteoclast differentiation and bone remodeling, further highlighting the link between joint and bone health[172, 241].

A flipside of these findings is that they reveal the high degree of pleiotropy in many OA-associated pathways. For example, the effects of TGF-beta signaling encompass cartilage maintenance, bone remodeling and immune cell function[242]. Consequently, drugs targeting these pathways will need to act locally in the affected joint (e.g. administered via injection)[243] and/or target a protein downstream in the signaling chain with a more specific function to avoid systemic side effects.

2.4.6 Conclusion

The analyses outlined here present the first comprehensive evaluation of genetic overlap between BMD and radiographic OA. Our results lend further support to the hypothesis of common genetic factors underlying these two traits and establish *SMAD3* as a genome-wide significant risk locus for OA with a potential pleiotropic effect on BMD. Pinpointing the common biological pathways of these two complex traits will provide insight into the underlying mechanisms of OA, facilitating the identification of novel therapeutic targets or drug repurposing opportunities for its treatment.

Chapter 3 – Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia

3.1 Introduction

Schizophrenia (SCZ) is a psychiatric disorder characterised by an inability to distinguish what is real from what is not. The most common symptoms include delusions, hallucinations and paranoia (collectively referred to as positive symptoms), as well as loss of motivation, social withdrawal (negative symptoms) and cognitive impairment[244]. The lifetime prevalence for SCZ is around 1%[244], and initial symptoms commonly appear during adolescence or early adulthood[245]; however, a diagnosis is often only made following the progression to psychosis and subsequent hospitalisation.

SCZ patients are at an elevated risk of developing metabolic syndrome compared to the general population, and are also 1.5-2 times more likely to develop type 2 diabetes (T2D)[246], whose hallmarks include isnulin resistance, high blood sugar and decreased insulin secretion by pancreatic beta cells[247]. Several theories regarding the cause of this epidemiologic link exist, including the use of antipsychotic medication and/or shared genetic aetiology[246, 248-250]. In addition, environmental factors are thought to play a role in the observed comorbidity. For example, patients with severe mental illness often lead a more sedentary life and are more likely to smoke compared to the general population[250] – both risk factors for T2D.

3.1.1 Metabolic effects of psychotropic drugs

Although psychosis is episodic in nature, negative symptoms of SCZ tend to be chronic, and require long-term management consisting of pharmacological intervention, psychotherapy and social support. Antipsychotic drugs are classified as first or second generation. The former includes compounds such as haloperidol and chlorpromazine, discovered serendipitously during the 1950s. While these drugs were effective in reducing psychotic symptoms, they also caused severe side effects including metabolic perturbation and medication-induced motor disorders (known as extrapyramidal symptoms). Second generation antipsychotics (also known as "atypicals"), such as clozapine, risperidone or olanzapine, are less likely to cause these motor control side effects, however, they are more likely to cause metabolic imbalances and often lead to significant weight gain[251]. Such perturbations have been aggregated into the umbrella term "metabolic syndrome", which encompasses cardiovascular, anthropometric and physiological measures such as hyptertension, obesity and insulin resistance[252].

Several studies have found an association between psychotropic medication and T2D risk[253-255], but it is still unclear to what extent interactions between different medications, life-style and inter-patient variability affect this association[250]. It is conceivable that the metabolic effects of antipsychotics are at least partly mediated by genetic predisposition. So far, studies on the genetics of antipsychotic response have been small (n < 400) and unable to identify replicating associations[256, 257].

3.1.2 Impaired metabolic regulation in drug-naïve SCZ patients

While the epidemiologic link between SCZ and T2D is often attributed to the side effects of psychotropic medication, there is evidence that metabolic dysregulation in SCZ may precede pharmacologic treatment. Proteomic studies have revealed perturbed expression of proteins involved in glucose metabolism in brain tissue and elevated insulin levels in peripheral blood of first-episode SCZ patients compared to controls[258, 259]. More recently, a large study following over 2.5 million Danish individuals found that antipsychotic-naïve SCZ patients were three times more likely to develop T2D than the general population, with antipsychotic drug use further increasing that risk[260]. This, along with findings from a systematic review and meta-analysis[261], suggests that impaired glucose homeostasis may already be present in drug-naïve SCZ patients.

3.1.3 Genetic basis of SCZ and T2D

It is also plausible that the observed overlap between SCZ and T2D is due to common susceptibility variants[248]. Both diseases are highly polygenic, with heritability estimates around 80% for SCZ[244] and 35% for T2D[247]. Efforts to aggregate genetic data for GWAS and meta-analysis by the DIAGRAM consortium for T2D[11, 262-264] and the Psychiatric Genomics Consortium (PGC) for SCZ[30, 265, 266] have successfully identified a substantial number of risk loci for both disorders, which explain roughly 20% of heritability for T2D[267] and 7% for SCZ[30]. Functional analyses showed that risk variants for SCZ are enriched for enhancers mapping to pancreatic beta cells[30]. Furthermore, variants mapping to central nervous system pathways have been associated with BMI – a key risk factor for T2D[41].

Genetic research into the shared pathobiology of SCZ and T2D has been limited to date, and has mainly focused on patients with one of the two disorders[248]. If SCZ without T2D comorbidity and SCZ with T2D are partly underpinned by different genetic aetiologies, such study designs will fail to identify risk factors predisposing to the latter.

3.1.4 Chapter overview

Here, I investigate the presence of shared genetic risk factors for T2D and SCZ using genotype data from a Greek cohort comprising three patient groups: SCZ only (n=924), T2D only (n=822), and comorbid SCZ and T2D (n=505). Samples from two separate Greek cohorts were used as population-based controls (n=1,125). I used genome-wide summary statistics from two large-scale GWAS of SCZ and T2D from the PGC and DIAGRAM consortia, respectively, to perform genetic overlap analyses. First, I assess the genetic overlap between the two disorders using polygenic risk scores; next, I conduct genome-wide comparisons between all three patient groups, as well as population controls; finally, I use summary statistics from published GWAS to search for genetic risk factors shared between SCZ and T2D.

3.1.5 Publication note and contributions

All analyses outlined in this chapter are my own work, with the following exceptions: Bram Prins conducted the individual-level QC in GOMAP up to the step of sample relatedness, as well as the case-case GWAS in GOMAP. The work outlined in this chapter has been peer reviewed and published in *Translational Psychiatry*[268].

3.2 Methods

3.2.1 Sample description

The GOMAP (Genetic Overlap between Metabolic and Psychiatric disorders) study comprises a collection of 2,880 samples from four different patient categories: T2D patients, SCZ patients, individuals with both SCZ and T2D (referred to from here on as SCZplusT2D), and individuals with a different psychiatric diagnosis (this last group was not used in analyses reported here). SCZ patients with and without T2D were recruited at the Dromokaitio Psychiatric Hospital and Dafni Psychiatric Hospital in Athens. SCZ diagnosis was determined by structured clinical interview of the Diagnostic and Statistical Manual of Mental Disorders 4th edition (DSM-IV)[269]. T2D participants were recruited from the diabetes outpatient clinics at Hippokrateio General Hospital and Laiko General Hospital. T2D status was assessed in all participants based on criteria outlined by the American Diabetes Association[270]. All participants gave written informed consent. A detailed description of sample collection has been previously published[271].

For the risk score and summary statistics-based analyses I used summary data from the DIAGRAMv3 meta-analysis of T2D[263] (http://diagram-consortium.org/downloads.html), and the PGC meta-analysis for SCZ[30] (https://www.med.unc.edu/pgc/results-and-downloads). The DIAGRAMv3 study included 12,171 T2D cases and 56,862 controls of mostly European descent. Each contributing study had performed imputation based on the HapMap3 reference panels, resulting in up to 2.5 million variants in the meta-analysis. The SCZ study consisted of 46 European and 3 Asian case-control datasets, amounting to a total of 34,241 cases and 45,604 controls, as well as three parent-offspring trio collections (1,235 trios). Genotypes in each dataset were imputed to the 1000 Genomes reference panel and approximately 9 million variants were used for the combined meta-analysis.

Sample Group	Pre-QC	Post-QC
SCZ	977	924
T2D	885	822
SCZplusT2D	542	505
Other	342	331
Total	**2,747**	**2,582**

Table 3.1. *Sample numbers in the three phenotype groups in GOMAP before and after QC*

3.2.2 Quality control

A total of 2,747 GOMAP samples and 538,448 markers were successfully genotyped on the Illumina HumanCoreExome 12v1.0 BeadChip (Illumina, San Diego, CA, USA) at the Wellcome Trust Sanger Institute, Hinxton, UK (Table 3.1). Quality control (QC) of genotype data was performed following a standard protocol[272] using the PLINK[199] software package. Individuals were removed if they had a call rate below 90%, discordant values for genotyped and reported sex or had heterozygosity rates deviating more than three standard deviations from the mean. For duplicates and related sample pairs (pi_hat>0.2) I excluded one and retained the other at random.

In order to identify potential ethnic outliers, I performed multidimensional scaling (MDS) in PLINK[199] on GOMAP together with different reference datasets. Prior to this, I filtered the variants in each dataset for MAF > 0.001 and excluded variants in complex regions with extended LD or long-range translocations, as these might bias MDS analysis (Table 3.2). I then pruned each dataset using the --indep flag in PLINK[199] with a window size of 50kb shifted by 5kb at the end of each iteration, and a variance inflation factor cut-off of 1.25. The variance inflation factor is equal to $1/(1-R^2)$, where R^2 denotes the coefficient of multiple correlation when one SNP is regressed on all other SNPs in the current window. I merged the pruned datasets and computed pi_hat estimates which served as the input for MDS.

I initially used the 1000 Genomes populations as a reference dataset (Figure 3.1). When using the position of the rightmost 1000 Genomes sample along the component 1 axis as a cut-off for inclusion (Figure 3.1.a), this would have resulted in the exclusion of 150 GOMAP samples. This sample spread is also observable when performing MDS on GOMAP alone, with two distinct clusters forming outside the main cluster (Figure 3.2).

Chromosome	Start position	End position
1	48287980	52287979
2	86088342	101041482
2	134666268	138166268
2	183174494	190174494
3	47524996	50024996
3	83417310	86917310
3	88917310	96017310
5	44464243	50464243
5	97972100	100472101
5	128972101	131972101
5	135472101	138472101
6	25392021	33392022
6	56892041	63942041
6	139958307	142458307
7	55225791	66555850
8	7962590	11962591
8	42880843	49837447
8	111930824	114930824
10	36959994	43679994
11	46043424	57243424
11	87860352	90860352
12	33108733	41713733
12	111037280	113537280
20	32536339	35066586

Table 3.2. *Chromosomal regions excluded from MDS analysis (aligned to GRCh38)*

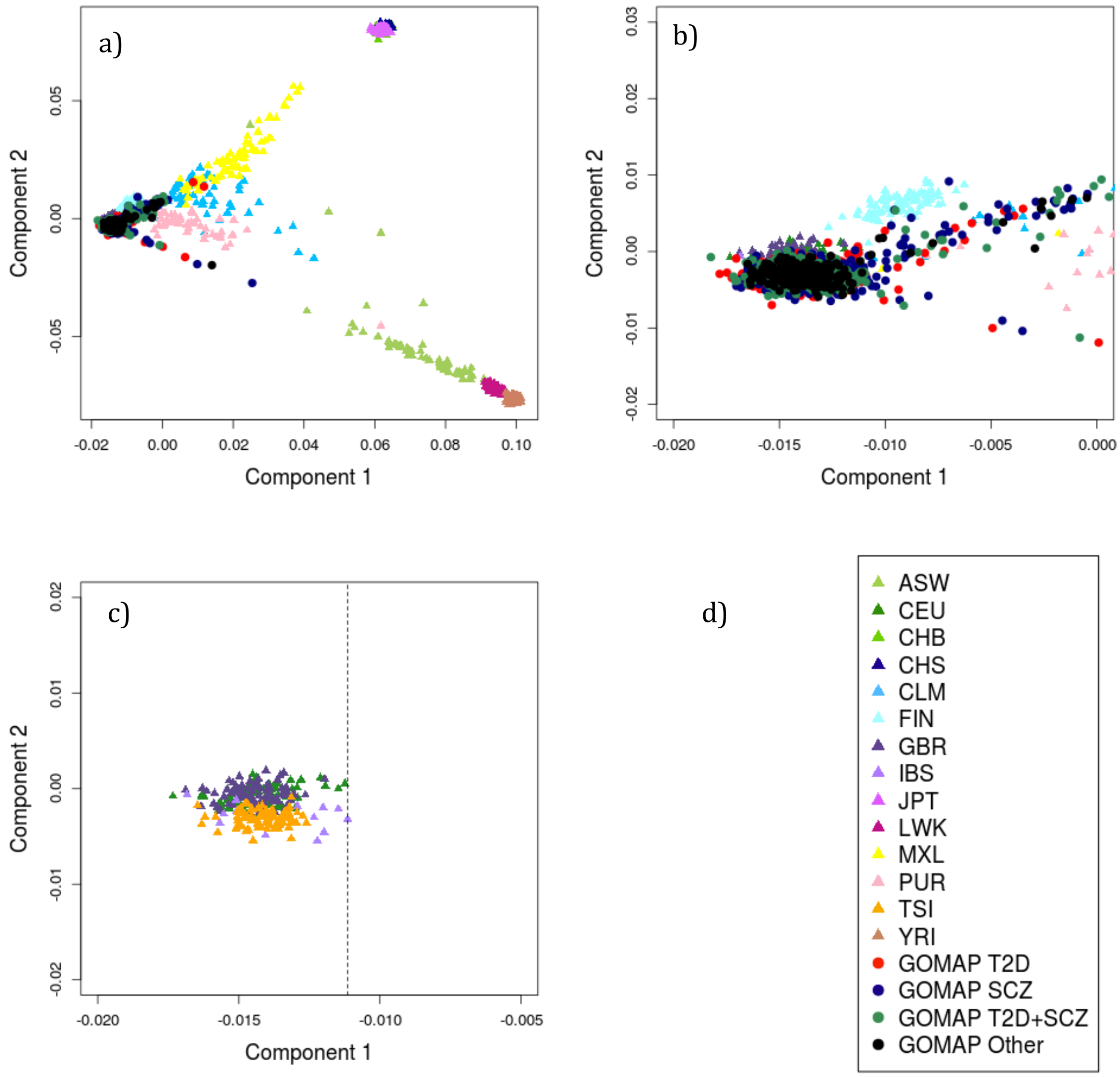

Figure 3.1. *MDS of GOMAP and 1000 Genomes. MDS analysis of GOMAP combined with 1000 Genomes. GOMAP samples are represented by circles, coloured in by diagnostic category. a) The main three clusters correspond to European (middle left), Asian (top right) and African (bottom right) populations. b) Zoom in of the European cluster. c) Zoom of the European cluster with only Italian (TSI), Spanish (IBS), British (GBR) and Central European (CEU) populations shown. Vertical dashed line marks potential inclusion threshold based on the rightmost individual. d) Legend of 1KG populations and GOMAP diagnostic categories*

Based on their parental country of origin all individuals in these two clusters are Greek. While self-reported ancestry information can be inaccurate (e.g. due to uncertain biological parentage), it is unlikely that the observed spread is solely due to such confounding factors. A more plausible explanation is that it simply reflects genetic variation within Greece: although all participants were recruited in Athens, they come from a diverse range of regions, which would have experienced different degrees of isolation and admixture over time. The majority of samples falling into the smallest cluster self-identified as Pontic Greeks; the area of Pontus lies in modern-day Turkey and has been the scene of many migratory events, especially in the 20th century[273].

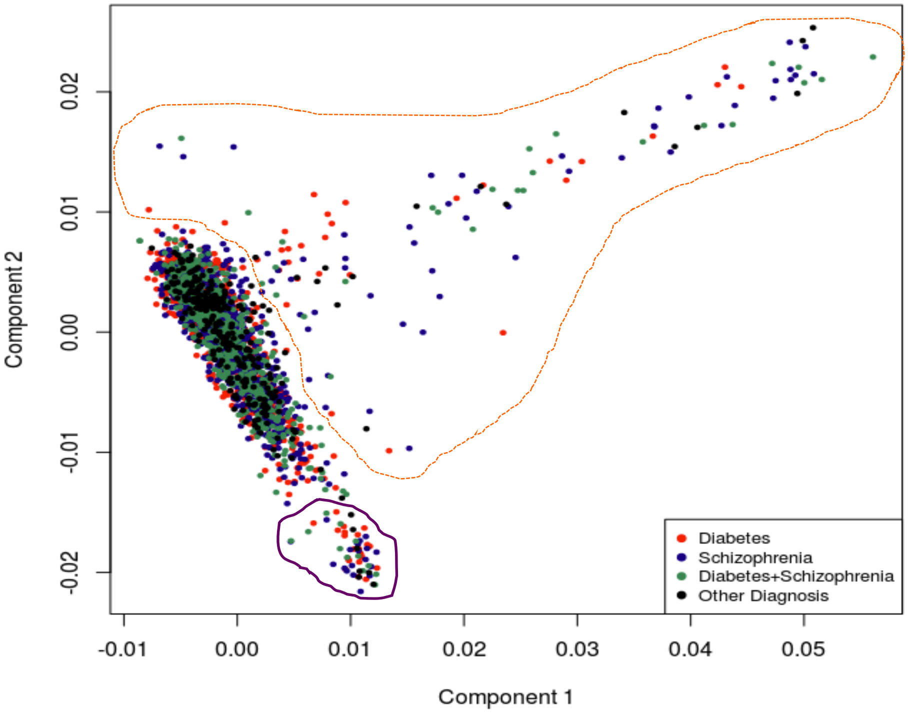

Figure 3.2. *MDS analysis of GOMAP only. Samples are coloured in according to diagnostic category. Two distinct batches of samples form clusters proximal to the major cluster (yellow dashed line: n = 136; purple dashed line: n = 60).*

I next chose three Greek sample cohorts as a reference: TEENAGE[274], a collection of adolescents from the general Greek population, HELIC-Pomak[272] and HELIC-MANOLIS[275], two Greek isolated population collections. The rationale was that since we were not planning to use GOMAP in conjunction with other non-Greek samples (e.g. as controls), these datasets would be more suited to detect subtle sub-structure within GOMAP. GOMAP, TEENAGE and

HELIC-MANOLIS formed a tight cluster, with HELIC-Pomak forming three distinct clusters. The Pomaks are a Muslim minority, inhabiting villages in a mountainous region in Northern Greece. They are both geographically and religiously isolated from the general Greek population. Compared to TEENAGE and MANOLIS, they show longer runs of homozygosity as well as a higher inbreeding coefficient, both of which indicate a larger degree of isolation[272]. I removed seven individuals from GOMAP as outliers based on the first and second MDS components (Figure 3.3).

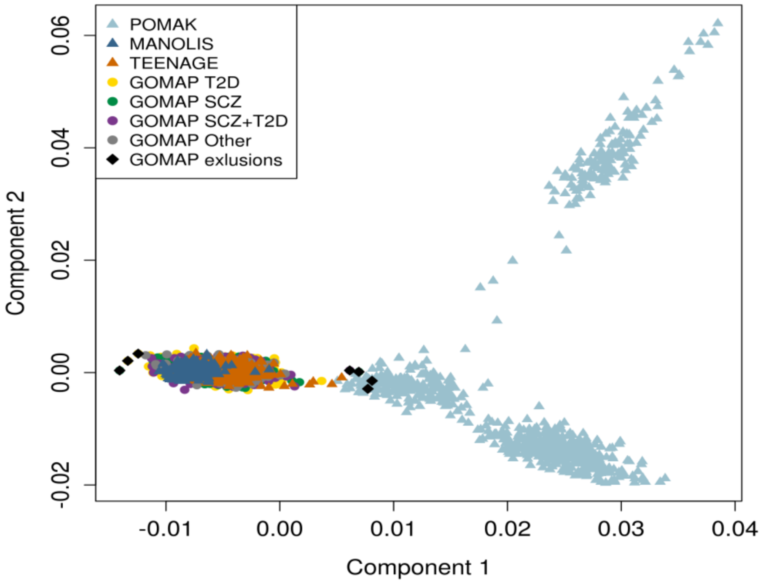

Figure 3.3. *MDS in GOMAP, TEENAGE and HELIC. Components 1 and 2 from multidimensional scaling (MDS) analysis of GOMAP (pre-QC), HELIC-POMAK, HELIC-MANOLIS and TEENAGE. Each data point represents one individual. Black diamond shapes depict individuals excluded as ethnic outliers.*

A total of 2,611 samples passed QC (Table 3.3). After removal of individuals failing QC, variants were filtered for call rates lower than 98%, a Hardy-Weinberg Equilibrium deviation p-value < 1x10^{-4} and cluster separation scores below 0.4. In addition, I removed X-chromosomal markers not within the pseudo-autosomal region with heterozygous haploid genotypes in males. A total of 524,271 autosomal and X-chromosomal markers passed QC (Table 3.4).

Total samples before QC	**2,747**
QC step	**Exclusions**
Call rate < 90%	5
Sex mismatch	42
Heterozygosity outliers (± 3 SD), MAF ≥1%	32
Heterozygosity outliers (± 3SD), MAF <1%	16
Related and duplicated samples	61
Sample ID mismatch	17
Ethnic outliers	7
Total unique exclusions	**138**
Total samples left	**2,582**

Table 3.3. *Number of individuals excluded during QC in GOMAP. "Total samples" refers to SCZ, T2D and SCZplusT2D sample groups.*

Total variants before QC	**538,403**
QC step	**Exclusions**
Non-autosomal, non-chrX nonPAR	2,512
Call rate < 98%	8,949
HWE deviation p < 1x10-4	829
cluster separation score < 0.4	1,126
chrX-nonPAR and heterozygous haploid	716
Total exclusions	14,132
Total variants left	**524,271**

Table 3.4. *Number of variants excluded during QC in GOMAP*

Since GOMAP is a cases-only sample collection, I selected two independent Greek sample collections, TEENAGE[276] and ARGO, as control datasets. ARGO comprises osteoarthritis cases and healthy controls from Larissa, Greece. Samples from all three collections formed a single cluster in MDS analysis (Figure 3.4).

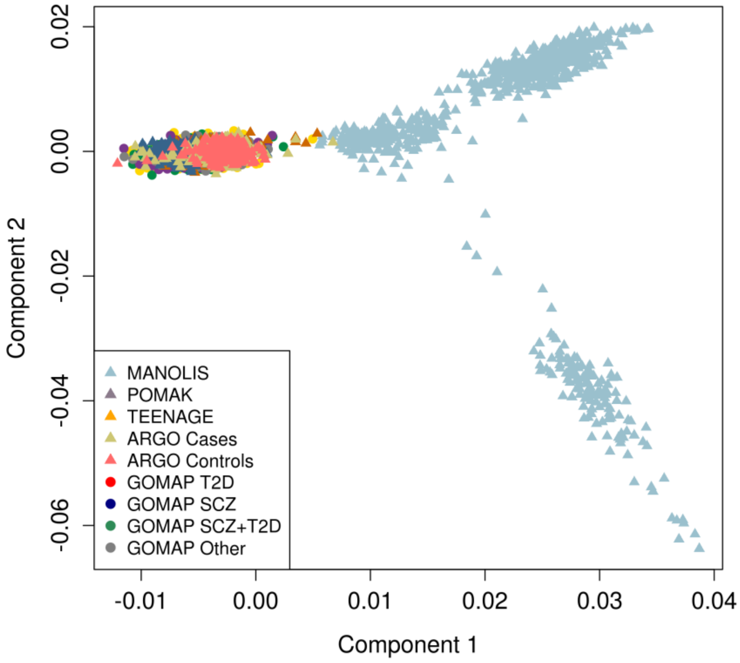

Figure 3.4. *Components 1 and 2 from MDS analysis of GOMAP (post-QC), HELIC-POMAK, HELIC-MANOLIS, ARGO and TEENAGE. Each data point represents one individual.*

3.2.3 Imputation

Following QC I merged GOMAP with 413 samples from TEENAGE[274] and 712 from ARGO, an in-house Greek sample collection. I performed pre-phasing of the merged dataset in SHAPEIT[277] and imputed the phased haplotypes with IMPUTE2[191] using a combined reference panel consisting of UK10K[278], 1000 Genomes[201] and HELIC-MANOLIS[275]. I filtered imputed genotypes for Hardy-Weinberg equilibrium deviation (p-value < 1x10^{-4}), IMPUTE2 info scores < 0.4, and a minor allele frequency (MAF) < 1%. A total of 14,528,340 markers passed imputation QC.

3.2.4 GWAS

Power to detect genetic associations in GOMAP was estimated using the software package QUANTO[279], using the following parameters: MAF=0.45; disease prevalence=0.01 (equivalent to the prevalence of SCZ); sample size=950. I carried out a GWAS for each case-

case and case-control combination in GOMAP using the ‘method --expected’ option in SNPTEST version 2.5[133], which performs an additive association test. I adjusted for the first ten MDS components from the MDS analysis including GOMAP, TEENAGE, ARGO, HELIC-MANOLIS and HELIC-Pomak.

3.2.5 Polygenic risk scores

I used summary statistics from DIAGRAM and PGC (the “base” datasets) to construct T2D and SCZ polygenic risk scores, respectively, in GOMAP (the “target” dataset). The risk score analyses are divided into two stages: first, I computed scores using only established risk variants for each disease (see section 3.2.5.2); next, I relaxed the inclusion criteria incrementally by using all variants falling below a given p-value threshold in the respective base dataset (see section 3.2.5.3).

Before conducting risk score analyses I harmonized the data between DIAGRAM/PGC and GOMAP. I converted chromosome positions in DIAGRAMv3 from NCBI build 36 to the Genetic Reference Consortium human build 37 (GRCh37), in order to match GOMAP. I then matched variants between GOMAP and DIAGRAMv3 and PGC-SCZ, respectively, based on chromosome position.

3.2.5.1 *Risk score construction*

I used PRSice version 1.25[90] to calculate the risk scores in GOMAP and test for an association between scores and phenotype. For each variant the number of risk alleles in the target data (GOMAP) is multiplied by the log(OR) from the base data (DIAGRAM or PGC). The total score for an individual is the average score across all SNPs in the set. Following the approach described by Purcell et al.[89], two logistic regression models are used to obtain the variance in phenotype explained (Nagelkerke’s pseudo R^2):

Full model:

$$\text{Phenotype} \sim \text{Score} + C1 + C2 + C3 + C4 + C5 + C6 + C7 + C8 + C9 + C10$$

Null model:

$$\text{Phenotype} \sim C1 + C2 + C3 + C4 + C5 + C6 + C7 + C8 + C9 + C10$$

In the full model, phenotypes are regressed on risk scores adjusting for the first ten MDS components (C1-C10); in the null model, phenotypes are regressed on MDS components only. The final pseudo R^2 estimate is obtained by:

$$R^2_{final} = R^2_{full} - R^2_{null}$$

A p-value for association of score with phenotype was obtained from the full model. Risk score analysis was carried out in each pairwise comparison between the three disease groups and controls in GOMAP.

3.2.5.2 *Established variant risk scores*

For SCZ, I obtained odds ratios (ORs) of 125 autosomal risk variants from the psychiatric genomics consortium (PGC)[30] (Appendix B). I excluded three X-chromosomal markers of the original 128 independent variants identified by Ripke et al.[30], since calculating scores for non-autosomal alleles is not straightforward.

I used 73 variants identified in a trans-ethnic meta-analysis[11] for the T2D risk score. In order to match the ancestry of the base data as closely to GOMAP as possible, I looked up summary statistics of all independent variants (76 in total) identified in the trans-ethnic study[11] in the DIAGRAMv3 stage 1 meta-analysis[263](Appendix C). Three of the 76 variants were not present in the DIAGRAMv3 data and therefore excluded.

To assess whether the sample size difference between the single-disease and comorbid groups in GOMAP affected the power to detect associations between phenotype and risk scores, I randomly down-sampled the SCZ-only and T2D-only group to 500 individuals each and performed risk score analyses with this reduced set. I repeated this process 5,000 times and computed average pseudo R^2 and p-values to compare to the full analysis.

3.2.5.3 *Genome-wide risk scores*

In addition to calculating risk scores based on established genome-wide significant risk variants, I also performed PRS at ten cumulative p-value thresholds, including all independent variants that fall below a given threshold. I used PRSice[90] for this, a pipeline automating data preparation in PLINK[199] and risck score regression in R. First, p-value informed LD clumping was performed on the intersection of SNPs between the base summary statistics (DIAGRAMv3[263] and PGC-SCZ[30]) and target data (GOMAP), using an r^2 threshold of 0.1 and a window size of 250kb. Next, alleles were matched between the base and target data and ambiguous variants (A/T and G/C variants, which preclude the distinction between flipped alleles and different strand alignments between datasets) removed to produce a final list of clumped variants used for the risk scores. Score calculations and regression analyses are conducted following the same procedure as outlined for the established risk variants. I performed risk score analyses at ten cumulative p-value thresholds, meaning that all variants below a given threshold in the base data were included in the score: $p<5x10^{-8}$, p<0.001, p<0.005, p<0.05, p<0.1, p<0.2, p<0.3, p<0.4, p<0.5, p<1.

3.2.6 Summary statistics-based overlap analyses

I obtained genome-wide summary data for T2D from the DIAGRAMv3 meta-analysis[263], and for SCZ from the PGC meta-analysis[30]. To assess the genetic overlap between the two datasets I performed four complementary analyses – LD score regression[79], extent of shared signals analysis[200], Bayesian colocalization analysis[78] and gene and pathway analysis[202] – which have been described in chapter 2 (sections 2.2.4, 2.2.5, 2.2.6, and 2.2.7)[168].

3.3 Results

3.3.1 GWAS

I performed six case-case and case-control genome-wide association studies in GOMAP and population controls. There was no indication of inflation of test statistics, with lambda values ranging from 0.99 to 1.04 (Figure 3.5; Figure 3.6). Power to detect genome-wide significant associations of small to moderate effects was low given the limited sample size (Figure 3.7).

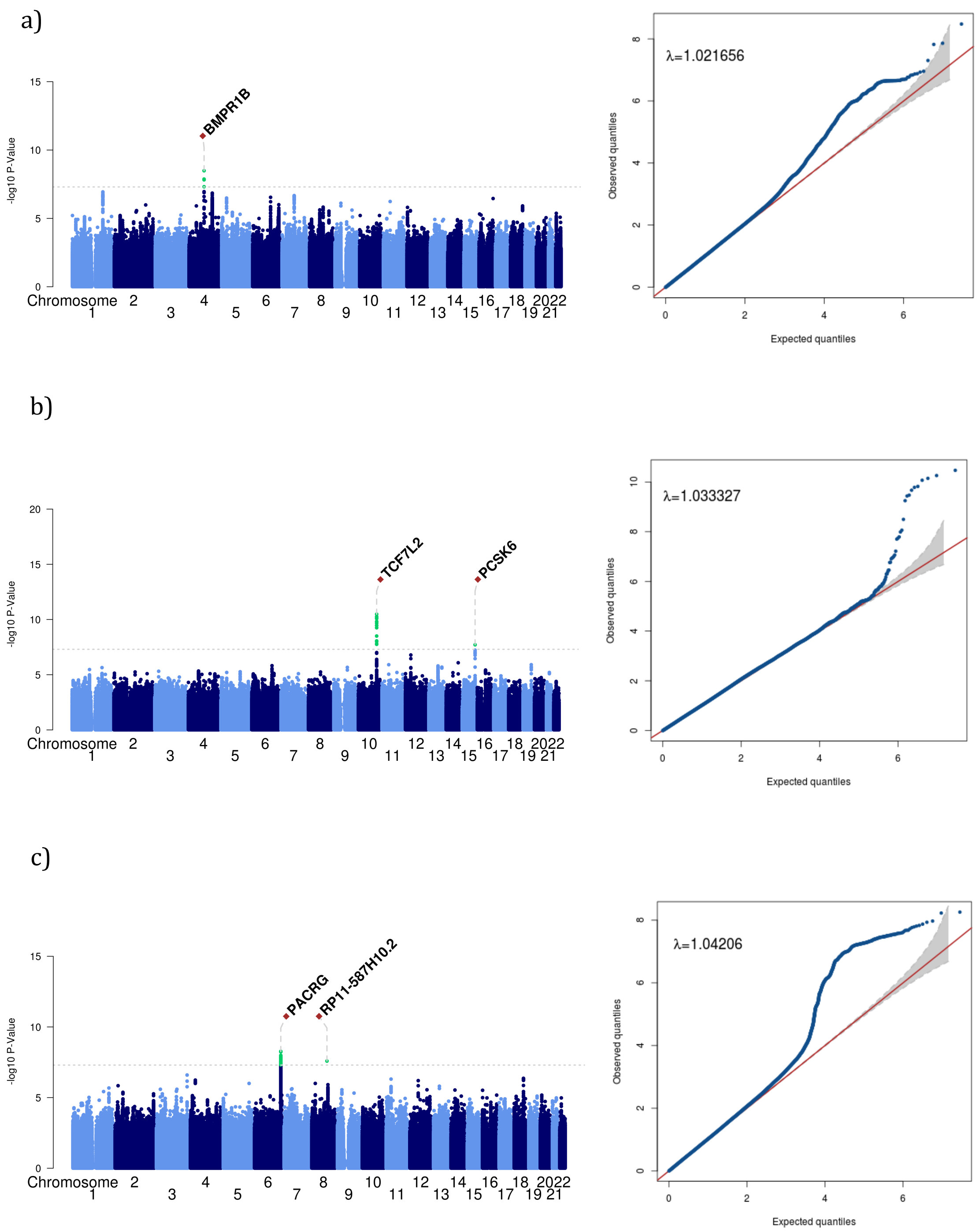

Figure 3.5. *Case-control GWAS in GOMAP. In the Manhattan plots (left), the -log10 of each variant p-value is plotted against its chromosomal location. In the QQ plots (right), the observed -log10 p-value is plotted against its expected value. a) SCZ vs controls, b) T2D vs controls, c) SCZplusT2D vs controls*

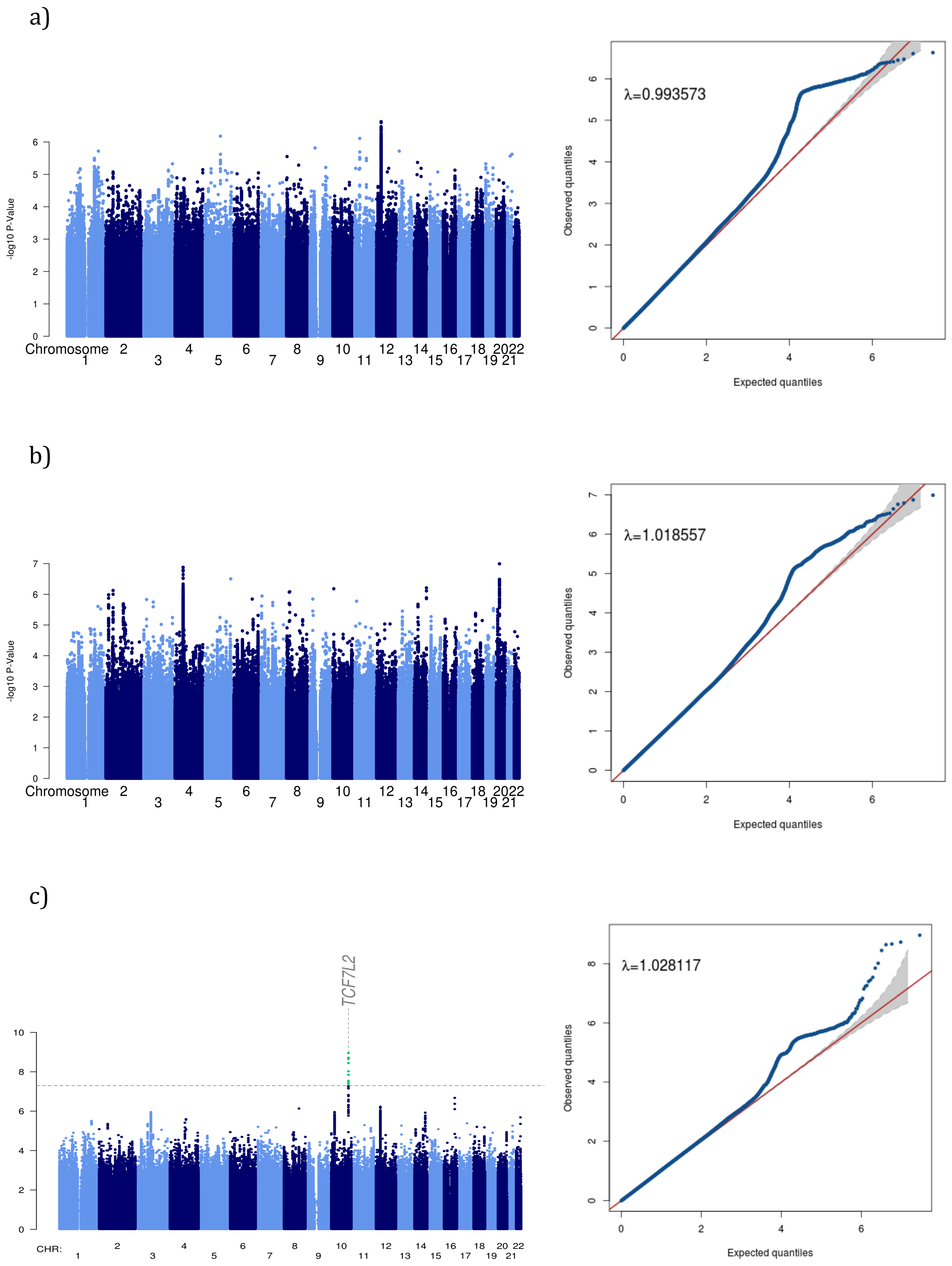

Figure 3.6. *Case-case GWAS in GOMAP. In the Manhattan plots (left), the -log10 of each variant p-value is plotted against its chromosomal location. In the QQ plots (right), the observed -log10 p-value is plotted against its expected value. a) SCZ vs SCZplusT2D, b) T2D vs SCZplusT2D, c) SCZ vs T2D*

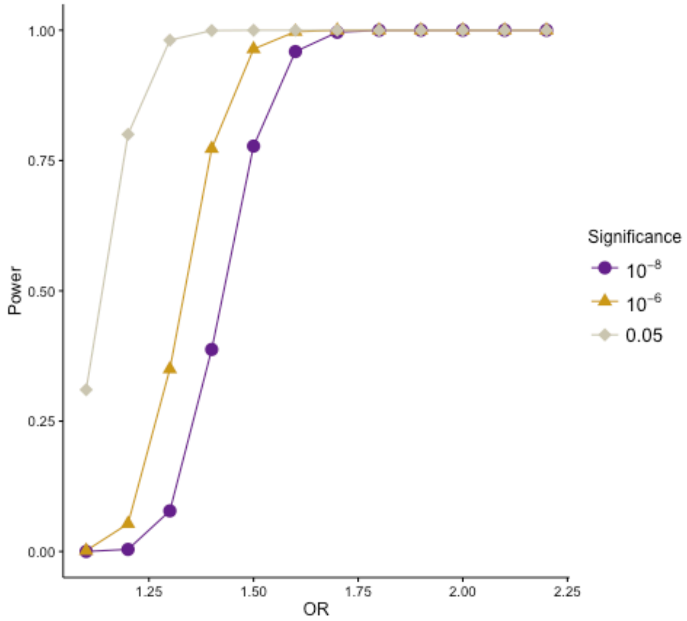

Figure 3.7. *Power calculations for GOMAP assuming a disease prevalence of 1% (equivalent to population risk of SCZ), a minor allele frequency of 45% and a sample size of 950. Estimated power is plotted against effect size (odds ratio) for three different significance thresholds.*

I identified two genome-wide significant signals in the SCZplusT2D vs controls analysis (Figure 3.5; Table 3.5). The most strongly associated variant resides within an intron of the *PACRG* gene (chr6:163319442_G/A, effect allele (EA) G, effect allele frequency (EAF) 0.91, OR 3.81 [95% CI 3.32-4.29], p-value=5.46x10^{-9}). The second signal is located in an intron of *RP11-587H10.2* on chromosome 8 (rs1449245, EA A, EAF 0.79, OR 1.96 [95% CI 1.77-2.20], p-value=2.58x10^{-8}).

