THE UNIVERSITY
of EDINBURGH

Genetic analysis using family-based populations

Réka Nagy

PhD

The University of Edinburgh

2017

Abstract

Most human traits are influenced by a combination of genetic and environmental effects. Heritability expresses the proportion of trait variance that can be explained by genetic factors, and the 1980s heralded the beginning of studies that aimed to pinpoint genetic loci that contribute to trait variation, also known as quantitative trait loci (QTLs). Subsequently, the availability of cheap, high-resolution genotyping chips ushered in the era of genome-wide association studies (GWAS). These genetic studies have discovered many associations between single-nucleotide polymorphisms (SNPs) and complex traits, but these associations do not explain the genetic component of these traits entirely. This is known as the 'missing heritability' problem.

Within this thesis, 40 medically-relevant human complex traits are studied in order to identify new QTLs. These traits include eye biometric traits, blood biochemical traits and anthropometric traits measured in approximately 28,000 individuals belonging to family-based samples from the general Scottish population (the Generation Scotland study) or from population isolates from Croatian (Korčula, Vis) or Scottish (Shetland, Orkney) islands. These individuals had been genotyped using commercially-available arrays, and unobserved genotypes were imputed using the Haplotype Reference Consortium (HRC) dataset.

In parallel to standard GWAS, these traits are analysed using two other statistical genetics approaches: variance component linkage analysis and regional heritability (RH) mapping. Each study is analysed separately, in order to detect study-specific genetic effects that may not generalise across populations. At the same time, because most traits are available in several studies, this also enables meta-analysis, which boosts the power of discovery and can reveal cross-study genetic effects.

These methods are a priori complementary to each other, exploiting different aspects of human genetic variation, such as the segregation of variants within families (identity by descent, IBD), or the presence of the same variant throughout the general population (identity by state, IBS). The strengths and weaknesses of these methods are systematically assessed by applying them to real and simulated datasets.

Lay Summary

‘You have your mother’s eyes’ or ‘You are growing up to be just as tall as your father’ are phrases most of us hear while growing up. They are also testaments to the fact that our traits are heritable – that is, we inherit these (and many more) characteristics from our parents. Similarly, certain diseases, such as Huntington’s disease or some types of cancer are known to ‘run in the family’, where members of a family with a history of the disease are more likely to develop it during their lifetimes than the general population.

This heritable information is encoded in our genes and it is passed down from parents to their offspring through DNA, a double helix-shaped molecule made up of the bases adenine, thymine, cytosine and guanine (A, T, C, G). Strikingly, while the human genome consists of 3 billion base pairs, the genomes of any two people are 99.9% identical. The remaining 0.1% is responsible for most of the diversity we see in humans – this is because here, mutations have arisen, changing one base to a different one, like typos in a book. If they become widespread in a population, these mutations are called polymorphisms. Some mutations have no effect (that we currently know of), while others can either have a benign effect (influencing eye colour for example) or a serious health effect, either causing disease or increasing its likelihood of developing.

Most human traits and diseases are complex, which means that they are influenced by many genes. Similarly, one gene can influence several traits – one example of this is the melanin gene, which partly determines your hair, eye and skin colour. If a gene acquires a mutation, this can alter its effect on a trait – for example, individuals who carry certain mutations in the *FTO* gene have higher rates of obesity than non-carriers. We have already discovered many associations between polymorphisms and traits, but they do not explain the full spectrum of human diversity, which is known as the ‘missing heritability problem’. The aim of this thesis is to uncover some of this missing heritability by applying different statistical tools to study a variety of medically-relevant traits within large family-based datasets. At the same time, this also enables the systematic evaluation of the relative strengths and weaknesses of these tools.

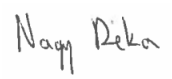
By discovering new connections between our genome and our traits, we can improve the accuracy of trait prediction based on genetic information alone. This means that we may be able to diagnose a disease at a more early stage (possibly even before any visible symptoms have appeared), which can have direct consequences over disease management and prevention. Additionally, it can lead to the discovery of new drug targets, paving the way for the development of novel treatments.

Declaration

I declare that I composed this thesis, the work presented within this thesis is my own, and the contributions of others are clearly indicated in the text. This work has not been submitted for any other degree or professional qualification.

Réka Nagy

November 2017



Acknowledgements

I would firstly like to extend my deepest gratitude to my supervisor, Dr. Veronique Vitart for her unending support, and for being patient when I asked her to explain the same thing for the *n*th time, as well as my supervisors Dr. Pau Navarro and Prof. Chris Haley for their guidance and advice. I would additionally like to thank Prof. Caroline Hayward and Dr. Shona Kerr for their mentorship and encouragement. With your help, I have learned a lot not only about science, but also about surviving in academia and beyond. To my friends and family, you should know that your support and encouragement was worth more than I can express on paper. Finally, thank you to Lucija and Jonathan for occasionally playing the role of rubber duck when my work needed debugging and the rest of the QTL group for fuelling my brain with carbohydrates throughout my PhD. The cake was *not* a lie.