Three further signals reached genome-wide significance in other analyses (Table 3.5): an intronic SNP in *TCF7L2* (rs7903146, EA T, EAF 0.38), a well-established T2D risk gene[263], in the T2D vs controls (OR 1.66 [95% CI 1.50-1.80], p-value=3.31x10^{-11}) and T2D vs SCZ analyses (OR 1.53 [95% CI 1.39-1.67], p-value=1.09x10^{-9}); an intronic SNP in *BMPR1B* (rs17616243, EA T EAF 0.16, OR 2.03 [95% CI 1.79-2.27], p-value=3.26x10^{-9}) in the SCZ vs controls GWAS; and an intronic SNP in *PCSK6* in the T2D vs controls GWAS (rs6598475, EA T, EAF 0.36, OR 1.56 [95% CI 1.40-1.72], p-value=1.95x10^{-8}).

SNP	GWAS	EA	NEA	EAF	OR (95% CI)	Info	P
chr6:163319442	SCZplusT2D vs Controls	G	A	0.91	3.81 (3.32-4.29)	0.56	5.46E-09
rs1449245	SCZplusT2D vs Controls	A	G	0.79	1.96 (1.71-2.2)	0.85	2.58E-08
rs7903146	T2D vs Controls	T	C	0.38	1.66 (1.5-1.81)	1.00	3.31E-11
rs7903146	T2D vs SCZ	T	C	0.38	1.53 (1.39-1.67)	1.00	1.09E-09
rs17616243	SCZ vs Controls	T	C	0.16	2.03 (1.79-2.27)	0.72	3.26E-09
rs6598475	T2D vs Controls	T	G	0.36	1.56 (1.4-1.72)	0.93	1.95E-08

Table 3.5. *Top SNPs of genome-wide significant signals in the GOMAP GWAS analyses. EA=effect allele; NEA=non-effect allele; EAF=effect allele frequency; OR=odd ratio; CI=confidence interval*

3.3.2 Genetic risk scores

I performed genetic risk score analyses of SCZ and T2D for each pairwise case-case and case-control combination in GOMAP (Figure 3.8). In the case-control analyses, risk scores for SCZ and T2D were significantly associated with these respective disorders (SCZ R^2=1.7%, p-value=5.25x10^{-9}; T2D R^2=6.8%, p-value=6.12x10^{-27}), serving as a positive control for the validity of the included variants and patient groups. Conversely, risk scores for one disorder were not associated with the other in the case-control comparisons. In the comorbid sample both SCZ and T2D risk scores were significantly associated with phenotype (SCZ risk score p-value=7.17x10^{-5}; T2D risk score p-value=4.14x10^{-4}), with R^2 values lower than those in the single-disease groups (SCZ risk score R^2=1%; T2D R^2=0.8%).

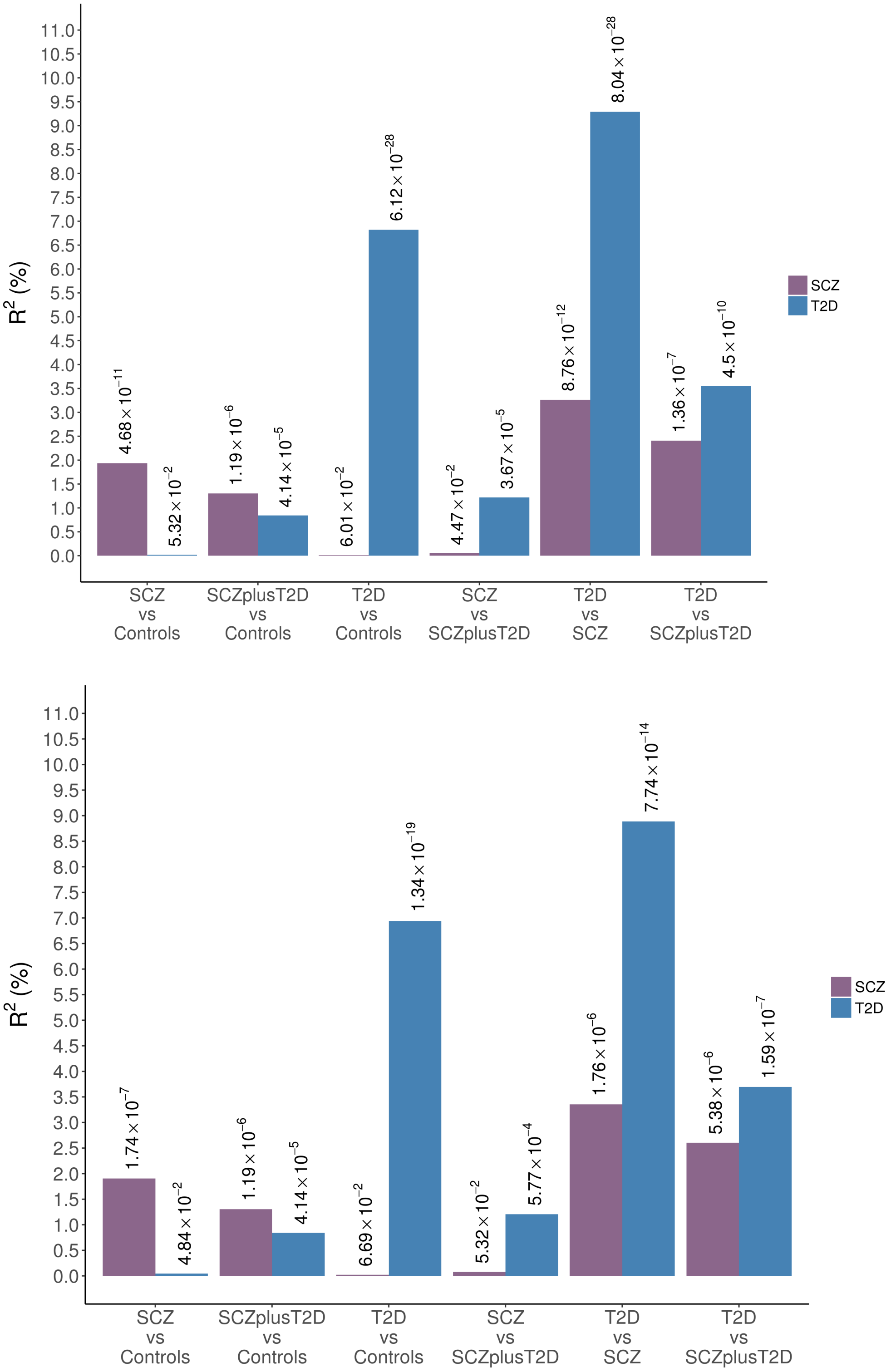

Figure 3.8. *Genetic risk scores of established risk variants for SCZ and T2D in GOMAP. For each analysis Nagelkerke's pseudo R^2 values are plotted and p-values for association between score and phenotype are denoted above each bar. Risk scores are shown for the full GOMAP data (top) and for GOMAP with the SCZ and T2D groups each down-sampled to 500 cases (bottom).*

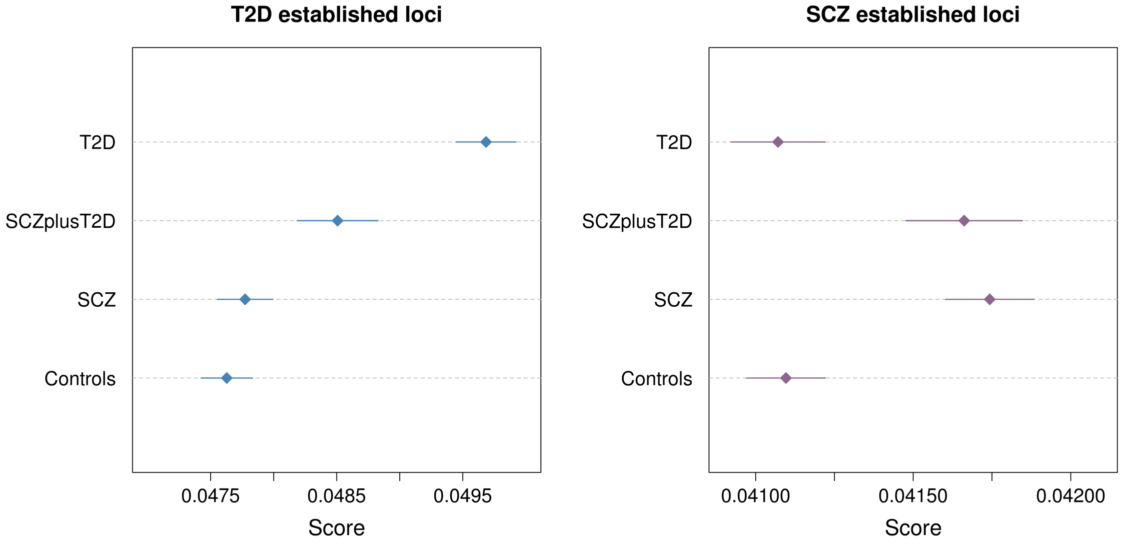

Figure 3.9. *Mean risk scores and 95% confidence intervals for established SCZ and T2D loci in each sample group in GOMAP. Risk scores are constructed based on the effect sizes of 73 and 125 variants from DIAGRAMv3 and PGC-SCZ, respectively. Scores are the weighted sum of risk alleles present in an individual divided by the number of variants included in the score.*

In the comparison between T2D and SCZ cases, risk scores for T2D explained 9.3% of variance (p-value=8.04x10^{-28}) and risk scores for SCZ explained 3.4% of variance (p-value=8.06x10^{-12}). These R^{2} values may be higher than in the case-control analyses due to the fact that controls are population based and not ascertained for either SCZ or T2D status; it is therefore plausible that a subset of controls carries risk alleles for these disorders. In the comparison of individuals with SCZ to those with SCZ and T2D, SCZ risk scores and their R^{2} values were not significantly associated with disease. This is expected, as both sample groups are likely to be enriched for SCZ risk alleles. Interestingly, the R^{2} estimate of the T2D variant risk scores in the T2D vs SCZplusT2D analysis was intermediate in magnitude to that measured in the SCZ vs SCZplusT2D and the SCZ vs T2D analyses. This can be recapitulated by examining the average T2D scores across the different sample groups (Figure 3.9): the average score of the SCZplusT2D sample is higher than for the SCZ-only sample but lower than for the T2D-only sample, indicating that the comorbid group is enriched for T2D risk alleles compared to the SCZ-only group.

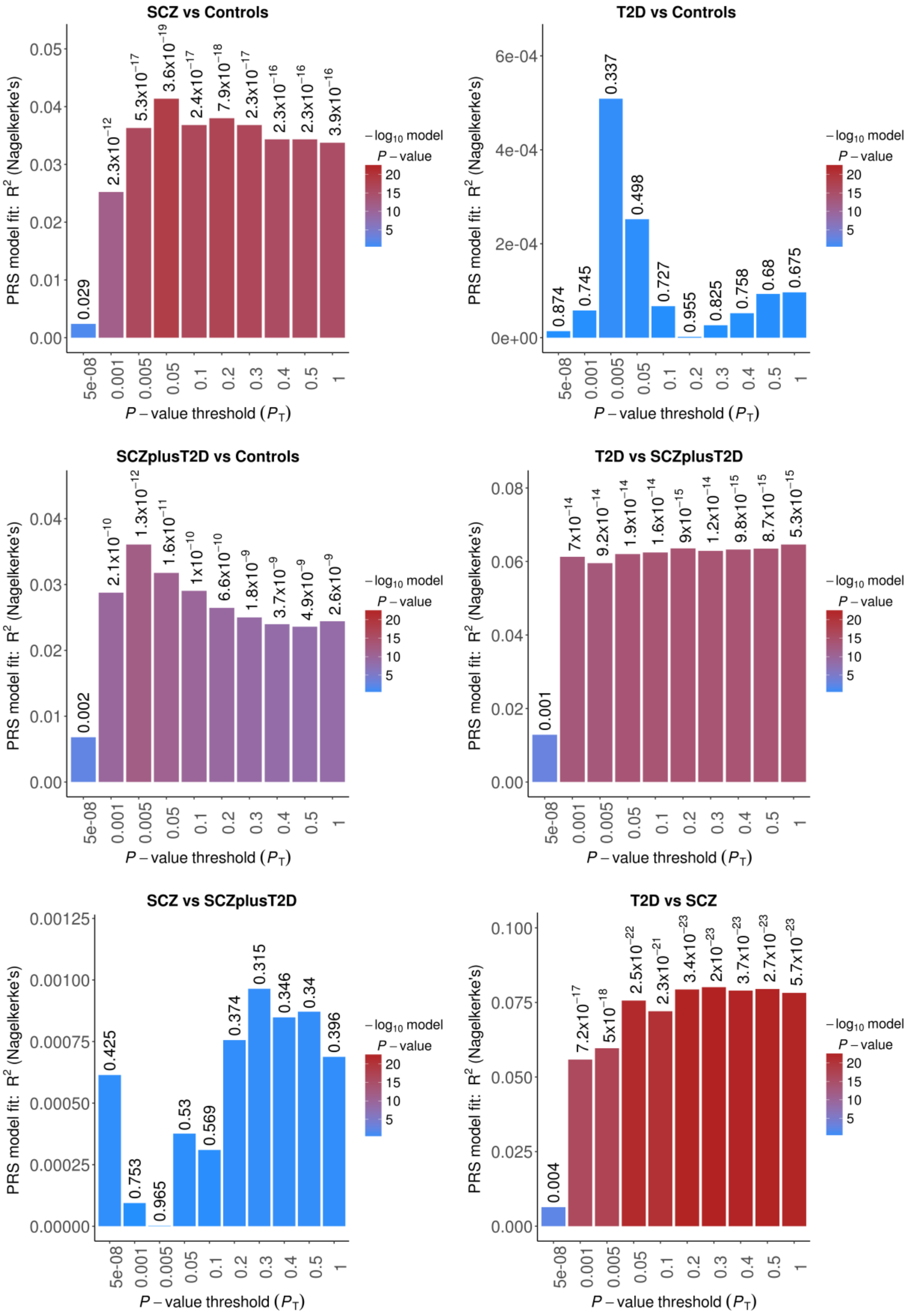

Figure 3.10. *Polygenic risk score analyses for SCZ in GOMAP. Nagelkerke's pseudo-R2 estimates are plotted at ten cumulative p-value thresholds. Note that the y-axis scales differ between plots.*

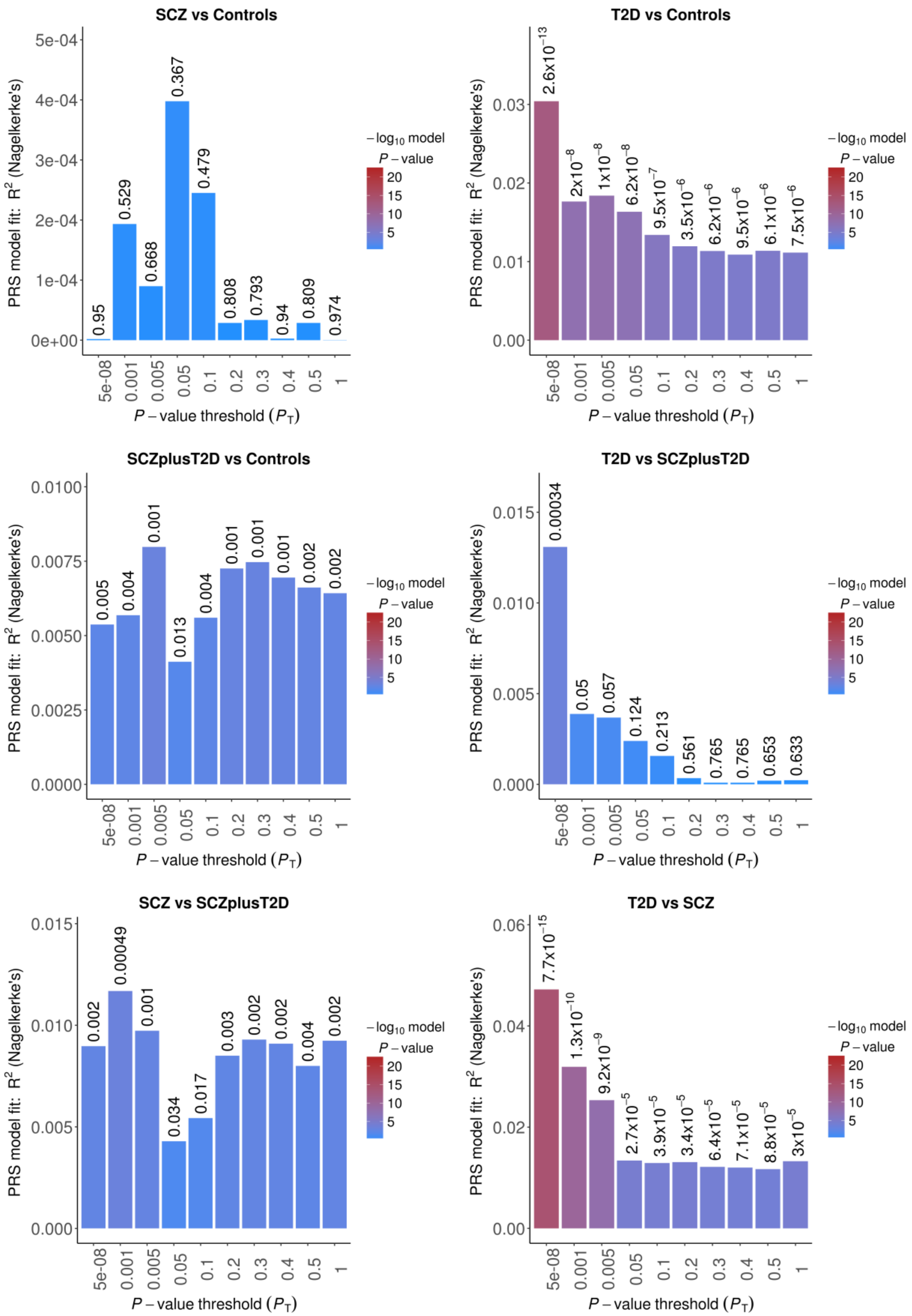

Figure 3.11. *Polygenic risk score analyses for T2D in GOMAP. Nagelkerke's pseudo-R2 estimates are plotted at ten cumulative p-value thresholds. Note that the y-axis scales differ between plots*

To determine whether the observed strength of association of the risk scores was influenced by the difference in sample size among the single-disease and comorbid groups, I repeated the risk score analyses with equally-sized (n=500), randomly down-sampled T2D- and SCZ-only cases. Risk scores significantly associated with phenotype using the full dataset remained significant even with the decreased sample size ($p<0.05$) (Figure 3.8).

It has been shown that the inclusion of variants not reaching genome-wide significance can enhance the power of genetic risk scores[89]. I constructed polygenic scores at ten cumulative p-value thresholds using the same base datasets (DIAGRAMv3 and PGC-SCZ) as for the established variant scores. For the SCZ scores, the most stringent threshold ($p<5x10^{-8}$) resulted in lower levels of association and pseudo-R^2 estimates than the established variant score, most likely due to the fact that some of the variants included in the latter had $p>5x10^{-8}$ in the PGC-SCZ discovery data, which was used here, and will have therefore been excluded. At more permissive p-value thresholds the strength of association increased by several orders of magnitude compared to the established variant scores for all but the SCZ vs SCZplusT2D and T2D vs Controls analyses (Figure 3.10). While pseudo-R^2 also increased at the first increments of variant inclusion, they plateaued or even decreased slightly for thresholds with $p>0.005$. While more relaxed thresholds will include more variants with true effects, they will inevitably also add more null variants contributing to noise. Unlike the SCZ score, T2D scores demonstrated decreasing levels of association as more variants were included in the risk score (Figure 3.11).

3.3.3 Summary statistics-based overlap analyses

I investigated the genetic overlap between summary data from the DIAGRAMv3 meta-analysis for T2D[263] and the PGC meta-analysis for SCZ[30] using both genome-wide and regional approaches.

3.3.3.1 *LD score regression*

There was no significant correlation between these datasets on a genome-wide scale (r^2=-0.01, SE=0.04, p-value=0.82), as previously reported elsewhere[79].

3.3.3.2 *Colocalisation analysis*

I employed a Bayesian colocalisation analysis to search for genomic regions that potentially exert pleiotropic effects. For each region, the method returns posterior probabilities for the five tested hypotheses, as well as the maximum absolute Z-scores found in each of the two input datasets; in some cases, there is more than one variant with the same Z-score (i.e. effect estimate) in a region.

There were no regions with a high posterior probability (>0.9) of containing one causal variant common to both diseases. Five regions had a high posterior probability of harbouring two distinct causal variants (Table 3.6). The first of these regions is located on chromosome 2 and includes nominally significant SCZ variant (top variant in PGC: rs10189857, p=5.14x10^{-7})[30] in an intron of *BCL11A*, and a T2D risk locus upstream of the same gene (top variant in DIAGRAM: rs243021, p=3x10^{-15})[280].

The second region falls within the major histocompatibility complex on chromosome 6, which is known to harbour several SCZ and T2D loci[30, 263]. This region contained three variants with the same effect size for T2D, one of which lies in an intron of *SLC44A* (rs9267658, OR 0.89, 95% CI 0.85-0.94, p=2.2x10^{-5}). The strongest SCZ signal occurred at rs3117574 (OR 0.85, 95% CI 0.82-0.89, p=6.71x10^{-19}), a variant in the 5' untranslated region of *MSH5*, a protein involved in meiotic recombination and DNA mismatch repair. Both *SLC44A* and *MSH5* have been previously associated with SCZ in the Japanese population[281].

The third region resides on chromosome 7, harbouring both a known T2D locus downstream of *KLF14* (top variant in DIAGRAM: rs10954284, p=1.20x10^{-8}) and a known SCZ variant at rs7801375 (PGC p=2.26x10^{-8})[30].

The fourth region, identified on chromosome 8, does not contain any known T2D or SCZ associated variants. The strongest signals in that region occur at rs11993663 for SCZ (PGC p=1.46x10^{-7}) and rs17150816 for T2D (DIAGRAM p=1.60x10^{-5}).

					DIAGRAMv3				PGC-SCZ			
Variants	chr	start	stop	PP4	max |Z|	Top variant	Gene	Dist.	max |Z|	Top variant	Gene	Dist.
1419	2	60293221	62425639	0.92	5.68	rs243019	BCL11A	92496	5.00	rs10189857	BCL11A	0
						rs9267576	C6orf48	4497				
160	6	31704294	32634467	0.96	4.63	rs9267658	SLC44A4	0	8.89	rs3117574	MSH5, MSH5-SAPCD1	0
						rs3130285	TNXB	0				
2158	7	130424544	132805104	1.00	6.51	rs10954284	KLF14	44870	5.61	rs7801375	PLXNA4	240828
1301	8	9641034	10462806	0.92	5.31	rs17150816	MSRA	121041	5.26	rs11993663	MSRA	0
						rs8026735	VPS13C	26113				
2213	15	61266132	63214206	0.99	5.64				6.09	rs11632947	RP11-507B12.2	0
						rs875513	VPS13C	18729				

Table 3.6. *Genomic regions with a high posterior probability of harbouring two distinct causal variants for T2D and SCZ (PP4>0.9). For each region, the variants with the highest absolute Z score for T2D and SCZ are given. Chromosome positions are aligned to GRCh build 37. Gene=closest protein coding gene; Gene dist.=distance in bp to closest protein coding gene*

Finally, a region identified on chromosome 15 encompasses a known SCZ locus in the *VPS13C* gene (top variant in PGC: rs12903146, p=3.00x10^{-10}), as well as the *C2CD4A-C2CD4B* locus, which has been associated with T2D in East Asian populations and also replicated in Europeans (top variant in DIAGRAM: rs8026735, p=2.50x10^{-7})[282].

3.3.4 Extent of shared signals

I assessed the extent of shared association signals between DIAGRAMv3 and PGC-SCZ at ten different p-value thresholds (P_t) and found significant evidence for overlap (p_{perm}<0.05) at all but one P_t (Table 3.7). Of the 19 variants overlapping at P_t=0.001, five are located in known T2D loci, and four within known SCZ loci. One of the variants identified at this P_t, rs6488868, is a synonymous SNP in *SBNO1*, and in partial LD with both a known T2D (rs1727313, r^2=0.53) and a known SCZ (rs2851447, r^2=0.45) risk variant. The two risk variants lie in the 3'UTR and in an intron of *MPHOSPH9*, respectively, and are also in LD with each other (r^2=0.79). Other variants fall within or around several genes previously linked to SCZ or T2D, such as *CACNA1*, *HLA-B*, *PROX1* and *BCL11A*[30, 263] (Table 3.8).

P_t	Variants	χ^2	P	P_{perm}
0.5	58504	1.4	2.30E-01	2.32E-01
0.1	6247	39.7	3.00E-10	0.00E+00
0.05	2324	40.9	1.60E-10	0.00E+00
0.04	1749	53.5	2.50E-13	0.00E+00
0.03	1180	49	2.50E-12	0.00E+00
0.02	658	32.4	1.30E-08	0.00E+00
0.01	287	41.4	1.30E-10	0.00E+00
0.005	125	37.7	8.10E-10	0.00E+00
0.001	19	14.2	1.70E-04	8.30E-04
5.00E-04	10	13.8	2.00E-04	2.00E-03

Table 3.7. *Overlap analysis between DIAGRAM and PGC summary statistics. For each p-value threshold (P_t) the number of independent SNPs overlapping at this threshold is given, along with the resulting chi-squared statistic (χ^2), p-value (P) and empirical p-value obtained by permutations (P_{perm}).*

							DIAGRAMv3		PGC	
rsID	**chr:pos**	**A1**	**A2**	**Freq$_{CEU}$**	**Gene**	**Gene dist.**	**P**	**OR (95% CI)**	**P**	**OR (95% CI)**
rs340835	1:214163675	G	A	0.45	*PROX1*	0	1.10E-06	0.92 (0.95-0.89)	1.38E-02	1.03 (1.05-1.01)
rs6752494	2:200313323	T	C	0.17	*SATB2*	0	2.10E-04	0.87 (0.94-0.81)	1.64E-04	1.09 (0.96-0.88)
rs7652609	3:1812610	C	A	0.06	*CNTN4*	327887	8.90E-04	1.12 (1.06-1.18)	2.65E-02	0.96 (1.07-1.01)
rs7614727	3:52295895	C	T	0.43	*WDR82*	0	8.50E-04	0.97 (1.03-0.92)	6.71E-19	1.18 (0.89-0.81)
rs830644	3:71665559	T	C	0.57	*FOXP1*	32419	3.60E-04	0.90 (0.95-0.85)	2.32E-06	1.08 (1.11-1.05)
rs8192675	3:170724883	T	C	0.29	*SLC2A2*	0	7.00E-04	0.93 (0.97-0.90)	1.82E-04	0.95 (1.07-1.03)
rs7682321	4:19099432	A	G	0.26	*LCORL*	1075933	9.70E-04	0.92 (0.95-0.89)	4.32E-06	1.05 (0.97-0.93)
rs27419	5:52420938	C	T	0.68	*MOCS2*	15045	1.00E-04	1.00 (1.03-0.96)	1.19E-09	1.07 (1.09-1.05)
rs2523589	6:31327334	G	T	0.53	*HLA-B*	2369	4.20E-05	1.10 (1.07-1.14)	1.00E-01	0.98 (1.04-1.00)
rs7450789	6:111816975	G	T	0.86	*REV3L*	12057	5.00E-04	1.07 (1.03-1.12)	1.67E-05	1.06 (0.97-0.91)
rs7783665	7:110039196	G	A	0.42	*IMMP2L*	263914	4.10E-05	0.93 (0.97-0.90)	2.64E-05	0.95 (0.97-0.93)
rs6999153	8:9193501	G	A	0.56	*RP11-10A14.4*	167855	2.00E-04	1.07 (1.03-1.12)	2.21E-07	1.05 (0.97-0.93)
rs2796441	9:84308948	G	A	0.40	*TLE1*	4728	2.50E-04	0.98 (1.02-0.94)	1.46E-07	0.94 (1.08-1.04)
rs1079268	11:83255725	C	G	0.65	*DLG2*	0	9.60E-05	0.98 (1.01-0.94)	5.14E-07	1.06 (1.08-1.04)
rs6488868	12:123799974	A/G	T	0.70	*SBNO1*	0	5.10E-04	0.93 (0.96-0.89)	9.92E-06	0.95 (0.97-0.93)
rs275746	15:40154938	G	C	0.08	*GPR176*	0	7.60E-04	1.09 (1.05-1.13)	8.26E-04	1.04 (0.98-0.94)
rs1163439	15:80432222	A	G	0.66	*ZFAND6*	1487	7.30E-06	1.07 (1.03-1.11)	1.47E-04	0.96 (1.06-1.02)
rs2838968	21:47038365	C	T	0.18	*PCBP3*	25243	1.00E-03	1.26 (1.13-1.40)	1.24E-01	1.04 (1.01-0.91)
rs5757761	22:40054948	T	C	0.46	*CACNA1I*	0	7.10E-04	1.08 (1.04-1.12)	9.69E-07	1.06 (1.08-1.04)

Table 3.8. *Variants overlapping at P_t=0.001 in DIAGRAM and PGC. Chromosome positions are aligned to GRCh build 37. A1=effect allele; Freq$_{CEU}$=frequency of A1 in the 1000 Genomes CEU population; Gene=closest protein coding gene; Gene dist.=distance in bp to closest protein coding gene*

Gene	Chr	Start	Stop	Strand	$Q_{DIAGRAMv3}$	Q_{PGC}	Known
PROX1	1	214161278	214214853	+	2.24E-02	9.89E-03	T2D
UBE2D3*	4	103715540	103790050	-	2.16E-02	6.52E-04	No
CISD2*	4	103749224	103813964	+	2.18E-02	6.61E-04	No
SLC9B1*	4	103806205	103947552	-	2.06E-02	1.30E-06	No
SLC9B2*	4	103946647	103998480	-	2.04E-02	9.32E-04	SCZ
SSR1*	6	7281283	7313541	-	3.70E-02	4.12E-02	T2D
CDKAL1	6	20534688	21232635	+	4.73E-06	2.63E-02	T2D
HLA-B	6	31321649	31324989	-	4.67E-02	1.69E-10	T2D;SCZ
MICB	6	31462054	31478901	+	3.68E-02	2.63E-06	No
MCCD1	6	31496739	31498008	+	2.94E-02	2.37E-05	No
DDX39B	6	31497996	31510252	-	2.06E-02	6.80E-04	No
ATP6V1G2	6	31512228	31514625	-	2.06E-02	2.61E-03	No
NFKBIL1	6	31514628	31526606	+	2.24E-02	2.57E-03	No
NEU1	6	31826829	31830709	-	2.28E-02	8.92E-11	SCZ
SLC44A4	6	31830969	31846823	-	2.02E-02	4.73E-11	No
EHMT2	6	31847536	31865464	-	2.02E-02	4.16E-10	No
ZBTB12	6	31867394	31869769	-	8.57E-03	2.43E-04	No
PRRT1	6	32116140	32119720	-	2.06E-02	8.02E-03	No
HLA-DRB1*	6	32546546	32557613	-	2.33E-02	1.21E-06	T2D; SCZ
KCNJ11*	11	17406795	17410878	-	3.99E-02	1.56E-02	T2D
EHBP1L1	11	65343509	65360121	+	2.94E-02	2.63E-02	No
KCNK7	11	65360326	65363467	-	4.18E-02	4.32E-03	No
TSPAN8*	12	71518877	71551779	-	2.02E-02	2.49E-02	T2D
ZFAND6*	15	80351910	80430735	+	1.75E-02	7.25E-03	T2D
UNC45A	15	91473410	91497323	+	2.71E-02	3.56E-02	No
RCCD1	15	91498106	91506355	+	2.02E-02	4.98E-02	T2D
UBE2Z	17	46985731	47006422	+	4.88E-02	5.51E-03	No
SNF8	17	47007458	47022484	-	4.84E-02	3.35E-03	No
GIP	17	47035918	47045955	-	2.66E-02	9.32E-04	No

Table 3.9. *Genes significantly associated (q-value<0.05) in DIAGRAM and PGC after FDR correction. Q-values shown are from the 'lenient' gene annotation allowing a 20kb window around the transcription start and stop sites. Genes marked with an asterisk were also significantly associated using strict gene annotations. Start and stop positions are aligned to GRCh build 37. "Known" refers to previously reported associations with SCZ and/or T2D in a given gene.*

3.3.4.1 *Gene and pathway analysis*

I tested for enrichment of association signals in genes and pathways in the DIAGRAM and PGC summary statistics. I did not identify any pathways that were significantly associated (q-value < 0.05) with both SCZ and T2D. In the gene-level analysis, 29 genes had a q-value < 0.05 in both datasets (Table 3.9). Ten of the genes have been previously associated with SCZ and/or T2D. Of note, variants in or in close proximity to *ZFAND6*, *PROX1*, and *HLA-B* were also found to overlap at P_t=0.001. *SLC44A4*, which is strongly associated with SCZ (q-value=4.73x10^{-11}), falls within the region on chromosome 6 identified in the colocalisation analysis.

3.4 Discussion

I investigated the genetic overlap between SCZ and T2D, using summary statistics from large-scale meta-analyses and genome-wide genotype data from a dedicated collection of individuals with SCZ, T2D or both disorders. The work presented here benefits from clinically ascertained diagnoses and robust base datasets used to construct the risk scores.

3.4.1 GWAS

Due to the limited sample size and, consequently, low power to detect genetic associations in GOMAP, I did not expect to identify novel genome-wide significant loci, but rather to harness the presence of the comorbid patient group for risk score analyses. The two genome-wide significant signals identified in the SCZplusT2D vs controls GWAS map to introns of *PACRG* and *RP11-587H10.2*. *PACRG* has been associated with the risk of leprosy[283], while *RP11-587H10.2*, a long non-coding RNA, is of unknown function. Replication of these newly arising signals in independent datasets is required to establish or refute them as true associations.

3.4.2 Polygenic risk scores

The main novel finding of this project arises from the risk score analyses, which demonstrated that the SCZplusT2D sample is enriched for both SCZ and T2D risk alleles compared to controls; this is in line with the increased prevalence of T2D among schizophrenia patients being at least partly due to genetic predisposition[248, 249]. Patients suffering from both diseases had SCZ risk scores comparable to the SCZ-only group but fell between the SCZ-only and T2D-only groups for T2D risk scores. This implies that patients with comorbid SCZ and T2D have almost the same SCZ risk allele profile as SCZ patients without T2D but carry fewer risk-increasing variants for T2D than T2D patients without comorbid SCZ. Two conclusions might be drawn from this: first, at least part of the risk for T2D in SCZ patients is driven by genetic predisposition to T2D, rather than antipsychotic use alone; and second, the comorbid group appears to have a less strong T2D genetic risk profile compared to T2D-only patients. This is in line with environmental factors, including response to antipsychotic treatment and sedentary lifestyle, contributing to T2D risk. Such factors might exacerbate an otherwise moderate genetic predisposition to T2D.

To my knowledge, three other studies have to date compared risk scores for T2D and SCZ[89, 284, 285]. Purcell et al. first performed SCZ risk scores analysis in a T2D sample but did not identify a significant correlation between scores and phenotype[89], potentially due to the relatively low sample sizes available at the time (~3,300 cases for SCZ; ~1,900 cases for T2D). More recently, a study investigating the genetic liability to SCZ in immune-related disorders found a weak association between SCZ risk scores and T2D[285]. The investigators used an earlier release of the PGC-SCZ summary data[265] with lower sample numbers than currently available. One study has previously reported an association between T2D risk scores and self-reported diabetes (any type) in individuals with psychosis, but did not detect an association when repeating the analysis for SCZ risk scores[284].

It should be noted that although both LD score regression and PRS can provide useful insights into the shared genetic architecture of two traits, they suffer from a bias towards common variants resulting from the pre-analysis pruning step. While common variants do make up a large proportion of the heritability of common disorders, low-frequency and rare polymorphisms of larger effects offer important insights into disease biology and affected pathways, as they often occur within coding regions. As briefly touched upon in Chapter 1, a non-significant genetic correlation estimate does not necessarily imply that two traits share no pleiotropic variants. For example, it is possible that they share specific affected genes or pathways rather than having a similar pathobiology overall.

3.4.3 Shared risk loci

The SNP-based overlap analysis highlighted one region where a known T2D and a known SCZ signal map to the same locus in the *MPHOSPH*9 gene[11, 30], which codes for a phosphoprotein highly expressed in the cerebellum. This gene has been previously associated with multiple sclerosis[286]; however, its function is not well understood. I also identify *PROX1* as a potentially pleiotropic locus based on the gene-based analysis and the SNP-based overlap test. *PROX1* has been previously implicated in both T2D and SCZ, and acts either as a transcriptional activator and repressor depending on the cellular context. It has been implicated in murine beta-cell development[287], as well as in neurogenesis in humans[288]. While functional investigation of the genes identified here is necessary, an emerging hypothesis is that pleiotropic loci might influence T2D and SCZ by acting in different biological pathways.

3.4.4 Conclusions and future directions

The work in this chapter lends further support to the theory that the observed comorbidity between SCZ and T2D is in part mediated by genetics. It also highlights several genes and loci with putative pleiotropic effects. The greatest limitation of this study is the lack of power in the GOMAP data, as well as its observational nature, which precludes any analyses on outcomes associated with disease progression, such as response to medication, change in BMI or metabolic measures. Furthermore, potential confounding factors, such as smoking status or diet and exercise, might have biased the results, and these issues are further discussed under 5.2.

Establishing large SCZ sample collections with in-depth phenotype data is challenging considering the nature of the disease; conceivably, people suffering from delusions and/or paranoia may be less likely to agree to share their genetic material along with detailed health data. This is also evident in the SCZ sample numbers in the full UK Biobank dataset: only 728 out of roughly 500,000 individuals had a SCZ diagnosis code, compared to 14,803 with T2D. Based on the prevalence estimates, one would expect almost 10-times as many SCZ patients, whereas the number of T2D patients matches prevalence rates more closely.

There are several ways in which the finding of a genetic component to SCZ and T2D comorbidity could be used for further research and clinical management of SCZ:

First, if the finding that T2D risk variants are enriched in SCZ patients with T2D is replicated by independent studies, this could make a case for stratifying patients according to risk profiles and targeting treatments accordingly. The incorporation of genetic variants into risk scores based on non-genetic factors, such as BMI or family history, has been shown to lead to improved predictive accuracy[289-292]. However, even if SCZ patients at high risk of developing T2D could be confidently identified early on, choosing an antipsychotic with minimal metabolic impact might not be straightforward, since most antipsychotic drugs have some metabolic side effects[293], and patients might not respond to the specific medication chosen or might suffer from non-metabolic side effects (e.g. motor control impairment). Nevertheless, investigating the shared genetic basis of SCZ and T2D is important to assess the validity of current diagnostic boundaries.

Second, elucidating common affected genes and pathways between SCZ and T2D could also aid drug development and repurposing efforts. The discovery of new therapeutic agents for psychiatric conditions has been stagnant for over three decades [294], owing to the lack of clear molecular targets and the difficulty of obtaining relevant tissue samples. As evidence for a link between SCZ and T2D independent of antipsychotic side effects accumulates[259-261, 295-298], identifying specific shared pathways and their involvement in disease mechanisms could reveal targets for intervention. The summary statistics-based overlap analyses in this chapter highlighted several genes with evidence of association in both SCZ and T2D. Their exact function and relevance to both disorders will need to be explored. If their association with SCZ and T2D, respectively, is due to biological pleiotropy and not confounding or statistical artefacts, they might highlight potential aetiopathological mechanisms underlying metabolic abnormalities in drug-naïve SCZ patients[298]. While past repositioning efforts of metabolic drugs for SCZ have shown limited success[299], the identification of shared genes might reveal novel molecular targets.

Third, SCZ is a heterogeneous disorder, and it is conceivable that impaired glycaemic control might present a distinct subtype of the disease. Pleiotropic genes whose expression in peripheral blood can be linked to disease status could potentially be used as biomarkers for early detection/classification of disease.

Future studies with larger sample sizes and detailed phenotype information (ideally including longitudinal medication data) will be necessary to precisely disentangle the shared genetic basis of SCZ and T2D.