The CROATIA study was supported through grants from the Medical Research Council UK and the Ministry of Science, Education and Sport in the Republic of Croatia (number 108-1080315-0302). ORCADES was supported by the Chief Scientist Office of the Scottish Government and the Royal Society. CROATIA and ORCADES have all been funded as part of the European Union framework program 6 EUROSPAN project (contract no. LSHG-CT-2006-018947). Recruitment, data and sample collection and genotyping in Shetland were funded by a core award from the Medical Research Council to the MRC Human Gnetics Unit. Generation Scotland received core funding from the Chief Scientist Office of the Scottish Government Health Directorate CZD/16/6 and the Scottish Funding Council HR03006. Genotyping of the GS:SFHS samples was carried out by staff at the Genetics Core Laboratory at the Clinical Research Facility, University of Edinburgh, Scotland and was funded by the UK's Medical Research Council.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

%	percent
µg	microgram
µl	microliter
µmol	micromole
BMI	body mass index
BP	blood pressure
bpm	beats per minute
CAD	coronary artery disease
CCT	central corneal thickness
cM	centiMorgan
cm	centimetre
CRP	C-reactive protein
DEXA	dual energy X-ray absorptiometry
DNA	deoxyribonucleic acid
EA	effect allele
EAF	effect allele frequency
EHR	electronic health record
FEF	forced expiratory flow
FEV	forced expiratory volume
FVC	forced vital capacity
g	gram
GGT	gamma-glutamlytransferase
GP	general practitioner
GRM	genetic relationship matrix
GS	Generation Scotland
GS:SFHS	Generation Scotland:Scottish Family Health Study
GWAS	genome-wide association study
GWC	GWAS catalog
GWS	genome-wide significant/genome-wide significance
h2	heritability
HBD	homozygous by descent
HDL	high-density lipoprotein
HRC	Haplotype Reference Consortium
HWE	Hardy-Weinberg equilibrium
IBD	identical by descent
IBS	identical by state
Indel	insertion/deletion
IOP	intra-ocular pressure
IU	international unit
kb	kilobase
kg	kilogram

L	litre
LD	linkage disequilibrium
LDL	low-density lipoprotein
LMM	linear mixed model
LOD	log of odds
LRT	likelihood ratio test
m	metre
MAF	minor allele frequency
Mb	megabase
MCMC	Markov chain Monte Carlo
MEs	Mendelian errors
mg	milligram
mm	millimetre
mmHg	millimetres of mercury
mmol	millimole
n	number
NA	not available
ng	nanogram
NHGRI	National Human Genome Research Institute
ORCADES	Orkney Complex Disease Study
p	probability
PLIER	probe logarithmic intensity error
pmol	picomole
PRIMUS	Pedigree Reconstruction/Identification of Maximum Unrelated Set
QC	quality control
QTL	quantitative trait locus
REML	restricted maximum likelihood
RH	regional heritability
s	second
SD	standard deviation
SE	standard error
SNP	single nucleotide polymorphism
SOLAR	Sequential Oligogenic Linkage Analysis Routines
U	units
VHSS	Viking Health Study - Shetland
YOB	year of birth

Chapter 1 Introduction

1.1 Complex Traits

Most of our traits, and our likelihood of developing certain diseases, are at least partially under genetic control. The proportion of trait variation that can be explained by genetic variation is called heritability. In this thesis, heritability refers to the narrow-sense heritability (h^2), which is due to additive genetic effects – those transmitted from parents to offspring. Narrow-sense heritability is an important population parameter that informs on the limits of genetic prediction in humans and achievement of selection in animal breeding programmes.

Few traits are monogenic, whereby the heritability of these traits can be explained by variation at one genetic locus. Often, these mutations affect a protein, destabilising it by altering its three-dimensional structure, rendering its active sites non-functional or truncating it prematurely. Such traits are also called Mendelian traits (or disorders, in the case of disease). For example, expansions of the CAG triplet repeat in the *HTT* gene gives rise to Huntington's disease and homozygote carriers of certain mutations in the *CFTR* gene are highly likely to develop cystic fibrosis.

In contrast, most common traits are polygenic, as they are influenced by tens, hundreds or even thousands of loci across the genome, with complex interplay with environmental factors. These loci may have one or more causal variants, and these variants vary in terms of allele frequency, effect size and penetrance. For example mutations in the *ABO* gene explain around 30% of the heritability of the quantitative trait "von Willebrand factor levels" [1], but in most cases, a single quantitative trait locus (QTL) only explains a small proportion of trait heritability. Height is under strong genetic control, its heritability is estimated to be around 80% and hundreds of loci have been reported to associate with this trait, but most height QTLs only alter stature individually by a fraction of a millimetre [2].

At the completion of the Human Genome Project, before genetic mapping studies were in full swing, several models were proposed to explain the genetic basis of complex traits. Initially, the "common disease-common variant" and "common disease-rare variant" models were in competition [3–5], but today it is recognised that the variance of most complex traits arises through some combination of common and rare variants, with the latter often assumed to have a stronger effect.

The "common disease-common variant" model [6] posits that the variants that influence common traits must themselves be common. Several factors back this theory, such as the rapid

expansion of modern humans from a founder population of relatively small size. The allelic diversity at neutral loci in such a population would be low. Additionally, variants that influence late-onset, common diseases such as type 2 diabetes or dementia do not have such a high impact on reproductive fitness, and therefore they are not under strong purifying selection so can drift to higher frequency. If a risk allele for a certain disease was at a high frequency in the past, because it conferred a selective advantage to an environmental factor that is no longer present, it would take a very long time before it is diluted out of a population by new alleles. For example, the ancestral E4 allele of the *APOE* gene is common in human populations today, and it is a risk factor for coronary artery disease and Alzheimer's disease [7]. While this allele confers a selective advantage when food is scarce, as it leads to increased lipid absorption, the exposure to a 'western' diet that is rich in fats and carbohydrates but low in fibre, combined with a sedentary lifestyle, leads to carriers of the E4 allele having an increased risk of developing coronary artery disease and Alzheimer's disease [8]. While many common variants have now been associated with quantitative traits and common diseases, with a few exceptions, most have a small effect on the trait.

The competing "common disease, rare variant" model posits that causal variants are rare and have large effects in the people who carry them, but individually have a small effect on the population-wide trait variance due to their low allele frequency. Under this model, only a small number of individuals suffering from a common disease carry a specific risk allele, but there can be a high number of rare risk alleles segregating in the population, most of which are likely to have arisen within the last few generations [9]. For example, most breast cancer-causing mutations within the *BRCA2* gene are rare, and only rare mutations of the *LDLR* (low-density lipoprotein receptor) gene have been found to affect premature coronary artery disease through familial hypercholesterolemia [4].

Recently, an 'omnigenic' model was proposed by Boyle *et al.* [10] based on the suggestion that over 100000 variants could contribute independent effects to human height, but most of these would make only a one seventh of a millimetre difference. These variants are spread throughout the genome, possibly implicating every gene and pathway active in the tissues relevant to the trait. This is similar to the 'infinitesimal model' of Gibson [11], who postulated that "ultimately, every gene contributes to every trait, but with effect sizes so small that it would take samples greater than the population size of the species to detect them".

Today, we know that most common traits are highly polygenic and their variance is due to combinations of common and rare causal variants. The heritability of a trait depends primarily on additive variance, where the changes in the phenotype of an individual scale linearly with

the number of causal alleles they carry. In addition to additive variance, however, other types of genetic effects may also contribute to the architecture of complex traits, although their effects are often negligible compared to additive effects [12]. Dominance effects occur when one allele completely (or partially) masks the effect of the other allele, so that the phenotype of a heterozygote is not exactly at the midpoint between opposite homozygotes. Epistasis effects describe interactions between different loci that lead to phenotype values that are different to the sum of the effects at the two loci.

1.2 Mapping Quantitative Trait Loci

The aim of genetic mapping is to identify loci contributing to phenotypic variation, establishing new links between genetic variation and traits in order to gain a better understanding of the biological pathways that influence a trait. Genetic mapping can be done through several different statistical approaches that leverage the genetic similarities between individuals who are also similar at the phenotype level. The genetic architecture of a complex trait under study is a strong determinant of the success of different mapping strategies. The allele frequencies and effect sizes of the causal variants, as well as the presence (or absence) of composite large QTL effects from numerous clustered small effect variants all affect the power of a mapping strategy.

Within this thesis, genome-wide association studies (GWAS), variance component linkage analyses and regional heritability (RH) mapping are performed using large, well-phenotyped family-based datasets in order to recover some of the missing heritability of medically-relevant complex traits. This also allows for the systematic comparison of these methods, evaluating their strengths and weaknesses in detecting variation contributing to complex traits.

The availability of genotyping arrays ushered in the era of GWAS, first proposed in 1996 [13]. This method regresses the phenotype value onto the genotype of each genotyped (or imputed) single nucleotide polymorphism (SNP), while correcting for sources of potential confounding, for example covariates such as age and sex or excess genetic sharing due to relatedness [14], in order to identify genomic regions where SNP genotypes correlate with the phenotype. Even if the causal variant is not typed, it may be tagged by nearby variants that are in linkage disequilibrium (LD) with it [15], so the power to detect a QTL with this method depends on the allele frequency of the causal variant, as rarer variants will be in LD with fewer ‘tagging’ SNPs. GWAS utilise SNP identity-by-state (IBS). If two individuals carry the same allele at a SNP, these alleles are considered IBS, regardless of whether they have the same ancestral origin.

Over 2500 GWAS have been published, identifying over 25000 SNP-trait associations [16]. While these associations have led to a greater understanding of the biological pathways that underlie complex traits, they fail to explain the full heritability of complex traits, leading to the 'missing heritability' problem [17].

Several complementary hypotheses propose that the remaining heritability may be hidden in rare variants of large effect or in common variants of very small effect or low penetrance [11]. Studies have calculated the joint amount of heritability contributed by all genotyped, or imputed, SNPs, and they have shown that there is a sizeable contribution to heritability from genetic factors that GWAS are underpowered to detect because their effect sizes are too small [18, 19].

GWAS are underpowered to study rare variants at the single-SNP level because by definition, these variants occur too infrequently to allow association tests of individual variants. Instead, rare variants may be studied with the help of burden or candidate gene set tests, by aggregating them into groups [20], but rare variants are often not included on genotyping chips and sequencing is required to detect them. Additionally, there is uncertainty about which variants to include in these groups, as well as the weights that should be assigned to different variants [21]. Genetic studies of complex diseases or traits that use whole-exome sequencing or exome "chips" designed to enrich for rare variants have largely been unsuccessful in identifying associations with novel genes however, generally due to lack of power, but also because many causal variants lie outside of gene coding regions [20].

Most GWAS are also underpowered to detect associations with common variants that have a small effect on the trait, as they require much larger sample sizes to be detected – in order to be well-powered to detect SNPs with such effect sizes, sample sizes larger than the current human population may be required. For example, height GWAS have been performed on increasingly larger sample sizes and these studies have identified an increasing number of associations. The largest height GWAS that had been performed at the commencement of this PhD used data from 250000 individuals, identifying 423 independent genome-wide significant loci that together only explained 16% of the heritability of height [22]. A follow-up study using 700000 individuals identified associations with an additional 83 uncommon and rare variants, 24 of which affect height by more than 1cm, effect sizes which are larger than those observed in associations with more common variants [23]. With these new findings, 27.4% of the heritability of height is accounted for, and studies on 2 million individuals are currently underway, but the full spectrum of variants that affect the heritability of height might never be discovered.