Chapter 4 – Multi-trait association analyses of high-depth sequencing data in population isolates

4.1 Introduction

4.1.1 Advantages of population isolates in genetic studies

The majority of genetic association studies to date have been carried out in cosmopolitan populations of mostly European descent. This is partly due to the possibility of collecting large sample sizes leading to increased statistical power and more opportunities for replication. Furthermore, the genetic make-up of samples drawn from the general population can be expected to be representative of that population, and findings from association studies will be more generalisable. Nevertheless, isolated populations afford a number of potential advantages for the study of complex traits:

Isolated populations arise from one or multiple founding events, such as migration and subsequent settlement at a geographically remote location or drastic reduction in population size due to adverse conditions (e.g. natural disaster, famine, epidemic)[300]. Restricted gene flow and endogamy over multiple generations then lead to an increase in genetic homogeneity and random allele frequency fluctuations (genetic drift)[301]. As a result, functional variants may rise in frequency[302] and can thus be more easily identified in association studies. This, together with the lower degree of genetic heterogeneity, can lead to increased power to detect trait-associated variants.[303] For example, a study in 2,575 Greenlandic individuals found a protein-altering variant in *TBC1D4* with large effects on glycaemic traits and T2D risk (OR=10.3)[7]. While the deleterious variant was observed at a frequency of 17% in the study population, it is only found in one individual in all 1000 Genomes samples.

Environmental factors such as diet will be more homogeneous within isolates than in the general populations, minimising the possible confounding of association results. While this is also important for single-trait studies, multi-trait analyses relying on the covariance of several phenotypes are especially susceptible to such biases. Extreme environmental conditions may exert selective pressures leading to a change in allele frequency for variants affecting the pertinent phenotype. For example, indigenous people of the Tibetan Plateau were found to have significantly higher frequency of variants in *EPAS1* compared to Han Chinese and other populations not residing at high altitude[304-306]. *EPAS1* is a transcription factor involved in increased erythrocyte production in response to hypoxic conditions, and the alleles found in Tibetans are associated with increased haemoglobin concentrations[304], suggesting that the observed difference in allele frequency is a result of natural selection.

4.1.2 Leveraging proteomics data for locus discovery

When conducting multi-trait analyses, it can be advantageous to utilise intermediate trait measurements, such as metabolite or inflammatory markers, rather than disease endpoints. If one assumes that morbidity results in part from the perturbation of multiple proteins acting in biological pathways, then studying these proteins should yield a more fine-grained resolution of the phenotypic variance explained by genotypes.

Until recently, biomarker GWAS have focused on a small number of traits with known or at least hypothesised involvement in disease. As many fields of genetic research have moved into the "high-throughput" era, assays for the quantification of hundreds of biomolecules on large sample sizes are now available[307, 308].

The availability of these tools offers unprecedented breadth of molecular information. To maximise the insights that can be gleaned from these data, it will be necessary to look at them not as independent traits, but in the context of biological networks. Most studies of plasma proteins to date have conducted univariate GWAS of each protein separately[32, 309-312]. The few multivariate biomarker studies that have been carried out mostly involved a small number of biomarkers selected based on their shared biological function[131, 134, 313]. In possibly the largest multi-trait biomarkers study to date, Inouye et al. used data of more than 100 metabolite measures (broken up into clusters to facilitate analysis)[314]. Multivariate GWAS of these trait clusters lead to a nearly two-fold increase of detected association signals, several of which were confirmed as expression QTLs for metabolites in their respective clusters. This exemplifies the power advantages of performing multi-trait association studies on proteomics data.

4.1.3 Multivariate analysis in the context of sample relatedness

To date there has been one study using multivariate association analysis in an isolated population[313]. As outlined in Chapter 1, there are numerous methods available for the analysis of multi-trait data. When individual-level data are available, and the traits to be analysed have been measured on the same set of samples, multivariate approaches that explicitly model the inter-trait correlation structure are preferable. A further consideration is the degree of relatedness between study samples. Several methods allow for the inclusion of covariates. Population structure and cryptic relatedness can thus be accounted for by including principal components in the analysis. However, in the case of isolates, the first ten or so PCs usually included as covariates might not capture the extensive degree of relatedness among samples[315]. In standard univariate GWAS, the use of mixed models has gained some popularity: phenotypes are modelled as dependent on both fixed effects (genotypes, covariates) and random effects (a relatedness matrix).

4.1.4 Chapter overview

In this chapter I perform multi-trait GWAS in the HELIC-MANOLIS cohort comprising samples from a Greek population isolate. All samples have high-depth (22x) sequencing data available, as well as 57 quantitative trait measurements. In addition to these traits, I also use expression data from 275 plasma proteins measured on the metabolic, cardiovascular II and cardiovascular III panels of the Olink platform.

I first describe the dataset as well as a phenotype imputation procedure I used to estimate missing phenotype values across all traits in MANOLIS. I then outline how I selected trait groups for analysis, followed by multivariate GWAS and comparison to univariate (single-trait) results. The initial focus of this project was to use multivariate GWAS to identify variants associated with osteocalcin, due to its relevance for both cardiometabolic and musculoskeletal physiology – two major themes of this thesis. I therefore performed two multivariate GWAS on osteocalcin and manually selected traits based on their correlation with and/or biological link to osteocalcin. I then used a clustering approach to identify additional trait groups among the phenotypes available in MANOLIS.

Finally, I discuss the findings from these analyses, current challenges in the field, as well as ongoing and future work.

4.1.5 Contributions

All analyses outlined in this chapter have been carried out by me with the following exception: univariate GWAS of unimputed phenotypes were carried out by Young-Chan Park for Olink traits, and Karoline Kuchenbäcker for non-Olink traits. Transformation of non-Olink traits had been previously performed by Karoline Kuchenbäcker, and transformation of Olink traits by Young-Chan Park (I repeated this transformation after recovering previously excluded values based on the assay limit of detection, as described in section 4.2.4). The effective number of variants in HELIC-MANOLIS 22x sequencing data was calculated by Young-Chan Park.

4.2 Methods

4.2.1 Datasets

The work described in this chapter is based on the Hellenic Isolates Cohort (HELIC) MANOLIS (n=1,457) sample collection from the Mylopotamos villages Anogia, Zoniana, Livadia and Gonies (estimated total population size of 6,000) in Crete, Greece. Samples were whole-genome sequenced to high-depth (average of 22x) on the Illumina HiSeqX platform. Quality control had been performed previously[316].

4.2.2 Phenotypes

I used information from 57 quantitative traits in MANOLIS assessed at the time of sample collection. These can be broadly classified as metabolic, haematological and anthropometric. Phenotype QC and transformation had been previously carried out on all traits[275]. Briefly, trait values lying more than 3, 4 or 5 SDs away from the mean were set to missing; sex was adjusted for if it was significantly associated with phenotype (Wilcoxon rank sum, $p<0.05$); traits were transformed to follow a standard normal distribution (inverse normal or log normal transformation), and residuals from age-, age^2- and (for some traits) BMI-adjusted regressions were used for association analyses.

In addition, I used protein expression data of 275 blood biomarkers measured on 1,325 MANOLIS samples using the Olink platform (Olink Bioscience, Uppsala, Sweden). Protein measurements were performed on three Olink panels (cardiovascular disease (CVD) II, CVDIII and metabolism (META)) using a proximity extension assay[317]. Since only 1,325 individuals in MANOLIS had Olink measurements, I used this subset of the full 1,457 samples for all analyses outlined in this chapter.

4.2.3 Phenotype imputation

Multivariate mixed models rely on complete phenotype data, meaning that a sample will be excluded if it is missing information for at least one of the analysed traits. This can lead to substantial sample loss and, consequently, a drop in power if the missingness patterns of analysed traits do not overlap completely. To circumvent this problem, I used a Bayesian phenotype imputation tool, PHENIX[318], to recapitulate missing trait values. PHENIX jointly models multiple phenotypes in an LMM where the random effects are defined by the relatedness matrix, which can be estimated from genetic data. I performed imputation across all 57 quantitative traits and 274 Olink protein biomarkers available in HELIC-MANOLIS.

I did not exclude any traits based on missingness prior to imputation; my rationale was that while phenotypes with high missingness (e.g. more than 40%) might not be imputed accurately, they should also not decrease the imputation quality of other traits as they will add comparatively little information to the imputation algorithm.

By default, PHENIX masks about 5% of non-missing values for each phenotype and computes the correlation between the true value and the imputed one (referred to from here on as “imputation accuracy”). To allow for comparisons across multiple imputation runs and (potentially) different imputation software packages, I randomly sampled the values to mask for each trait in R and hard-coded them in the PHENIX script. The authors of PHENIX suggest to use the correlation between true and imputed values in a similar fashion to the IMPUTEv2 info score, and propose an exclusion threshold of 0.36. While this threshold might seem lenient, it is worthwhile to keep in data and then apply careful post-association analysis filtering to traits that had low imputation accuracy and/or high missingness. I therefore only excluded imputed values for traits whose imputation accuracy was below 0.4, and did not filter based on missingness.

4.2.4 Inclusion of Olink below-LOD measurements

Olink assays are reported as normalised protein expressions, and need to undergo a series of pre-processing and QC steps before being used for analyses. This includes the exclusion of measurements that fall below the limit of detection (LOD)[319]. The LOD value differs for each assay (i.e. biomarker) and is calculated as:

$$LOD_{Assay} = Average(Expression_{NegativeControl}) + 3 * SD_{Assay}$$

However, analysts have found that this exclusion criterion is conservative and that the inclusion of below-LOD data points increases power by preserving sample size without sacrificing specificity (Anders Mälarstig and Arthur Gilly, personal communication). To explore whether phenotype imputation accuracy could be improved by the inclusion of more samples per trait (in effect, complete data for all Olink traits), I recovered below-LOD values for Olink traits and repeated the phenotype imputation for Olink and non-Olink traits. Since PHENIX expects normally distributed phenotypes, I also repeated the transformation of Olink proteins prior to imputation. I followed the same transformation procedure that had been previously applied: I regressed each Olink protein on age, age^2, sex, sample plate number (to adjust for potential batch effects), and season at sample collection (summer or winter; this is to account for seasonal expression differences of some proteins); I then standardised all proteins using an inverse normal transformation.

4.2.5 Selecting trait groups for analysis

When dealing with high-dimensional datasets comprising hundreds of phenotypes, the selection of appropriate trait groups to analyse together is not straightforward. One obvious solution is to go by prior knowledge, such as biological pathways or epidemiology. However, this approach is restrictive as it inevitably limits the number of trait groups that will be selected, and will miss potentially interesting trait pairings whose connection has not been previously studied. Especially for biomarkers, which often act in multiple pathways and show high degrees of interconnectedness, a hypothesis-free approach to identify trait groups is preferable. Here, I used both approaches:

First, I chose traits known to be biologically linked to or highly correlated with osteocalcin, which was the initial focus of this project (see section 4.1.4). I then used network analysis to define trait groups that satisfied a given threshold of inter-trait correlation.

Cluster index	Included traits	Number of traits
1	Hip, Waist, Weight, Height	4
2	TR, Fe_iron	2
3	LYMPC, GRANPC	2
4	WBC, GRAN, MID	3
5	HGB, HCT	2
6	MPV, PDW, LPCR	3
7	MCV, RDW, MCH	3
8	PLT, PCT	2
9	CD84, CD40L	2
10	LPL, PRELP, HO1, MERTK, XCL1	5
11	TM, TRAILR2, PGF, TNFRSF10A, TNFRSF11A	5
12	PRTN3, MMP9, MPO, PGLYRP1, RNASE3, AZU1, RETN	6
13	JAMA, PECAM1, CASP3	3
14	CPB1, CPA1	2
15	CCDC80, MEPE, CHRDL2	3
16	PDGFSUBUNITA, PAI	2
17	FKBP4, THOP1, QDPR, BAG6, ENO2, KYAT1	6
18	BGP, MEPE, COL1A1, OPN, ROR1	5
19	BGP, leptin, adiponectin, RI, RG, BMI	6

Table 4.1. *Trait clusters taken forward to analysis. All clusters except cluster 18 and 19 were identified through network analysis in the igraph R package.*

For the osteocalcin analyses, I initially chose adiponectin, leptin, random glucose and random insulin, as these traits are known to be either directly or indirectly regulated by osteocalcin levels[320-323]. I chose random glucose and insulin measurements instead of fasting ones, as only a small subset of people in MANOLIS had fasting measurements available. All of those four traits are weakly correlated with osteocalcin.

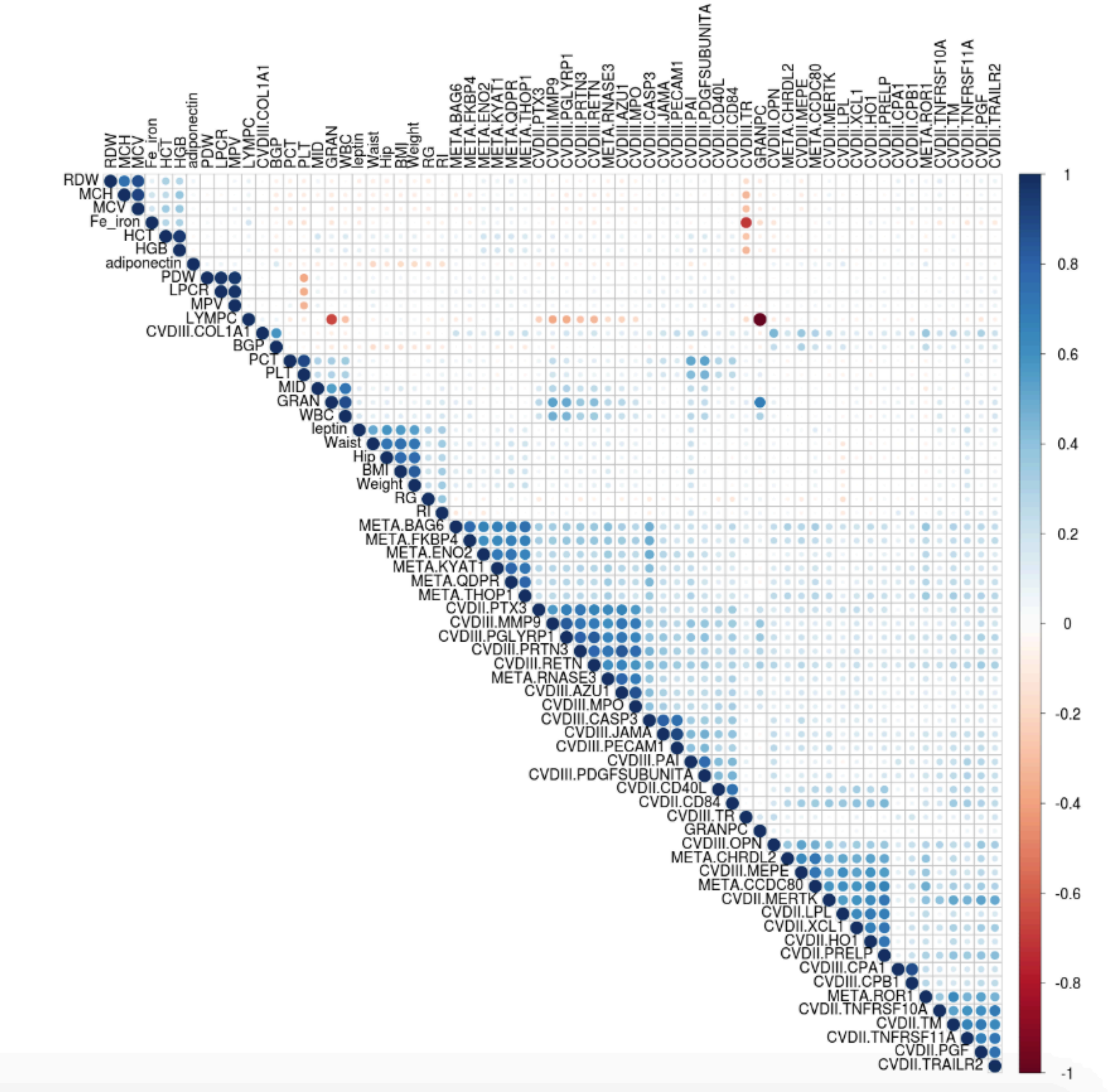

Figure 4.1. *Phenotype correlations between the 66 individual traits analysed in at least one trait cluster.*

I subsequently also searched for suitable traits across Olink and standard traits that were significantly ($p<3x10^{-4}$; equivalent to $0.05/M_{eff} = 0.05/165$) correlated with osteocalcin at an absolute correlation value of at least 0.2. All ten traits that passed this filter were Olink proteins. I sorted these traits by their correlation p-value and took forward the 4 most significant ones (COL1A1, OPN, MEPE and ROR1).

The power of multivariate association models is in part dependent on the correlation structure of the included traits, as well as on the effect sizes of a variant on those traits. Generally, trait correlations of less than -0.3 or more than 0.7 result in the greatest gain in power compared to conducting multiple univariate analyses and pooling the results in a meta-analytic approach[71]. For the hypothesis-free selection of trait groups, I computed a correlation matrix in R of all Olink and standard traits in MANOLIS, using directly measured as well as imputed values for all traits that passed imputation QC, and directly measured values only for those that did not. I then used the igraph R package (v 1.0.0) to build networks of trait clusters. As input igraph requires a data frame containing a list of edges (i.e. trait pairs) pre-filtered for a certain correlation threshold. To choose an appropriate threshold, I looked at the number and size of trait clusters at incremental absolute correlation values, ranging from 0 to 1. In terms of both computational burden and interpretability of results, it is preferable to have a higher number of trait clusters each consisting of a relatively small number (n<10) of traits. I therefore chose an absolute correlation cut-off of 0.7, which resulted in 27 trait clusters ranging in size from 2 to 44 traits. I filtered this list manually and excluded clusters where traits were derivations of one another (e.g. systolic blood pressure and systolic blood pressure adjusted for BMI) or traits were measurements of the same value (e.g. random and fasting glucose, where the latter is a subset of the former); I also excluded the cluster of 44 traits, as analysis and interpretation would have been intractable. After applying these exclusion criteria, there were 19 trait clusters left, each containing between 2 and 6 traits (Table 4.1). In total, 66 traits across Olink and the standard traits were included in at least one cluster (Appendix D; Appendix E).

4.2.6 Multi- and univariate GWAS

I conducted 19 multivariate GWAS using the GEMMA software v0.94[75]. GEMMA implements both univariate and multivariate mixed models, which can account for relatedness between samples through the inclusion of a random effects term (the relatedness matrix). While there are other implementations of multivariate mixed models[101, 128], I chose GEMMA due to its computational efficiency (implemented in C++), as well as to facilitate comparisons with previous analyses carried out in HELIC.

In addition to the multi-trait GWAS, I also performed univariate association studies on each individual trait included in any of the multivariate trait groups to obtain marginal p-values and effect sizes. This is helpful in determining whether a multivariate signal is driven by a subset of traits. I used imputed QCed phenotype data for all analyses. The relatedness matrix specified had been previously computed using GEMMA v0.94 on all 1,457 individuals in MANOLIS and genotypes filtered for MAF>5% and HWE p-value>1x10^{-5}. The p-values reported here are based on the score test (option "-lmm 3").

I filtered all association results for MAF>0.1%, sample missingness<1% and HWE p-value>1x10^{-5}. I used PeakPlotter with a p-value threshold of 2.66x10^{-10} and default settings otherwise to obtain a list of independent signals (https://github.com/wtsi-team144/peakplotter). Briefly, PeakPlotter sorts all variants satisfying the set significance threshold by their p-values, and then iterates over them, retaining only the top signal within each 500kb window.

4.1.1.1 *GEMMA version issues*

I initially used the latest stable release (v0.97) of GEMMA to run both multi- and univariate GWAS. Since the univariate GWAS had previously been conducted on the unimputed phenotypes by my colleagues, I cross-checked my results with theirs. I noticed that the results differed considerably from the unimputed GWAS (Figure 4.2). This could be expected for traits where missingness is high and imputation accuracy low, leading to spurious associations. However, I observed these discrepancies even for traits with low missingness and good imputation accuracy. A comparison of imputed and unimputed GWAS runs using both mine and a colleague's pipeline revealed that the variation in results was due to different GEMMA versions used for the imputed and unimputed GWAS. The latter had used a pre-release version (v0.94, available from http://www.xzlab.org/software.html), while I had used the preview version of the latest stable release (v0.97, October 2017, https://github.com/genetics-statistics/GEMMA/releases/tag/v0.97). Since Plink[199] gave results in line with those of GEMMAv0.94 (Table 4.2), I repeated all analyses with this GEMMA version.

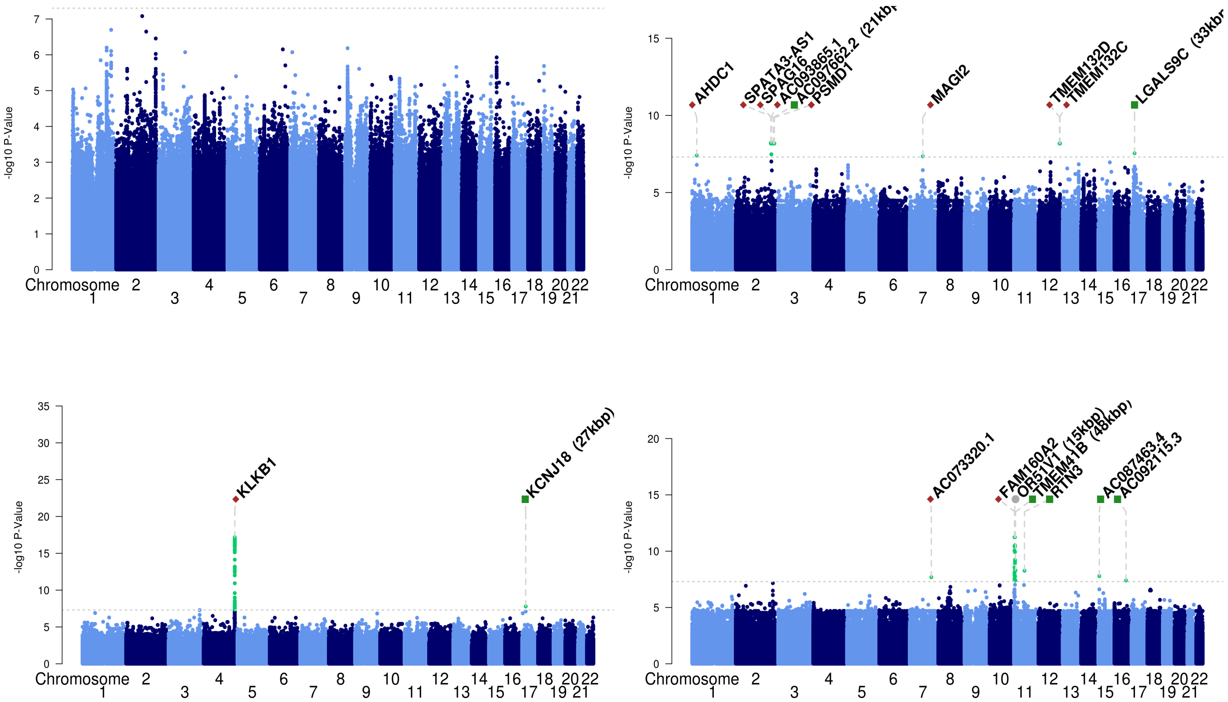

Figure 4.2. *Discrepancies between GEMMA versions 0.97 (top) and 0.94 (bottom). Results shown are from univariate GWAS of CCDC80 (left) and HCT (right). All results were filtered for MAF>0.001.*

Software	beta	SE	P
GEMMA v0.97-preview	0.073	0.040	6.95E-02
GEMMA v0.97-release	0.073	0.040	6.95E-02
GEMMA v0.94	-0.351	0.039	1.25E-17
Plink v1.9	-0.359	0.038	1.00E-20

Table 4.2. *Association summary statistics for variant rs1973612 (chr4:186248013, T/C, MAF=0.48) from a univariate GWAS of CCDC80 levels using unimputed trait values.*

4.2.7 Significance threshold and effective number of traits

One of the (many) unanswered questions in multi-trait analyses is how to correct for multiple testing and choose an adequate significance threshold. In frequentists statistics, significance is often set at a level that controls the family-wise error rate (i.e. the probability of falsely rejecting the null hypothesis at least once in a group of tests when it is in fact true for all tests) at 0.05. Bonferroni correction (0.05 divided by the number of tests) is a simple and straightforward way to adjust the significance threshold when dealing with independent tests. This, however, is not the case in GWAS: alleles of nearby variants will co-occur more often than those of variants on different chromosomes. Likewise, different phenotypes correlate with each other at varying degrees. Simply dividing by the number of variants and traits tested will therefore lead to an overly conservative significance estimate. Instead, one can adjust for the effective number of tests:

$$P = \frac{0.05}{N_{eff} * M_{eff}}$$

where N_{eff} the effective number of variants and M_{eff} is the effective number of traits analysed. N_{eff} can be computed based on the LD structure across the genome by only retaining independent variants that are not correlated with each other. Similarly, M_{eff} can be determined by considering the inter-trait correlation structure. One approach is to simply take all individual traits tested in at least one analysis and compute M_{eff} based on these. In practice, this approach is conservative in a multi-variate setting as traits analysed together in one GWAS do not contribute to the multiple testing burden. However, as I also report the marginal, single trait GWAS, it is an appropriate adjustment in this context.

Here, I calculated the effective number of traits across all OLINK and standard traits in MANOLIS using three different methods. The first determines $\mathrm{M_{eff}}$ from the eigenvalues of the phenotype correlation matrix and resulted in an estimate of 33.5 effective traits[324]. The second is based on the same approach, but estimates $\mathrm{M_{eff}}$ based only on the integral part of the eigenvalues[325]. This approach estimated $\mathrm{M_{eff}}$=37. The third method conducts a PCA and declares $\mathrm{M_{eff}}$ as the number of PCs cumulatively explaining a 90% or 95% of the total variance (39 PCs in this case; Figure 4.3).

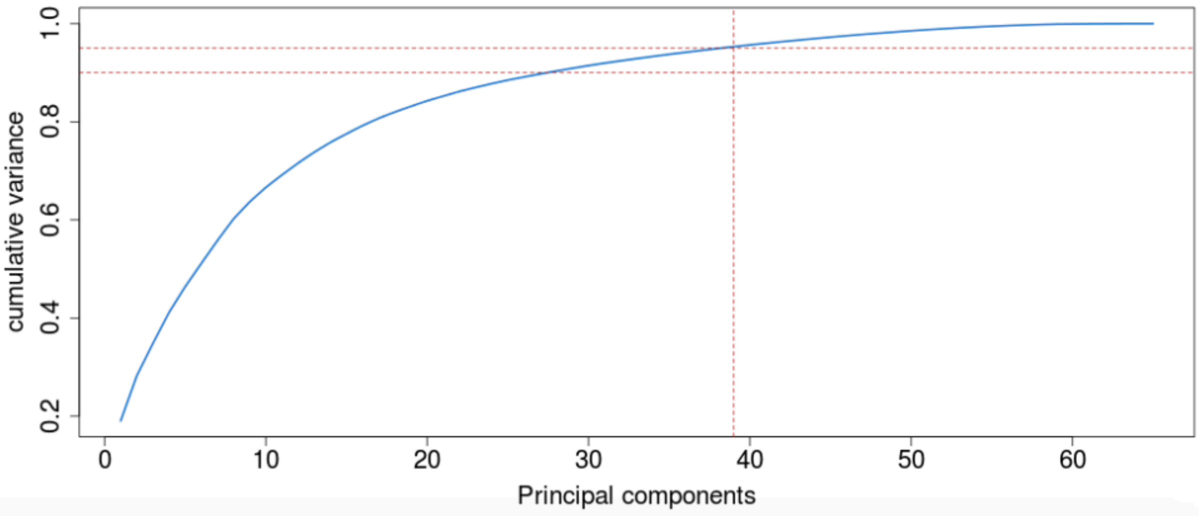

Figure 4.3. *Cumulative variance explained by principal components derived from 66 phenotypes. Horizontal dashed lines mark 90 and 95% of cumulative variance, respectively. The vertical dashed line marks the number of principal components that cumulatively explain 95% of variance.*

For the traits included in the multivariate GWAS analyses presented in this chapter, I set M_{eff}=37. The effective number of variants in the 15x MANOLIS data was previously calculated by filtering for MAC>10 and LD-pruning in Plink with the "--indep" option and a 50kb window-size, a 5 variant increment and variance inflation factor of 2. The variance inflation factor is equal to 1/(1-R^2), where R^2 denotes the coefficient of multiple correlation when one SNP is regressed on all other SNPs in the current window. This resulted in a N_{eff} estimate of 5,078,182, and the significance for all GWAS outlined in this chapter is therefore: 0.05/(5,078,182*40)=2.66x10^{-10}. From here on, I will refer to this threshold as the "study-wide" significance threshold, and to p<5x10^{-8} as the "genome-wide" one.

4.3 Results

4.3.1 Phenotype imputation

I imputed missing phenotype values for all OLINK and standard traits with PHENIX. The average imputation accuracy across all traits was 0.74. Imputation accuracy was only weakly correlated with missingness (Figure 4.4; Pearson's correlation=-0.13, p=0.017), and several traits with high missingness (>20% of samples missing) had imputation accuracy exceeding 0.9. Conversely, I found strong associations between imputation accuracy and the number of additional traits a given phenotype is highly correlated with (absolute r^2>0.7) (Figure 4.4). This underpins the advantage of imputation algorithms that model the inter-trait covariance, rather than considering each trait separately (such as the BSLMM algorithm implemented in GEMMA[75]).

I also repeated the imputation after including Olink trait values that fell below the LOD cut-off and had therefore been excluded. Imputation accuracy was on average lower for the run including below-LOD values (Figure 4.4). This is in line with the finding that accuracy depends more strongly on the inter-trait correlation structure than on the degree of missingness (Figure 4.4). Even if, as others have found (Anders Mälarstig, personal communication), the LOD cut-off is too stringent and leads to the exclusion of some high-quality trait measurements, one would still expect a large number of the below-LOD trait values to be inaccurate. Feeding these into the imputation algorithm will then cause a drop in overall accuracy as the low-quality values will affect all traits correlated with the one under consideration.

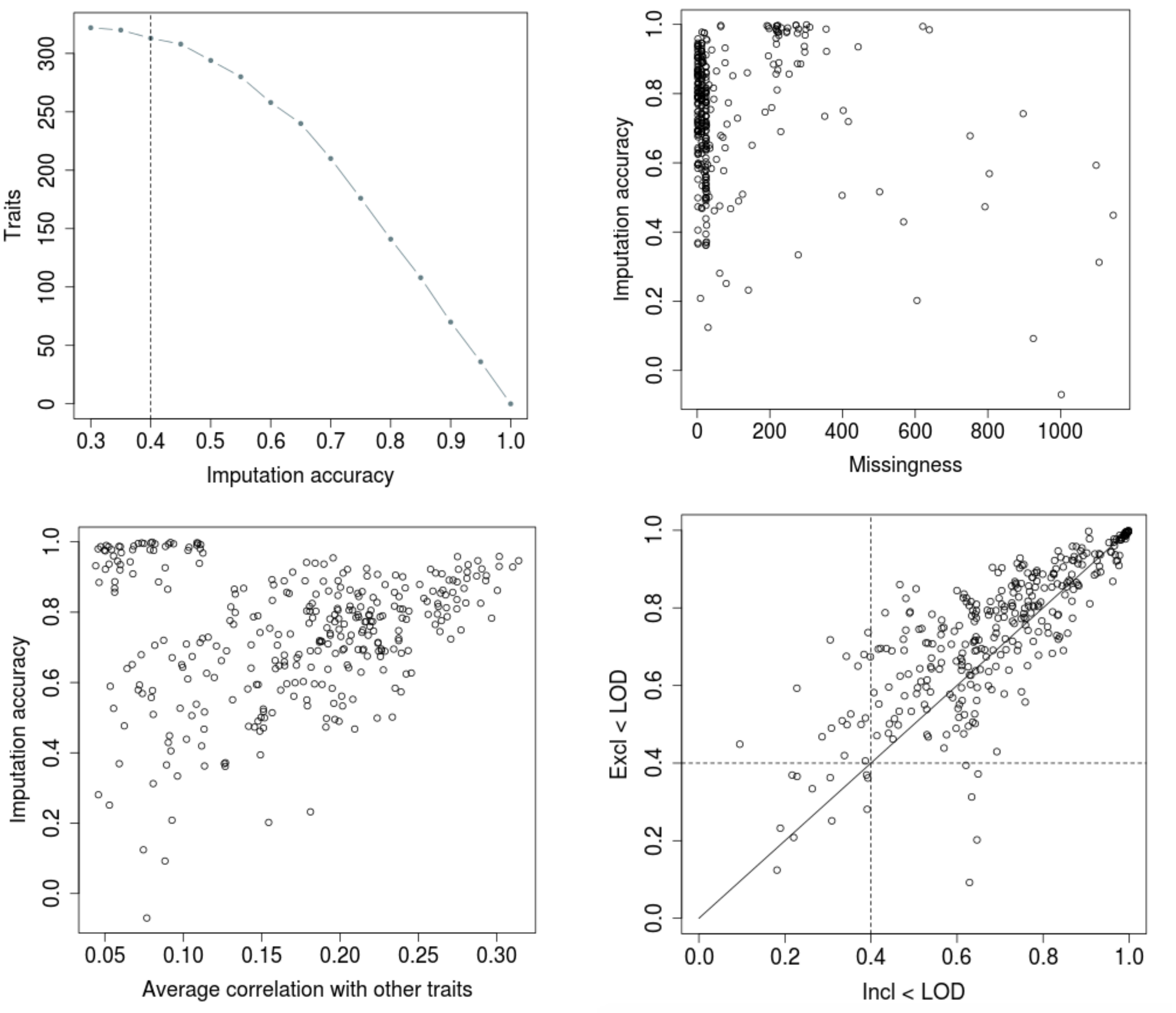

Figure 4.4. *Phenotype imputation of 330 traits in MANOLIS. Number of traits with imputation accuracy greater than a given threshold (top left). Imputation accuracy plotted against sample missingness (top right) and the average correlation of each trait with all other traits in the datasets (bottom left). Comparison of imputation accuracy when Olink trait values below the LOD were excluded vs when they were included (bottom right).*

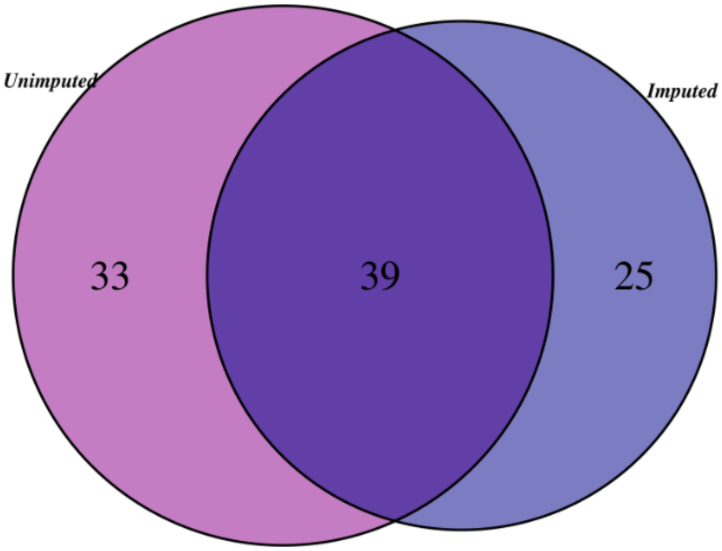

Figure 4.5. *Venn diagram of unique independent genome-wide ($p<5x10^{-8}$) signals in the single-trait GWAS using unimputed and imputed phenotype data.*

To assess the effect of phenotype imputation on association analysis, I compared the results of the single-trait GWAS for all 66 traits performed using unimputed values with the results from the imputed GWAS. I looked at the signal concordance between the imputed and unimputed runs at genome-wide significance ($p<5x10^{-8}$) after filtering for MAF>0.1%. In total, there were 72 independent genome-wide significant signals in the unimputed GWAS, and 64 in the imputed ones. Of these, 33 were unique to the unimputed, and 25 were unique to the imputed analyses (Figure 4.5). The majority (52% of the unimputed and 68% of the imputed) of these "private" loci were driven by rare variants (MAF<1%).

Out of the 66 traits, 27 had discrepant signals between the two GWAS runs, and 16 traits out of these had a pre-imputation missingness greater than 14% (equivalent to approximately 190 samples). Even though imputation accuracy for those 27 traits was high (>80% for all but 7 traits), inaccurately imputed phenotype values for even a small subset of samples can induce spurious effects at rare variants if these samples happen to be carriers of the rare allele. This might explain the observed discrepancies, but further inspection of these signals, as well as comparisons at more lenient p-value thresholds – or even across all included variants – are needed to fully evaluate the strengths and weaknesses of the phenotype imputation tool used here.

4.3.2 Multivariate GWAS

I conducted 19 multivariate GWAS different trait combinations of 66 traits. The number of traits included in each analysis ranged from 2 to 6 (Table 4.1). Study-wide significance was set at $p<2.66\text{x}10^{-10}$. There was no sign of inflation for any of the trait clusters after filtering for MAF>0.1% (lambda values ranging from 0.96 to 1.00); however, five trait clusters had a large excess of spurious signals across the genome (see section 4.3.2.1).

4.3.2.1 *Problematic trait groups*

Filtering for MAF>0.1% and "low-quality" variants (sample missingness>1% and/or HWE $\text{p-value}<1\text{x}10^{-5}$) should eliminate most spurious associations. However, the GWAS results from trait clusters 2, 7, 8, 11 and 17 showed a high degree of noise, as evident from inspection of their Manhattan plots (e.g. Figure 4.6). The lambda values for these clusters ranged from 0.97 to 1.00, showing no sign of systematic inflation and ruling out possible causes such as cryptic relatedness or population structure (which should have been accounted for by the inclusion of the relatedness matrix in the mixed model). When applying a MAF>1% filter, the vast majority of noisy signals disappear (Figure 4.6). This inflation in the low MAF range did not occur in the GWAS of any of the marginal traits for these clusters.