Gibson [11] proposed that hundreds of thousands of common and rare variants contribute to a complex trait and GWAS are detecting those with the largest effect sizes from among a range of normally distributed effects. With such a high number of small-effect variants, the utility of conducting GWAS on larger and larger scales can be called into question, as the variants identified with these are likely to act through complex regulatory networks rather than affecting the trait directly, and these variants would contribute little to understanding the underlying biology of complex traits and diseases [10, 24]. Corroborating this statement, many of the GWAS hits reported in this thesis as well as in the literature are in gene deserts or in or near genes that have no immediately obvious relevance to the trait they associate with.

Some of the missing heritability may lie in structural variation [25], in the form of copy number variants (CNVs) [26], insertions or deletions (indels), inversions and translocations. Such variation cannot always be studied with the help of SNP arrays because SNP sites are not always altered directly, and it might be hard to pinpoint them through sequencing as repetitive regions, for example, are notoriously hard to sequence [27].

Allelic heterogeneity may also be a reason for the missing heritability – one SNP might (weakly) tag several independent variants but only a diluted effect may be observed through this tagging SNP, especially if the independent variants have opposite effects on the trait, or if their effects are very small to begin with [28]. GWAS are not well-powered to detect such loci, and while resequencing efforts targeting promising genes flagged with GWAS might uncover additional causal variants within the same gene [24, 29], other statistical methods may be better able to detect such effects without the need for additional sequencing. It is suggested that a combination of single-SNP and region-based analyses may provide more robust results than relying on either method in isolation [30]. For example, a linkage study of human longevity identified several broad linkage peaks that were then fine-mapped with the help of association studies, identifying the *APOE* locus to be a longevity gene [31].

Before SNP genotyping arrays became available, inheritance-informative microsatellite data were used to conduct classical linkage studies. Linkage analysis relies on establishing whether DNA segments between pairs of individuals are identical-by-descent (IBD, that is, alleles that are IBS but additionally also inherited from the same common ancestor), and uses this IBD sharing information to determine the location of QTLs. This is based on the principle that relatives who are phenotypically similar are more likely to have inherited a DNA segment harbouring a QTL from a common ancestor than relatives who are phenotypically dissimilar. Because genotypes are only available for 1-4 generations within a family, co-inherited DNA segments are large as the small number of meioses only allow for a few recombination events

to take place. A consequence of this is a lower mapping resolution, because regions flagged by linkage analysis are broad, often spanning several megabases (Mb) of a chromosome and encompassing tens or hundreds of genes. This makes it difficult to pinpoint not just the causal variant (or variants), but a causal gene as well.

Linkage studies have been successful in locating highly penetrant variants that cause Mendelian disorders, for example linking the *HTT* gene to Huntington's disease [32] and the *CFTR* gene to cystic fibrosis [33]. An early success story of complex trait linkage mapping was the identification of an obesity-related linkage peak on the short arm of chromosome 2, which was subsequently replicated in other populations [34]. The peak region contains multiple genes, two of which were good functional candidates, and resequencing efforts at the *POMC* (pro-opiomelanocortin) gene revealed rare (minor allele frequency < 0.01) coding variants [35]. These variants failed to explain the entire linkage signal, however, suggesting that there may be additional causal variants in the same, or another, gene that contribute to this signal – this is called allelic heterogeneity. Taking it one step further, functional follow-up of candidate variants may confirm that the candidate variants identified through statistical testing are indeed causal. An example of this is the identification of mutations within the *F7* gene that affects the levels of Factor VII, a blood clotting protein. The locus containing the *F7* gene was identified through a linkage study and, through resequencing, several variants within this gene were identified, cumulatively accounting for the entire QTL [36]. Functional follow-up studies, using *in vitro* expression assays, later demonstrated that these variants do indeed modulate Factor VII levels [37]. These examples are the exception, rather than the rule, in the field of complex trait linkage studies, as many linkage peaks do not replicate and they often do not contain genes that are obvious functional candidates. As a consequence, most of these linkage peaks are not followed up to the point of identifying the underlying causal variant, and even if one causal variant is identified, it often fails to explain the full QTL effect [38], as was the case with the obesity-linked locus described above. Linkage studies have more power to detect loci at which allelic heterogeneity is present, compared to GWAS, because they assess the proportion of variance explained by all causal variants in an analysed region – this means that if a QTL signal is due to the effects of several independent variants of small effect, these might be missed in an association study where the effect of each SNP is assessed individually [38]. It should be noted that these QTLs still need to explain a large amount of trait variance to be detected with linkage analysis, and many complex traits might not have such QTLs.

The power to detect a QTL with linkage analysis can be boosted by using families consisting of many individuals and leveraging LD information between all genotyped SNPs in order to

construct haplotype patterns that can be used to accurately estimate IBD. Additionally, progress has been made with IBD inference methodologies that allow for the estimation of IBD sharing probabilities not just within families, but also between nominally unrelated individuals (those that are not connected in social pedigrees) who nonetheless shared a common ancestor and so may share some regions of their genome IBD [39]. This can also be an economical means of boosting the power of a linkage study by increasing the number of pairs that share a region IBD without the need to recruit additional family members. These IBD estimation methods rely on Hidden Markov Models (HMMs) and they can also be used to phase genotype data to facilitate parent-of-origin studies, or to impute missing genotypes based on a reference set [40]. IBD sharing estimation through HMMs was also applied to exome sequencing data in order to map recessive mutations leading to Mendelian diseases [41], which indicates that if IBD sharing can be inferred with high confidence, it could be used to flag regions carrying rare risk alleles of strong effect. In addition to applying IBD estimation methods to exome sequencing data, linkage analysis has recently also been re-emerging as a means to prioritise disease causing variants using whole-genome sequencing data, although such studies have mostly involved Mendelian diseases segregating in specific families [42–44].

Regional heritability (RH) mapping relies on estimating the heritability contributed by genetic regions consisting of groups of SNPs [45], drawing from the principles used by Yang *et al.* who demonstrate, by calculating the joint heritability explained by all genotyped SNPs, that the heritability is not 'missing', but merely 'hidden' in genotyped variants of small effect [18]. These regions can be defined in several ways. For example, one can use a sliding window consisting of an arbitrary number of SNPs, but these SNPs may be separated by recombination hotspots, so it might not make sense to analyse them together. Alternatively, genetic maps can be used to define recombination hotspots, and allocating SNPs this way leads to regions that are more biologically meaningful. Regions defined this way tag the effects of genotyped variants, so they are able to identify the same signals that are genome-wide significant in a GWAS, but they also capture the effect of ungenotyped and rare variants that might affect the trait but to an extent too small to be detected by single-SNP GWAS [46]. This has been demonstrated in a study that employed GWAS and RH mapping to evaluate the genetic architecture of nematode resistance in Scottish Blackface lambs, as well as in a study of blood lipid traits in isolated human populations. Both of these studies found that RH mapping identified additional suggestively significant loci compared to GWAS [47, 48], and the human study indicates that this method might be less likely to detect false positive associations. Lack of replication of hits identified with RH, but not GWAS, is generally due to a lack of a

widespread use of this method, and performing RH mapping of complex traits in several cohorts should provide a systematic assessment of this method.

By substituting the sliding window approach with the haplotype-based approach of defining regions, some IBD sharing between pairs of individuals could be captured in addition to regions where causal variants are shared IBS. This has not been analysed previously, and by comparing the results of RH mapping to those generated by linkage analysis, it will be possible to assess how well the results obtained with these two methods match. The haplotype-based approach could also facilitate meta-analysis of RH results because the start and end positions of each region are constant between different studies, regardless of the number of SNPs used by that study, as long as the study participants belong to the same general ancestry group.

It is important to note that, regardless of the mapping strategy used, if a QTL is identified, direct causal links cannot always be drawn between mutations, the genes they are found in/near, and the complex traits they associate with. This is because these mutations do not always alter protein structure or gene expression directly, or they are present within genes that encode transcription factors that may have an effect on many genes, rather than a gene that directly affects the studied trait. Reported associations are often found in intergenic regions or within gene introns, and while they might affect the trait indirectly (by altering transcription factor binding sites, modulating enhancer activity or affecting splicing, for example), these effects would need to be validated through functional follow-up studies.

1.3 Family-based Studies and Population Isolates

There are several advantages to using family-based datasets [30], and specifically population isolates [49], in genetic studies. For example, they help control for phenotypic and genetic heterogeneity, as well as population stratification, and they often have large, well-documented genealogies that are particularly useful for linkage analyses [50].

Family-based studies have helped identify the causal genes that underlie monogenic disorders or contribute to the variance of complex traits [30], primarily through linkage analysis. In addition to being used to study the relationships between genotype and phenotype, datasets of related individuals continue to inform sequencing strategies, as sequencing a small number of unrelated individuals enables the imputation of rare variants into their genotyped or even ungenotyped relatives [51]. Haplotypes can be imputed thanks to the availability of long-range phasing methodologies. Once the haplotypes of genotyped relatives have been phased, the haplotypes of ungenotyped individuals can be imputed if the social pedigree (family tree,

genealogy) indicates that the haplotype must necessarily have passed through the ungenotyped relative.

This is more cost-effective than sequencing several related individuals whose underlying haplotypes may not provide much additional information. Imputed genotypes have successfully been used in many large GWAS. For example, sequencing the genomes of 2636 Icelandic individuals allowed the imputation of 20 million variants into 104220 genotyped individuals, down to a minor allele frequency (MAF) of 0.1%, and GWAS of this dataset revealed novel associations with causal variants affecting atrial fibrillation and gallstone disease [52]. Imputed genotypes are also useful in GWAS meta-analyses, as they ensure that the maximum number of genotypes is tested in every participating cohort, even if they were genotyped using different platforms. In addition to imputation, phased data can be used to study parent-of-origin effects, and such studies have identified genomic regions that are imprinted [53].

Population isolates derive from a limited number of founders so they have lower haplotype diversity and minimise the number of risk alleles entering the population. Population isolates may also harbour rare variants that are absent from the general population. A sequencing study of Icelandic individuals found that most variants that are common in this population (minor allele frequency > 2%) are also present in SNP databases, but this was true for only 20% of rare (minor allele frequency < 0.5%) Icelandic variants present in coding regions [54].

The higher prevalence of variants that are linked to recessive diseases or complex disorders causes a rise in the prevalence of these disorders in an isolate population. For example, the Greenlandic population has seen a marked increase in type 2 diabetes cases in the last few decades. Genetic studies in this population have revealed a variant within the *TBC1D4* gene that causes carriers to have higher glucose levels and confers resistance to insulin-stimulated glucose uptake. This variant was present at a MAF of 17% in the Greenlandic population, while it was absent in most other populations, including Chinese, European and African populations, and only one heterozygote carrier was found in a Japanese individual in the 1000 Genomes data [55]. In the Genome Aggregation Database (GnomAD, a follow-on from the Exome Aggregation Consortium ExAC [56]), in addition to this Japanese individual, only two non-Finnish Europeans are heterozygote carriers of this variant out of 123,000 sequenced individuals.