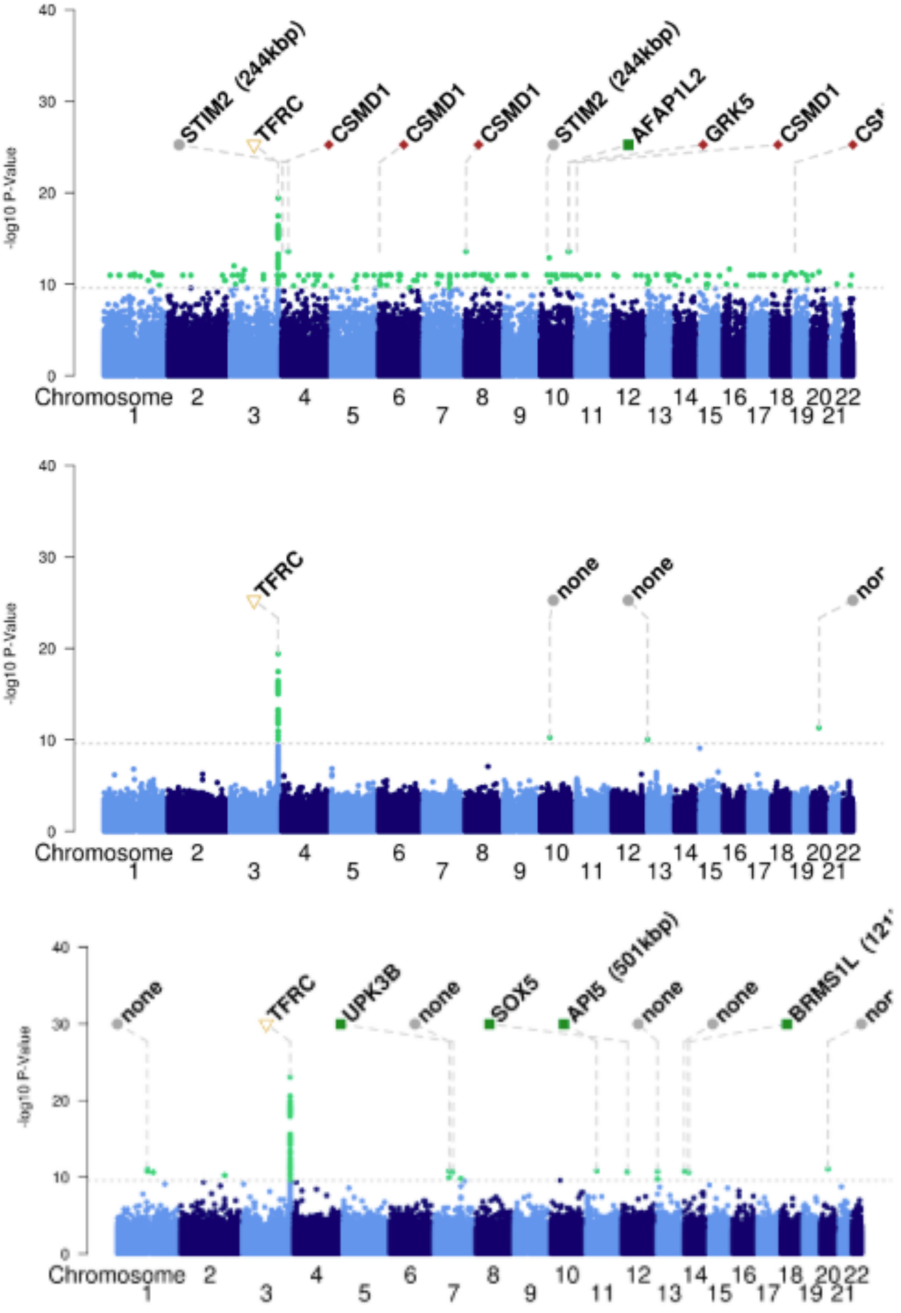

Figure 4.6. *Example of p-value inflation in the low allele frequency range. The results shown are from a multivariate GWAS using imputed (top and middle) and unimputed (bottom) phenotypes of transferrin receptor and iron levels (TR and Fe_iron). There is a systematic excess of signal when only filtering for MAF>0.001 in the imputed data (top) which disappears when filtering for MAF>0.01 (middle). This low-MAF inflation is not observed in the unimputed GWAS filtered for MAF>0.001 (bottom)*

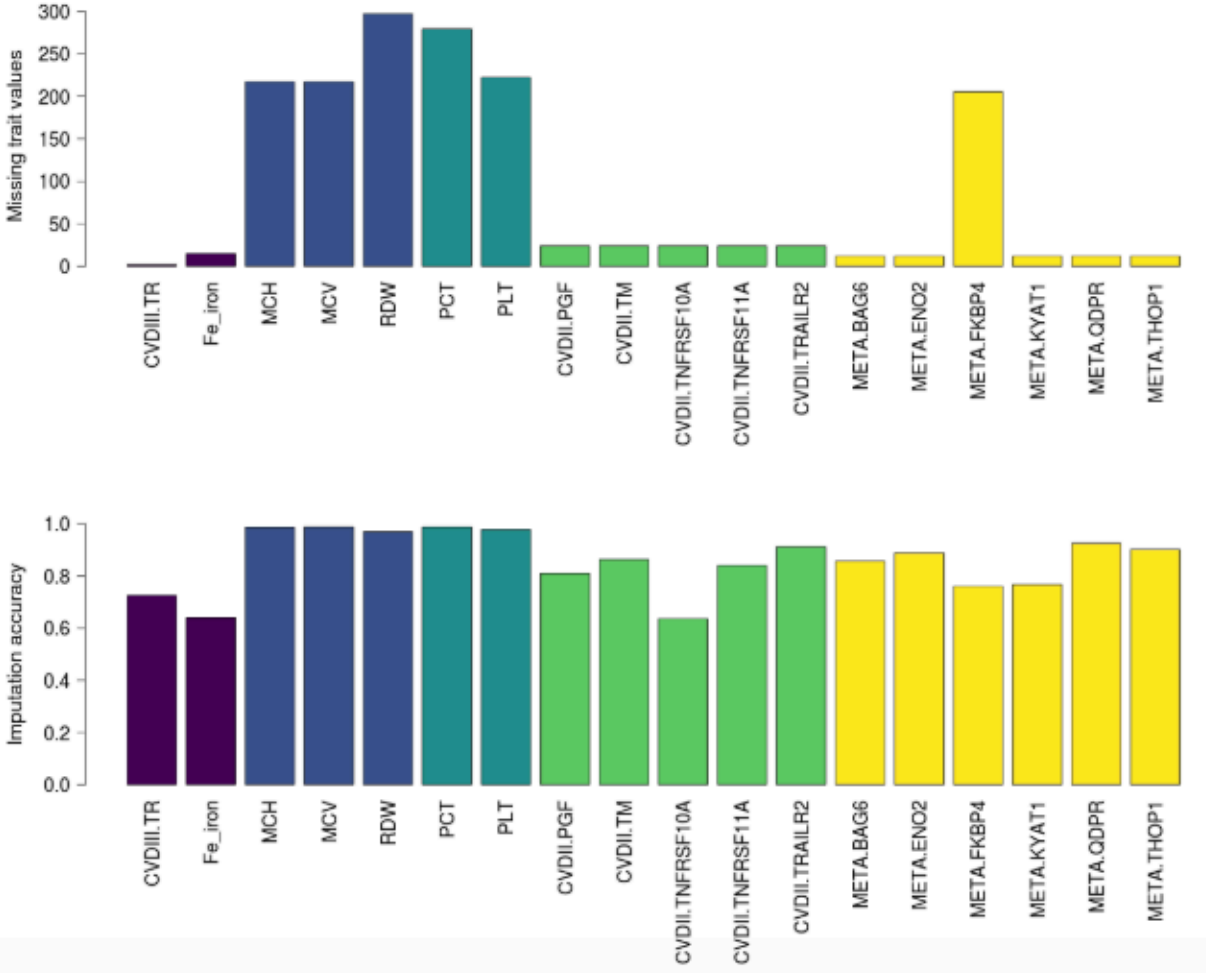

Figure 4.7. *Missing trait values before imputation (top panel) and imputation accuracy (bottom panel) for each of the traits contained in the trait clusters with inflation in the low MAF range. Bars are coloured in by cluster membership. For Olink traits, trait names are preceded by their respective panel.*

I next looked at phenotype missingness and imputation quality for these traits. One thing that stood out was that the individual traits within each cluster had an almost complete overlap in sample missingness (Figure 4.7). For the blood trait clusters (PLT-PCT and MCV-RDW-MCH) this could be due to inadequate sample quality, while for the Olink traits it might reflect batch effects (note that all traits within each cluster are from the same panel, except FKBP4). Although imputation accuracy was high for most of these traits (Figure 4.7), given that almost the same individuals within each cluster will have imputed phenotype values, this could lead to spurious associations with rare variants. Indeed, when re-running those clusters with unimputed phenotypes, most of the low-MAF associations disappear (Figure 4.6).

4.3.2.2 *Summary of multivariate GWAS results*

In total, there were 82 independent study-wide significant signals in the multivariate GWAS, 68 of which were not found in the univariate analyses. All 15 study-wide significant signals found in imputed univariate analyses showed even stronger association in the multivariate GWAS.

Of the new signals, 37 consisted of a single variant (i.e. no peak of variants). All of these single-variant signals also had a MAF<1%; while it is possible that some of these signals constitute real effects, it is likely that they are false positives, possibly arising due to inaccurately imputed phenotypes. I therefore did not take these 37 variants forward for further inspection at this stage. In the sections below, I briefly describe newly arising loci of potential biological interest.

4.3.2.3 *Newly discovered loci*

I performed two multivariate GWAS of osteocalcin: one with adiponectin, leptin, BMI, random glucose and random insulin (trait cluster 19), and another with OPN, MEPE, COL1A1 and ROR1 (trait cluster 18). Osteocalcin is a bone matrix-derived protein found in plasma that plays a role in both bone maintenance and glucose metabolism[320, 322, 326, 327]. So far, studies of the genetics underlying variation in circulating osteocalcin levels have not yielded replicating associations[328].

rsID	**chr:pos**	**A1,A1**	**MAF**	**Gene**	**P_{MV}**	**P_{BGP}**	**P_{COL1A1}**	**P_{MEPE}**	**P_{OPN}**	**P_{ROR1}**
rs7679698	chr4:87936047	G,A	0.405	*MEPE (-89kb)*	5.85E-16	8.32E-01	5.88E-01	3.29E-01	4.14E-10	5.88E-01
rs142201367	chr4:186235350	Indel	0.484	*KLKB1*	4.92E-26	7.93E-01	3.38E-01	3.54E-21	3.52E-05	4.46E-01
rs991353408	chr17:50304344	C,G	0.019	*TMEM92 (-22kb)*	1.09E-11	2.19E-01	1.58E-06	6.16E-01	5.02E-01	2.64E-01

Table 4.3. *Study-wide significant peaks in multivariate GWAS of BGP (osteocalcin), MEPE, COL1A1, and ROR1. Chromosome positions are aligned to GRCh build 38. A1 is the effect and minor allele. Closest gene within a 1Mb range is given for each signal, followed by the multivariate (P_{MV}) and univariate p-values.*

GWAS of trait cluster 19 did not result in any study-wide significant peaks, whereas trait cluster 18 yielded three study-wide significant signals (Table 4.3). Two of these lie on chromosome 4 and were also detected in the single-trait GWAS: the first signal (top variant: rs142201367, chr4:186235350, p=2.52x10^{-24}) lies 89,320 bp downstream of *MEPE* and was also study-wide significant in the MEPE GWAS (p=3.54x10^{-21}), and the second (top variant: rs7679698, chr4:87936047, p=1.98x10^{-16}) in the OPN GWAS, albeit with a different top variant (rs2126651, chr4:87922658, p=4.14x10^{-10}, LD with rs7679698: r^2=0.77). While both of these signals could be detected in the single-trait GWAS, the strength of association was several orders of magnitude larger in the multi-trait analysis.

The third signal lies on chromosome 17 and was not observed in any of the marginal GWAS. The top SNP (rs991353408; p=1.05x10^{-10}) lies in an intergenic region. The closest protein-coding gene is *TMEM92* (22kb upstream), a trans-membrane protein identified as a putative therapeutic target in prostate cancer[329]. Interestingly, the gene encoding COL1A1 lies 103kb upstream of rs991353408. This variant was also nominally significant in the COL1A1 GWAS (p=1.58x10^{-06}), but not in any of the other marginal GWAS (p>0.1 for BGP, MEPE, and OPN), making it plausible that the association is driven by a variant affecting *COL1A1* protein expression.

rsID	chr:pos	A1,A1	MAF	Gene	P$_{MV}$	P$_{MCH}$	P$_{MCV}$	P$_{RDW}$
rs1022764735	3:48791054	T,C	0.002	*PRKAR2A*	3.95E-12	6.70E-01	9.48E-01	4.20E-03
rs543237404	4:78208876	A,C	0.002	*FRAS1*	7.03E-11	3.04E-01	4.96E-01	4.94E-01
novel	11:131106748	G,C	0.002	*SNX19 (-190kb)*	4.91E-14	2.01E-01	3.54E-01	7.23E-03
novel	16:1726791	G,C	0.002	*MAPK8IP3*	1.10E-11	9.90E-01	8.37E-01	1.55E-02
rs867469149	16:7015419	T,A	0.002	*RBFOX1*	6.26E-11	6.51E-01	8.77E-01	5.06E-03
rs546767097	22:37390261	A,G	0.003	*ELFN2*	8.98E-13	1.75E-01	2.93E-01	3.00E-02

Table 4.4. *Study-wide significant peaks in multivariate GWAS of MCH, MCV and RDW. Chromosome positions are aligned to GRCh build 38. A1 is the effect and minor allele. Closest gene within a 1Mb range is given for each signal, followed by the multivariate (P$_{MV}$) and univariate p-values.*

Trait clusters 3-8 are all comprised of haematological measurements routinely obtained as part of full blood counts and used as diagnostic markers for a variety of health outcomes. Six signals (Table 4.4) were found in the multivariate GWAS of MCV, RDW and MCH –

measures used, for example, to distinguish between different types of anemia. All but one of the signals fall within introns of protein-coding genes, the exception being an intergenic signal at chr11:131106748. None of the genes in or around these signals have any direct documented link with blood biomarkers.

rsID	**chr:pos**	**A1,A1**	**MAF**	**Gene**	P_{MV}	P_{LPCR}	P_{MPV}	P_{PDW}
rs180950569	4:132130160	C,T	0.002	*none*	2.57E-10	2.81E-03	3.17E-03	1.06E-01
rs551490559	4:126614700	A,C	0.002	*none*	6.33E-12	3.14E-03	1.61E-03	1.30E-01
rs573664301	14:21356884	C,T	0.002	*SUPT16H*	1.62E-12	1.22E-08	1.84E-08	5.14E-06

Table 4.5. *Study-wide significant peaks in multivariate GWAS of LPCR, MPV and PDW. Chromosome positions are aligned to GRCh build 38. A1 is the effect and minor allele. Closest gene within a 1Mb range is given for each signal, followed by the multivariate (P_{MV}) and univariate p-values.*

Three signals reached study-wide significance in the cluster comprised of LPCR, MPV and PDW – indices of platelet reactivity used as markers for thrombosis[330] (Table 4.5). Two signals are intergenic; the third signal maps to an intron of *SUPT16H*, which encodes a chromatin factor that facilitates transcription by disassembling nucleosomes.

One signal was found in the analysis of PLT and PCT, two measurements of platelet abundance that can indicate bone marrow problems or excessive platelet destruction. The top variant (rs1249792881, chr2:34210095, p=1.26x10^{-10}) was also nominally significant in the PCT GWAS (p=8.27x10^{-6}), and lies in an intron of *LINC01317*, a long non-coding RNA of unknown function.

The CPA1-CPB1 GWAS contained a signal at a regulatory variant 5.6 kb upstream of *CPA1* (top variant: rs13240039, p=6.34x10^{-11}). CPA1 and CPB1 are pancreatic secretory enzymes that have been associated with pancreatic cancer and chronic pancreatitis[331]. Variants in the *CPA1* locus have also been associated with waist circumference adjusted for BMI[332].

rsID	chr:pos	A1,A1	MAF	Gene	P_{MV}	P_{THOP1}	P_{QDPR}	P_{KYAT1}	P_{FKBP4}	P_{ENO2}	P_{BAG6}
novel	14:72504434	T,C	0.002	*RGS6*	3.96E-13	1.37E-01	1.56E-02	1.05E-02	5.77E-03	4.50E-01	3.33E-02
novel	14:88517661	indel	0.003	*PTPN21*	2.68E-11	7.55E-02	1.34E-01	2.89E-03	1.82E-01	8.64E-01	3.32E-01
novel	18:24355834	A,T	0.002	*OSBPL1A*	6.77E-11	1.49E-02	2.58E-03	2.28E-01	8.05E-01	9.34E-03	9.24E-03
rs1021533786	17:71954309	G,A	0.002	*SOX9 (166kb)*	2.12E-10	6.86E-02	7.84E-02	2.83E-02	7.44E-01	2.40E-01	1.17E-02
rs1272582211; rs2741991	19:2802094	C,T	0.269	*THOP1*	1.52E-14	6.09E-06	6.45E-01	7.76E-01	5.35E-01	9.49E-01	6.58E-01
rs1471795955	9:127430680	C,G	0.002	*ZNF79*	4.43E-13	4.15E-01	1.33E-01	3.11E-03	3.85E-01	9.97E-02	2.44E-01
rs189262291	9:86991776	G,A	0.002	*GAS1RR*	1.38E-16	9.04E-03	5.52E-02	1.93E-03	7.91E-01	9.67E-02	5.11E-02

Table 4.6. *Study-wide significant peaks in multivariate GWAS of THOP1, QDPR, KYAT1, FKBP4, ENO2 and BAG6. Chromosome positions are aligned to GRCh build 38. A1 is the effect and minor allele. Closest gene within a 1Mb range is given for each signal, followed by the multivariate (P_{MV}) and univariate p-values.*

Trait cluster 17, comprised of FKBP4, THOP1, QDPR, BAG6, ENO2 and KYAT1, yielded a study-wide significant signal in *THOP1,* as well as five additional intronic signals and one in an intergenic region (Table 4.6). The proteins in this cluster broadly reflect immune- and neuronal functions. THOP1 is an enzyme involved in the cleavage of neuropeptides and the generation of amyloid-forming polypeptides. The *THOP1* locus has previously been associated with total cholesterol levels in East Asians[333].

Analysis of trait cluster 10 (LPL, PRELP, HO1, MERTK and XCL1) led to a signal at an intron of *LPL* (rs116135967, chr8:19902655, EA=C, EAF=0.29% p=8.61x10^{-20}), which also reached genome-wide significance in the marginal LPL analysis (3.34x10^{-08}), but none of the other included traits (p>0.1).

Multivariate GWAS of cluster 11 (TM, TRAILR2, PGF, TNFRSF10A and TNFRSF11A), which broadly represents angiogenesis- and apoptosis-related proteins(Appendix E), identified a cis pQTL for TM in the *THBD* gene, which encodes thrombomodulin (TM) and is involved in blood clotting[334]. The top SNP at this signal is a missense variant (rs1042579, chr20:23048087, EA=A, EAF=8.4%) leading to the replacement of an alanine amino acid by valine (p.Ala473Val). It has been associated with hemolytic-uremic syndrome[335]. The variant also reached genome-wide significance in the TM univariate GWAS (p=1.16x10^{-11}), but is not associated with any of the other individual traits in the cluster (p>0.1).

Another signal occurred at an intron of *ANKS1B* (rs150347635, chr12:99608373, EA=T, EAF=0.02%, $p_{\text{multivariate}}$=4.06x10^{-15}). The top SNP is also suggestively associated with PGF levels (p=6.43x10^{-08}), but not with any of the other traits included in that cluster (p>0.1). *ANKS1B* is involved in the maintenance of endothelial permeability[336].

Trait cluster 12 was the largest trait cluster, consisting of six protein measurements (PRTN3, MMP9, MPO, PGLYRP1, RNASE3, AZU1, RETN and PTX3). Multivariate GWAS lead to the identification of five putative cis-pQTLs for proteins contained in the cluster (Table 4.7). The first is driven by a rare splice acceptor variant, rs35897051, in the myeloperoxidase (*MPO*) gene and was suggestively significant in the univariate MPO analysis (p=1.02x10^{-05}). The second falls into a regulatory region approximately 7kb upstream of *PGLYRP1* and was suggestively significant at $P<10^{-2}$ in the PGLYRP1 and MMP9 analyses. The third is also a regulatory region variant 1kb upstream of *RETN* and reached genome-wide significance in the marginal RETN GWAS. The fourth signal lies in an intron of *PRTN3* and was also genome-wide significant for the protein product of that gene. Lastly, multivariate GWAS identified a variant overlapping a promoter region of *RNASE3*the same signal was suggestively significant in the RNAS3 GWAS.

One pQTL signal was identified in the analysis of trait cluster 13 (JAMA, PECAM1 and CASP3) in an intron of the *ABO* locus, specifically the *ABO-201* transcript. The top variant is an indel (chr9:133263362, alleles=GCGCCCACCACTA/G, MAF=48%, $p=1.26x10^{-36}$; closest 1000 Genomes variant: rs8176686, chr9:133263373) and also reached suggestive significance in the PECAM1 marginal GWAS ($p=7.39x10^{-07}$), but had p>0.1 in the JAMA and CASP3 analyses. The *ABO* locus encodes a glycosyltransferase which is abundantly expressed in several tissues and whose activity determines an individual's blood group. The enzyme is responsible for modifying cell surface antigens bound by platelet glycoproteins such as PECAM1[337] or platelet receptor such as JAMA[338].

rsID	Chr:pos	A1,A2	MAF	Gene	P_{MV}	P_{RNASE3}	P_{RETN}	P_{PTX3}	P_{PRTN3}	$P_{PGLYRP1}$	P_{MPO}	P_{MMP9}	P_{AZU1}
rs35897051	14:20917832	T,C	0.05	*RNASE3*	9.86E-17	2.29E-07	7.93E-01	3.74E-01	6.74E-01	6.68E-01	4.43E-01	3.50E-01	9.02E-01
rs12982353	17:58270865	G,T	0.01	*MPO*	2.24E-12	1.83E-01	5.36E-01	8.48E-01	2.86E-01	1.41E-01	1.02E-05	7.96E-02	6.98E-01
rs147307766	17:59127420	T,G	0.01	*SKA2*	4.97E-11	2.87E-01	5.19E-01	6.62E-01	4.73E-01	1.39E-01	2.06E-04	3.45E-01	9.27E-01
rs34861192	19:46026198	A,G	0.21	*PGLYRP1 (-7kb)*	3.75E-18	5.57E-01	4.40E-02	4.35E-01	9.91E-03	1.80E-01	2.72E-01	5.10E-03	2.99E-01
rs566673173	19:7668689	A,G	0.02	*RETN (0.1kb)*	4.24E-36	8.46E-01	2.50E-09	9.65E-01	1.44E-01	5.38E-01	6.86E-01	9.96E-01	4.91E-01
rs6510982	19:845535	C,G	0.21	*PRTN3*	9.78E-31	5.52E-01	5.35E-01	1.89E-01	3.07E-08	6.88E-01	5.74E-01	7.96E-01	9.88E-01

Table 4.7. *Study-wide significant peaks in the multivariate GWAs of PRTN3, MMP9, MPO, PGLYRP1, RNASE3, AZU1, RETN, and PTX3. Chromosome positions are aligned to GRCh build 38. A1 is the effect and minor allele. Closest gene within a 1Mb range is given for each signal, followed by the multivariate (P_{MV}) and univariate p-values.*

4.4 Discussion

In this chapter I have outlined a framework for multivariate GWAS analysis in datasets with high-dimensional phenotype data. Compared to the single-trait analyses, multi-variate GWAS uncovered 9 new study-wide significant signals with putative cis-effects for one of the analysed proteins, as well as several other loci of potential biological relevance. Multivariate analysis also successfully recapitulated univariate signals, with an average 11-fold increase in association strength. Overall, this highlights the advantages of leveraging inter-trait correlations for locus discovery. There are, however, a few pitfalls to multivariate GWAS, and additional work will be necessary to gain a more comprehensive picture of those. I discuss this further in the below sections, and also outline possible approaches for follow-up of signals as well as future analyses.

4.4.1 Advantages and challenges of multivariate GWAS

Jointly analysing multiple phenotypes in a multivariate GWAS framework can increase power to detect associations, as well as refine existing ones. In line with previous reports, all marginal signals are recapitulated by multivariate analysis with increased strength of association[128], supporting multivariate GWAS as a robust tool for the identification of trait-associated loci. Even for cis-loci primarily affecting one trait, leveraging the correlation between the trait of interest and others in the dataset can result in a stronger signal. The number of independent study-wide significant signals increased almost 5-fold compared to univariate GWAS, however, this number should be taken with a grain of salt: more than half of these new signals consisted of a single variant. While some of these might constitute real associations (for example, if the variant is rare and not in LD with any nearby variants), they could also be statistical artefacts and need to be investigated further. Moreover, 48 of the 68 new multivariate signals arise from two trait clusters (clusters 4 and 17, Table 4.1). These trait clusters were among the five that had to be re-run with unimputed phenotypes due to signal inflation in the lower MAF range. Discarding the imputed phenotypes removed the majority, but not all, of these spurious associations, and the results from these GWAS should therefore be interpreted with caution.

Since multivariate analyses in a GWAS setting are only beginning to gain traction, there is no standardised analysis pipeline. Here, I chose a mixed model in order to account for between-sample relatedness; Shen et al. chose to first adjust phenotypes for relatedness in a mixed model and use the adjusted residuals to perform MANOVA, while Inouye et al. used canonical correlation analysis (CCA). While one can certainly argue about the relative advantages of these methods, they are all statistically sound, yet result in different outputs. CCA returns trait loadings for each variant that capture the extent to which each trait contributes to the observed genetic effect. Multivariate linear models, on the other hand, return matrices of beta estimates and covariances, respectively. Replication and meta-analysis are therefore not as simple as in a univariate setting, and further work is needed to establish a clear framework for such studies.

4.4.2 Caveats of phenotype imputation

The requirement of multivariate models for complete phenotype data will in many cases necessitate the imputation of missing values, unless sample sizes are very large and missingness patterns between phenotypes do not lead to significant samples loss. Furthermore, performing a complete cases analysis (i.e. excluding all samples with missing trait values) can lead to biased results if missingness is informative and the variables associated with missingness are not included in the analysis. As shown here, even when imputation accuracy is good, careful inspection of the results is necessary to guard against spurious associations. In this chapter, I used PHENIX[318] to impute missing trait values, but a number of other methods exist. These can be broadly categorised as single or multiple imputation methods. Single imputation, of which PHENIX is an example, imputes each missing value only once based on a pre-specified model that incorporates known information (e.g. other phenotypes, genetic data, or relatedness). This leads to overly small standard errors and biased results, since the uncertainty implicit in the imputation procedure is not accounted for. Once missing values are imputed, they are treated no differently from measured values.

Conversely, in multiple imputation each missing value is imputed multiple times from its predictive distribution based on the observed data[136, 339]. After each imputation ‘cycle’, the model of interest (e.g. logistic regression for binary traits, or linear regression for quantitative ones) is fitted to the complete dataset; the imputation and model fitting steps are repeated a given number of times. After each iteration, the obtained test statistics will differ slightly, since the imputed values will not be exactly the same. The overall association estimates and standard errors can be calculated by averaging test statistics across all iterations[136]. This way, the obtained results should accurately reflect the uncertainty of imputing missing data points. In recent years, multiple imputation in epidemiological research has gained popularity as a more robust tool for handling missing data compared to traditional approaches such as complete case analysis or last value carried forward approach (for data with multiple time points)[136].

Multiple imputation by chained equation (MICE) is an R package implementing a multiple imputation framework that iterates over each phenotype in the data, fitting a univariate model to predict missing values based on all other phenotypes/variables in the dataset[340]. A downside of this approach is that, unlike PHENIX, it does not take into account phenotypic correlations, which can lead to less accurate imputation estimates[318]. On the other hand, MICE allows for more flexibility with regards to the distribution of individual phenotypes. If one only wants to obtain imputation estimates for missing values, the association analysis after each imputation step can be omitted and the final reported imputed value for each missing observation is the average across all repetitions (five by default). Thus, MICE could have been used to obtain a complete dataset of HELIC phenotypes to use for multivariate GWAS in GEMMA.

Another caveat of MICE is that by default rests on the assumption that missing data are missing at random (see section 1.4.6.1). In practice, it is impossible to test whether this assumption holds. In case of the Olink protein measurements, MAR might be plausible because it is unlikely that the level of protein expression is predictive of missingness – assuming assay quality and calibration is more or less uniform across all proteins. However, if for example very low protein levels were more likely to be excluded based on LOD, MAR would be violated and MICE might lead to bias. Simulations have shown that under missingness not at random, both PHENIX and MICE lose accuracy, but PHENIX still outperforms MICE (Ref [318], Supplementary Figure 6).

In summary, although multiple imputation has several advantages over single imputation, it also requires a more detailed assessment of the data and model specifications. Phenotype imputation is only beginning to gain traction in genetic research, and regardless of the exact method used careful inspection of results and/or sensitivity analyses[341] should be carried out to guard against false inference.

4.4.3 Selection of biologically meaningful trait groups

In this chapter I have shown that choosing trait groups based solely on the inter-trait correlation can aid the detection of pQTL signals. This is an encouraging lesson for future work done in high-dimensional datasets with hundreds (or even thousands) of proteomic measurements. The approach I used here has the advantage of being easy to implement and straightforward to interpret: all traits included in one cluster will have satisfied the specified correlation threshold. A downside is that, depending on the dataset and phenotypes, there might be a trade-off between choosing an appropriate correlation threshold and getting very large clusters of more than a dozen traits. In this case, individual clusters could be further broken up by changing the correlation threshold for each one individually.

Other approaches may also be used, such as the unsupervised clustering algorithm employed by Inouye et al.[314] to form trait groups from 130 metabolites[342]. In order to maximise biological relevance of the resulting clusters, one could also incorporate additional data such as Gene Ontology annotations or pathways.

4.4.4 Interpretation of multivariate signals

Compared to other multifactorial or disease traits, proteins and RNA expression levels are more proximal to the genetic code, and thus allow for a more direct investigation of genetic effects when used as quantitative traits in GWAS. While this may increase power to detect associations, it also makes the subsequent interpretation of these more challenging. Connections between proteins might not be immediately obvious, such as here for the cluster surrounding FKB4. Proteins themselves are often pleiotropic, meaning they act in different pathways and their functional consequences are therefore context-specific.

A caveat of proteomic analyses is that, similar to gene expression studies, relationships between genotype and protein levels are only reflective of the tissue the samples were drawn from. This is of special importance when studying pleiotropy, which can be tissue- or even developmental stage-specific.

Relating QTL associations to disease states constitutes a further challenge. One question to ask is whether perturbed expression levels precede onset of disease or whether they are in fact a result of it. In the absence of longitudinal data, where this can be explicitly modelled, Mendelian randomisation might be employed to infer causality. This is made more complex when trying to disentangle cross-phenotype effects: does a QTL only truly affect expression of one of the proteins in question, which then in turn has a knock-on effect on other proteins? Or does the causal locus perhaps affect transcription of several genes, whose protein products are involved in a disease-relevant pathway? To truly answer these questions, one would need access to samples of diseased and healthy tissues (ideally from the same individual, to minimise inter-individual variability). As outlines in Chapter 1, this option is not unrealistic for disorders where the affected tissue is both known and readily available, such as osteoarthritis. For many others, however, this is still a challenge.

A more realistic approach might be to overlay association results with other omics information from publicly available datasets such as the Roadmap Epigenomics Consortium, which contains data on epigenetic modifications and gene expression. Colocalisation analysis between pQTL and disease associations might also shed light onto the mechanism of action for some loci. This type of analysis has been widely used to relate gene expression QTLs to GWAS hits, and could also help to refine cross-phenotype effects. The GTEx Project comprises gene expression data across over 40 tissues[234, 343] and thus constitutes a comprehensive resource for functional analysis of GWAS data[344].

4.4.5 Future work

4.1.1.2 *Follow-up and replication*

While the results presented in this chapter are encouraging, further work is required to robustly establish and characterise newly found association signals, as well as to fine-tune phenotype imputation and post-association QC procedures.

So far, I have only compared multivariate signals with univariate results from the traits included in at least one cluster. In order to determine whether they are truly novel, they will need to be cross-checked with univariate GWAS results of all Olink and standard traits in MANOLIS.

Novel multivariate signals will need to be replicated in an independent cohort. Ideally, replication would involve both multi- and univariate analyses of the same trait clusters. However, in practice this is often not feasible, and several associations originally reported from multivariate analyses[131] have then been replicated in univariate GWAS of one or several of the traits[345]. A potential avenue for seeking replication in this context is the HELIC-Pomak cohort, for which we are exploring the option of Olink measurements. Another option is the replication in a general population sample with proteomic data, such as the INTERVAL study[32]. A caveat of this approach is that effect sizes and lead SNPs at an associated locus might differ from those in MANOLIS and/or Pomak, reflecting the variation in genetic architecture of complex traits in population isolates.

4.1.1.3 *Further analyses*

In addition to follow-up of the signals identified here, there are several other analyses that could be explored. For one, the 19 trait clusters analysed here are by no means an exhaustive list of possible groupings, and there are undoubtedly additional trait combinations and clustering methods that can be explored. For example, the cluster comprising 44 traits returned by the igraph algorithm and excluded from further analysis could be broken down further by inter-trait correlation and biological functions.

A comparison of other phenotype imputation tools might also reveal performance advantages under different scenarios (e.g. high missingness, weak inter-trait correlation, low heritability of traits).

Finally, to fully harness the potential of population isolates coupled with high-depth sequencing data, multi-trait burden analysis could be carried out to identify rare variants associated with multiple traits.

Chapter 5 – Discussion and future outlook

5.1 Thesis summary

Together, the results presented in this thesis highlight the potential of the joint analysis of phenotypes to increase power to detect novel associations, prioritise sub-genome-wide significant variants for replication and investigate the genetic basis of comorbidities. While results from Chapters 2 and 3 focus on two trait pairs that have established medical and epidemiologic connections, Chapter 4 makes a case for selecting trait groups based on their correlation rather than prior documented links.

In Chapter 2, I applied four overlap analysis tools on summary statistics from an OA and a BMD GWAS, respectively. I followed up variants with evidence of cross-trait association in independent OA datasets, which lead to the identification of a new OA risk locus at *SMAD3*. This gene has previously been studied as a candidate risk locus for osteoarthritis due to its role in cartilage maintenance, but could not be robustly associated with the disorder. This exemplifies the potential of identifying novel trait loci by using summary statistics of related phenotypes for variant prioritisation. Despite the caveats this approach suffers from (inability to control for confounding factors or stratify samples due to lack of individual-level data), I identified several loci in or near genes relevant to OA and/or BMD through association or functional studies. As multivariate analyses on high-dimensional datasets are becoming the gold-standard for pleiotropy research, these results highlight that using data from existing studies can be a fast and efficient way to assess the genetic overlap between two traits.

In Chapter 3, I investigated the genetic contribution to SCZ and T2D comorbidity. I used results from published GWAS for each of these disorders, in conjunction with individual-level data from a cohort comprising patients with T2D and/or SCZ. Polygenic risk score analyses showed that patients with both disorders had a higher burden of T2D risk variants than patients with only SCZ. This further supports the hypothesis that the observed comorbidity of SCZ and T2D is not solely due to environmental factors. It also shows that risk scores constructed based on well-powered GWAS summary statistics can be successfully applied to smaller datasets.

In Chapter 4, I assessed a framework for multi-trait analyses in an isolated population with 15x sequencing data. This project serves as a proof-of-concept for statistical trait clustering in high-dimensional data. It also exemplifies the added power for cis-signal identification afforded by including correlated traits with the “driver” trait of the association. The results from uni- and multivariate GWAS demonstrate the utility of phenotype imputation to recapitulate missing data points, while also pointing to filtering/QC criteria (e.g. sample missingness not at random) that could be applied to avoid false positive associations.

5.2 Limitations of this thesis

There are several limitations to the type of data used in this thesis which warrant a cautious interpretation of the obtained results. Chapter 2 focuses on the shared genetics of OA and BMD, a disease and a quantitative trait known to be inversely correlated. The data used relies on summary statistics from published GWAS; while individual-level information was available for the arcOGEN OA dataset, this was not the case for the GEFOS discovery data, which consisted of a meta-analysis of 17 studies. Since it was not possible to obtain information on OA disease status in GEFOS, and BMD measurements were not available in arcOGEN, the possibility of inflated overlap estimates due to the presence of OA cases in the BMD data cannot be excluded. Several of the studies participating in GEFOS are aimed at a range of complex disorders[13], including OA (e.g. the Rotterdam study); furthermore, the average age in most GEFOS discovery studies exceeded 50 years, making it likely that the proportion of OA cases present in the GEFOS meta-analysis is higher than in the general population.

Furthermore, arcOGEN cases were ascertained for severe OA, with about two thirds of participants having undergone joint replacement surgery[38]. This selection process might lead to an artificial enrichment for high BMD that is not due to shared genetics: if, in fact, BMD lies on a causal pathway to OA, then selecting for severe OA might also to some extent select for higher BMD. An overlap analysis with a BMD dataset might then result in significant overlap, despite the fact that there is only mediated, not biological pleiotropy between the two traits[58]. Another possibility is that people with OA who also happen to be genetically predisposed to higher BMD have more severe symptoms, e.g. more pain due to bone spurs; it is conceivable that those people are more likely to participate in genetic studies, which would lead to inflated overlap estimates.

The GOMAP study presented in Chapter 3 suffers from similar limitations. The finding that patients comorbid for SCZ and T2D have a higher burden of T2D risk variants than controls or patients with SCZ supports the idea that the increased prevalence of T2D in SCZ patients is not purely environmental. However, the increased genetic risk of T2D in the comorbid patient group might simply reflect the population risk of the condition: if a certain proportion of the general population carries a higher load of risk variants for T2D, then so will a proportion of patients who have SCZ. By ascertaining for T2D status among SCZ patients, one also indirectly selects for SCZ patients likely to carry more T2D risk variants.

It is furthermore possible that the comorbid patients in GOMAP represent a "high risk" subgroup of SCZ patients, having both a genetic predisposition as well as the added metabolic burden of antipsychotic medication. This may make them susceptible to more severe metabolic side effects and/or earlier onset of T2D compared to SCZ patients with no genetic predisposition. Similar to the possibility of ascertainment bias in arcOGEN, this elevated disease burden might motivate patients to participate in a study aimed at better understanding their specific health issues.

Another consideration is the potential for disease misclassification in SCZ[346, 347] (and psychiatry in general[348]). The DSM-IV diagnostic criteria for SCZ overlap with other psychiatric disorders, such as bipolar disorder[349, 350], schizoaffective disorder or psychotic depression. Such blurred boundaries between different diagnoses could affect genetic overlap estimates with other disorders (such as PRS analyses). For these effects to be significant, misclassification rates must be relatively high (>10%)[351]. Longitudinal studies examining the reclassification of initial SCZ and bipolar diagnoses found that approximately 15% and 4-6%, respectively, were later revised[352, 353]. If the misdiagnosis rate in the GOMAP and/or PGC SCZ cases was similar, this could have led to over- or underestimate of genetic overlap between SCZ and T2D.

The phenotypes used in chapter 4 are less likely to suffer from the above-mentioned biases, as they are quantitative, do not rely on diagnostic classification systems, and (in case of the Olink protein measurements) are more proximal to the genetic code than disease traits. Nevertheless, measurement errors due to assay performance or sample degradation, as well as environmental confounding cannot be completely ruled out. It has been previously shown that long time storage (over six months) of serum samples significantly affected read intensity of proteins[354]. While the measurement method used by the authors differs from that of Olink (mass spectrometry vs proximity extension assay), it is possible that the four years between the HELIC sample collection and the application of the Olink assay had confounding effects on the protein measurements.

5.3 Interpreting multi-trait associations

An important question moving forward in multi-trait and -omics analyses is what a statistical association at a locus means biologically. The functional follow-up and characterisation of an association with a single trait already requires substantial resources. With the inclusion of several traits and data sources, the interpretation of a significant p-value becomes even more complicated. The first step is to determine which trait(s) a signal is likely driven by, which can be achieved by inspecting the marginal betas and p-values from univariate analyses. This might already offer some hints as to which of a group of traits are truly affected by the associated variant(s). If a variant shows at least nominal association with more than one trait, the possibility of only one phenotype driving the signal can already be excluded. This leaves three broad scenarios that could underlie a multi-trait association:

First, the variant might truly influence all traits, for example by altering transcription of a relevant gene or multiple genes that each in turn affect one of the studied traits.

Second, the association might arise due to mediated pleiotropy, meaning there exists a causal chain between the included traits. This can be formally tested in a statistical framework through MR analysis, although results should be interpreted with caution (see section 1.4.7). As outlined in Chapter 1, there are a number of caveats to this approach, such as collider bias when stratifying study samples by a factor that is itself associated with the instrumental variables. MR studies have helped to untangle epidemiological associations between several complex trait pairs[355-358], such as BMI and OA[61].

Third, the multi-trait association signal might be an artefact caused by confounding factors. For example, environmental factors might induce a correlation between two traits, leading to a variant causal for trait one to also be associated with trait two.

Of the above scenarios, only the first can be classified as "true" biological pleiotropy, and this classification itself has several sub-categories depending on the causal mechanism underlying an association[57]. Determining whether an association signifies biological pleiotropy is not straightforward, as it requires in-depth information on the functional consequences of a variant for the analysed traits. In the absence of such data, the broader functional consequences of a variant may be examined instead:

For example, MR analysis has been adapted to link mRNA expression levels to complex traits[159]. This approach, SMR, does not require that expression and phenotype data are measured in the same individuals. Publicly available functional datasets derived from multiple cell types and tissues, such as GTEx[359, 360], Roadmap[361] or Blueprint[362], can be used instead. While such resources are useful to predict functional consequences of GWAS loci, they also have several shortcomings[363]: since phenotype information is not available, it is not possible to distinguish whether differences in functional measurements (e.g. gene expression) are due to normal variation or disease. The absence of individual-level sample information also precludes the adjustment for environmental confounding factors. Lastly, gene/protein expression and epigenomic marks change over time, and the age or even time of year of sample collection might themselves affect the traits measured (especially in post-mortem samples, which GTEx is comprised of).

While statistical frameworks might be used to gain initial insights as to whether an association independently affects all analysed traits, they ultimately do not replace functional follow-up of a signal. It is therefore important that datasets with comprehensive phenotype and omics measurements are set up, and to fully harness such data collaborative analysis approaches will be paramount. In the below sections I discuss the prospects for the evolving field of multi-omic/multi-trait research, as well as some of the challenges that still lie ahead.

5.4 The measured man – the growing field of phenomics

We now have access to an unprecedented breadth of information describing the state of an individual's health, from ICD-10 codes to biomarkers measurements and imaging data. Biobanks with links to electronic health records are slowly becoming the norm rather than the exception. With this wealth of data at our fingertips, we can for the first time comprehensively characterise the human "phenome". This brings with it a new set of challenges quite different from the ones posed by studying genomic variation.

5.4.1 Prioritising phenotypes to analyse

Sequencing costs have dropped drastically, and other high-throughput omics technologies such as RNA-seq and protein expression assays are following this trend. Nevertheless, extensive molecular phenotyping in large sample sizes is often still prohibitively costly, necessitating a prioritisation step when deciding which traits to assay. Houle et al. distinguish between intensive and extensive phenotyping[364]. The former refers to very detailed characterisation of one or a small number of phenotypes across multiple time points, tissues or even cell types, whereas the latter refers to efforts to obtain data on as many phenotypes as possible. Extensive phenotyping has the advantage that it can be carried out on very large sample sizes. An example of this kind of study set up is the UK Biobank study, where hundreds of phenotypes, including questionnaire data and clinical records (hospital episode statistics) are available for approximately 500,000 participants from the general UK population.

When deciding between an extensive or intensive phenotyping approach, a key consideration is how much additional information can be gleaned from each phenotype that is measured[56]. For example, the Olink protein expression panels each contain 92 proteins chosen based on their (presumed) relevance to a broad biological domain or disease category (e.g. cardiovascular, immuno-oncology, neurology, or metabolism). As a consequence, within-panel correlations of protein measurements are relatively high (Figure 5.1). On the one hand, this is useful as it allows in-depth investigation of the pathways represented in a panel. On the other hand, if the main goal is to obtain a comprehensive "snapshot" of the human proteome, maximising the number of non-correlated measurements would be more cost-effective, albeit with the downside of decreased resolution.

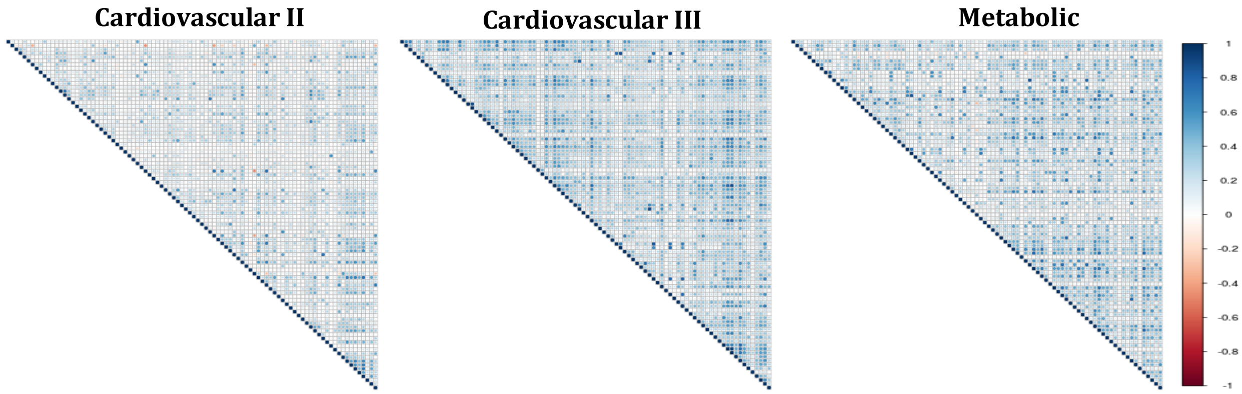

Figure 5.1. *Correlograms of protein levels included in each of the three Olink panels measured in the HELIC-MANOLIS cohort, as outlined in Chapter 4.*

Even when assay cost is not a primary issue, specimen availability poses another potential restriction on how many traits can be measured. Especially for existing collections, there may be a limited quantity of tissue samples – most commonly serum or plasma – that can be used for molecular assays.

A compromise between an intensive and extensive phenotyping approach can be achieved by extensively phenotyping a large study cohort, and then performing more in-depth molecular assays on a subset of individuals. This approach has been used in the LifeLines cohort, a multi-generational Dutch study of over 167,000 individuals. A sub-sample of these comprised of 1,539 individuals (LifeLines DEEP) were taken forward for detailed multi-omics measurements, including RNA-seq, proteomics, gut microbiome characterisation and methylation[365].

5.4.2 Rethinking diagnostic criteria

The question of how to define a trait has perhaps been most relevant to health-related phenotypes, many of which have conventionally been modelled as categorical variables, e.g. disease vs no disease. However, this may lead to a loss of information as the underlying biological changes will almost always be quantitative. The use of endophenotypes as well as analyses of disease subtypes have been proposed to help bridge the gap between genetic variation and current disease classifications. In psychiatry, where current diagnostic boundaries are widely deemed inadequate[346, 366, 367], the investigation of disease sub-categories has provided evidence in support of continuous disease models[368]. For disorders where diagnostic criteria rely on clinical tests and are relatively robust, patient stratification can elucidate genetic heterogeneity and refine treatment approaches. For example, osteoarthritis patients can be stratified by affected joint site[38] or presence of radiographic features[369, 370]. It should be noted that stratification or subtype analysis requires large sample sizes to prevent a loss of power, and thus only make such approaches practical in big sample collections, such as the UK Biobank[172]. Another caveat is that existing datasets might lack the detailed information needed to perform such analyses.

Of course, endopheno- or subtype definitions are often based on the same framework as disease classifications, and may therefore themselves be inaccurate. Bilder et al. suggested "dynamic phenotyping" in the context of behavioural traits, referring to the "iterative refinement of phenotype assays based on prior genotype-phenotype associations"[371]. The possibility of such fine-tuning of disease definitions based on molecular data has been employed in oncology (e.g. estrogen-sensitive/insensitive breast cancer[372]), and is now also on the horizon for immune-mediated and other disorders[373, 374]. Nevertheless, a lot of work still lies ahead to establish and then successfully mine the deeply phenotyped datasets necessary for such studies.

5.5 Study designs

5.5.1 Experimental setup

To study the genetic effects on individual diseases, ascertained samples comprising healthy and affected individuals are most frequently used. This type of sample selection does not work well when the aim is to establish a "phenomic" dataset. Instead, cohort studies, ideally with longitudinal data, can be used. The set-up and maintenance of such projects is time-consuming and expensive. Even the initial planning of which type of data to collect requires considerable resources, as discussed in section 5.3. Unlike our DNA sequence, which is assigned at birth and does not change over time, phenotypes can vary over developmental stages and across different environments. The genetic effects on phenotypic variation therefore have a temporal component that is disregarded unless longitudinal data is available and different time points modelled explicitly[128]. While for some types of data it is possible to obtain measurements retrospectively from stored samples (e.g. proteomics data from blood collected at the initial assessment), some need to be assessed at the time point of interest (e.g. imaging data). For clinical phenotypes, longitudinal data is often available through electronic health records. The downside is that this information does not routinely encompass molecular phenotypes and can be noisy due to the lack of standardisation (e.g. subjectivity of physicians, irregularity in hospital or doctor visits). The establishment of birth cohorts or prospective studies, where a group of individuals is followed over a period of several years or decades, can eliminate this problem, but both approaches are resource-intensive.

Another aspect of experimental design is which genotyping method to use. High-throughput sequencing of whole exomes or genomes is now feasible for large sample sizes, but is still more costly and time-intensive than array-based methodologies. WES has been successfully used to study rare deleterious mutations in protein-coding genes[375, 376]. On the other hand, the majority of associated loci for complex traits localise to intergenic regions whose function has not been clearly characterised. Deep phenotyping coupled to WGS could help to bridge this knowledge gap. To reduce costs, array-based genotyping can also be used in conjunction with WGS on a subset of individuals, allowing for imputation of untyped variants based on the whole-genome sequenced haplotypes[363]. This approach was successfully used in Iceland to recapitulate rare variants in over 100,000 individuals based on sequencing data of another 2,636[207]. External reference panels can further boost imputation accuracy: the Haplotype Reference Consortium spans over 32,000 predominately European samples, with plans to include more diverse ancestries in future releases[103].

5.5.2 Choice of population

The overwhelming majority of genetic association studies to date have been carried out in individuals of European descent. However, studying genotype-phenotype relations across different ethnicities has a number of advantages. The most obvious one is that combining data across ancestries increases sample size and therefore power. One caveat of trans-ancestry GWAS or meta-analyses could be that allelic heterogeneity could offset the aforementioned gain in power. This concern was alleviated by findings that risk alleles and effect sizes of most GWAS hits appear to be shared across ancestries[11, 12, 377-379], suggesting that the true (often unknown) causal variants at these loci arose before migratory events that separated populations, and that these causal variants are likely common[380]. At the same time, several studies have identified population-specific risk variants[7, 262]. Such loci not only add to our understanding of the genetic architecture of complex traits, but might also reveal gene-gene or gene-environment interactions.

Allele frequencies and linkage disequilibrium vary between different populations, which can be harnessed for fine-mapping of association signals and to pinpoint causal variants[381]. For example, trait-associated loci discovered in European samples might be followed up in individuals of African ancestry, who exhibit much shorter LD blocks[28].

Several complex disorders have markedly different prevalence estimates between different ethnic groups[382], which are not entirely due to environmental factors[383]. Consequently, it might be easier to collect large numbers of affected individuals by selecting an ancestral group where the disease or trait of interest is observed at a higher rate. Furthermore, contrasting phenotypic profiles of different ethnic groups could help to disentangle causal chains and reduce the risk of collider bias in MR study designs. Perhaps most importantly, a more complete understanding of ancestry-specific genetic effects will inform risk prediction and clinical management[384].

5.6 Methodological challenges

5.6.1 A case for quality over quantity

Method development is a necessary aspect of modern genetic research. However, one aspect that in my opinion has not garnered enough attention is the maintenance and evaluation (both in terms of statistics and performance) of existing methods. So far, there have been few comparison studies of multivariate and/or univariate methods involving simulations and/or real data to evaluate power and type I error rates under different scenarios[67, 69-71, 385]. For univariate GWAS, which is now an established tool in genetic research, there are a handful of widely used software implementations of the most common statistical models (linear/logistic regression, mixed models) that are being updated and maintained regularly[199, 386, 387]. This is not (yet) the case for multi-variate GWAS, both because they are a comparatively new field and because they require the consideration of additional factors, such as phenotype covariance. Despite this, there are already a number of software implementations with flexible parameter settings, efficient runtimes, and good scalability[75, 101, 125]. It would perhaps benefit the scientific community if more focus was put on further developing and properly maintaining those.

5.6.2 Replication and meta-analysis

One caveat of multivariate methods is that they do not give a single effect estimate per variant, hampering effect size-based meta-analysis across studies. To overcome this, Shen et al. suggested to transform all analysed traits into one phenotypic score based on which a beta coefficient of association with genotype can then be calculated[388]. A dedicated method for multivariate meta-analysis was recently developed that extends the classical random effects meta-analysis to incorporate a vector of beta estimates and corresponding covariance matrix[116]. Finally, Ried et al. showed that deriving average principal components across multiple studies adequately models multiple anthropometric measurements[124]. While this approach is quite elegant, it requires all participating studies to share the PCA results from their data. This is only practical in the case of multi-centre studies or large consortia, where the pooling of data is planned from the early stages of a project. However, association analyses are often conducted in one dataset and replication is then sought independently through collaborators. In this case, it would be necessary to have summary statistics that can be combined across studies without prior data harmonisation.

5.6.3 Data harmonisation and sharing

As sample sizes and number of phenotypes grow, so does the need for digital storage and computational speed. Analyses are increasingly carried out in collaborative efforts spanning different institutions and geographic region. Options for fast and secure data access and sharing across analysis groups is therefore another important consideration. Some research groups and consortia have already embraced cloud computing and are leading by example, showcasing infrastructures for distributing and analysing large-scale datasets[389].

With the increased sharing and combining of datasets derived from different cohorts, scrutiny will also need to be applied to how different traits were measured[363]. Technological advances have led to a shift from array-based to sequencing methods for genotyping, gene expression, and epigenomic assays. For protein expression studies, combining data from different assays can also introduce considerable noise due to varying specificity/sensitivity. Similar problems also exist for microbiome analyses. These disparities between datasets are likely to grow due to the speed with which assaying technologies are improved and developed. The use of imputation algorithms to recapitulate unmeasured datapoints, as well as correction for methodological noise through non-genetic principal components[390] can help to alleviate these problems. Prospective data harmonisation through centralised assay and analysis pipelines may also be put in place[363].

5.7 Concluding remarks

Not too long ago, the central dogma of molecular biology, “DNA makes RNA makes protein”, painted a fairly linear picture of genetic causality. Along this line of thinking, a theory formed that the genome could be partitioned into relevant regions comprised of protein-coding genes and close-by regulatory elements, and largely irrelevant regions of “junk DNA”, artefacts of our evolutionary journey. This idea had to be revised considerably with the finding that so-called junk DNA contained an abundance of functional elements with important consequences for our phenotypic makeup. Just as the definition of what constitutes a gene has changed with our advanced understanding of molecular mechanisms, our definitions of what constitutes disease are likely to shift. Current diagnostic criteria for some disorders are inadequate, and early detection as well as effective treatment options limited for many others. In light of the growing body of work on multi-trait genetics, the question of how to leverage information on pleiotropy for clinical use arises.

As we uncover more and more connections across different levels of molecular function, the hope is that (both research and) medicine will slowly move away from binary nosologic categories, and towards a more comprehensive understanding of human health.

Appendices

Appendix A. *Replication of 143 variants identified in the OA-BMD analyses. EA=effect allele; NEA=non-effect allele, Meta-analysis 1=deCODE and UKBB; Meta-analysis 2=arcOGEN, deCODE and UKBB*

Variant charactersitics						arcOGEN			UKBB			deCODE			Meta-analysis 1			Meta-analysis 2		
SNP	Gene	Dist.	chr:pos	EA	NEA	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P
rs494157	GRIK3	13	1:37512995	A	G	-0.03	0.02	1.32E-01	0.02	0.02	2.96E-01	0.00	0.02	9.21E-01	-0.01	0.01	5.03E-01	-0.0029	0.012	8.05E-01
rs2282231	MACF1	0	1:39569571	T	C	0.04	0.03	1.67E-01	0.04	0.02	7.06E-02	0.02	0.02	3.71E-01	-0.03	0.02	5.63E-02	0.0328	0.014	1.87E-02
rs1726486	PPIEL	0	1:40022777	G	C	0.04	0.03	1.02E-01	0.03	0.02	1.53E-01	0.03	0.02	1.56E-01	0.03	0.02	4.56E-02	0.0358	0.014	1.02E-02
rs1208602	RNA5SP45	30	1:41903038	T	C	0.06	0.02	4.57E-03	-0.01	0.02	6.30E-01	-0.03	0.02	2.17E-01	0.02	0.01	2.03E-01	0.0052	0.012	6.62E-01
rs7545984	RNF220	0	1:45003893	C	T	0.22	0.06	7.22E-04	-0.03	0.06	6.74E-01	0.13	0.07	5.05E-02	0.04	0.05	3.80E-01	0.0999	0.037	7.44E-03
rs7554123	RNF220	0	1:45004204	T	G	0.23	0.06	3.67E-04	-0.02	0.06	6.93E-01	0.13	0.07	4.15E-02	-0.04	0.05	3.52E-01	0.1061	0.037	4.49E-03
rs3131780	DAB1	0	1:58170218	G	T	0.08	0.02	6.70E-04	0.03	0.02	2.17E-01	-0.01	0.02	6.65E-01	0.01	0.01	5.83E-01	0.0283	0.012	2.10E-02
rs2052959	WLS	0	1:68609491	G	T	0.02	0.02	4.49E-01	-0.02	0.02	2.79E-01	-0.02	0.02	4.46E-01	-0.02	0.01	1.87E-01	-0.0087	0.012	4.83E-01
rs1212998	RP11-76N22.1	4	1:88960005	T	C	-0.09	0.03	1.25E-02	0.00	0.03	8.76E-01	-0.17	0.17	5.15E-01	-0.01	0.03	8.76E-01	-0.037	0.024	1.16E-01
rs967554	RP5-936J12.1	165	1:10306417	G	A	-0.05	0.03	4.25E-02	-0.07	0.02	2.46E-03	-0.04	0.02	4.03E-02	-0.06	0.02	2.67E-04	-0.0546	0.013	3.07E-05
rs1903787	COL11A1	0	1:10344449	G	A	-0.07	0.02	4.96E-03	-0.05	0.02	2.45E-02	-0.05	0.02	4.11E-02	-0.05	0.02	1.68E-03	-0.0533	0.013	3.23E-05
rs1741294	KRT8P45	7	1:15705180	T	C	0.08	0.03	1.28E-02	0.02	0.03	5.30E-01	0.06	0.03	5.47E-02	-0.04	0.02	8.34E-02	0.0487	0.017	4.77E-03
rs2494454	RP11-416K24.2	7	1:18231238	C	T	0.05	0.02	2.30E-02	0.04	0.02	6.40E-02	0.00	0.02	9.19E-01	0.02	0.01	1.70E-01	0.0277	0.012	1.68E-02
rs1734343	PDIA6	0	2:10937037	G	T	-0.05	0.02	1.53E-02	0.01	0.02	5.96E-01	0.00	0.02	9.00E-01	0.00	0.02	7.45E-01	-0.014	0.012	2.54E-01
rs7586601	AC074117.10	0	2:27584666	A	G	-0.06	0.02	6.99E-03	-0.03	0.02	1.74E-01	-0.03	0.02	8.07E-02	0.03	0.01	2.91E-02	-0.0387	0.012	9.80E-04
rs1019603	MRPL33	23	2:27971195	T	C	0.06	0.02	4.67E-03	0.00	0.02	9.54E-01	-0.01	0.02	5.00E-01	0.01	0.01	6.49E-01	0.0131	0.012	2.56E-01
rs1247413	AC093166.3	0	2:11250010	C	T	-0.06	0.02	5.36E-03	0.02	0.02	3.23E-01	-0.04	0.02	8.29E-02	-0.01	0.01	5.24E-01	-0.0245	0.012	4.37E-02
rs4848209	ANAPC1	0	2:11262663	G	A	0.07	0.02	2.88E-03	-0.02	0.02	3.87E-01	0.03	0.02	1.04E-01	0.01	0.01	5.54E-01	0.0257	0.012	3.50E-02
rs1017046	AC093901.1	112	2:11905623	G	T	-0.07	0.02	2.09E-03	0.07	0.02	1.46E-03	0.01	0.02	5.93E-01	0.04	0.01	1.04E-02	0.0065	0.012	5.97E-01
rs1049657	AC018737.3	165	2:12271895	A	G	-0.07	0.02	2.19E-03	0.01	0.02	6.26E-01	0.01	0.02	7.31E-01	-0.01	0.02	5.54E-01	-0.0158	0.013	2.30E-01
rs3116152	DIS3L2	0	2:23295014	C	G	0.06	0.02	4.13E-03	0.01	0.02	7.43E-01	0.03	0.02	1.04E-01	-0.02	0.01	1.39E-01	0.033	0.012	5.33E-03
rs1171405	ITPR1	0	3:4783407	C	T	0.06	0.02	1.01E-02	-0.05	0.02	1.33E-02	0.01	0.02	7.10E-01	-0.03	0.02	1.18E-01	0.0021	0.013	8.74E-01
rs7619689	TRANK1	0	3:36981363	C	A	0.07	0.02	7.92E-04	-0.01	0.02	7.76E-01	-0.03	0.02	1.22E-01	-0.02	0.01	1.97E-01	0.0089	0.012	4.52E-01
rs336601	LRRFIP2	0	3:37184473	A	G	-0.07	0.02	4.29E-03	0.01	0.02	7.66E-01	0.03	0.02	2.18E-01	-0.02	0.02	2.84E-01	-0.0085	0.013	5.15E-01
rs7632108	RP11-259K5.1	0	3:37258063	A	C	-0.07	0.02	1.44E-03	0.01	0.02	5.77E-01	0.02	0.02	2.20E-01	-0.02	0.01	2.15E-01	-0.0087	0.012	4.59E-01
rs2669904	ZBTB20	0	3:11436869	G	A	0.07	0.03	1.62E-02	0.00	0.03	9.17E-01	-0.01	0.02	6.33E-01	-0.01	0.02	7.70E-01	0.0159	0.015	3.05E-01
rs2877561	ILDR1	0	3:12171205	A	C	-0.06	0.02	1.88E-02	-0.03	0.02	1.22E-01	0.03	0.02	2.21E-01	0.00	0.02	8.26E-01	-0.0192	0.013	1.44E-01
rs4686872	RTP4	25	3:18706066	C	T	-0.07	0.03	1.97E-02	0.05	0.03	8.39E-02	0.01	0.03	7.59E-01	0.03	0.02	1.55E-01	-0.0008	0.016	9.57E-01
rs3755955	IDUA	0	4:994414	A	G	-0.02	0.03	5.89E-01	-0.02	0.03	3.67E-01	0.01	0.03	7.82E-01	0.01	0.02	6.57E-01	-0.011	0.016	5.04E-01
rs6826705	RP11-20I20.2	0	4:1122634	G	A	0.05	0.02	2.34E-02	0.02	0.02	2.98E-01	-0.01	0.02	5.19E-01	0.00	0.02	7.78E-01	0.0186	0.013	1.41E-01
rs3755920	CTBP1	0	4:1243617	C	T	0.06	0.02	7.97E-03	-0.01	0.02	6.85E-01	0.01	0.02	7.76E-01	0.00	0.01	9.19E-01	0.0165	0.012	1.64E-01
rs9291326	OCIAD1	0	4:48847859	A	G	-0.04	0.02	4.73E-02	-0.01	0.02	5.85E-01	-0.02	0.02	2.83E-01	0.02	0.01	2.35E-01	-0.0237	0.011	3.85E-02
rs465705	CTD-2029E14.1	52	5:3126445	A	G	-0.09	0.04	1.02E-02	0.01	0.03	6.58E-01	-0.02	0.03	4.34E-01	0.01	0.02	7.30E-01	-0.0284	0.018	1.13E-01
rs2398218	CTD-2324F15.2	0	5:6325532	C	T	0.10	0.04	5.70E-03	-0.06	0.03	8.36E-02	0.06	0.03	5.72E-02	0.00	0.02	9.50E-01	0.0319	0.02	1.17E-01
rs316405	ZNF131	0	5:43070866	C	T	0.07	0.02	1.57E-03	-0.02	0.02	2.30E-01	-0.01	0.02	5.80E-01	-0.02	0.01	2.11E-01	0.0075	0.012	5.19E-01
rs7735277	ITGA2	0	5:52318916	T	C	0.07	0.03	2.91E-03	0.02	0.02	3.35E-01	0.03	0.02	2.12E-01	-0.03	0.02	1.19E-01	0.0413	0.014	3.14E-03
rs1690325	LINC00461	0	5:87871578	A	C	0.12	0.05	1.90E-02	0.03	0.05	5.18E-01	0.05	0.04	2.18E-01	-0.04	0.03	1.74E-01	0.0602	0.026	1.87E-02
rs6894139	MEF2C-AS1	0	5:88327782	T	G	0.03	0.02	1.60E-01	0.02	0.02	2.72E-01	0.00	0.02	9.26E-01	-0.01	0.01	3.86E-01	0.0178	0.012	1.33E-01
rs1283614	MEF2C-AS1	0	5:88361042	T	C	0.12	0.04	3.77E-03	0.00	0.04	9.36E-01	-0.03	0.05	5.77E-01	0.01	0.03	7.54E-01	0.0371	0.025	1.45E-01
rs1524355	RNA5SP189	24	5:10528275	A	G	-0.08	0.02	1.69E-03	0.03	0.02	1.29E-01	-0.01	0.02	6.02E-01	-0.01	0.02	4.91E-01	-0.015	0.013	2.63E-01

Continued on next page

Variant characteristics						arcOGEN			UKBB			deCODE			Meta-analysis 1			Meta-analysis 2		
SNP	Gene	Dist.	chr:pos	EA	NEA	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P
rs7965	MRPL22	0	5:154346325	A	G	0.11	0.04	3.77E-03	-0.04	0.03	2.39E-01	0.04	0.03	1.85E-01	0.00	0.02	8.93E-01	0.0271	0.018	1.38E-01
rs2431613	CTB-181F24.1	20	5:164627303	G	A	-0.14	0.05	1.98E-03	0.02	0.04	6.44E-01	0.03	0.04	4.57E-01	0.03	0.03	3.93E-01	-0.0258	0.026	3.11E-01
rs630765	RP1-151F17.2	5	6:16771615	G	A	-0.26	0.07	5.46E-04	-0.08	0.07	2.43E-01	0.07	0.07	3.38E-01	-0.01	0.05	8.75E-01	-0.0854	0.041	3.89E-02
rs9466056	RP1-135L22.1	29	6:21384613	A	G	-0.06	0.02	6.29E-03	0.03	0.02	1.73E-01	-0.02	0.02	2.53E-01	0.00	0.01	8.88E-01	-0.0165	0.012	1.69E-01
rs9350354	RP1-135L22.1	38	6:21393645	G	A	-0.07	0.02	1.56E-03	0.05	0.02	8.21E-03	-0.04	0.02	3.45E-02	0.00	0.01	8.45E-01	-0.018	0.012	1.28E-01
rs7738847	CASC15	0	6:21814752	T	C	0.03	0.02	2.31E-01	0.00	0.02	8.93E-01	0.00	0.02	9.15E-01	0.00	0.01	9.92E-01	0.0074	0.012	5.25E-01
rs915894	NOTCH4	0	6:32190390	G	T	0.03	0.02	1.26E-01	0.03	0.02	2.07E-01	-0.04	0.02	1.00E-01	0.00	0.01	7.51E-01	0.0067	0.012	5.79E-01
rs10948155	SUPT3H	89	6:44687957	C	T	0.10	0.02	1.30E-05	0.04	0.02	4.84E-02	0.04	0.02	6.10E-02	0.04	0.01	7.08E-03	0.0573	0.012	3.24E-06
rs12190136	SUPT3H	10	6:44767072	G	A	-0.07	0.02	1.23E-03	-0.03	0.02	1.62E-01	-0.03	0.02	8.73E-02	-0.03	0.01	2.96E-02	-0.0419	0.012	3.33E-04
rs12212190	SUPT3H	0	6:44926271	C	T	0.10	0.02	3.90E-05	0.04	0.02	4.33E-02	0.07	0.02	1.62E-03	0.06	0.02	4.89E-04	0.0681	0.013	1.98E-07
rs1329710	SUPT3H	0	6:44969125	T	C	0.07	0.02	6.64E-04	0.02	0.02	2.20E-01	0.04	0.02	2.64E-02	-0.03	0.01	1.82E-02	0.0456	0.012	1.14E-04
rs9381373	SUPT3H	0	6:45175908	T	C	0.07	0.02	8.29E-04	0.03	0.02	1.63E-01	0.04	0.02	2.51E-02	-0.04	0.01	1.14E-02	0.0464	0.012	7.56E-05
rs11964690	SUPT3H	0	6:45258068	C	T	0.09	0.02	4.14E-05	0.04	0.02	3.27E-02	0.07	0.02	1.21E-03	0.06	0.02	2.69E-04	0.0682	0.013	1.03E-07
rs17053584	SGK1	0	6:134622514	C	G	-0.21	0.07	3.71E-03	0.02	0.07	7.77E-01	0.02	0.06	7.12E-01	-0.02	0.04	6.35E-01	-0.0411	0.038	2.75E-01
rs4896622	AIG1	0	6:143488835	G	A	0.06	0.03	1.20E-02	0.00	0.02	9.07E-01	0.03	0.02	1.81E-01	0.02	0.02	3.16E-01	0.0315	0.014	2.65E-02
rs9384514	IYD	24	6:150751032	G	T	0.10	0.03	1.16E-04	0.04	0.02	9.89E-02	-0.01	0.02	7.66E-01	0.02	0.02	3.53E-01	0.0398	0.014	4.27E-03
rs4869742	CCDC170	0	6:151907748	C	T	0.00	0.02	9.86E-01	0.00	0.02	9.30E-01	0.03	0.02	1.26E-01	0.02	0.02	2.27E-01	0.0131	0.013	3.06E-01
rs10943837	RP11-801I18.1	9	6:NA	T	C	0.04	0.02	9.01E-02	NA	NA	NA	-0.02	0.02	4.49E-01	0.02	0.02	4.29E-01	0.0084	0.015	5.78E-01
rs11773376	BRAT1	0	7:2594338	T	C	-0.05	0.03	9.97E-02	0.01	0.03	7.38E-01	-0.02	0.03	3.71E-01	0.01	0.02	6.62E-01	-0.0189	0.015	2.22E-01
rs7788807	FOXK1	0	7:4768038	C	T	0.13	0.05	6.61E-03	-0.04	0.04	3.36E-01	-0.11	0.04	1.45E-02	-0.08	0.03	5.55E-03	-0.0231	0.024	3.41E-01
rs10282661	C7orf10	0	7:40685811	T	C	0.03	0.02	1.25E-01	0.00	0.02	8.15E-01	0.02	0.02	4.41E-01	-0.01	0.02	4.66E-01	0.0183	0.013	1.45E-01
rs11976696	EGFR	0	7:55232333	G	A	-0.06	0.03	2.93E-02	0.01	0.02	6.01E-01	-0.03	0.02	2.81E-01	-0.01	0.02	6.48E-01	-0.0213	0.014	1.22E-01
rs17158899	SEMA3A	0	7:83967998	T	C	0.19	0.05	2.46E-05	0.02	0.04	6.13E-01	0.02	0.04	6.94E-01	-0.02	0.03	5.20E-01	0.065	0.024	5.87E-03
rs1524928	SHFM1	18	7:96356794	T	C	-0.05	0.02	3.69E-02	-0.01	0.02	5.96E-01	0.00	0.02	7.97E-01	0.00	0.01	8.63E-01	-0.0146	0.012	2.06E-01
rs2395962	NRCAM	0	7:108023548	G	A	0.08	0.02	4.85E-04	-0.03	0.02	1.38E-01	0.02	0.02	2.70E-01	0.00	0.02	7.99E-01	0.0228	0.013	8.76E-02
rs9640740	IFRD1	0	7:112064376	A	G	0.05	0.02	2.25E-02	-0.02	0.02	3.77E-01	0.02	0.02	3.52E-01	0.00	0.02	9.74E-01	0.0161	0.013	2.08E-01
rs10227242	RP11-328J2.1	13	7:119332021	T	C	-0.15	0.06	1.21E-02	-0.13	0.06	1.45E-02	-0.01	0.07	9.47E-01	0.08	0.04	5.38E-02	-0.1063	0.035	2.39E-03
rs2457400	RP11-26J3.1	0	8:80733475	C	T	-0.04	0.02	8.09E-02	-0.01	0.02	7.55E-01	-0.01	0.02	8.00E-01	-0.01	0.02	6.85E-01	-0.0175	0.014	1.93E-01
rs10808475	TRPS1	0	8:116601071	C	T	0.06	0.02	2.63E-03	0.03	0.02	1.03E-01	-0.02	0.02	3.63E-01	0.01	0.01	6.37E-01	0.0235	0.012	4.33E-02
rs7022051	WNK2	0	9:96002983	T	C	-0.08	0.03	1.49E-02	0.02	0.03	5.76E-01	0.02	0.03	5.14E-01	-0.02	0.02	3.88E-01	-0.0107	0.018	5.60E-01
rs10512249	PTCH1	0	9:98256309	A	G	-0.07	0.04	6.25E-02	-0.05	0.03	1.78E-01	-0.08	0.03	1.06E-02	0.06	0.02	2.10E-03	-0.065	0.018	3.25E-04
rs11012679	CACNB2	41	10:18388401	C	A	0.25	0.12	3.31E-02	0.02	0.10	8.78E-01	-0.19	0.07	4.61E-02	-0.12	0.06	3.77E-02	-0.0466	0.052	3.72E-01
rs2246047	RP11-490O24.2	19	10:34178920	T	G	0.08	0.03	1.17E-03	0.05	0.02	3.02E-02	-0.01	0.02	6.93E-01	-0.02	0.02	2.00E-01	0.0361	0.013	5.44E-03
rs734949	KCNMA1	0	10:79387112	C	T	0.06	0.02	2.83E-03	0.01	0.02	4.66E-01	0.02	0.02	2.57E-01	0.02	0.01	1.76E-01	0.0315	0.011	5.99E-03
rs7071206	KCNMA1	3	10:79401316	C	T	0.08	0.03	3.22E-03	0.02	0.02	3.32E-01	0.03	0.03	2.82E-01	0.03	0.02	1.47E-01	0.0412	0.015	4.53E-03

Continued on next page

Variant characteristics						arcOGEN			UKBB			deCODE			Meta-analysis 1			Meta-analysis 2		
SNP	Gene	Dist.	chr:pos	EA	NEA	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P
rs716255	KCNMA1	21	10:79419679	T	A	0.09	0.03	1.26E-03	0.03	0.03	2.78E-01	0.04	0.03	1.72E-01	0.03	0.02	8.72E-02	0.05	0.015	1.30E-03
rs1225421	KCNMA1	45	10:79443520	T	C	0.09	0.03	8.49E-04	0.02	0.03	3.93E-01	0.02	0.03	3.90E-01	-0.02	0.02	2.22E-01	0.04	0.015	4.67E-03
rs7894701	DLG5-AS1	0	10:79688069	G	A	0.13	0.05	3.38E-03	-0.15	0.04	7.32E-04	0.00	0.04	9.57E-01	-0.06	0.03	3.94E-02	-0.01	0.024	8.24E-01
rs946328	TLX1	0	10:102893345	T	C	-0.09	0.02	4.64E-04	0.02	0.02	3.36E-01	0.07	0.03	6.59E-03	-0.04	0.02	1.50E-02	0.00	0.014	9.69E-01
rs884672	RP11-127O4.2	33	10:106342559	T	C	-0.08	0.04	3.66E-02	-0.09	0.04	2.41E-02	0.00	0.04	9.20E-01	0.04	0.03	1.31E-01	-0.05	0.023	1.57E-02
rs4536164	LGR4	0	11:27467109	A	C	0.07	0.02	1.98E-03	0.02	0.02	2.81E-01	0.03	0.02	1.44E-01	-0.03	0.02	7.63E-02	0.04	0.013	1.46E-03
rs1787663	SCYL1	0	11:65294830	A	C	-0.06	0.02	7.77E-03	-0.07	0.02	8.78E-04	-0.05	0.02	4.28E-02	0.06	0.02	9.89E-05	-0.06	0.013	2.42E-06
rs1201342	C11orf80	0	11:66608521	G	C	-0.08	0.03	4.18E-03	-0.06	0.02	1.47E-02	-0.04	0.02	9.11E-02	-0.05	0.02	2.73E-03	-0.06	0.014	5.39E-05
rs1182628	LRP5	0	11:68146661	C	T	-0.11	0.03	6.72E-05	0.00	0.02	8.52E-01	-0.02	0.02	4.18E-01	-0.01	0.02	6.35E-01	-0.04	0.015	1.16E-02
rs1713482	RP11-111M22.2	0	11:76103445	T	G	-0.07	0.05	1.58E-01	-0.04	0.04	2.93E-01	-0.06	0.04	1.90E-01	0.05	0.03	7.72E-02	-0.06	0.026	2.43E-02
rs2513508	C11orf30	0	11:76234528	C	T	0.04	0.02	6.29E-02	-0.03	0.02	1.01E-01	0.01	0.02	7.46E-01	-0.01	0.01	3.84E-01	0.00	0.012	8.09E-01
rs7104420	FAT3	0	11:92572940	G	A	0.07	0.02	1.69E-03	-0.01	0.02	5.47E-01	0.03	0.02	1.93E-01	0.01	0.01	6.24E-01	0.03	0.012	3.49E-02
rs1104920	RN7SKP15	54	12:28013563	G	A	0.12	0.03	1.59E-05	0.05	0.02	4.41E-02	0.03	0.03	2.34E-01	0.04	0.02	2.34E-02	0.06	0.015	2.15E-05
rs7956544	RP11-734E19.1	10	12:77140628	C	T	-0.07	0.02	2.62E-03	-0.03	0.02	1.09E-01	0.00	0.03	9.74E-01	-0.02	0.02	1.77E-01	-0.04	0.013	4.02E-03
rs9526582	ARL11	0	13:50207220	G	T	-0.04	0.02	5.81E-02	0.00	0.02	7.99E-01	0.00	0.02	7.86E-01	0.00	0.01	9.81E-01	-0.01	0.011	3.28E-01
rs978332	DLEU1	0	13:51136739	C	T	0.07	0.02	1.60E-03	0.00	0.02	9.82E-01	0.02	0.02	3.63E-01	0.01	0.02	5.14E-01	0.03	0.013	2.29E-02
rs803798	LINC00348	0	13:71604672	T	C	-0.07	0.03	6.66E-03	0.05	0.02	1.51E-02	0.01	0.02	6.51E-01	-0.03	0.02	3.65E-02	0.00	0.014	7.82E-01
rs9318939	RNU6-67P	93	13:83965649	A	G	0.06	0.04	8.04E-02	0.02	0.03	5.69E-01	-0.01	0.03	7.95E-01	0.00	0.02	8.59E-01	0.02	0.019	2.90E-01
rs9556714	MBNL2	7	13:98053846	A	G	0.29	0.10	2.66E-03	-0.10	0.09	2.18E-01	-0.05	0.07	4.72E-01	0.07	0.05	1.64E-01	0.01	0.046	8.14E-01
rs9554402	PSMA6P4	83	13:98211440	A	T	0.29	0.10	3.07E-03	-0.08	0.10	4.06E-01	-0.05	0.07	4.78E-01	0.06	0.05	2.59E-01	0.02	0.048	6.50E-01
rs1957764	STXBP6	0	14:25383572	G	A	-0.06	0.03	3.45E-02	0.02	0.03	3.98E-01	0.03	0.03	2.69E-01	0.03	0.02	1.68E-01	0.00	0.016	9.51E-01
rs1860710	PAPLN	1	14:73703112	T	C	0.06	0.04	1.04E-01	0.00	0.03	9.28E-01	0.08	0.04	2.81E-02	-0.03	0.02	1.71E-01	0.04	0.02	4.12E-02
rs1711184	RP11-299L17.3	0	14:81916033	T	C	0.10	0.05	3.54E-02	0.01	0.05	7.80E-01	-0.06	0.04	1.91E-01	0.03	0.03	3.71E-01	0.01	0.026	7.17E-01
rs1013327	TTC7B	0	14:91199626	G	A	-0.13	0.04	2.31E-03	-0.08	0.04	4.95E-02	0.00	0.04	9.01E-01	-0.04	0.03	1.97E-01	-0.06	0.023	5.92E-03
rs1286077	RPS6KA5	0	14:91446384	C	T	0.10	0.03	5.75E-04	0.06	0.03	1.74E-02	0.00	0.03	9.48E-01	0.04	0.02	6.42E-02	0.06	0.016	4.95E-04
rs1286063	RPS6KA5	0	14:91452841	C	T	0.09	0.03	8.78E-04	0.06	0.03	1.67E-02	0.00	0.04	9.56E-01	0.04	0.02	4.79E-02	0.06	0.017	3.55E-04
rs1286147	RPS6KA5	0	14:91467567	C	A	0.09	0.03	1.23E-03	0.06	0.03	2.05E-02	-0.04	0.02	9.62E-01	0.02	0.02	1.37E-01	0.04	0.014	3.77E-03
rs7174138	GABRG3	0	15:27703715	G	A	-0.08	0.03	4.37E-03	0.02	0.03	4.90E-01	0.00	0.02	9.02E-01	0.01	0.02	7.05E-01	-0.02	0.015	2.44E-01
rs3087970	RP11-489D6.2	0	15:33595935	G	C	-0.07	0.03	3.53E-02	-0.03	0.03	2.64E-01	-0.04	0.03	2.49E-01	-0.03	0.02	9.85E-02	-0.04	0.017	1.23E-02
rs714784	AP4E1	0	15:51296407	T	C	-0.06	0.02	4.88E-03	0.03	0.02	1.47E-01	0.00	0.02	9.54E-01	-0.01	0.01	3.73E-01	-0.01	0.011	5.06E-01
rs8039089	CYP19A1	0	15:51561328	G	T	-0.04	0.02	1.02E-01	0.02	0.02	3.23E-01	-0.04	0.02	3.30E-02	-0.01	0.01	3.41E-01	-0.02	0.012	9.27E-02
rs1051870	SMAD3	0	15:67365622	G	A	-0.07	0.02	6.34E-04	-0.04	0.02	5.97E-02	-0.08	0.02	7.03E-05	-0.06	0.01	7.90E-06	-0.06	0.011	2.15E-08
rs1290107	SMAD3	0	15:67370389	G	A	-0.08	0.02	2.92E-04	-0.05	0.02	2.28E-02	-0.10	0.02	1.39E-05	-0.07	0.01	2.46E-07	-0.08	0.012	3.12E-10
rs8031440	SMAD3	0	15:67483979	A	G	-0.07	0.03	4.01E-03	-0.01	0.02	5.26E-01	-0.01	0.02	7.36E-01	0.01	0.02	4.88E-01	-0.03	0.014	3.25E-02