In isolated populations, individuals tend to marry within their communities. This is also known as endogamy and it means that there is little genetic admixture from outside sources, which reduces genetic heterogeneity. In addition to reduced genetic complexity, population isolates

tend to be geographically isolated, share the same culture, are exposed to the same environmental effects and consume the same types of food and drink, which reduces phenotypic noise originating from environmental effects. It also implies that some genetic variants might be enriched in these populations if they improve adaptation to the local environment, and such variants could point to novel biological processes, such as the high-altitude adaptation conferred to Tibetans by mutations in the *EPAS1* gene, a transcription factor that modulates oxygen-regulated genes [57]. For example, a study of the Inuit isolate population in Greenland revealed novel genetic variants involved in fat metabolism that also had an influence on height and weight. [58]. A study of 6307 Sardinian individuals revealed mutations in two different genes, one of which is imprinted, that decreased height by 4.2cm and 1.83 cm (when maternally inherited) in heterozygote carriers. This could be a human example of the 'island effect' that selects for smaller size in mammals [59].

Taken together, these factors mean that population isolates have reduced genetic complexity and less phenotypic noise, so they are expected to lend themselves particularly well to genetic analyses, not just of Mendelian disorders but also complex traits [49]. Additionally, family-based studies are enriched for genetic effects and so have higher power to detect causal loci than studies consisting of an equivalent number of unrelated individuals. Finally, family-based studies also allow for the identification and correction of genotyping errors.

Genetic analysis in a family-based study presents a challenge in that the structure due to relatedness must be accounted for, as the genotypes of individuals are not completely independent of each other, and even within one individual, there could be some correlation between the alleles if inbreeding is present [50]. Failure to correct for this can lead to false positive associations due to an enrichment of the variant between, or within, individuals, rather than an actual link between the variant and the phenotype [60]. This means that often, statistical methods designed to be used in unrelated individuals are not appropriate for studying related individuals without some modification. For example, association analyses using related individuals can be done but this should be through a linear mixed model that includes a genetic relationship matrix as an additional random effect, and variance component based methods, such as the linkage analysis and RH mapping performed in this thesis, should include a genetic relationship matrix as an additional variance component in order to account for relatedness.

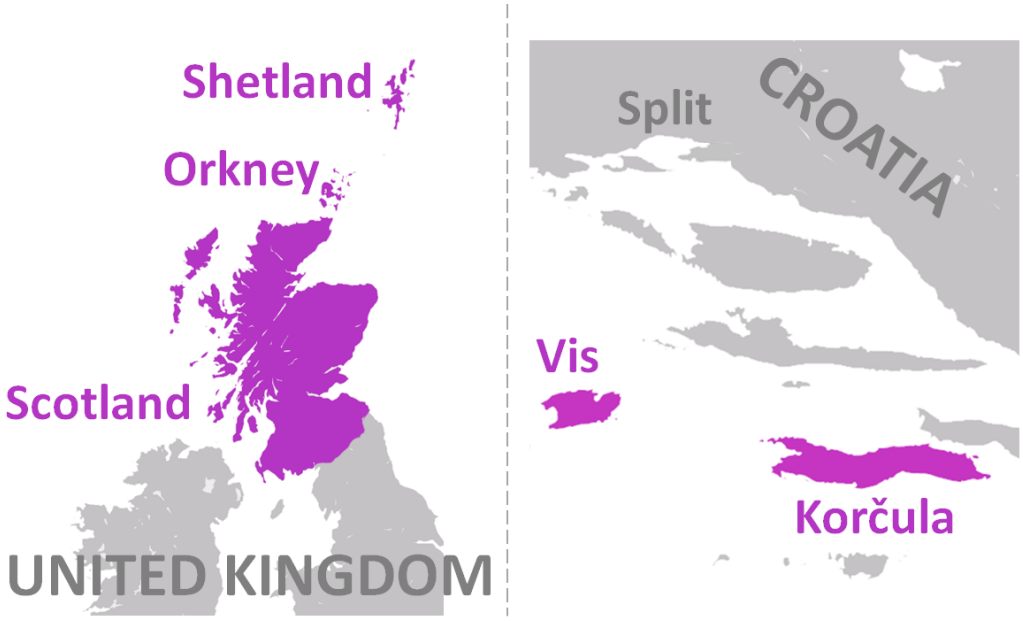
Within this thesis, genotype and phenotype data from five studies were used, and the location of these studies is indicated in Figure 1. Four of these studies consist of populations living on Croatian and Scottish islands: genotype, phenotype and pedigree data were collected from individuals on the islands of Vis (data from 960 individuals) and Korčula (3000 individuals)

in Croatia, and the Orkney (2100 individuals) and Shetland (2100 individuals) archipelagos in the north of Scotland.

The final dataset is from a large family-based study consisting of individuals from the general Scottish population, the Generation Scotland: Scottish Family Health Study (GS:SFHS) [61]. This study consists of genotype, phenotype and pedigree data on ~20000 healthy adult volunteers from the Scottish mainland [62], and most participants have additionally consented to linking their electronic health record (EHR) data, allowing the analysis of phenotypes that were not directly measured as part of the study [63]. Recruitment for this cohort was through primary care and participants were encouraged to recruit family members, so this is also a family-based cohort, but not a population isolate.

Figure 1 - The location of the population isolates used in this thesis.

The Scottish islands are indicated on the left, while the Croatian islands are indicated on the right. Additionally, Scotland, the location the individuals in the GS:SFHS cohort were recruited from, is also indicated.



1.4 Thesis Aims

The aim of this thesis is to provide a better understanding of the variation that underlies complex traits. A particular focus was to provide a comprehensive assessment of the performance of different statistical methods (GWAS, linkage analysis and RH mapping) for the analysis of complex traits with varying degrees of heritability measured in studies where family structure is present. GWAS results were used as a baseline against which the results of

linkage and RH analyses were compared, in order to assess the strengths and weaknesses of each method, in terms of their ability to detect signals. The five family-based studies used within this thesis varied in size and family structure, providing different genetic settings to carry out such analyses, but shared a subset of traits measured in a similar manner allowing comparisons of findings. The main aim is to investigate whether methods leveraging haplotype sharing, which should be increased in family-based samples, help pinpoint genetic loci influencing traits not revealed by standard single SNP GWAS. This would help showcase the advantages of using population isolates and justify the more widespread use of methods complementary to GWAS.

The original contributions of this study to the field are that there are currently no studies that comprehensively compare the performance of these statistical methods on such a large scale. The large scale of this project is due to multiple factors. One of these factors is the use of datasets consisting of thousands or tens of thousands of related individuals and families of varying sizes and complexities. Another factor is the number of phenotypes analysed and the use of whole-genome genotyping and imputed data also contribute to the large scale of this project.

Within this thesis, IBD sharing is computed between pairs of related individuals in these family-based population cohorts (as determined from social pedigrees) and this is then used to compute linkage to complex, medically-relevant traits. Additionally, a 'pedigree-free' linkage method is developed, which aims to leverage IBD sharing between pairs of distantly-related individuals who appear unrelated in the social pedigree, in order to boost the power of linkage, taking the process from the family to the population level, which has not been done to this extent in previous linkage studies.

This study also refined the RH mapping methodology by using regions defined by recombination hotspots rather than windows consisting of an arbitrary number of SNPs. This leads to more biologically relevant regions, and also enables cross-study meta-analysis, as the window positions are independent of the number of genotyped SNPs available in a study.

Meta-analysis was conducted on the results of GWAS using imputed genotypes, RH results as well as pedigree-based and pedigree-free variance component linkage results in order to identify variation that contributes to complex traits in multiple studies.

Chapter 3, Chapter 4 and Chapter 5 describe the three statistical methodologies used and present the results obtained by applying these methods to each cohort individually as well as those obtained by meta-analysing the results. In Chapter 6, these methods are applied to

simulated phenotypes and the Orkney genetic data in order to obtain more easily-quantifiable metrics of their ability to detect a causal variant of small to large effect in the presence of varying degrees of environmental and genetic noise. In Chapter 7, the results obtained from real data are systematically compared in order to highlight the similarities and differences in the signals detected by these methods. Final conclusions are drawn in Chapter 8, where the implications of this study are also discussed.

Chapter 2 Materials and Methods

2.1 Population Studies

This thesis investigates five family-based European population cohorts from Scotland and Croatia. Recruitment, as well as phenotype collection and SNP genotyping was undertaken by colleagues and collaborators as described in the sections below.

2.1.1 The Croatia Studies

In the Adriatic Sea, along the Croatian Dalmatian coast, there are fifteen islands where the population exceeds 1000 individuals. Each island has its own specific history and founder population due to having been isolated from the population they derive from for several centuries, or, in some cases, millennia, and they have been studied for decades [64]. A pilot study was initiated in 2002 by collecting 100 individuals in each of ten villages on these islands to determine their suitability for recruitment into cohorts aimed at studying complex trait genetics. These villages differed in terms of population genetic and ethnic history, founding time, bottleneck events and admixture and showed high levels of genetic structure and differentiation, mostly due to their isolation and endogamy [65].

The islands of Vis (study referred to as Vis in this thesis) and Korčula (study referred to as Korčula in this thesis) were selected for further recruitment. 1008 participants were recruited from the villages of Komiza (546 participants) and Vis (414 participants) on the island of Vis between 2003 and 2004. Recruitment in Korčula occurred in three phases, with 968 participants recruited between 2007 and 2008, 878 participants recruited in 2012 and 986 participants recruited in 2014, for a total of 2832 individuals. Blood DNA, plasma and serum was collected from participants who also filled in questionnaires relating to general health, medical history, lifestyle and diet and were subjected to anthropometric and physical measurements.

Ethical approval was given for recruitment of all Croatia study participants by ethics committees in Scotland and Croatia and participants gave informed consent prior to participation [66, 67].

2.1.2 The ORCADES Study

The Orkney Complex Disease Study (ORCADES, referred to as ORCADES or Orkney in this thesis) is based on an isolated population from the 10 inhabited Orkney isles in the north of Scotland and was designed as a sister study to the Croatia studies. Compared to the Scottish

mainland, genetic diversity is reduced, due to isolation and endogamy [68], as was seen in the Croatia cohorts.

Individuals were eligible to participate in this study if they had at least two grandparents born in the Northern Isles of Orkney and were residing in Orkney at the time of recruitment, which took place between 2005 and 2011. 2080 participants were recruited into this cohort, giving a fasting blood sample and attending a cardiovascular measurement clinic in addition to being subjected to anthropometric measurements, dual energy X-ray absorptiometry (DEXA) scans, cognitive tests and eye measurements.

Ethical approval was given for recruitment of all ORCADES study participants by ethics committees in Scotland in 2004 and participants gave informed consent prior to participation [69].

2.1.3 The VHSS Study

The Viking Health Study – Shetland (VHSS, referred to as Shetland in this thesis) is based on an isolated population from the Shetland isles in the north of Scotland. Study design was similar to that of the ORCADES study, with eligibility to participate based on volunteers having at least two grandparents originating from Shetland. Recruitment of 2105 study participants occurred between March 2013 and March 2015 at a dedicated research centre in Lerwick, where participants attended a measurement clinic and gave fasting blood samples.

2.1.4 Generation Scotland

The Generation Scotland: Scottish Family Health Study (GS:SFHS, referred to as Generation Scotland or GS in this thesis) is a family-based genetic epidemiology cohort with 24000 participants recruited from the general Scottish population through referrals from general medical practitioners and individuals with at least one first-degree relative already in the study. Participants were recruited from the Glasgow and Tayside areas of Scotland between 2006 and 2011, and participants from Ayrshire, Arran and Northeast Scotland were recruited during 2011 [62, 70]. Participants provided blood and urine samples and attended a measurement clinic for anthropometric, respiratory function and cardiometabolic measurements. Participants were also assessed for psychiatric and emotional disorders using the structured clinical interview for DSM-IV and filled in questionnaires relating to medical history, general health, lifestyle and diet.