Continued on next page

Variant characteristics						arcOGEN			UKBB			deCODE			Meta-analysis 1			Meta-analysis 2		
SNP	Gene	Dist.	chr:pos	EA	NEA	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P	BETA	SE	P
rs8032675	MAP2K5	0	15:67959464	T	C	-0.06	0.02	8.03E-03	-0.02	0.02	3.69E-01	-0.01	0.02	6.56E-01	0.01	0.01	3.40E-01	-0.0266	0.012	2.52E-02
rs11645645	CLCN7	0	16:1504149	C	T	-0.02	0.02	3.23E-01	-0.02	0.02	2.67E-01	-0.04	0.02	4.45E-02	-0.03	0.01	2.01E-02	-0.0285	0.011	1.27E-02
rs4985155	PDXDC1	0	16:15129459	G	A	0.02	0.02	4.58E-01	0.01	0.02	5.79E-01	0.00	0.02	8.11E-01	0.01	0.01	5.74E-01	0.0107	0.012	3.81E-01
rs881803	MYH11	0	16:15802334	T	C	0.05	0.02	4.20E-02	-0.01	0.02	6.05E-01	0.00	0.02	8.36E-01	0.01	0.01	6.13E-01	0.0075	0.012	5.22E-01
rs9935327	MYH11	0	16:15813766	A	C	-0.17	0.08	3.16E-02	0.05	0.08	5.48E-01	-0.10	0.07	2.21E-01	0.04	0.05	4.55E-01	-0.0778	0.044	7.43E-02
rs7342689	RBBP6	17	16:24601666	T	G	0.09	0.03	2.74E-03	-0.02	0.03	4.24E-01	0.00	0.04	9.34E-01	0.01	0.02	4.96E-01	0.0212	0.018	2.31E-01
rs10852527	HNRNPA1P4	121	16:51558328	A	G	-0.06	0.02	9.87E-03	0.01	0.02	5.73E-01	-0.01	0.02	6.10E-01	0.00	0.01	9.51E-01	-0.0168	0.012	1.70E-01
rs739414	RP11-	2	16:73097956	C	T	0.07	0.03	4.18E-03	0.04	0.02	9.34E-02	0.01	0.02	5.47E-01	0.02	0.02	1.10E-01	0.0377	0.013	4.33E-03
rs11654856	WSCD1	0	17:5997476	A	G	-0.05	0.03	6.31E-02	-0.03	0.02	2.31E-01	-0.03	0.02	1.55E-01	0.03	0.02	5.61E-02	-0.036	0.014	9.44E-03
rs7217473	CCL15-CCL14	0	17:34319064	T	C	0.15	0.05	1.22E-03	0.06	0.04	1.66E-01	0.00	0.05	9.17E-01	-0.03	0.03	2.74E-01	0.0704	0.026	6.64E-03
rs4793022	RP11-	1	17:41798855	G	C	-0.05	0.02	2.25E-02	-0.03	0.02	8.84E-02	0.01	0.02	5.43E-01	-0.01	0.01	4.35E-01	-0.0233	0.012	5.57E-02
rs17762165	CRHR1	0	17:43778602	T	C	0.01	0.03	6.16E-01	0.05	0.02	3.43E-02	0.00	0.03	8.53E-01	-0.03	0.02	7.64E-02	0.0244	0.014	7.91E-02
rs4792891	MAPT	0	17:43973498	G	T	-0.03	0.02	2.61E-01	0.02	0.02	2.40E-01	0.00	0.02	9.59E-01	0.01	0.01	4.05E-01	0.0013	0.012	9.11E-01
rs11655490	SLC39A11	25	17:71113564	A	G	0.08	0.03	1.12E-03	0.00	0.02	8.49E-01	0.00	0.02	8.67E-01	0.00	0.02	9.88E-01	0.0245	0.014	7.73E-02
rs12150506	MAPT	0	17:NA	A	G	0.02	0.03	5.39E-01	NA	NA	NA	0.01	0.02	8.05E-01	-0.01	0.02	8.03E-01	0.0105	0.017	5.46E-01
rs11665347	APCDD1	0	18:10478812	A	G	-0.10	0.02	8.22E-06	-0.02	0.02	4.54E-01	-0.03	0.02	1.21E-01	0.03	0.02	8.77E-02	-0.0486	0.013	1.12E-04
rs16974630	APCDD1	9	18:10499123	T	C	-0.09	0.02	1.03E-04	-0.02	0.02	3.13E-01	-0.01	0.02	8.01E-01	0.01	0.01	3.89E-01	-0.0342	0.012	5.56E-03
rs11659747	RP11-19F9.1	457	18:35710345	T	C	0.05	0.03	1.53E-01	0.04	0.03	1.94E-01	0.02	0.03	5.03E-01	-0.03	0.02	1.71E-01	0.0341	0.018	5.51E-02
rs17069898	TNFRSF11A	0	18:60029281	A	G	-0.07	0.02	1.46E-03	-0.04	0.02	2.49E-02	0.00	0.04	9.81E-01	0.04	0.02	4.24E-02	-0.0495	0.014	3.34E-04
rs9945428	FBXO15	0	18:71788676	C	A	-0.05	0.02	4.69E-02	0.02	0.02	2.82E-01	-0.01	0.02	7.65E-01	0.01	0.02	5.68E-01	-0.0083	0.013	5.29E-01
rs2238686	FOSB	0	19:45973056	C	T	0.10	0.03	3.71E-03	0.00	0.03	9.86E-01	-0.01	0.03	6.79E-01	-0.01	0.02	7.42E-01	0.0223	0.018	2.05E-01
rs11881883	FBXO46	0	19:46221726	A	G	0.14	0.05	9.73E-03	-0.01	0.05	8.29E-01	-0.01	0.05	8.20E-01	0.01	0.04	7.46E-01	0.0359	0.03	2.34E-01
rs16995654	PLCB4	0	20:9191578	T	C	-0.13	0.05	7.07E-03	-0.08	0.05	1.00E-01	-0.01	0.05	8.59E-01	0.04	0.03	1.91E-01	-0.0713	0.027	8.92E-03
rs6048946	CST3	6	20:23602307	T	G	-0.06	0.03	1.39E-02	0.01	0.02	7.35E-01	-0.01	0.02	7.39E-01	0.00	0.02	9.97E-01	-0.0186	0.014	1.83E-01
rs6060300	EDEM2	0	20:33780970	C	T	0.03	0.03	3.23E-01	-0.03	0.03	2.65E-01	-0.01	0.03	7.35E-01	-0.02	0.02	2.99E-01	-0.005	0.015	7.42E-01
rs6094511	EYA2	0	20:45559727	A	G	0.03	0.03	3.42E-01	0.01	0.03	6.43E-01	0.03	0.03	4.13E-01	-0.02	0.02	3.65E-01	0.0235	0.018	2.01E-01
rs4925370	OSBPL2	0	20:60867591	G	A	0.06	0.02	7.54E-03	0.03	0.02	1.74E-01	0.04	0.02	5.95E-02	0.03	0.01	2.24E-02	0.0395	0.012	7.37E-04
rs2838665	TSPEAR	17	21:46148874	A	G	-0.06	0.02	1.16E-02	0.00	0.02	9.70E-01	0.00	0.03	8.82E-01	0.00	0.02	9.42E-01	-0.021	0.014	1.33E-01
rs5993790	AC000067.1	14	22:19668888	A	G	-0.10	0.03	2.57E-03	0.01	0.03	8.48E-01	0.00	0.03	9.70E-01	0.00	0.02	8.81E-01	-0.024	0.017	1.63E-01
rs767919	TTC28	0	22:28490909	G	A	0.07	0.05	1.24E-01	0.01	0.04	8.00E-01	-0.02	0.05	7.43E-01	0.00	0.03	9.72E-01	0.0229	0.027	3.91E-01
rs5757762	CACNA1I	0	22:40064581	A	G	-0.08	0.02	3.32E-04	0.01	0.02	6.91E-01	-0.01	0.02	6.45E-01	0.00	0.01	9.44E-01	-0.0228	0.012	4.71E-02
rs714031	CACNA1I	0	22:40070234	T	C	0.08	0.02	2.01E-04	-0.01	0.02	6.07E-01	0.00	0.02	8.56E-01	0.01	0.01	6.14E-01	0.02	0.012	9.41E-02
rs5766632	UPK3A	5	22:45696695	G	A	0.10	0.03	4.46E-05	0.03	0.02	1.27E-01	0.08	0.02	4.71E-04	0.06	0.02	1.26E-04	0.0716	0.013	6.52E-08
rs9616477	RPL35P8	41	22:49569291	T	C	0.06	0.03	6.75E-02	0.00	0.03	9.81E-01	-0.03	0.03	2.69E-01	0.02	0.02	4.07E-01	0.0046	0.018	7.94E-01

Appendix B. *Established risk variants for SCZ used for genetic risk score analysis. Shown are 125 autosomal variants associated with schizophrenia in Ref. 11 that were used to construct genetic risk scores in GOMAP. For each variant, chromosome position, effect (EA) and alternative (NEA) allele, as well as odds ratios with 95% confidence interval and p-values are given.*

Variant	Chr	Pos (hg18)	EA	NEA	OR (95% CI)	P
rs115329265	6	28712247	A	G	1.21 (1.18-1.25)	3.86E-32
rs11191419	10	104612335	A	T	0.91 (0.88-0.93)	9.24E-18
rs2007044	12	2344960	A	G	0.91 (0.89-0.93)	2.63E-17
rs1702294	1	98501984	T	C	0.89 (0.86-0.92)	2.79E-17
chr2_200825237_I	2	200825237	AT	A	0.91 (0.88-0.93)	1.78E-14
rs2851447	12	123665113	C	G	0.91 (0.89-0.94)	2.19E-14
chr7_2025096_I	7	2025096	A	ACT	0.92 (0.90-0.94)	6.12E-14
chr10_104957618_I	10	104957618	CA	C	0.84 (0.80-0.89)	1.04E-13
rs12887734	14	104046834	T	G	1.09 (1.07-1.11)	1.17E-13
rs4391122	5	60598543	A	G	0.92 (0.90-0.95)	1.73E-13
rs4129585	8	143312933	A	C	1.08 (1.06-1.10)	2.03E-13
rs13240464	7	110898915	T	C	1.08 (1.06-1.11)	6.16E-13
rs9636107	18	53200117	A	G	0.93 (0.91-0.95)	9.09E-13
rs35518360	4	103146890	A	T	0.87 (0.83-0.90)	9.56E-13
rs8042374	15	78908032	A	G	1.09 (1.07-1.12)	1.87E-12
rs4702	15	91426560	A	G	0.92 (0.90-0.95)	2.30E-12
rs11682175	2	57987593	T	C	0.93 (0.91-0.95)	2.54E-12
rs10791097	11	130718630	T	G	1.08 (1.06-1.10)	2.88E-12
rs6704768	2	233592501	A	G	0.93 (0.91-0.95)	3.15E-12
rs75968099	3	36858583	T	C	1.08 (1.06-1.10)	3.39E-12
rs72934570	18	53533189	T	C	0.87 (0.82-0.91)	3.67E-12
rs55661361	11	124613957	A	G	0.92 (0.90-0.95)	3.68E-12
rs12826178	12	57622371	T	G	0.85 (0.80-0.89)	5.30E-12
rs9607782	22	41587556	A	T	1.09 (1.07-1.12)	6.76E-12
rs11693094	2	185601420	T	C	0.93 (0.91-0.95)	7.13E-12
rs75059851	11	133822569	A	G	1.10 (1.07-1.12)	1.23E-11
rs6434928	2	198304577	A	G	0.93 (0.90-0.95)	1.48E-11
chr18_52749216_D	18	52749216	I2	D	1.08 (1.05-1.10)	1.75E-11
chr11_46350213_D	11	46350213	I2	D	0.90 (0.88-0.93)	1.97E-11
chr22_39987017_D	22	39987017	TA	T	0.93 (0.91-0.95)	2.20E-11
rs7893279	10	18745105	T	G	1.12 (1.09-1.15)	3.56E-11
rs2535627	3	52845105	T	C	1.07 (1.05-1.09)	3.96E-11
rs17194490	3	2547786	T	G	1.10 (1.07-1.13)	4.87E-11
rs7432375	3	136288405	A	G	0.93 (0.91-0.95)	5.27E-11
chr3_180594593_I	3	180594593	TA	T	0.91 (0.89-0.94)	5.35E-11
rs6065094	20	37453194	A	G	0.93 (0.91-0.95)	5.52E-11
rs7907645	10	104423800	T	G	1.14 (1.10-1.18)	5.82E-11
rs950169	15	84706461	T	C	0.92 (0.90-0.95)	7.62E-11
rs12704290	7	86427626	A	G	0.90 (0.87-0.93)	1.04E-10
rs36068923	8	111485761	A	G	0.92 (0.89-0.94)	1.05E-10
rs12691307	16	29939877	A	G	1.07 (1.05-1.09)	1.30E-10
rs12129573	1	73768366	A	C	1.07 (1.05-1.09)	2.35E-10
rs7405404	16	13749859	T	C	1.08 (1.06-1.11)	3.93E-10
rs2514218	11	113392994	T	C	0.93 (0.91-0.95)	4.09E-10

rs11210892	1	44100084	A	G	0.93 (0.91-0.95)	4.97E-10
rs4766428	12	110723245	T	C	1.07 (1.05-1.09)	7.09E-10
chr6_84280274_D	6	84280274	GC	G	1.07 (1.05-1.09)	8.57E-10
rs140505938	1	150031490	T	C	0.91 (0.88-0.94)	9.34E-10
rs2973155	5	152608619	T	C	0.93 (0.91-0.96)	1.02E-09
rs12903146	15	61854663	A	G	1.07 (1.05-1.09)	1.04E-09
rs4523957	17	2208899	T	G	1.07 (1.05-1.09)	1.04E-09
rs1498232	1	30433951	T	C	1.07 (1.05-1.09)	1.28E-09
rs111294930	5	152177121	A	G	1.09 (1.06-1.12)	1.31E-09
rs6002655	22	42603814	T	C	1.07 (1.05-1.09)	1.48E-09
rs2332700	14	72417326	C	G	1.08 (1.05-1.10)	1.69E-09
rs6984242	8	60700469	A	G	0.94 (0.92-0.96)	1.76E-09
rs77502336	11	123394636	C	G	1.07 (1.05-1.09)	2.01E-09
chr1_8424984_D	1	8424984	GA	G	1.07 (1.05-1.09)	2.03E-09
rs6466055	7	104929064	A	C	1.07 (1.05-1.09)	2.46E-09
rs11139497	9	84739941	A	T	1.07 (1.05-1.09)	3.09E-09
rs11027857	11	24403620	A	G	1.06 (1.04-1.09)	3.21E-09
rs2053079	19	30987423	A	G	0.93 (0.91-0.95)	3.79E-09
rs4648845	1	2387101	T	C	1.07 (1.05-1.09)	4.03E-09
rs77149735	1	243555105	A	G	1.33 (1.23-1.42)	4.40E-09
rs3849046	5	137851192	T	C	1.06 (1.04-1.09)	4.83E-09
rs2239063	12	2511831	A	C	1.07 (1.05-1.09)	5.39E-09
rs9922678	16	9946319	A	G	1.07 (1.05-1.09)	6.72E-09
rs8082590	17	17958402	A	G	0.94 (0.91-0.96)	6.84E-09
rs2905426	19	19478022	T	G	0.94 (0.92-0.96)	6.92E-09
rs3735025	7	137074844	T	C	1.07 (1.04-1.09)	7.75E-09
rs75575209	2	58138192	A	T	0.90 (0.86-0.93)	1.01E-08
rs10520163	4	170626552	T	C	1.06 (1.04-1.08)	1.02E-08
chr2_146436222_I	2	146436222	TC	T	1.08 (1.06-1.11)	1.07E-08
rs59979824	2	193848340	A	C	0.94 (0.91-0.96)	1.08E-08
rs78322266	18	53063676	T	G	1.19 (1.13-1.25)	1.10E-08
rs11685299	2	225391296	A	C	0.94 (0.92-0.96)	1.11E-08
rs1106568	4	176861301	A	G	0.93 (0.91-0.96)	1.15E-08
rs12325245	16	58681393	A	T	0.92 (0.89-0.95)	1.15E-08
rs215411	4	23423603	A	T	1.07 (1.04-1.09)	1.22E-08
rs1501357	5	45364875	T	C	0.93 (0.90-0.95)	1.24E-08
rs16867576	5	88746331	A	G	1.10 (1.07-1.13)	1.36E-08
rs2693698	14	99719219	A	G	0.94 (0.92-0.96)	1.38E-08
rs55833108	10	104741583	T	G	1.08 (1.05-1.10)	1.42E-08
rs9841616	3	181167585	A	T	0.92 (0.89-0.95)	1.65E-08
rs117074560	6	96459651	T	C	0.86 (0.80-0.91)	1.66E-08
rs10803138	1	243555219	A	G	0.93 (0.91-0.96)	1.79E-08
rs7819570	8	89588626	T	G	1.08 (1.05-1.11)	1.90E-08
rs73229090	8	27442127	A	C	0.91 (0.87-0.94)	1.95E-08
rs10043984	5	137712121	T	C	1.07 (1.05-1.09)	2.18E-08
rs12522290	5	152797656	C	G	1.09 (1.06-1.11)	2.23E-08
rs7801375	7	131567263	A	G	0.92 (0.89-0.95)	2.26E-08
rs832187	3	63833050	T	C	0.94 (0.92-0.96)	2.58E-08
chr2_149429178_D	2	149429178	AT	A	0.86 (0.80-0.91)	2.62E-08
rs10503253	8	4180844	A	C	1.07 (1.05-1.10)	2.69E-08
chr1_243881945_I	1	243881945	AT	A	1.07 (1.04-1.09)	3.11E-08
rs8044995	16	68189340	A	G	1.08 (1.05-1.11)	3.27E-08

rs6704641	2	200164252	A	G	1.08 (1.05-1.11)	3.40E-08
rs715170	18	53795514	T	C	0.94 (0.91-0.96)	3.47E-08
rs79212538	5	151993104	T	G	1.15 (1.10-1.20)	3.84E-08
rs11740474	5	153680747	A	T	0.94 (0.92-0.96)	3.94E-08
rs2068012	14	30190316	T	C	0.93 (0.91-0.96)	4.14E-08
rs2909457	2	162845855	A	G	0.94 (0.92-0.96)	4.38E-08
rs56205728	15	40567237	A	G	1.07 (1.05-1.09)	4.92E-08
rs1023500	22	42340844	T	C	1.08 (1.05-1.10)	5.04E-08
rs12148337	15	70589272	T	C	1.06 (1.04-1.08)	5.33E-08
rs4330281	3	17859366	T	C	0.94 (0.92-0.96)	5.51E-08
rs9420	11	57510294	A	G	1.06 (1.04-1.09)	6.65E-08
rs1339227	6	73155701	T	C	0.94 (0.92-0.96)	6.86E-08
rs679087	12	29917265	A	C	0.94 (0.92-0.96)	7.06E-08
rs190065944	15	78859610	A	G	1.08 (1.05-1.11)	7.22E-08
rs10860964	12	103596455	T	C	1.06 (1.04-1.08)	9.92E-08
rs4388249	5	109036066	T	C	1.07 (1.05-1.10)	1.03E-07
rs4240748	12	92246786	C	G	0.94 (0.92-0.96)	1.03E-07
rs6670165	1	177280121	T	C	1.07 (1.05-1.10)	1.16E-07
rs7267348	20	48131036	T	C	0.94 (0.91-0.96)	1.18E-07
rs3768644	2	72361505	A	G	0.91 (0.88-0.95)	1.30E-07
rs14403	1	243663893	T	C	0.93 (0.91-0.96)	1.31E-07
rs76869799	1	97834525	C	G	0.85 (0.79-0.91)	1.44E-07
rs7523273	1	207977083	A	G	1.06 (1.04-1.08)	1.61E-07
rs12421382	11	109378071	T	C	0.94 (0.92-0.96)	1.72E-07
rs324017	12	57487814	A	C	0.94 (0.92-0.96)	2.13E-07
rs56873913	19	50091199	T	G	1.07 (1.04-1.09)	2.19E-07
chr7_24747494_D	7	24747494	C	CTA	1.09 (1.06-1.13)	3.59E-07
chr5_140143664_I	5	140143664	CATTGAAAGAAA	C	1.05 (1.03-1.08)	3.60E-07
rs211829	7	110048893	T	C	1.06 (1.04-1.08)	5.47E-07

Appendix C. *Established risk variants for T2D used for genetic risk score analysis. Shown are 74 autosomal variants associated with T2D in Ref. 10 that were used to construct genetic risk scores in GOMAP. For each variant, chromosome position, effect (EA) and alternative (NEA) allele, as well as odds ratios with 95% confidence interval and p-values are given.*

Variant	**Chr**	**Pos (hg18)**	**EA**	**NEA**	**OR (95% CI)**	**P**
rs7903146	10	114758349	T	C	1.40 (1.35-1.46)	5.50E-65
rs7756992	6	20679709	G	A	1.20 (1.16-1.25)	1.30E-22
rs1111875	10	94462882	C	T	1.15 (1.11-1.18)	1.10E-15
rs10811661	9	22134094	T	C	1.18 (1.13-1.23)	1.50E-13
rs3802177	8	118185025	G	A	1.16 (1.11-1.22)	2.10E-11
rs4402960	3	185511687	T	G	1.13 (1.09-1.17)	2.70E-11
rs9936385	16	53819169	C	T	1.13 (1.09-1.18)	4.70E-11
rs849135	7	28196413	G	A	1.12 (1.08-1.16)	3.40E-10
rs1801282	3	12393125	C	G	1.16 (1.11-1.22)	5.00E-09
rs13233731	7	130437689	G	A	1.10 (1.06-1.13)	4.30E-08
rs17791513	9	81905590	A	G	1.21 (1.13-1.30)	1.00E-07
rs2261181	12	66212318	T	C	1.16 (1.10-1.23)	1.00E-07
rs12571751	10	80942631	A	G	1.09 (1.06-1.13)	1.80E-07
rs4458523	4	6289986	G	T	1.09 (1.06-1.13)	1.90E-07
rs1552224	11	72433098	A	C	1.13 (1.08-1.19)	4.90E-07
rs17168486	7	14898282	T	C	1.13 (1.08-1.18)	6.90E-07
rs516946	8	41519248	C	T	1.10 (1.06-1.15)	7.30E-07
rs10830963	11	92708710	G	C	1.11 (1.07-1.16)	7.30E-07
rs1359790	13	80717156	G	A	1.10 (1.06-1.14)	9.20E-07
rs12427353	12	121426901	G	C	1.12 (1.07-1.17)	1.00E-06
rs6878122	5	76427311	G	A	1.13 (1.07-1.18)	1.20E-06
rs10203174	2	43690030	C	T	1.15 (1.08-1.21)	1.50E-06
rs7593730	2	161171454	C	T	1.11 (1.06-1.15)	1.50E-06
rs2943640	2	227093585	C	A	1.09 (1.05-1.12)	1.80E-06
rs4430796	17	36098040	G	A	1.13 (1.07-1.19)	2.40E-06
rs7955901	12	71433293	C	T	1.09 (1.05-1.13)	3.20E-06
rs5215	11	17408630	C	T	1.08 (1.05-1.12)	4.40E-06
rs9505118	6	7290437	A	G	1.08 (1.05-1.12)	6.10E-06
rs11634397	15	80432222	G	A	1.09 (1.05-1.13)	7.30E-06
rs11717195	3	123082398	T	C	1.09 (1.05-1.14)	9.70E-06
rs243088	2	60568745	T	A	1.09 (1.05-1.13)	1.00E-05
rs7845219	8	95937502	T	C	1.08 (1.04-1.12)	1.40E-05
rs702634	5	53271420	A	G	1.08 (1.04-1.12)	1.80E-05
rs163184	11	2847069	G	T	1.09 (1.05-1.13)	1.90E-05
rs3130501	6	31136453	G	A	1.09 (1.05-1.13)	2.00E-05
rs7202877	16	75247245	T	G	1.15 (1.08-1.22)	2.30E-05
rs12899811	15	91544076	G	A	1.09 (1.04-1.13)	3.30E-05
rs6813195	4	153520475	C	T	1.08 (1.04-1.12)	6.10E-05

rs2075423	1	214154719	G	T	1.08 (1.04-1.12)	6.70E-05
rs7178572	15	77747190	G	A	1.08 (1.04-1.12)	1.00E-04
rs12970134	18	57884750	A	G	1.08 (1.04-1.12)	1.10E-04
rs6808574	3	187740523	C	T	1.08 (1.04-1.12)	1.30E-04
rs11063069	12	4374373	G	A	1.10 (1.05-1.15)	1.50E-04
rs10842994	12	27965150	C	T	1.09 (1.04-1.13)	1.50E-04
rs6795735	3	64705365	C	T	1.07 (1.03-1.10)	2.30E-04
rs2796441	9	84308948	G	A	1.07 (1.03-1.12)	2.50E-04
rs10923931	1	120517959	T	G	1.10 (1.05-1.16)	3.10E-04
rs10401969	19	19407718	C	T	1.13 (1.05-1.21)	5.40E-04
rs2334499	11	1696849	T	C	1.07 (1.03-1.11)	7.30E-04
rs4275659	12	123447928	C	T	1.06 (1.03-1.10)	8.80E-04
rs7612463	3	23336450	C	A	1.10 (1.04-1.16)	9.80E-04
rs17106184	1	50909985	G	A	1.10 (1.04-1.17)	1.10E-03
rs7163757	15	62391608	C	T	1.06 (1.02-1.10)	1.30E-03
rs8108269	19	46158513	G	T	1.06 (1.02-1.11)	3.10E-03
rs4812829	20	42989267	A	G	1.07 (1.02-1.12)	9.10E-03
rs7041847	9	4287466	A	G	1.05 (1.01-1.09)	9.90E-03
rs11257655	10	12307894	T	C	1.06 (1.01-1.11)	1.30E-02
rs10278336	7	44245363	A	G	1.05 (1.01-1.09)	2.00E-02
rs459193	5	55806751	G	A	1.05 (1.01-1.09)	2.10E-02
rs780094	2	27741237	C	T	1.04 (1.00-1.08)	2.50E-02
rs3923113	2	165501849	A	C	1.04 (1.00-1.08)	3.10E-02
rs2028299	15	90374257	C	A	1.04 (1.00-1.09)	3.50E-02
rs831571	3	64048297	C	T	1.03 (0.99-1.08)	1.60E-01
rs1802295	10	70931474	T	C	1.02 (0.98-1.06)	2.80E-01
rs3786897	19	33893008	A	G	1.02 (0.98-1.06)	3.10E-01
rs7403531	15	38822905	T	C	1.02 (0.98-1.06)	3.60E-01
rs16861329	3	186666461	C	T	1.03 (0.97-1.09)	3.90E-01
rs6467136	7	127164958	A	G	1.01 (0.98-1.05)	5.30E-01
rs10886471	10	121149403	T	C	1.01 (0.97-1.05)	5.90E-01
rs6723108	2	135479980	T	G	1.01 (0.97-1.04)	7.10E-01
rs9470794	6	38106844	T	C	1.01 (0.95-1.08)	8.00E-01
rs17584499	9	8879118	T	C	1.00 (0.95-1.06)	9.40E-01
rs391300	17	2216258	C	T	1.00 (0.96-1.04)	9.50E-01

Appendix D. *Olink protein biomarkers included in at least one cluster for multivariate analysis in HELIC-MANOLIS. The "Panel" column defines whether a trait is a plasma protein measured on one of the Olink panels (META=metabolic; CVDII=cardiovascular II; CVDIII=cardiovascular III).* N_{MISS}*=sample missingness prior to phenotype imputation;* R^2_{Imp} *=imputation accuracy.*

Panel	**Protein**	**Trait description**	**Comments**	$\mathbf{N_{MISS}}$	$\mathbf{R^2_{Imp}}$
CVDII	CD40L	Cluster of Differentiation 40 Ligand	Expressed on T cell surfaces; modulates B cell function by binding CD40 on B cell surface	24	0.80
CVDII	CD84	Cluster of Differentiation 84	Membrane glycoportein; modulates immune cell function through ligand-receptor interactions (similar to CD40L)	24	0.81
CVDII	HO1	Heme oxygenase (decycling) 1	Enzyme involved in heme catabolism; anti-inflammatory effects through upregulation of IL-10 and IL-1RA expression	24	0.88
CVDII	LPL	Lipoprotein lipase	Enzyme found on luminal surface of endothelial cells; hydrolyses triglycerides found in lipoproteins	24	0.79
CVDII	MERTK	Proto-oncogene tyrosine-protein kinase MER	Transmembrane protein involved in several processes, including cell survival, migration and phagocytosis	24	0.88
CVDII	PGF	Placental growth factor	Key regulator of angiogenesis; associated with inflammation and neovascularisation in artherosclerosis	24	0.81
CVDII	PRELP	Prolargin	Extracellular matrix protein tethering basement membrane to connective tissue; interacts with type I and type II collagens	24	0.82
CVDII	PTX3	Pentraxin-related protein PTX3	Expressed in several cell types and released during inflammatory response. Involved in classical complement pathway activation and pathogen recognition.	33	0.82
CVDII	TM	Thrombomodulin	Cofactor for thrombin expressed on surface of endothelial cells; reduces blood coagulation	24	0.86
CVDII	TNFRSF10A	Death receptor 4	Cell surface receptor of the TNF-receptor superfamily; mediates apoptosis	24	0.64
CVDII	TNFRSF11A	Receptor activator of nuclear factor κ B	Cell surface receptor of the TNF-receptor superfamily; reeptor for RANK ligand (RANKL) and involved in several processes, including bone remodeling and immune function	24	0.84

CVDII	TRAILR2	Death receptor 5	Cell surface receptor of the TNF-receptor superfamily; mediates apoptosis	24	0.91
CVDII	XCL1	Chemokine (C motif) ligand	Small cytokine involved in immune function; involved in the activation of cytotoxic T cells	45	0.82
CVDIII	AZU1	Azurocidin	Antimicrobial serine protease expressed in neutrophil granules and involved in inflammation	2	0.91
CVDIII	CASP3	Caspase 3	Member of the caspase superfamily which plays a central role in apoptosis	2	0.94
CVDIII	COL1A1	Collagen, type I, alpha 1	Major component of type I collagen, the fibrillar collagen found in most connective tissues, including cartilage	2	0.74
CVDIII	CPA1	Carboxipeptidase A1	Pancreatic enzyme involved in blocking activation of precursor enzymes (zymogens)	2	0.87
CVDIII	CPB1	Carboxipeptidase B1	Pancreatic enzyme involved in blocking activation of precursor enzymes (zymogens)	2	0.89
CVDIII	JAMA	Junctional adhesion molecule A	Member of the immunoglobulin superfamily, involved in formation of tight juncitons between epithelial cells	2	0.94
CVDIII	MEPE	Matrix Extracellular Phosphoglycoprotein	Calcium-binding secreted phosphoprotein found in extracellular matrix of bone; regulates bone mineralisation	351	0.73
CVDIII	MMP9	Matrix metallopeptidase 9	Enzyme involved in extracellular matrix degradation	2	0.87
CVDIII	MPO	Myeloperoxidase	Lysosomal enzyme with antimicrobial function, stored in neutrophil granules	3	0.83
CVDIII	OPN	Osteopontin	Secreted phosphoprotein found in bone and other tissues; involved in bone remodeling and immune function	3	0.81
CVDIII	PAI	Plasminogen activator inhibitor-1	Serine protease inhibitor that blocks breakdown of blood clots (fibrinolysis)	2	0.77
CVDIII	PDGFSUBUNITA	Platelet-derived growth factor subunit A	Member of the PDGF family invovled in cell growth and division	2	0.78
CVDIII	PECAM1	Platelet endothelial cell adhesion molecule	Immunoglobulin molecule found on cell surface of certain immune cells; involved in angiogenesis and integrin activation	2	0.94
CVDIII	PGLYRP1	Peptidoglycan recognition protein 1	Protein with bactericidal function, found mainly in neutrophil granules	2	0.91

CVDIII	PRTN3	Proteinase 3	Serine protease primarily expressed in neutrophils	2	0.91
CVDIII	RETN	Resistin	Peptide hormone involved in innate immune response	2	0.87
CVDIII	TR	Transferrin receptor protein 1	Transmembrane glycoprotein involved in iron import into cells	2	0.73
META	BAG6	Large proline-rich protein BAG6	Cleaved by caspase 3 and involved in apoptosis	12	0.86
META	CCDC80	Coiled-coil domain-containing protein 80	Promotes cell adhesion and matrix assembly	16	0.91
META	CHRDL2	Chordin Like 2	Secreted protein expressed in osteoblasts and associated with TGF-beta activity; negatively regulates cartilage formation, implicated in tumor angiogenesis	46	0.78
META	ENO2	Enolase 2	Enzyme found in neuronal cells; used as biomarker in lung cancer	12	0.89
META	FKBP4	FK506-binding protein 4	Member of immunophilin family involved in immunoregulation and protein folding/trafficking	205	0.76
META	KYAT1	Kynurenine—oxoglutarate transaminase 1	Cytosolic enzyme whose activity produces reactive metabolites associated with nephro- and neurotoxicity	12	0.77
META	QDPR	Quinoid dihydropteridine reductase	Enzyme involved in phenylalanine metabolism	12	0.92
META	RNASE3	Ribonuclease III	Cleaves double-stranded RNA, involved in RNA silencing	16	0.85
META	ROR1	Tyrosine-protein kinase transmembrane receptor ROR1	Cell surface receptor tyrosine kinase involved in neurite growth regulation; putative role in metastasis of cancer cells	12	0.92
META	THOP1	Thimet oligopeptidase 1	Metallopeptidase cleaving cytosolic and short neuropeptides	12	0.90

Appendix E. *Quantitative traits included in at least one cluster for multivariate analysis in HELIC-MANOLIS. Presented are measurement units, exclusion criteria, transformation applied and missing samples before phenotype imputation as well as imputation accuracy (R^2_{Imp}). INT=inverse normal transformation*

Trait	**Trait description**	**Unit**	**Exclusions**	**Transform.**	**N_{MISS}**	**R^2_{Imp}**
Adiponectin	Adiponectin	μg*mL^{-1}	±4σ, after sex stratification	log-normal	196	0.91
BGP	Bone-growth protein (osteocalcin)	ng*mL^{-1}	±4σ	log-normal	65	0.68
BMI	Body-mass index	kg*m^{-2}	none	log-normal	231	0.99
Fe_iron	Iron	μg*dL^{-1}	±4σ, after sex stratification	none	15	0.64
GRAN	Granulocytes	10^{9}*L^{-1}	±4σ	log-normal	310	0.99
GRANPC	Granulocytes (%)	%	none	none	301	1.00
HCT	Haematocrit	hct	±4σ, after sex stratification	none	220	1.00
Height	Standing height	cm	none	none	210	0.88
HGB	Haemoglobin	g*dL^{-1}	±4σ, after sex stratification	none	218	1.00
Hip	Hip circumference	cm	none	log-normal	196	0.99
Leptin	Leptin	ng*mL^{-1}	±4σ, after sex stratification	log-normal	399	0.51
LPCR	Large platelet concentration ratio	%	±4σ	none	299	0.99
LYMPC	Lymphocytes (%)	%	>60	none	222	1.00
MCH	Mean corpuscular haemoglobin	pg	none	INT	217	0.99
MCV	Mean corpuscular volume	fL	none	INT	217	0.99
MID	Mid-range absolute count	10^{3}*L^{-1}	>1.1	none	285	0.89
MPV	Mean platelet volume	fL	±4σ	none	222	0.87
PCT	Plateletcrit	μg*L^{-1}	±4σ	none	279	0.99
PDW	Platelet distribution width	fL	none	INT	276	0.97
PLT	Platelets	10^{9}*L^{-1}	±4σ, after sex stratification	none	222	0.98
RDW	Red cell distribution width	fL	none	INT	297	0.97
RG	Random glucose	mmol*L^{-1}	>15	INT	19	0.98
RI	Random insulin	μIU*mL^{-1}	none	log-normal	20	0.98
Waist	Waist circumference	cm	none	none	192	1.00
WBC	White blood cells	10^{9}*L^{-1}	±4σ	none	225	0.97
Weight	Current weight	kg	none	INT	219	0.98

Bibliography

1. Ozaki K, Ohnishi Y, Iida A, Sekine A, Yamada R, Tsunoda T, et al. **Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction**. *Nature Genetics* 2002; 32:650-654.

2. Klein RJ. **Complement Factor H Polymorphism in Age-Related Macular Degeneration**. *Science* 2005; 308:385-389.

3. Hirschhorn JN, Daly MJ. **Genome-wide association studies for common diseases and complex traits.** *Nature reviews Genetics* 2005; 6:95-108.

4. Timpson NJ, Greenwood CMT, Soranzo N, Lawson DJ, Richards JB. **Genetic architecture: the shape of the genetic contribution to human traits and disease**. *Nature Reviews Genetics* 2017; 19(2):110-124.

5. Badano JL, Katsanis N. **Beyond Mendel: an evolving view of human genetic disease transmission**. *Nature Reviews Genetics* 2002; 3(10):779-789.

6. Frazer KA, Murray SS, Schork NJ, Topol EJ. **Human genetic variation and its contribution to complex traits**. *Nature Reviews Genetics* 2009; 10(4):241-251.

7. Moltke I, Grarup N, Jørgensen ME, Bjerregaard P, Treebak JT, Fumagalli M, et al. **A common Greenlandic TBC1D4 variant confers muscle insulin resistance and type 2 diabetes**. *Nature* 2014; 512(7513):190-193.

8. Zheng HF, Forgetta V, Hsu YH, Estrada K, Rosello - Diez A, Leo PJ, et al. **Whole - genome sequencing identifies EN1 as a determinant of bone density and fracture**. *Nature* 2015; 526:112-117.

9. Gudmundsson J, Sulem P, Steinthorsdottir V, Bergthorsson JT, Thorleifsson G, Manolescu A, et al. **Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes**. *Nature Genetics* 2007; 39:977-983.

10. Luo Y, de Lange KM, Jostins L, Moutsianas L, Randall J, Kennedy NA, et al. **Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7**. *Nat Genet* 2017; 49(2):186-192.

11. Mahajan A, Go MJ, Zhang W, Below JE, Gaulton KJ, Ferreira T, et al. **Genome-wide trans-ancestry meta-analysis provides insight into the genetic architecture of type 2 diabetes susceptibility.** *Nature genetics* 2014; 46:234-244.

12. Liu JZ, van Sommeren S, Huang H, Ng SC, Alberts R, Takahashi A, et al. **Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations.** *Nature Genetics* 2015; 47(9):979-986.

13. Estrada K, Styrkarsdottir U, Evangelou E, Hsu YH, Duncan EL, Ntzani EE, et al. **Genome-wide meta-analysis identifies 56 bone mineral density loci and reveals 14 loci associated with risk of fracture**. *Nature Genetics* 2012; 44:494-501.

14. Ott J, Kamatani Y, Lathrop M. **Family-based designs for genome-wide association studies.** *Nature Reviews Genetics* 2011; 12(7):465-474.

15. Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, Tanzi RE, et al. **A polymorphic DNA marker genetically linked to Huntington's disease**. *Nature* 1983; 306(5940):234-238.

16. Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, Schumm JW, et al. **Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker**. *Science (New York, NY)* 1985; 230(4729):1054-1057.

17. Altshuler D, Daly MJ, Lander ES. **Genetic mapping in human disease**. *Science (New York, NY)* 2008; 322(5903):881-888.
18. Siontis KC, Patsopoulos NA, Ioannidis JP. **Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies**. *Eur J Hum Genet* 2010; 18(7):832-837.
19. International Human Genome Sequencing C. **Finishing the euchromatic sequence of the human genome**. *Nature* 2004; 431(7011):931-945.
20. Altshuler D, Pollara VJ, Cowles CR, Van Etten WJ, Baldwin J, Linton L, et al. **An SNP map of the human genome generated by reduced representation shotgun sequencing**. *Nature* 2000; 407(6803):513-516.
21. Visscher PM, Hill WG, Wray NR. **Heritability in the genomics era — concepts and misconceptions**. *Nature Reviews Genetics* 2008; 9:255-266.
22. Bamshad MJ, Ng SB, Bigham AW, Tabor HK, Emond MJ, Nickerson DA, et al. **Exome sequencing as a tool for Mendelian disease gene discovery**. *Nat Rev Genet* 2011; 12(11):745-755.
23. Metzker ML. **Sequencing technologies - the next generation**. *Nat Rev Genet* 2010; 11(1):31-46.
24. Marchini J, Howie B. **Genotype imputation for genome-wide association studies.** *Nature reviews Genetics* 2010; 11:499-511.
25. International HapMap C. **The International HapMap Project**. *Nature* 2003; 426(6968):789-796.
26. Genomes Project C, Abecasis GR, Auton A, Brooks LD, DePristo MA, Durbin RM, et al. **An integrated map of genetic variation from 1,092 human genomes**. *Nature* 2012; 491(7422):56-65.
27. Astle WJ, Elding H, Jiang T, Allen D, Ruklisa D, Mann AL, et al. **The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease.** *Cell* 2016; 167:1415-1429.
28. Liu C-T, Raghavan S, Maruthur N, Kabagambe EK, Hong J, Ng MCY, et al. **Trans-ethnic Meta-analysis and Functional Annotation Illuminates the Genetic Architecture of Fasting Glucose and Insulin**. *The American Journal of Human Genetics* 2016; 99:56-75.
29. Wain LV, Verwoert GC, O'Reilly PF, Shi G, Johnson T, Johnson AD, et al. **Genome-wide association study identifies six new loci influencing pulse pressure and mean arterial pressure**. *Nat Genet* 2011; 43(10):1005-1011.
30. Ripke S, Neale BM, Corvin A, Walters JTR, Farh K-H, Holmans Pa, et al. **Biological insights from 108 schizophrenia-associated genetic loci**. *Nature* 2014; 511:421-427.
31. Turcot V, Lu Y, Highland HM, Schurmann C, Justice AE, Fine RS, et al. **Protein-altering variants associated with body mass index implicate pathways that control energy intake and expenditure in obesity**. *Nature Genetics* 2018; 50:26-41.
32. Sun BB, Maranville JC, Peters JE, Stacey D, Staley JR, Blackshaw J, et al. **Genomic atlas of the human plasma proteome**. *Nature* 2018; 558(7708):73-79.
33. Wood AR, Esko T, Yang J, Vedantam S, Pers TH, Gustafsson S, et al. **Defining the role of common variation in the genomic and biological architecture of adult human height**. *Nat Genet* 2014; 46(11):1173-1186.
34. Natarajan P, Peloso GM, Zekavat SM, Montasser M, Ganna A, Chaffin M, et al. **Deep-coverage whole genome sequences and blood lipids among 16,324 individuals**. *Nat Commun* 2018; 9(1):3391.

35. Bis JC, Jian X, Kunkle BW, Chen Y, Hamilton-Nelson KL, Bush WS, et al. **Whole exome sequencing study identifies novel rare and common Alzheimer's-Associated variants involved in immune response and transcriptional regulation**. *Mol Psychiatry* 2018.
36. Sivakumaran S, Agakov F, Theodoratou E, Prendergast JG, Zgaga L, Manolio T, et al. **Abundant Pleiotropy in Human Complex Diseases and Traits**. *The American Journal of Human Genetics* 2011; 89:607-618.
37. Hobbs BD, de Jong K, Lamontagne M, Bossé Y, Shrine N, Artigas MS, et al. **Genetic loci associated with chronic obstructive pulmonary disease overlap with loci for lung function and pulmonary fibrosis.** *Nature genetics* 2017; 49:426-432.
38. Zeggini E, Panoutsopoulou K, Southam L, Rayner NW, Day-Williams AG, Lopes MC, et al. **Identification of new susceptibility loci for osteoarthritis (arcOGEN): A genome-wide association study**. *The Lancet* 2012; 380:815-823.
39. Anttila V, Winsvold BS, Gormley P, Kurth T, Bettella F, McMahon G, et al. **Genome-wide meta-analysis identifies new susceptibility loci for migraine**. *Nature Genetics* 2013; 45:912-917.
40. Freilinger T, Anttila V, de Vries B, Malik R, Kallela M, Terwindt GM, et al. **Genome-wide association analysis identifies susceptibility loci for migraine without aura**. *Nature Genetics* 2012; 44:777-782.
41. Locke AE, Kahali B, Berndt SI, Justice AE, Pers TH, Day FR, et al. **Genetic studies of body mass index yield new insights for obesity biology**. *Nature* 2015; 518:197-206.
42. Smoller JW. **Identification of risk loci with shared effects on five major psychiatric disorders: a genome-wide analysis**. *The Lancet* 2013; 381:1371-1379.
43. Consortium C-DGotPG, Lee SH, Ripke S, Neale BM, Faraone SV, Purcell SM, et al. **Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs.** *Nature Genetics* 2013; 45:984-994.
44. Li YR, Li J, Zhao SD, Bradfield JP, Mentch FD, Maggadottir SM, et al. **Meta-analysis of shared genetic architecture across ten pediatric autoimmune diseases**. *Nature Medicine* 2015; 21:1018-1027.
45. Stuart PE, Nair RP, Tsoi LC, Tejasvi T, Das S, Kang HM, et al. **Genome-wide Association Analysis of Psoriatic Arthritis and Cutaneous Psoriasis Reveals Differences in Their Genetic Architecture.** *American journal of human genetics* 2015; 97:816-836.
46. Sudlow C, Gallacher J, Allen N, Beral V, Burton P, Danesh J, et al. **UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age.** *PLoS medicine* 2015; 12:e1001779.
47. Matise TC, Ambite JL, Buyske S, Carlson CS, Cole SA, Crawford DC, et al. **The Next PAGE in Understanding Complex Traits: Design for the Analysis of Population Architecture Using Genetics and Epidemiology (PAGE) Study**. *American Journal of Epidemiology* 2011; 174:849-859.
48. Muñoz M, Pong-Wong R, Canela-Xandri O, Rawlik K, Haley CS, Tenesa A. **Evaluating the contribution of genetics and familial shared environment to common disease using the UK Biobank**. *Nature Genetics* 2016; 48:980-983.
49. Bush WS, Oetjens MT, Crawford DC. **Unravelling the human genome-phenome relationship using phenome-wide association studies.** *Nature Reviews Genetics* 2016; 17:129-145.

50. Verma A, Verma SS, Pendergrass SA, Crawford DC, Crosslin DR, Kuivaniemi H, et al. **eMERGE Phenome-Wide Association Study (PheWAS) identifies clinical associations and pleiotropy for stop-gain variants**. *BMC Medical Genomics* 2016; 9:32.
51. Krapohl E, Euesden J, Zabaneh D, Pingault J-B, Rimfeld K, von Stumm S, et al. **Phenome-wide analysis of genome-wide polygenic scores**. *Molecular Psychiatry* 2016; 21:1188-1193.
52. Liu J, Ye Z, Mayer JG, Hoch BA, Green C, Rolak L, et al. **Phenome-wide association study maps new diseases to the human major histocompatibility complex region.** *Journal of medical genetics* 2016; 53:681-689.
53. Paaby AB, Rockman MV. **The many faces of pleiotropy.** *Trends in genetics* 2013; 29:66-73.
54. Pendergrass SA, Brown-Gentry K, Dudek S, Frase A, Torstenson ES, Goodloe R, et al. **Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network.** *PLoS Genetics* 2013; 9:e1003087.
55. Stearns FW. **One hundred years of pleiotropy: A retrospective**. *Genetics* 2010; 186:767-773.
56. Wagner GP, Zhang J. **Fundamental concepts in genetics: The pleiotropic structure of the genotype–phenotype map: the evolvability of complex organisms**. *Nature Reviews Genetics* 2011; 12:204-213.
57. Hodgkin J. **Seven types of pleiotropy**. *The International Journal of Developmental Biology* 1998; 505:501-505.
58. Solovieff N, Cotsapas C, Lee PH, Purcell SM, Smoller JW. **Pleiotropy in complex traits: challenges and strategies.** *Nature Reviews Genetics* 2013; 14:483-495.
59. Gage SH, Davey Smith G, Ware JJ, Flint J, Munafò MR. **G = E: What GWAS Can Tell Us about the Environment**. *PLOS Genetics* 2016; 12:e1005765.
60. Timpson NJ, Lawlor DA, Harbord RM, Gaunt TR, Day INM, Palmer LJ, et al. **C-reactive protein and its role in metabolic syndrome: mendelian randomisation study.** *Lancet (London, England)* 2005; 366:1954-1959.
61. Panoutsopoulou K, Metrustry S, Doherty SA, Laslett LL, Maciewicz RA, Hart DJ, et al. **The effect of FTO variation on increased osteoarthritis risk is mediated through body mass index: a Mendelian randomisation study.** *Annals of the Rheumatic Diseases* 2014; 73:2082-2086.
62. Liu JZ, Almarri Ma, Gaffney DJ, Mells GF, Jostins L, Cordell HJ, et al. **Dense fine-mapping study identifies new susceptibility loci for primary biliary cirrhosis**. *Nature Genetics* 2012; 44:1137-1141.
63. Liu JZ, Hov JR, Folseraas T, Ellinghaus E, Rushbrook SM, Doncheva NT, et al. **Dense genotyping of immune-related disease regions identifies nine new risk loci for primary sclerosing cholangitis.** *Nature genetics* 2013; 45:670-675.
64. Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, et al. **Population structure, differential bias and genomic control in a large-scale, case-control association study.** *Nature genetics* 2005; 37:1243-1246.
65. Zhao W, Rasheed A, Tikkanen E, Lee J-J, Butterworth AS, Howson JMM, et al. **Identification of new susceptibility loci for type 2 diabetes and shared etiological pathways with coronary heart disease**. *Nature Genetics* 2017.

66. Lie BA, Todd JA, Pociot F, Nerup J, Akselsen HE, Joner G, et al. **The predisposition to type 1 diabetes linked to the human leukocyte antigen complex includes at least one non-class II gene.** *American journal of human genetics* 1999; 64:793-800.
67. Galesloot TE, Van Steen K, Kiemeney LaLM, Janss LL, Vermeulen SH. **A comparison of multivariate genome-wide association methods**. *PLoS ONE* 2014; 9:1-8.
68. Zhu H, Zhang S, Sha Q. **Power Comparisons of Methods for Joint Association Analysis of Multiple Phenotypes**. *Human Heredity* 2015; 80:144-152.
69. Aschard H, Vilhjálmsson BJ, Greliche N, Morange P-EE, Trégouët D-AA, Kraft P. **Maximizing the power of principal-component analysis of correlated phenotypes in genome-wide association studies.** *American Journal of Human Genetics* 2014; 94:662-676.
70. Majumdar A, Haldar T, Witte JS. **Determining Which Phenotypes Underlie a Pleiotropic Signal**. *Genetic Epidemiology* 2016; 40:366-381.
71. Porter HF, O'Reilly PF. **Multivariate simulation framework reveals performance of multi-trait GWAS methods.** *Scientific Reports* 2017; 7:38837.
72. Klei L, Luca D, Devlin B, Roeder K. **Pleiotropy and principal components of heritability combine to increase power for association analysis.** *Genetic epidemiology* 2008; 32:9-19.
73. Liu J, Pei Y, Papasian CJ, Deng H-W. **Bivariate association analyses for the mixture of continuous and binary traits with the use of extended generalized estimating equations.** *Genetic Epidemiology* 2009; 33:217-227.
74. O'Reilly PF, Hoggart CJ, Pomyen Y, Calboli FCF, Elliott P, Jarvelin M-R, et al. **MultiPhen: joint model of multiple phenotypes can increase discovery in GWAS.** *PloS one* 2012; 7:e34861.
75. Zhou X, Stephens M. **Efficient multivariate linear mixed model algorithms for genome-wide association studies.** *Nature Methods* 2014; 11:407-409.
76. Cichonska A, Rousu J, Marttinen P, Kangas AJ, Soininen P, Lehtimäki T, et al. **metaCCA: summary statistics-based multivariate meta-analysis of genome-wide association studies using canonical correlation analysis**. *Bioinformatics* 2016:btw052.
77. Giambartolomei C, Vukcevic D, Schadt EE, Franke L, Hingorani AD, Wallace C, et al. **Bayesian Test for Colocalisation between Pairs of Genetic Association Studies Using Summary Statistics**. *PLoS Genetics* 2014; 10.
78. Pickrell J, Berisa T, Segurel L, Tung JY, Hinds D. **Detection and interpretation of shared genetic influences on 42 human traits**. In: *Nature Genetics*; 2015. pp. 019885.
79. Bulik-Sullivan B, Finucane HK, Anttila V, Gusev A, Day FR, Loh P-R, et al. **An atlas of genetic correlations across human diseases and traits.** *Nature Genetics* 2015; 47:1236-1241.
80. Bhattacharjee S, Rajaraman P, Jacobs KB, Wheeler Wa, Melin BS, Hartge P, et al. **A subset-based approach improves power and interpretation for the combined analysis of genetic association studies of heterogeneous traits**. *American Journal of Human Genetics* 2012; 90:821-835.
81. Liley J, Wallace C. **A pleiotropy-informed Bayesian false discovery rate adapted to a shared control design finds new disease associations from GWAS summary statistics.** *PLoS genetics* 2015; 11:e1004926.
82. Province MA, Borecki IB. **A correlated meta-analysis strategy for data mining "OMIC" scans.** *Pacific Symposium on Biocomputing* 2013:236-246.

83. Han B, Pouget JG, Slowikowski K, Stahl E, Lee CH, Diogo D, et al. **A method to decipher pleiotropy by detecting underlying heterogeneity driven by hidden subgroups applied to autoimmune and neuropsychiatric diseases**. *Nature Genetics* 2016; 48:803-810.
84. Lin D-Y, Sullivan PF. **Meta-analysis of genome-wide association studies with overlapping subjects.** *American journal of human genetics* 2009; 85:862-872.
85. Talmud PJ, Hingorani AD, Cooper JA, Marmot MG, Brunner EJ, Kumari M, et al. **Utility of genetic and non-genetic risk factors in prediction of type 2 diabetes: Whitehall II prospective cohort study**. *BMJ* 2010; 340:b4838-b4838.
86. Dudbridge F. **Power and Predictive Accuracy of Polygenic Risk Scores**. *PLoS Genetics* 2013; 9:e1003348.
87. Evans DM, Visscher PM, Wray NR. **Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk.** *Human molecular genetics* 2009; 18:3525-3531.
88. Palla L, Dudbridge F. **A Fast Method that Uses Polygenic Scores to Estimate the Variance Explained by Genome-wide Marker Panels and the Proportion of Variants Affecting a Trait**. *The American Journal of Human Genetics* 2015; 97:250-259.
89. Purcell SM, Wray NR, Stone JL, Visscher PM, O'Donovan MC, Sullivan PF, et al. **Common polygenic variation contributes to risk of schizophrenia and bipolar disorder.** *Nature* 2009; 460:748-752.
90. Euesden J, Lewis CM, O'Reilly PF, Reilly PFO. **PRSice: Polygenic Risk Score software**. *Bioinformatics* 2014; 31:1466-1468.
91. Lee SH, Yang J, Goddard ME, Visscher PM, Wray NR. **Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood**. *Bioinformatics* 2012; 28:2540-2542.
92. Loh P-R, Bhatia G, Gusev A, Finucane HK, Bulik-Sullivan BK, Pollack SJ, et al. **Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis**. *Nature Genetics* 2015; 47:1385-1392.
93. Furlotte NA, Eskin E. **Efficient multiple-trait association and estimation of genetic correlation using the matrix-variate linear mixed model.** *Genetics* 2015; 200:59-68.
94. Bulik-Sullivan BK, Loh P-R, Finucane HK, Ripke S, Yang J, Patterson N, et al. **LD Score regression distinguishes confounding from polygenicity in genome-wide association studies.** *Nature Genetics* 2015; 47:291-295.
95. Zheng J, Erzurumluoglu AM, Elsworth BL, Kemp JP, Howe L, Haycock PC, et al. **LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis**. *Bioinformatics* 2017; 33:272-279.
96. Berisa T, Pickrell JK. **Approximately independent linkage disequilibrium blocks in human populations.** *Bioinformatics* 2016; 32:283-285.
97. Tang CS, Ferreira MaR. **A gene-based test of association using canonical correlation analysis.** *Bioinformatics* 2012; 28:845-850.
98. Seoane Ja, Campbell C, Day INM, Casas JP, Gaunt TR. **Canonical Correlation Analysis for Gene-Based Pleiotropy Discovery**. *PLoS Computational Biology* 2014; 10:e1003876.
99. Wang Y, Liu A, Mills JL, Boehnke M, Wilson AF, Bailey-Wilson JE, et al. **Pleiotropy Analysis of Quantitative Traits at Gene Level by Multivariate Functional Linear Models.** *Genetic Epidemiology* 2015; 39:259-275.

100. Lutz SM, Fingerlin TE, Hokanson JE, Lange C. **A general approach to testing for pleiotropy with rare and common variants**. *Genetic Epidemiology* 2016; 41:163-170.
101. Casale FP, Rakitsch B, Lippert C, Stegle O. **Efficient set tests for the genetic analysis of correlated traits**. *Nature Methods* 2015; 12:755-758.
102. Mardis ER. **DNA sequencing technologies: 2006–2016**. *Nature Protocols* 2017; 12:213-218.
103. McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, Teumer A, et al. **A reference panel of 64,976 haplotypes for genotype imputation**. *Nature Genetics* 2016; 48:1279-1283.
104. Asimit J, Zeggini E. **Rare Variant Association Analysis Methods for Complex Traits**. *Annual Review of Genetics* 2010; 44:293-308.
105. Lee S, Abecasis GR, Boehnke M, Lin X. **Rare-variant association analysis: Study designs and statistical tests**. *American Journal of Human Genetics* 2014; 95:5-23.
106. Wu Michael C, Lee S, Cai T, Li Y, Boehnke M, Lin X. **Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test**. *The American Journal of Human Genetics* 2011; 89:82-93.
107. Madsen BE, Browning SR. **A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic**. *PLoS Genetics* 2009; 5:e1000384.
108. Wu B, Pankow JS. **Sequence Kernel Association Test of Multiple Continuous Phenotypes.** *Genetic Epidemiology* 2016; 40:91-100.
109. Lee S, Won S, Kim YJ, Kim Y, Kim B-J, Park T. **Rare variant association test with multiple phenotypes**. *Genetic Epidemiology* 2016.
110. Wang Z, Wang X, Sha Q, Zhang S. **Joint Analysis of Multiple Traits in Rare Variant Association Studies**. *Annals of Human Genetics* 2016; 80:162-171.
111. Kaakinen M, Magi R, Fischer K, Heikkinen J, Jarvelin MR, Morris AP, et al. **A rare-variant test for high-dimensional data**. *Eur J Hum Genet* 2017; 25(8):988-994.
112. Kaakinen M, Magi R, Fischer K, Heikkinen J, Jarvelin MR, Morris AP, et al. **MARV: a tool for genome-wide multi-phenotype analysis of rare variants**. *BMC Bioinformatics* 2017; 18(1):110.
113. Cotsapas C, Voight BF, Rossin E, Lage K, Neale BM, Wallace C, et al. **Pervasive sharing of genetic effects in autoimmune disease**. *PLoS Genetics* 2011; 7:e1002254.
114. Zhu X, Feng T, Tayo BO, Liang J, Young JH, Franceschini N, et al. **Meta-analysis of Correlated Traits via Summary Statistics from GWASs with an Application in Hypertension**. *The American Journal of Human Genetics* 2015; 96:21-36.
115. Turley P, Walters RK, Maghzian O, Okbay A, Lee JJ, Fontana MA, et al. **Multi-trait analysis of genome-wide association summary statistics using MTAG**. *Nat Genet* 2018; 50(2):229-237.
116. Vuckovic D, Gasparini P, Soranzo N, Iotchkova V. **MultiMeta: An R package for meta-analyzing multi-phenotype genome-wide association studies**. *Bioinformatics* 2015; 31:2754-2756.
117. Evangelou E, Ioannidis JPA. **Meta-analysis methods for genome-wide association studies and beyond**. *Nature Reviews Genetics* 2013; 14:379-389.
118. Wang Z, Zhu B, Zhang M, Parikh H, Jia J, Chung CC, et al. **Imputation and subset-based association analysis across different cancer types identifies multiple independent risk loci in the TERT-CLPTM1L region on chromosome 5p15.33**. *Human Molecular Genetics* 2014; 23:6616-6633.

119. Park H, Li X, Song YE, He KY, Zhu X. **Multivariate Analysis of Anthropometric Traits Using Summary Statistics of Genome-Wide Association Studies from GIANT Consortium.** *PloS one* 2016; 11:e0163912.
120. Shriner D. **Moving toward System Genetics through Multiple Trait Analysis in Genome-Wide Association Studies.** *Frontiers in Genetics* 2012; 3:1.
121. Yang Q, Wang Y. **Methods for Analyzing Multivariate Phenotypes in Genetic Association Studies**. *Journal of Probability and Statistics* 2012:1-13.
122. Ott J, Rabinowitz D. **A principal-components approach based on heritability for combining phenotype information.** *Human Heredity* 1999; 49:106-111.
123. Lange C, van Steen K, Andrew T, Lyon H, DeMeo DL, Raby B, et al. **A family-based association test for repeatedly measured quantitative traits adjusting for unknown environmental and/or polygenic effects**. *Statistical Applications in Genetics and Molecular Biology* 2004; 3:Article17.
124. Ried JS, Jeff M J, Chu AY, Bragg-Gresham JL, van Dongen J, Huffman JE, et al. **A principal component meta-analysis on multiple anthropometric traits identifies novel loci for body shape.** *Nature communications* 2016; 7:13357.
125. Ferreira MAR, Purcell SM. **A multivariate test of association**. *Bioinformatics* 2009; 25:132-133.
126. Zhang Z, Ersoz E, Lai C-Q, Todhunter RJ, Tiwari HK, Gore MA, et al. **Mixed linear model approach adapted for genome-wide association studies**. *Nature Genetics* 2010; 42:355-360.
127. Yu J, Pressoir G, Briggs WH, Vroh Bi I, Yamasaki M, Doebley JF, et al. **A unified mixed-model method for association mapping that accounts for multiple levels of relatedness**. *Nature Genetics* 2006; 38:203-208.
128. Korte A, Vilhjálmsson BJ, Segura V, Platt A, Long Q, Nordborg M. **A mixed-model approach for genome-wide association studies of correlated traits in structured populations**. *Nature Genetics* 2012; 44:1066-1071.
129. Joo JWJ, Kang EY, Org E, Furlotte N, Parks B, Hormozdiari F, et al. **Efficient and Accurate Multiple-Phenotype Regression Method for High Dimensional Data Considering Population Structure.** *Genetics* 2016; 204:1379-1390.
130. Hartley SW, Monti S, Liu C-T, Steinberg MH, Sebastiani P. **Bayesian methods for multivariate modeling of pleiotropic SNP associations and genetic risk prediction.** *Frontiers in Genetics* 2012; 3:176.
131. Stephens M. **A unified framework for association analysis with multiple related phenotypes.** *PloS one* 2013; 8:e65245.
132. Hartley SW, Sebastiani P. **PleioGRiP: genetic risk prediction with pleiotropy.** *Bioinformatics* 2013; 29:1086-1088.
133. Marchini J, Howie B, Myers S, McVean G, Donnelly P. **A new multipoint method for genome-wide association studies by imputation of genotypes.** *Nature genetics* 2007; 39:906-913.
134. Shim H, Chasman DI, Smith JD, Mora S, Ridker PM, Nickerson DA, et al. **A multivariate genome-wide association analysis of 10 LDL subfractions, and their response to statin treatment, in 1868 Caucasians.** *PLoS ONE* 2015; 10:e0120758.
135. Magi R, Suleimanov YV, Clarke GM, Kaakinen M, Fischer K, Prokopenko I, et al. **SCOPA and META-SCOPA: software for the analysis and aggregation of genome-wide**

association studies of multiple correlated phenotypes. *BMC Bioinformatics* 2017; 18(1):25.

136. Sterne JA, White IR, Carlin JB, Spratt M, Royston P, Kenward MG, et al. **Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls**. *BMJ* 2009; 338:b2393.

137. Pedersen AB, Mikkelsen EM, Cronin-Fenton D, Kristensen NR, Pham TM, Pedersen L, et al. **Missing data and multiple imputation in clinical epidemiological research**. *Clin Epidemiol* 2017; 9:157-166.

138. Little RJ, D'Agostino R, Cohen ML, Dickersin K, Emerson SS, Farrar JT, et al. **The prevention and treatment of missing data in clinical trials**. *N Engl J Med* 2012; 367(14):1355-1360.

139. Gomer B. **MCAR, MAR, and MNAR Values in the Same Dataset: A Realistic Evaluation of Methods for Handling Missing Data**. *Multivariate Behav Res* 2019:1.

140. Hayati Rezvan P, Lee KJ, Simpson JA. **The rise of multiple imputation: a review of the reporting and implementation of the method in medical research**. *BMC Med Res Methodol* 2015; 15:30.

141. Yucel RM. **State of the Multiple Imputation Software**. *J Stat Softw* 2011; 45(1).

142. Davey Smith G, Hemani G. **Mendelian randomization: genetic anchors for causal inference in epidemiological studies.** *Human Molecular Genetics* 2014; 23:R89-98.

143. Thomas DC, Conti DV. **Commentary: The concept of 'Mendelian randomization'**. *International Journal of Epidemiology* 2004; 33:21-25.

144. Lawlor DA, Harbord RM, Sterne JAC, Timpson N, Davey Smith G. **Mendelian randomization: Using genes as instruments for making causal inferences in epidemiology**. *Statistics in Medicine* 2008; 27:1133-1163.

145. White J, Sofat R, Hemani G, Shah T, Engmann J, Dale C, et al. **Plasma urate concentration and risk of coronary heart disease: a Mendelian randomisation analysis.** *The lancet Diabetes & endocrinology* 2016.

146. Palmer TM, Lawlor DA, Harbord RM, Sheehan NA, Tobias JH, Timpson NJ, et al. **Using multiple genetic variants as instrumental variables for modifiable risk factors.** *Statistical Methods in Medical Research* 2012; 21:223-242.

147. Swerdlow DI, Kuchenbaecker KB, Shah S, Sofat R, Holmes MV, White J, et al. **Selecting instruments for Mendelian randomization in the wake of genome-wide association studies.** *International Journal of Epidemiology* 2016; 45:1600-1616.

148. Didelez V, Sheehan N. **Mendelian randomization as an instrumental variable approach to causal inference**. *Statistical Methods in Medical Research* 2007; 16:309-330.

149. Burgess S, Thompson SG. **Use of allele scores as instrumental variables for Mendelian randomization**. *Int J Epidemiol* 2013; 42(4):1134-1144.

150. Schmidt AF, Swerdlow DI, Holmes MV, Patel RS, Fairhurst-Hunter Z, Lyall DM, et al. **PCSK9 genetic variants and risk of type 2 diabetes: a mendelian randomisation study**. *Lancet Diabetes Endocrinol* 2017; 5(2):97-105.

151. Greco M FD, Minelli C, Sheehan Na, Thompson JR. **Detecting pleiotropy in Mendelian randomisation studies with summary data and a continuous outcome**. *Statistics in Medicine* 2015; 34:2926-2940.

152. Bowden J, Del Greco M F, Minelli C, Davey Smith G, Sheehan N, Thompson J. **A framework for the investigation of pleiotropy in two-sample summary data Mendelian randomization.** *Statistics in medicine* 2017.

153. Bowden J, Smith GD, Burgess S. **Mendelian randomization with invalid instruments: Effect estimation and bias detection through Egger regression**. *International Journal of Epidemiology* 2015; 44:512-525.

154. Dai JY, Peters U, Wang X, Kocarnik J, Chang-Claude J, Slattery ML, et al. **Diagnostics for Pleiotropy in Mendelian Randomization Studies: Global and Individual Tests for Direct Effects**. *Am J Epidemiol* 2018; 187(12):2672-2680.

155. Schmidt AF, Dudbridge F. **Mendelian randomization with Egger pleiotropy correction and weakly informative Bayesian priors**. *Int J Epidemiol* 2018; 47(4):1217-1228.

156. Verbanck M, Chen C-Y, Neale B, Do R. **Detection of widespread horizontal pleiotropy in causal relationships inferred from Mendelian randomization between complex traits and diseases**. *Nature Genetics* 2018.

157. Burgess S, Butterworth A, Thompson SG. **Mendelian randomization analysis with multiple genetic variants using summarized data**. *Genetic Epidemiology* 2013; 37:658-665.

158. Rees JMB, Wood AM, Burgess S. **Extending the MR-Egger method for multivariable Mendelian randomization to correct for both measured and unmeasured pleiotropy**. *Statistics in Medicine* 2017.

159. Zhu Z, Zhang F, Hu H, Bakshi A, Robinson MR, Powell JE, et al. **Integration of summary data from GWAS and eQTL studies predicts complex trait gene targets.** *Nature genetics* 2016; 48:481-487.

160. Zhu Z, Zheng Z, Zhang F, Wu Y, Trzaskowski M, Maier R, et al. **Causal associations between risk factors and common diseases inferred from GWAS summary data**. *Nat Commun* 2018; 9(1):224.

161. Cai N, Bigdeli TB, Kretzschmar W, Li Y, Liang J, Song L, et al. **Sparse whole-genome sequencing identifies two loci for major depressive disorder**. *Nature* 2015; 523:588-591.

162. Ripke S, Wray NR, Lewis CM, Hamilton SP, Weissman MM, Breen G, et al. **A mega-analysis of genome-wide association studies for major depressive disorder**. *Molecular Psychiatry* 2013; 18:497-511.

163. Brainstorm C, Anttila V, Bulik-Sullivan B, Finucane HK, Walters RK, Bras J, et al. **Analysis of shared heritability in common disorders of the brain.** *Science* 2018; 360(6395).

164. Zhang Y, Jordan JM. **Epidemiology of osteoarthritis.** *Clinics in geriatric medicine* 2010; 26:355-369.

165. Glyn-Jones S, Palmer AJR, Agricola R, Price AJ, Vincent TL, Weinans H, et al. **Osteoarthritis**. *The Lancet* 2015; 386:376-387.

166. Hochberg MC, Yerges-Armstrong L, Yau M, Mitchell BD. **Genetic epidemiology of osteoarthritis: recent developments and future directions.** *Current opinion in rheumatology* 2013; 25:192-197.

167. Kellgren JH, Lawrence JS. **Radiological assessment of osteo-arthrosis**. *Ann Rheum Dis* 1957; 16(4):494-502.

168. Hackinger S, Trajanoska K, Styrkarsdottir U, Zengini E, Steinberg J, Ritchie GRS, et al. **Evaluation of shared genetic aetiology between osteoarthritis and bone mineral density identifies SMAD3 as a novel osteoarthritis risk locus.** *Human Molecular Genetics* 2017; 19:324-331.

169. Cibrián Uhalte E, Wilkinson JM, Southam L, Zeggini E. **Pathways to understanding the genomic aetiology of osteoarthritis**. *Human Molecular Genetics* 2017; 26(R2):R193-R201.
170. Panoutsopoulou K, Southam L, Elliott KS, Wrayner N, Zhai G, Beazley C, et al. **Insights into the genetic architecture of osteoarthritis from stage 1 of the arcOGEN study.** *Annals of the rheumatic diseases* 2011; 70:864-867.
171. Evangelou E, Kerkhof HJ, Styrkarsdottir U, Ntzani EE, Bos SD, Esko T, et al. **A meta-analysis of genome-wide association studies identifies novel variants associated with osteoarthritis of the hip.** *Annals of the Rheumatic Diseases* 2014; 73:2130-2136.
172. Zengini E, Hatzikotoulas K, Tachmazidou I, Steinberg J, Hartwig FP, Southam L, et al. **Genome-wide analyses using UK Biobank data provide insights into the genetic architecture of osteoarthritis.** *Nature Genetics* 2018:1.
173. Zengini E, Finan C, Wilkinson JM. **The Genetic Epidemiological Landscape of Hip and Knee Osteoarthritis: Where Are We Now and Where Are We Going?** *The Journal of Rheumatology* 2016; 43:260-266.
174. Casalone E, Tachmazidou I, Zengini E, Hatzikotoulas K, Hackinger S, Suveges D, et al. **A novel variant in GLIS3 is associated with osteoarthritis**. *Ann Rheum Dis* 2018; 77(4):620-623.
175. Richards JB, Zheng H-F, Spector TD. **Genetics of osteoporosis from genome-wide association studies: advances and challenges.** *Nature reviews Genetics* 2012; 13:576-588.
176. Kemp JP, Morris JA, Medina-Gomez C, Forgetta V, Warrington NM, Youlten SE, et al. **Identification of 153 new loci associated with heel bone mineral density and functional involvement of GPC6 in osteoporosis**. 2017.
177. Foss MV, Byers PD. **Bone density, osteoarthrosis of the hip, and fracture of the upper end of the femur.** *Annals of the Rheumatic Diseases* 1972; 31:259-264.
178. Nevitt MC, Zhang Y, Javaid MK, Neogi T, Curtis JR, Niu J, et al. **High systemic bone mineral density increases the risk of incident knee OA and joint space narrowing, but not radiographic progression of existing knee OA: the MOST study.** *Annals of the rheumatic diseases* 2010; 69:163-168.
179. Hart DJ, Mootoosamy I, Doyle DV, Spector TD. **The relationship between osteoarthritis and osteoporosis in the general population: the Chingford Study.** *Annals of the Rheumatic Diseases* 1994; 53:158-162.
180. Hannan MT, Anderson JJ, Zhang Y, Levy D, Felson DT. **Bone mineral density and knee osteoarthritis in elderly men and women. the framingham study**. *Arthritis & Rheumatism* 1993; 36:1671-1680.
181. Hardcastle Sa, Dieppe P, Gregson CL, Smith GD, Tobias JH. **Osteoarthritis and bone mineral density : are strong bones bad for joints ?** *BoneKEy Reports* 2015; 4:1-8.
182. Hardcastle Sa, Dieppe P, Gregson CL, Arden NK, Spector TD, Hart DJ, et al. **Individuals with high bone mass have an increased prevalence of radiographic knee osteoarthritis**. *Bone* 2015; 71:171-179.
183. Bettica P, Cline G, Hart DJ, Meyer J, Spector TD. **Evidence for increased bone resorption in patients with progressive knee osteoarthritis: Longitudinal results from the Chingford study**. *Arthritis & Rheumatism* 2002; 46:3178-3184.
184. Hart DJ, Cronin C, Daniels M. **The relationship of bone density and fracture to incedent and progressive radiographic osteoarthritis of the knee**. *Arthritis and Rheumatism* 2002; 46:92-99.

185. Zhang Y, Hannan MT, Chaisson CE, McAlindon TE, Evans SR, Aliabadi P, et al. **Bone mineral density and risk of incident and progressive radiographic knee osteoarthritis in women: the Framingham Study.** *The Journal of Rheumatology* 2000; 27:1032-1037.
186. Bergink AP, Uitterlinden AG, Van Leeuwen JPTM, Hofman A, Verhaar JAN, Pols HAP. **Bone mineral density and vertebral fracture history are associated with incident and progressive radiographic knee osteoarthritis in elderly men and women: the Rotterdam Study.** *Bone* 2005; 37:446-456.
187. Geusens PP, van den Bergh JP. **Osteoporosis and osteoarthritis: shared mechanisms and epidemiology.** *Current opinion in rheumatology* 2016; 28:97-103.
188. Balooch G, Balooch M, Nalla RK, Schilling S, Filvaroff EH, Marshall GW, et al. **TGF-beta regulates the mechanical properties and composition of bone matrix.** *Proceedings of the National Academy of Sciences of the United States of America* 2005; 102:18813-18818.
189. Panoutsopoulou K, Zeggini E. **Advances in osteoarthritis genetics.** *Journal of Medical Genetics* 2013; 50:715-724.
190. Yerges-Armstrong LM, Yau MS, Liu Y, Krishnan S, Renner JB, Eaton CB, et al. **Association analysis of BMD-associated SNPs with knee osteoarthritis**. *Journal of Bone and Mineral Research* 2014; 29:1373-1379.
191. Howie BN, Donnelly P, Marchini J, Hardy J, Abecasis G. **A Flexible and Accurate Genotype Imputation Method for the Next Generation of Genome-Wide Association Studies**. *PLoS Genetics* 2009; 5:e1000529.
192. Servin B, Stephens M. **Imputation-based analysis of association studies: candidate regions and quantitative traits**. *PLoS Genet* 2007; 3(7):e114.
193. Li Y, Willer C, Sanna S, Abecasis G. **Genotype imputation**. *Annu Rev Genomics Hum Genet* 2009; 10:387-406.
194. Medina-Gomez C, Kemp JP, Dimou NL, Kreiner E, Chesi A, Zemel BS, et al. **Bivariate genome-wide association meta-analysis of pediatric musculoskeletal traits reveals pleiotropic effects at the SREBF1/TOM1L2 locus**. *Nature Communications* 2017; 8:121.
195. Franklin J, Ingvarsson T, Englund M, Lohmander S. **Association between occupation and knee and hip replacement due to osteoarthritis: a case-control study.** *Arthritis research & therapy* 2010; 12:R102.
196. Styrkarsdottir U, Thorleifsson G, Helgadottir HT, Bomer N, Metrustry S, Bierma-Zeinstra S, et al. **Severe osteoarthritis of the hand associates with common variants within the ALDH1A2 gene and with rare variants at 1p31.** *Nature genetics* 2014; 46:498-502.
197. Zhu Z, Anttila V, Smoller JW, Lee PH. **Statistical power and utility of meta-analysis methods for cross-phenotype genome-wide association studies**. *PLOS ONE* 2018; 13(3):e0193256.
198. Revelle W. **psych: Procedures for Personality and Psychological Research**. In. 1.8.3 ed. Evanston, Illinois, USA: Northwestern University; 2018.
199. Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ. **Second-generation PLINK: rising to the challenge of larger and richer datasets**. *GigaScience* 2015; 4:1-16.
200. Elliott KS, Chapman K, Day-Williams A, Panoutsopoulou K, Southam L, Lindgren CM, et al. **Evaluation of the genetic overlap between osteoarthritis with body mass index and height using genome-wide association scan data**. *Annals of the Rheumatic Diseases* 2013; 72:935-941.
201. Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, et al. **A global reference for human genetic variation**. *Nature* 2015; 526:68-74.

202. de Leeuw CA, Mooij JM, Heskes T, Posthuma D. **MAGMA: generalized gene-set analysis of GWAS data.** *PLoS computational biology* 2015; 11:e1004219.

203. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, et al. **Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.** *Proceedings of the National Academy of Sciences* 2005; 102:15545-15550.

204. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. **Gene Ontology: tool for the unification of biology**. *Nature Genetics* 2000; 25:25-29.

205. Benjamini Y, Hochberg Y. **Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing**. *Journal of the Royal Statistical Society Series B (Methodological)* 1995; 57:289-300.

206. Willer CJ, Li Y, Abecasis GR. **METAL: Fast and efficient meta-analysis of genomewide association scans**. *Bioinformatics* 2010; 26:2190-2191.

207. Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, Gylfason A, et al. **Large-scale whole-genome sequencing of the Icelandic population**. *Nature Genetics* 2015; 47:435-444.

208. Steinberg J, Ritchie GRS, Roumeliotis TI, Jayasuriya RL, Clark MJ, Brooks RA, et al. **Integrative epigenomics, transcriptomics and proteomics of patient chondrocytes reveal genes and pathways involved in osteoarthritis.** *Scientific Reports* 2017; 7(1):8935.

209. Mankin HJ, Dorfman H, Lippiello L, Zarins A. **Biochemical and Metabolic Abnormalities in Articular Cartilage from Osteo-Arthritic Human Hips: II. CORRELATION OF MORPHOLOGY WITH BIOCHEMICAL AND METABOLIC DATA**. *The Journal of Bone and Joint Surgery* 1971; 53:523-537.

210. Pearson RG, Kurien T, Shu KSS, Scammell BE. **Histopathology grading systems for characterisation of human knee osteoarthritis – reproducibility, variability, reliability, correlation, and validity**. *Osteoarthritis and Cartilage* 2011; 19:324-331.

211. Rivadeneira F, Styrkársdottir U, Estrada K, Halldórsson BV, Hsu Y-H, Richards JB, et al. **Twenty bone-mineral-density loci identified by large-scale meta-analysis of genome-wide association studies**. *Nature Genetics* 2009; 41:1199-1206.

212. Hochberg MC, Lethbridge-Cejku M, Tobin JD. **Bone mineral density and osteoarthritis: Data from the Baltimore Longitudinal Study of Aging**. *Osteoarthritis and Cartilage* 2004; 12:45-48.

213. Harada A, Okuizumi H, Miyagi N, Genda E. **Correlation between bone mineral density and intervertebral disc degeneration.** *Spine* 1998; 23:857-861.

214. Grams AE, Rehwald R, Bartsch A, Honold S, Freyschlag CF, Knoflach M, et al. **Correlation between degenerative spine disease and bone marrow density: a retrospective investigation**. *BMC Medical Imaging* 2016; 16:17.

215. Pye SR, Reid DM, Adams JE, Silman AJ, O'Neill TW. **Radiographic features of lumbar disc degeneration and bone mineral density in men and women.** *Annals of the Rheumatic Diseases* 2006; 65:234-238.

216. Miyakoshi N, Itoi E, Murai H, Wakabayashi I, Ito H, Minato T. **Inverse Relation Between Osteoporosis and Spondylosis in Postmenopausal Women As Evaluated by Bone Mineral Density and Semiquantitative Scoring of Spinal Degeneration**. *Spine* 2003; 28:492-495.

217. Rand T, Seidl G, Kainberger F, Resch A, Hittmair K, Schneider B, et al. **Impact of spinal degenerative changes on the evaluation of bone mineral density with dual energy X-ray absorptiometry (DXA).** *Calcified tissue international* 1997; 60:430-433.
218. Kemp JP, Medina-Gomez C, Estrada K, St Pourcain B, Heppe DHM, Warrington NM, et al. **Phenotypic dissection of bone mineral density reveals skeletal site specificity and facilitates the identification of novel loci in the genetic regulation of bone mass attainment.** *PLoS genetics* 2014; 10:e1004423.
219. Magnusson K, Scurrah K, Ystrom E, Ørstavik RE, Nilsen T, Steingrímsdóttir ÓA, et al. **Genetic factors contribute more to hip than knee surgery due to osteoarthritis – a population-based twin registry study of joint arthroplasty**. *Osteoarthritis and Cartilage* 2016.
220. Valdes AM, Spector TD. **Genetic epidemiology of hip and knee osteoarthritis.** *Nature Reviews Rheumatology* 2011; 7:23-32.
221. Blagojevic M, Jinks C, Jeffery A, Jordan KP. **Risk factors for onset of osteoarthritis of the knee in older adults: a systematic review and meta-analysis.** *Osteoarthritis and Cartilage* 2010; 18:24-33.
222. Shimomura Y, Agalliu D, Vonica A, Luria V, Wajid M, Baumer A, et al. **APCDD1 is a novel Wnt inhibitor mutated in hereditary hypotrichosis simplex**. *Nature* 2010; 464:1043-1047.
223. Rodriguez-Fontenla C, Calaza M, Evangelou E, Valdes AM, Arden N, Blanco FJ, et al. **Assessment of osteoarthritis candidate genes in a meta-analysis of nine genome-wide association studies.** *Arthritis and Rheumatology* 2014; 66:940-949.
224. Styrkarsdottir U, Thorleifsson G, Sulem P, Gudbjartsson DF, Sigurdsson A, Jonasdottir A, et al. **Nonsense mutation in the LGR4 gene is associated with several human diseases and other traits.** *Nature* 2013; 497:517-520.
225. Fukuda T, Takeda S, Xu R, Ochi H, Sunamura S, Sato T, et al. **Sema3A regulates bone-mass accrual through sensory innervations.** *Nature* 2013; 497:490-493.
226. Ma P, Yang X, Kong Q, Li C, Yang S, Li Y, et al. **The ubiquitin ligase RNF220 enhances canonical Wnt signaling through USP7-mediated deubiquitination of beta-catenin**. *Molecular and cellular biology* 2014; 34(23):4355-4366.
227. Rivadeneira F, Mäkitie O. **Osteoporosis and Bone Mass Disorders: From Gene Pathways to Treatments.** *Trends in endocrinology and metabolism: TEM* 2016; 27:262-281.
228. Alliston T, Choy L, Ducy P, Karsenty G, Derynck R, Akiyoshi S, et al. **TGF-beta-induced repression of CBFA1 by Smad3 decreases cbfa1 and osteocalcin expression and inhibits osteoblast differentiation.** *The EMBO journal* 2001; 20:2254-2272.
229. Chen CG, Thuillier D, Chin EN, Alliston T. **Chondrocyte-intrinsic Smad3 represses Runx2-inducible matrix metalloproteinase 13 expression to maintain articular cartilage and prevent osteoarthritis**. *Arthritis and Rheumatism* 2012; 64:3278-3289.
230. van de Laar IMBH, Oldenburg RA, Pals G, Roos-Hesselink JW, de Graaf BM, Verhagen JMA, et al. **Mutations in SMAD3 cause a syndromic form of aortic aneurysms and dissections with early-onset osteoarthritis.** *Nature Genetics* 2011; 43:121-126.
231. Valdes AM, Spector TD, Tamm A, Kisand K, Doherty SA, Dennison EM, et al. **Genetic variation in the SMAD3 gene is associated with hip and knee osteoarthritis.** *Arthritis and Rheumatism* 2010; 62:2347-2352.
232. Raine EVA, Reynard LN, van de Laar IMBH, Bertoli-Avella AM, Loughlin J. **Identification and analysis of a SMAD3 cis-acting eQTL operating in primary osteoarthritis and in**

the aneurysms and osteoarthritis syndrome. *Osteoarthritis and cartilage* 2014; 22:698-705.

233. Aref-Eshghi E, Liu M, Razavi-Lopez SB, Hirasawa K, Harper PE, Martin G, et al. **SMAD3 Is Upregulated in Human Osteoarthritic Cartilage Independent of the Promoter DNA Methylation.** *The Journal of rheumatology* 2016; 43:388-394.

234. Lonsdale J, Thomas J, Salvatore M, Phillips R, Lo E, Shad S, et al. **The Genotype-Tissue Expression (GTEx) project**. *Nature Genetics* 2013; 45(6):580-585.

235. Ren G, Krawetz RJ. **Biochemical Markers for the Early Identification of Osteoarthritis: Systematic Review and Meta-Analysis**. *Mol Diagn Ther* 2018; 22(6):671-682.

236. Wang K, Xu J, Hunter DJ, Ding C. **Investigational drugs for the treatment of osteoarthritis**. *Expert Opin Investig Drugs* 2015; 24(12):1539-1556.

237. Barra F, Scala C, Mais V, Guerriero S, Ferrero S. **Investigational drugs for the treatment of endometriosis, an update on recent developments**. *Expert Opin Investig Drugs* 2018; 27(5):445-458.

238. Lynch TS, O'Connor M, Minkara AA, Westermann RW, Rosneck JT. **Biomarkers for Femoroacetabular Impingement and Hip Osteoarthritis: A Systematic Review and Meta-analysis**. *Am J Sports Med* 2018:363546518803360.

239. Panoutsopoulou K, Thiagarajah S, Zengini E, Day-Williams AG, Ramos YF, Meessen JM, et al. **Radiographic endophenotyping in hip osteoarthritis improves the precision of genetic association analysis**. *Ann Rheum Dis* 2017; 76(7):1199-1206.

240. Zhu Z, Li J, Ruan G, Wang G, Huang C, Ding C. **Investigational drugs for the treatment of osteoarthritis, an update on recent developments**. *Expert Opin Investig Drugs* 2018; 27(11):881-900.

241. Styrkarsdottir U, Lund SH, Thorleifsson G, Zink F, Stefansson OA, Sigurdsson JK, et al. **Meta-analysis of Icelandic and UK data sets identifies missense variants in SMO, IL11, COL11A1 and 13 more new loci associated with osteoarthritis**. *Nat Genet* 2018; 50(12):1681-1687.

242. Kwak EA, Lee NY. **Synergetic roles of TGF-beta signaling in tissue engineering**. *Cytokine* 2019; 115:60-63.

243. Tellegen AR, Rudnik-Jansen I, Pouran B, de Visser HM, Weinans HH, Thomas RE, et al. **Controlled release of celecoxib inhibits inflammation, bone cysts and osteophyte formation in a preclinical model of osteoarthritis**. *Drug Deliv* 2018; 25(1):1438-1447.

244. Owen MJ, Sawa A, Mortensen PB. **Schizophrenia**. *The Lancet* 2016; 388:86-97.

245. Lieberman JA, Perkins D, Belger A, Chakos M, Jarskog F, Boteva K, et al. **The early stages of schizophrenia: speculations on pathogenesis, pathophysiology, and therapeutic approaches.** *Biological psychiatry* 2001; 50:884-897.

246. Rouillon F, Sorbara F. **Schizophrenia and diabetes: Epidemiological data**. *European Psychiatry* 2005; 20:S345-S348.

247. Franks PW, McCarthy MI. **Exposing the exposures responsible for type 2 diabetes and obesity**. *Science* 2016; 354.

248. Lin PI, Shuldiner AR. **Rethinking the genetic basis for comorbidity of schizophrenia and type 2 diabetes**. *Schizophrenia Research* 2010; 123:234-243.

249. Suvisaari J, Keinänen J, Eskelinen S, Mantere O. **Diabetes and Schizophrenia.** *Current Diabetes Reports* 2016; 16:16.

250. Holt RIG, Mitchell AJ. **Diabetes mellitus and severe mental illness: mechanisms and clinical implications.** *Nature Reviews Endocrinology* 2015; 11:79-89.

251. Young SL, Taylor M, Lawrie SM. **"First do no harm." A systematic review of the prevalence and management of antipsychotic adverse effects**. *Journal of Psychopharmacology* 2015; 29:353-362.

252. Alberti KGMM, Zimmet P, Shaw J. **Metabolic syndrome-a new world-wide definition. A Consensus Statement from the International Diabetes Federation**. *Diabetic Medicine* 2006; 23:469-480.

253. Vancampfort D, Correll CU, Galling B, Probst M, De Hert M, Ward PB, et al. **Diabetes mellitus in people with schizophrenia, bipolar disorder and major depressive disorder: a systematic review and large scale meta-analysis.** *World psychiatry : official journal of the World Psychiatric Association (WPA)* 2016; 15:166-174.

254. Correll CU, Detraux J, De Lepeleire J, De Hert M. **Effects of antipsychotics, antidepressants and mood stabilizers on risk for physical diseases in people with schizophrenia, depression and bipolar disorder**. *World Psychiatry* 2015; 14:119-136.

255. Smith M, Hopkins D, Peveler RC, Holt RIG, Woodward M, Ismail K. **First- V. second-generation antipsychotics and risk for diabetes in schizophrenia: Systematic review and meta-analysis**. *British Journal of Psychiatry* 2008; 192:406-411.

256. Zhang J-P, Lencz T, Geisler S, DeRosse P, Bromet EJ, Malhotra AK. **Genetic variation in BDNF is associated with antipsychotic treatment resistance in patients with schizophrenia.** *Schizophrenia research* 2013; 146:285-288.

257. Lieberman JA, Stroup TS, McEvoy JP, Swartz MS, Rosenheck RA, Perkins DO, et al. **Effectiveness of antipsychotic drugs in patients with chronic schizophrenia.** *The New England journal of medicine* 2005; 353:1209-1223.

258. Prabakaran S, Swatton JE, Ryan MM, Huffaker SJ, Huang JT-J, Griffin JL, et al. **Mitochondrial dysfunction in schizophrenia: evidence for compromised brain metabolism and oxidative stress.** *Molecular Psychiatry* 2004; 9:684-697.

259. van Beveren NJM, Schwarz E, Noll R, Guest PC, Meijer C, de Haan L, et al. **Evidence for disturbed insulin and growth hormone signaling as potential risk factors in the development of schizophrenia**. *Translational Psychiatry* 2014; 4:e430.

260. Rajkumar AP, Horsdal HT, Wimberley T, Cohen D, Mors O, Børglum AD, et al. **Endogenous and Antipsychotic-Related Risks for Diabetes Mellitus in Young People With Schizophrenia: A Danish Population-Based Cohort Study**. *American Journal of Psychiatry* 2017; 174:686-694.

261. Pillinger T, Beck K, Gobjila C, Donocik JG, Jauhar S, Howes OD. **Impaired Glucose Homeostasis in First-Episode Schizophrenia: A Systematic Review and Meta-analysis.** *JAMA Psychiatry* 2017; 74:261.

262. Fuchsberger C, Flannick J, Teslovich TM, Mahajan A, Agarwala V, Gaulton KJ, et al. **The genetic architecture of type 2 diabetes.** *Nature* 2016; 536:41-47.

263. Morris ADP, Voight BF, Teslovich TM, Ferreira T, Segrè AV, Steinthorsdottir V, et al. **Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes**. *Nature genetics* 2012; 44:981-990.

264. Scott RA, Scott LJ, Mägi R, Marullo L, Gaulton KJ, Kaakinen M, et al. **An Expanded Genome-Wide Association Study of Type 2 Diabetes in Europeans**. *Diabetes* 2017; 66:2888-2902.

265. Ripke S, Sanders A, Kendler K, Levinson D, Sklar P, Holmans P, et al. **Genome-wide association study identifies five new schizophrenia loci**. *Nature Genetics* 2011; 43:969-976.

266. Ripke S, O'Dushlaine C, Chambert K, Moran JL, Kahler AK, Akterin S, et al. **Genome-wide association analysis identifies 13 new risk loci for schizophrenia**. *Nature Genetics* 2013; 45:1150-1159.

267. Morris AP. **Progress in defining the genetic contribution to type 2 diabetes susceptibility**. *Current Opinion in Genetics & Development* 2018; 50:41-51.

268. Hackinger S, Prins B, Mamakou V, Zengini E, Marouli E, Brcic L, et al. **Evidence for genetic contribution to the increased risk of type 2 diabetes in schizophrenia**. *Transl Psychiatry* 2018; 8(1):252.

269. Association. AP, Force. APAD-T. **Diagnostic and statistical manual of mental disorders : DSM-5.** 2013:947.

270. Association AD. **Standards of Medical Care in Diabetes—2014**. *Diabetes Care* 2014; 37:S14-S80.

271. Mamakou V, Hackinger S, Zengini E, Tsompanaki E, Marouli E, Serafetinidis I, et al. **Combination therapy as a potential risk factor for the development of type 2 diabetes in patients with schizophrenia: the GOMAP study**. *BMC Psychiatry* 2018; 18(1):249.

272. Panoutsopoulou K, Hatzikotoulas K, Xifara DK, Colonna V, Farmaki A-E, Ritchie GRS, et al. **Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants.** *Nature Communications* 2014; 5:5345.

273. Vergeti M. **Pontic Greeks from Asia Minor and the Soviet Union: Problems of Integration in Modern Greece**. *Journal of Refugee Studies* 1991; 4(4):382-394.

274. Ntalla I, Panoutsopoulou K, Vlachou P, Southam L, Rayner NW, Zeggini E, et al. **Replication of Established Common Genetic Variants for Adult BMI and Childhood Obesity in Greek Adolescents : The TEENAGE Study**. *Annals of Human Genetics* 2013; 77:268-274.

275. Southam L, Gilly A, Süveges D, Farmaki A-E, Schwartzentruber J, Tachmazidou I, et al. **Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits**. *Nature Communications* 2017; 8:15606.

276. Ntalla I, Giannakopoulou M, Vlachou P, Giannitsopoulou K, Gkesou V, Makridi C, et al. **Body composition and eating behaviours in relation to dieting involvement in a sample of urban Greek adolescents from the TEENAGE (TEENs of Attica : Genes & Environment) study**. *Public Health Nutrition* 2013; 17:561-568.

277. Delaneau O, Marchini J. **Integrating sequence and array data to create an improved 1000 Genomes Project haplotype reference panel.** *Nature communications* 2014; 5:3934.

278. Walter K, Min JL, Huang J, Crooks L, Memari Y, McCarthy S, et al. **The UK10K project identifies rare variants in health and disease**. *Nature* 2015; 526:82-90.

279. Gauderman WJ. **Sample size requirements for association studies of gene-gene interaction.** *American journal of epidemiology* 2002; 155:478-484.

280. Voight BF, Scott LJ, Steinthorsdottir V, Morris ADP, Dina C, Welch RP, et al. **Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis.** *Nature genetics* 2010; 42:579-589.

281. Yamada K, Hattori E, Iwayama Y, Toyota T, Iwata Y, Suzuki K, et al. **Population-dependent contribution of the major histocompatibility complex region to schizophrenia susceptibility**. *Schizophr Res* 2015; 168(1-2):444-449.

282. Yamauchi T, Hara K, Maeda S, Yasuda K, Takahashi A, Horikoshi M, et al. **A genome-wide association study in the Japanese population identifies susceptibility loci for type 2 diabetes at UBE2E2 and C2CD4A-C2CD4B.** *Nature genetics* 2010; 42:864-868.

283. Mira MT, Alcaïs A, Van Thuc N, Moraes MO, Di Flumeri C, Hong Thai V, et al. **Susceptibility to leprosy is associated with PARK2 and PACRG**. *Nature* 2004; 427:636-640.

284. Padmanabhan JL, Nanda P, Tandon N, Mothi SS, Bolo N, McCarroll S, et al. **Polygenic risk for type 2 diabetes mellitus among individuals with psychosis and their relatives.** *Journal of Psychiatric Research* 2016; 77:52-58.

285. Stringer S, Kahn RS, de Witte LD, Ophoff RA, Derks EM. **Genetic liability for schizophrenia predicts risk of immune disorders.** *Schizophrenia Research* 2014; 159:347-352.

286. (IMSGC) IMSGC. **IL12A, MPHOSPH9/CDK2AP1 and RGS1 are novel multiple sclerosis susceptibility loci.** *Genes and Immunity* 2010; 11:397-405.

287. Paul L, Walker EM, Drosos Y, Cyphert HA, Neale G, Stein R, et al. **Lack of Prox1 Downregulation Disrupts the Expansion and Maturation of Postnatal Murine β-Cells.** *Diabetes* 2016; 65:687-698.

288. Holzmann J, Hennchen M, Rohrer H. **Prox1 identifies proliferating neuroblasts and nascent neurons during neurogenesis in sympathetic ganglia**. *Developmental Neurobiology* 2015; 75:1352-1367.

289. Wang X, Strizich G, Hu Y, Wang T, Kaplan RC, Qi Q. **Genetic markers of type 2 diabetes: Progress in genome-wide association studies and clinical application for risk prediction**. *Journal of Diabetes* 2016; 8:24-35.

290. Keating BJ. **Advances in risk prediction of type 2 diabetes: integrating genetic scores with Framingham risk models.** *Diabetes* 2015; 64:1495-1497.

291. Talmud PJ, Cooper JA, Morris RW, Dudbridge F, Shah T, Engmann J, et al. **Sixty-five common genetic variants and prediction of type 2 diabetes**. *Diabetes* 2015; 64(5):1830-1840.

292. Läll K, Mägi R, Morris A, Metspalu A, Fischer K. **Personalized risk prediction for type 2 diabetes: the potential of genetic risk scores**. *Genetics in Medicine* 2017; 19:322-329.

293. De Hert M, Detraux J, van Winkel R, Yu W, Correll CU. **Metabolic and cardiovascular adverse effects associated with antipsychotic drugs**. *Nature Reviews Endocrinology* 2011; 8:114-126.

294. Fibiger HC. **Psychiatry, the pharmaceutical industry, and the road to better therapeutics**. *Schizophr Bull* 2012; 38(4):649-650.

295. Ryan MCM, Collins P, Thakore JH. **Impaired Fasting Glucose Tolerance in First-Episode, Drug-Naive Patients With Schizophrenia**. *American Journal of Psychiatry* 2003; 160:284-289.

296. Fehsel K, Löffler S. **First-episode psychosis and abnormal glycaemic control**. *The Lancet Psychiatry* 2017; 4:23-24.

297. Steiner J, Berger M, Guest PC, Dobrowolny H, Westphal S, Schiltz K, et al. **Assessment of Insulin Resistance Among Drug-Naive Patients With First-Episode Schizophrenia in the Context of Hormonal Stress Axis Activation**. *JAMA Psychiatry* 2017; 74(9):968-970.

298. Herberth M, Koethe D, Cheng TM, Krzyszton ND, Schoeffmann S, Guest PC, et al. **Impaired glycolytic response in peripheral blood mononuclear cells of first-onset antipsychotic-naive schizophrenia patients**. *Mol Psychiatry* 2011; 16(8):848-859.

299. Lago SG, Bahn S. **Clinical Trials and Therapeutic Rationale for Drug Repurposing in Schizophrenia**. *ACS Chem Neurosci* 2018.

300. Hatzikotoulas K, Gilly A, Zeggini E. **Using population isolates in genetic association studies**. *Briefings in Functional Genomics* 2014; 13(5):371-377.

301. Peltonen L, Palotie A, Lange K. **Use of population isolates for mapping complex traits**. *Nature Reviews Genetics* 2000; 1(3):182-190.

302. Xue Y, Mezzavilla M, Haber M, McCarthy S, Chen Y, Narasimhan V, et al. **Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations**. *Nature Communications* 2017; 8:15927-15927.

303. Andersen MK, Grarup N, Moltke I, Albrechtsen A, Hansen T. **Genetic architecture of obesity and related metabolic traits — recent insights from isolated populations**. *Current Opinion in Genetics and Development* 2018; 50:74-78.

304. Beall CM, Cavalleri GL, Deng L, Elston RC, Gao Y, Knight J, et al. **Natural selection on EPAS1 (HIF2alpha) associated with low hemoglobin concentration in Tibetan highlanders**. *Proc Natl Acad Sci U S A* 2010; 107(25):11459-11464.

305. Yi X, Liang Y, Huerta-Sanchez E, Jin X, Cuo ZX, Pool JE, et al. **Sequencing of 50 human exomes reveals adaptation to high altitude**. *Science* 2010; 329(5987):75-78.

306. Hackinger S, Kraaijenbrink T, Xue Y, Mezzavilla M, Asan, van Driem G, et al. **Wide distribution and altitude correlation of an archaic high-altitude-adaptive EPAS1 haplotype in the Himalayas**. *Hum Genet* 2016; 135(4):393-402.

307. Assarsson E, Lundberg M, Holmquist G, Björkesten J, Bucht Thorsen S, Ekman D, et al. **Homogenous 96-Plex PEA Immunoassay Exhibiting High Sensitivity, Specificity, and Excellent Scalability**. *PLoS ONE* 2014; 9(4):e95192-e95192.

308. Landegren U, Al-Amin RA, Björkesten J. **A myopic perspective on the future of protein diagnostics**. *New Biotechnology* 2018.

309. Folkersen L, Fauman E, Sabater-Lleal M, Strawbridge RJ, Frånberg M, Sennblad B, et al. **Mapping of 79 loci for 83 plasma protein biomarkers in cardiovascular disease**. *PLOS Genetics* 2017; 13:e1006706.

310. Suhre K, Arnold M, Bhagwat AM, Cotton RJ, Engelke R, Raffler J, et al. **Connecting genetic risk to disease end points through the human blood plasma proteome**. *Nature Communications* 2017; 8:14357-14357.

311. Benjamin EJ, Dupuis J, Larson MG, Lunetta KL, Booth SL, Govindaraju DR, et al. **Genome-wide association with select biomarker traits in the Framingham Heart Study**. *BMC Medical Genetics* 2007; 8:S11.

312. Mittelstrass K, Ried JS, Yu Z, Krumsiek J, Gieger C, Prehn C, et al. **Discovery of sexual dimorphisms in metabolic and genetic biomarkers**. *PLoS Genetics* 2011; 7.

313. Shen X, Klarić L, Sharapov S, Mangino M, Ning Z, Wu D, et al. **Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation**. *Nature Communications* 2017; 8(1):447-447.

314. Inouye M, Ripatti S, Kettunen J, Lyytikäinen L-P, Oksala N, Laurila P-P, et al. **Novel Loci for Metabolic Networks and Multi-Tissue Expression Studies Reveal Genes for Atherosclerosis**. *PLoS Genetics* 2012; 8(8):e1002907-e1002907.

315. Astle W, Balding DJ. **Population Structure and Cryptic Relatedness in Genetic Association Studies.** *Statistical Science* 2009; 24(4):451-471.

316. Gilly A, Suveges D, Kuchenbaecker K, Pollard M, Southam L, Hatzikotoulas K, et al. **Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits.** *Nat Commun* 2018; 9(1):4674.

317. Lundberg M, Eriksson A, Tran B, Assarsson E, Fredriksson S. **Homogeneous antibody-based proximity extension assays provide sensitive and specific detection of low-abundant proteins in human blood.** *Nucleic Acids Research* 2011; 39(15):e102-e102.

318. Dahl A, Iotchkova V, Baud A, Johansson Å, Gyllensten U, Soranzo N, et al. **A multiple-phenotype imputation method for genetic studies.** *Nature Genetics* 2016; Accepted f.

319. **Olink ® NPX Manager User Guide**. In.

320. Bilotta FL, Arcidiacono B, Messineo S, Greco M, Chiefari E, Britti D, et al. **Insulin and osteocalcin: further evidence for a mutual cross-talk.** *Endocrine* 2017.

321. Kalra SP, Dube MG, Iwaniec UT. **Leptin increases osteoblast-specific osteocalcin release through a hypothalamic relay.** *Peptides* 2009; 30(5):967-973.

322. Booth SL, Centi A, Smith SR, Gundberg C. **The role of osteocalcin in human glucose metabolism: marker or mediator?** *Nature Reviews Endocrinology* 2013; 9:43-55.

323. Otani T, Mizokami A, Hayashi Y, Gao J, Mori Y, Nakamura S, et al. **Signaling pathway for adiponectin expression in adipocytes by osteocalcin.** *Cellular Signalling* 2015; 27:532-544.

324. Cheverud JM. **A simple correction for multiple comparisons in interval mapping genome scans.** *Heredity* 2001; 87(1):52-58.

325. Li J, Ji L. **Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix.** *Heredity* 2005; 95(3):221-227.

326. Tangseefa P, Martin SK, Fitter S, Baldock PA, Proud CG, Zannettino ACW. **Osteocalcin-dependent regulation of glucose metabolism and fertility: Skeletal implications for the development of insulin resistance.** *Journal of Cellular Physiology* 2017.

327. Wei J, Karsenty G. **An overview of the metabolic functions of osteocalcin.** *Rev Endocr Metab Disord* 2015; 16(2):93-98.

328. Benjamin EJ, Dupuis J, Larson MG, Lunetta KL, Booth SL, Govindaraju DR, et al. **Genome-wide association with select biomarker traits in the Framingham Heart Study.** *BMC Medical Genetics* 2007; 8(Suppl 1):S11-S11.

329. Morgan R, Feng G, Pandha HS. **Abstract A193: Transmembrane protein TMEM92 as a novel target in prostate cancer.** *Molecular Cancer Therapeutics* 2013; 12(11_Supplement):A193-A193.

330. Kaito K, Otsubo H, Usui N, Yoshida M, Tanno J, Kurihara E, et al. **Platelet size deviation width, platelet large cell ratio, and mean platelet volume have sufficient sensitivity and specificity in the diagnosis of immune thrombocytopenia.** *Br J Haematol* 2005; 128(5):698-702.

331. Tamura K, Yu J, Hata T, Suenaga M, Shindo K, Abe T, et al. **Mutations in the pancreatic secretory enzymes <i>CPA1</i> and <i>CPB1</i> are associated with pancreatic cancer.** *Proceedings of the National Academy of Sciences* 2018; 115(18):201720588-201720588.

332. Graff M, Scott RA, Justice AE, Young KL, Feitosa MF, Barata L, et al. **Genome-wide physical activity interactions in adiposity — A meta-analysis of 200,452 adults.** *PLOS Genetics* 2017; 13(4):e1006528-e1006528.

333. Spracklen CN, Chen P, Kim YJ, Wang X, Cai H, Li S, et al. **Association analyses of East Asian individuals and trans-ancestry analyses with European individuals reveal new loci associated with cholesterol and triglyceride levels**. *Human Molecular Genetics* 2017; 26(9):1770-1784.

334. Ohlin AK, Norlund L, Marlar RA. **Thrombomodulin gene variations and thromboembolic disease.** *Thrombosis and haemostasis* 1997; 78(1):396-400.

335. Delvaeye M, Noris M, De Vriese A, Esmon CT, Esmon NL, Ferrell G, et al. **Thrombomodulin mutations in atypical hemolytic-uremic syndrome**. *N Engl J Med* 2009; 361(4):345-357.

336. Herberich SE, Klose R, Moll I, Yang W-J, Wüstehube-Lausch J, Fischer A. **ANKS1B Interacts with the Cerebral Cavernous Malformation Protein-1 and Controls Endothelial Permeability but Not Sprouting Angiogenesis**. *PLOS ONE* 2015; 10(12):e0145304-e0145304.

337. Stockelberg D, Hou M, Rydberg L, Kutti J, Wadenvik H. **Evidence for an expression of blood group A antigen on platelet glycoproteins IV and V**. *Transfusion Medicine* 1996; 6(3):243-248.

338. Sobocka MB, Sobocki T, Banerjee P, Weiss C, Rushbrook JI, Norin AJ, et al. **Cloning of the human platelet F11 receptor: a cell adhesion molecule member of the immunoglobulin superfamily involved in platelet aggregation**. *Blood* 2000; 95(8):2600-2609.

339. Lee KJ, Carlin JB. **Multiple imputation for missing data: fully conditional specification versus multivariate normal imputation**. *Am J Epidemiol* 2010; 171(5):624-632.

340. van Buuren S. **Multiple imputation of discrete and continuous data by fully conditional specification**. *Stat Methods Med Res* 2007; 16(3):219-242.

341. Carpenter JR, Kenward MG, White IR. **Sensitivity analysis after multiple imputation under missing at random: a weighting approach**. *Stat Methods Med Res* 2007; 16(3):259-275.

342. Langfelder P, Zhang B, Horvath S. **Defining clusters from a hierarchical cluster tree: the Dynamic Tree Cut package for R**. *Bioinformatics* 2008; 24(5):719-720.

343. Gamazon ER, Wheeler HE, Shah KP, Mozaffari SV, Aquino-Michaels K, Carroll RJ, et al. **A gene-based association method for mapping traits using reference transcriptome data.** *Nature genetics* 2015; 47:1091-1098.

344. Ongen H, Brown AA, Delaneau O, Panousis NI, Nica AC, Dermitzakis ET, et al. **Estimating the causal tissues for complex traits and diseases**. *Nature Genetics* 2017; 49:1676-1683.

345. Willer CJ, Schmidt EM, Sengupta S, Peloso GM, Gustafsson S, Kanoni S, et al. **Discovery and refinement of loci associated with lipid levels**. *Nat Genet* 2013; 45(11):1274-1283.

346. Hyman SE. **The diagnosis of mental disorders: the problem of reification.** *Annual review of clinical psychology* 2010; 6:155-179.

347. Craddock N, Owen MJ. **Rethinking psychosis: the disadvantages of a dichotomous classification now outweigh the advantages**. *World Psychiatry* 2007; 6(2):84-91.

348. Owen MJ. **New approaches to psychiatric diagnostic classification**. *Neuron* 2014; 84(3):564-571.

349. Craddock N, Owen MJ. **The beginning of the end for the Kraepelinian dichotomy**. *Br J Psychiatry* 2005; 186:364-366.

350. Joyce PR. **Age of onset in bipolar affective disorder and misdiagnosis as schizophrenia**. *Psychol Med* 1984; 14(1):145-149.
351. Wray NR, Lee SH, Kendler KS. **Impact of diagnostic misclassification on estimation of genetic correlations using genome-wide genotypes**. *Eur J Hum Genet* 2012; 20(6):668-674.
352. Laursen TM, Agerbo E, Pedersen CB. **Bipolar disorder, schizoaffective disorder, and schizophrenia overlap: a new comorbidity index**. *J Clin Psychiatry* 2009; 70(10):1432-1438.
353. Bromet EJ, Kotov R, Fochtmann LJ, Carlson GA, Tanenberg-Karant M, Ruggero C, et al. **Diagnostic shifts during the decade following first admission for psychosis**. *Am J Psychiatry* 2011; 168(11):1186-1194.
354. Ahmad S, Sundaramoorthy E, Arora R, Sen S, Karthikeyan G, Sengupta S. **Progressive degradation of serum samples limits proteomic biomarker discovery**. *Anal Biochem* 2009; 394(2):237-242.
355. Lotta LA, Scott RA, Sharp SJ, Burgess S, Luan J, Tillin T, et al. **Genetic Predisposition to an Impaired Metabolism of the Branched-Chain Amino Acids and Risk of Type 2 Diabetes: A Mendelian Randomisation Analysis**. *PLoS Med* 2016; 13(11):e1002179.
356. Corbin LJ, Richmond RC, Wade KH, Burgess S, Bowden J, Smith GD, et al. **Body mass index as a modifiable risk factor for type 2 diabetes: Refining and understanding causal estimates using Mendelian randomisation.** *Diabetes* 2016.
357. Timpson NJ, Lawlor DA, Harbord RM, Gaunt TR, Day IN, Palmer LJ, et al. **C-reactive protein and its role in metabolic syndrome: mendelian randomisation study**. *Lancet* 2005; 366(9501):1954-1959.
358. Katan MB. **Apolipoprotein E isoforms, serum cholesterol, and cancer.** *Lancet (London, England)* 1986; 1:507-508.
359. Gusev A, Ko A, Shi H, Bhatia G, Chung W, Penninx BWJH, et al. **Integrative approaches for large-scale transcriptome-wide association studies**. *Nature Genetics* 2016; 48:245-252.
360. Gamazon ER, Segrè AV, van de Bunt M, Wen X, Xi HS, Hormozdiari F, et al. **Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation**. *Nature Genetics* 2018; 50(7):956-967.
361. Amin V, Harris RA, Onuchic V, Jackson AR, Charnecki T, Paithankar S, et al. **Epigenomic footprints across 111 reference epigenomes reveal tissue-specific epigenetic regulation of lincRNAs**. *Nat Commun* 2015; 6:6370.
362. Stunnenberg HG, International Human Epigenome C, Hirst M. **The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery**. *Cell* 2016; 167(5):1145-1149.
363. Wijmenga C, Zhernakova A. **The importance of cohort studies in the post-GWAS era**. *Nature Genetics* 2018; 50(3):322-328.
364. Houle D, Govindaraju DR, Omholt S. **Phenomics: the next challenge**. *Nature Reviews Genetics* 2010; 11:855-866.
365. Tigchelaar EF, Zhernakova A, Dekens JA, Hermes G, Baranska A, Mujagic Z, et al. **Cohort profile: LifeLines DEEP, a prospective, general population cohort study in the northern Netherlands: study design and baseline characteristics**. *BMJ Open* 2015; 5(8):e006772.

366. O'Donovan MC, Owen MJ. **The implications of the shared genetics of psychiatric disorders**. *Nature Medicine* 2016; 22.
367. Flint J, Kendler KS. **The genetics of major depression.** *Neuron* 2014; 81:484-503.
368. Tesli M, Espeseth T, Bettella F, Mattingsdal M, Aas M, Melle I, et al. **Polygenic risk score and the psychosis continuum model**. *Acta Psychiatrica Scandinavica* 2014; 130(4):311-317.
369. Castaño-Betancourt MC, Evans DS, Ramos YFM, Boer CG, Metrustry S, Liu Y, et al. **Novel Genetic Variants for Cartilage Thickness and Hip Osteoarthritis.** *PLoS Genetics* 2016; 12:e1006260.
370. Castaño-Betancourt MC, Rivadeneira F, Bierma-Zeinstra S, Kerkhof HJM, Hofman A, Uitterlinden AG, et al. **Bone parameters across different types of hip osteoarthritis and their relationship to osteoporotic fracture risk.** *Arthritis and rheumatism* 2013; 65:693-700.
371. Bilder RM, Sabb FW, Cannon TD, London ED, Jentsch JD, Parker DS, et al. **Phenomics: the systematic study of phenotypes on a genome-wide scale**. *Neuroscience* 2009; 164(1):30-42.
372. Nielsen DL, Andersson M, Kamby C. **HER2-targeted therapy in breast cancer. Monoclonal antibodies and tyrosine kinase inhibitors**. *Cancer Treat Rev* 2009; 35(2):121-136.
373. Netea MG, Joosten LAB, Li Y, Kumar V, Oosting M, Smeekens S, et al. **Understanding human immune function using the resources from the Human Functional Genomics Project**. *Nature Medicine* 2016; 22(8):831-833.
374. Bakker OB, Aguirre-Gamboa R, Sanna S, Oosting M, Smeekens SP, Jaeger M, et al. **Integration of multi-omics data and deep phenotyping enables prediction of cytokine responses**. *Nature Immunology* 2018; 19(7):776-786.
375. Short PJ, McRae JF, Gallone G, Sifrim A, Won H, Geschwind DH, et al. **De novo mutations in regulatory elements in neurodevelopmental disorders**. *Nature* 2018; 555(7698):611-616.
376. Jin SC, Homsy J, Zaidi S, Lu Q, Morton S, DePalma SR, et al. **Contribution of rare inherited and de novo variants in 2,871 congenital heart disease probands**. *Nature Genetics* 2017; 49(11):1593-1601.
377. Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, Brown MA, et al. **10 Years of GWAS Discovery: Biology, Function, and Translation**. *The American Journal of Human Genetics* 2017; 101(1):5-22.
378. Demenais F, Margaritte-Jeannin P, Barnes KC, Cookson WOC, Altmüller J, Ang W, et al. **Multiancestry association study identifies new asthma risk loci that colocalize with immune-cell enhancer marks**. *Nature Genetics* 2018; 50:42-53.
379. Okada Y, Wu D, Trynka G, Raj T, Terao C, Ikari K, et al. **Genetics of rheumatoid arthritis contributes to biology and drug discovery**. *Nature* 2014; 506(7488):376-381.
380. Waters KM, Stram DO, Hassanein MT, Le Marchand L, Wilkens LR, Maskarinec G, et al. **Consistent Association of Type 2 Diabetes Risk Variants Found in Europeans in Diverse Racial and Ethnic Groups**. *PLoS Genetics* 2010; 6:e1001078.
381. Zaitlen N, Paşaniuc B, Gur T, Ziv E, Halperin E. **Leveraging Genetic Variability across Populations for the Identification of Causal Variants**. *The American Journal of Human Genetics* 2010; 86(1):23-33.

382. Spanakis EK, Golden SH. **Race/ethnic difference in diabetes and diabetic complications**. *Current diabetes reports* 2013; 13(6):814-823.
383. Cheng C-Y, Reich D, Haiman CA, Tandon A, Patterson N, Elizabeth S, et al. **African Ancestry and Its Correlation to Type 2 Diabetes in African Americans: A Genetic Admixture Analysis in Three U.S. Population Cohorts**. *PLoS ONE* 2012; 7(3):e32840-e32840.
384. Mak ACY, White MJ, Eckalbar WL, Szpiech ZA, Oh SS, Pino-Yanes M, et al. **Whole-Genome Sequencing of Pharmacogenetic Drug Response in Racially Diverse Children with Asthma**. *Am J Respir Crit Care Med* 2018; 197(12):1552-1564.
385. Guo X, Li Y, Ding X, He M, Wang X, Zhang H. **Association Tests of Multiple Phenotypes: ATeMP.** *PloS ONE* 2015; 10:e0140348.
386. Marchini J, Howie B, Myers S, McVean G, Donnelly P. **A new multipoint method for genome-wide association studies by imputation of genotypes**. *Nat Genet* 2007; 39(7):906-913.
387. Loh PR, Tucker G, Bulik-Sullivan BK, Vilhjalmsson BJ, Finucane HK, Salem RM, et al. **Efficient Bayesian mixed-model analysis increases association power in large cohorts**. *Nat Genet* 2015; 47(3):284-290.
388. Shen X, Klaric L, Sharapov S, Mangino M, Ning Z, Wu D, et al. **Multivariate discovery and replication of five novel loci associated with Immunoglobulin G N-glycosylation**. *Nat Commun* 2017; 8(1):447.
389. Brody JA, Morrison AC, Bis JC, O'Connell JR, Brown MR, Huffman JE, et al. **Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology**. *Nature Genetics* 2017; 49(11):1560-1563.
390. Fehrmann RS, Karjalainen JM, Krajewska M, Westra HJ, Maloney D, Simeonov A, et al. **Gene expression analysis identifies global gene dosage sensitivity in cancer**. *Nat Genet* 2015; 47(2):115-125.