Ethical approval for recruitment and Research Tissue Bank status was given by the NHS Tayside Committee on Medical Research Ethics (REC Reference Numbers 05/S1401/89 and

15/ES/0040, respectively), providing approval for a wide range of uses within medical research, including genetic analyses and linkage to electronic health records.

2.2 Traits and Covariates

Table 1 contains the list of traits that were analysed within this study, as well as their units, the covariates used and normalisations applied, if any. It also includes the number of individuals who possess each trait measure and all covariates used with that trait. For each trait, the covariates that are used match those previously used by research consortia in large meta-analyses. Some traits were only measured in a subset of a cohort, as indicated by the lower number of individuals with measurements for these traits.

Some traits were transformed to meet the requirement of a close to normal distribution of the residuals in the linear models used for analysis. This was achieved either by simple arithmetical transformation (e.g. taking the natural log of the trait) or "forced normalisation", using quantile normalisation [71] (in this thesis referred to as rank transformation), which was implemented using the rntransform() function in the GenABEL R package [72]. This method of trait normalisation is favoured by the GWAS community as it bypasses the sensitivity of models to outliers, kurtosis and skewness of analysed traits. Some trait measures displaying good distributional properties were standardised to facilitate effect size interpretation, using the formula $\frac{y-\mu_y}{\sigma_y}$, where y is the trait value, and μ and σ are this trait's mean and standard deviation, respectively. This operation is referred to as z-transformation within this thesis, and was implemented using the ztransform() function in the GenABEL R package.

Table 1 - List of traits, covariates, units and normalisations used in this thesis

This table lists the traits analysed within this thesis, as well as the number of individuals with measurements for each trait and all covariates used with that trait, within each cohort. NAs indicate that the trait was not analysed or available in a specific cohort. Rank and z-transformations were done by using the rntransform() and ztransform() functions, respectively, within the GenABEL R package, while ln indicates that the trait was log-transformed.

YOB, year of birth. [b] – Systolic and diastolic blood pressure measures were increased by 15 and 10, respectively, in individuals on anti-hypertensive medications. [a] - Area is a binary covariate indicating whether individuals are from Glasgow (1) or not (0). [d] – Removed individuals reported as being diabetic, as well as individuals with fasting glucose values > 7mmol/L. [s] – ever_smoke is a binary covariate indicating whether an individual was never a smoker (0) or is a current or former smoker (1).

Trait	Vis	Korčula	Orkney	Shetland	GS	Normalisation	Covariates	Units
Anthropometric Traits								
Body Mass Index (BMI)	946	2611	1941	2093	19900	Rank Transform	Sex+Age+Age*Age	kg/m^2
Body Fat Percentage	NA	NA	NA	NA	19480	No	Sex+Age	%
Height	946	2613	1941	2093	19965	No	Sex+Age	m
Waist	946	2574	1938	2093	19664	Rank Transform	Sex+Age+Age*Age+BMI	cm
Waist to Hip Ratio	945	2570	1936	2091	19645	Rank Transform	Sex+Age+Age*Age+BMI	NA
Biochemistry Traits								
Albumin	946	2676	2007	2093	NA	Rank Transform	Sex+Age	g/L
Serum Calcium	948	NA	2007	2093	NA	Rank Transform	Sex+Age	mmol/L
Total Cholesterol	928	2676	2007	2093	19259	Rank Transform	Sex+Age+Age*Age	mg/dL
Cortisol	922	NA	1999	NA	NA	Rank Transform	Sex+Age	µg/L
Creatinine	927	2579	2007	2093	16347	Rank Transform	Sex+Age	mg/dL
C-Reactive Protein (CRP)	900	NA	1991	2082	NA	No	Sex+Age	mg/L
D-Dimer	922	NA	1004	NA	NA	Rank Transform	Sex+Age	ng/dL

Trait	Vis	Korčula	Orkney	Shetland	GS	Normalisation	Covariates	Units
Serum Insulin	815	NA	1927	2092	NA	Rank Transform(ln)	Sex+Age+Age*Age+BMI	pmol/L
Glucose	940	2595	1930	2093	16076	Rank Transform	Sex+Age+Age*Age+BMI	mmol/L
Glucose (diabetics removed)[d]	818	2238	1845	2041	15141	Rank Transform	Sex+Age+Age*Age+BMI	mmol/L
Fibrinogen	918	NA	1003	2089	NA	Rank Transform	Sex+Age	g/L
Gamma-Glutamyltransferase (GGT)	920	NA	985	2093	NA	Rank Transform(ln)	Sex+Age*Sex+BMI	U/L
Glutamate Pyruvate Transaminase (GPT)	920	NA	984	2090	NA	Rank Transform	Sex+Age*Sex+BMI	U/L
HbA1c (Glycated Haemoglobin)[d]	815	2164	1828	2088	NA	Rank Transform	Sex+Age+BMI	% of total Haemoglobin
HDL-Cholesterol	928	2673	2006	2093	19223	Rank Transform	Sex+Age+Age*Age	mg/dL
LDL-Cholesterol	928	2640	2005	2093	NA	Rank Transform	Sex+Age+Age*Age	mg/dL
Serum Potassium	NA	NA	NA	NA	19020	Rank Transform	Sex+Age	mmol/L
Serum Sodium	NA	NA	NA	NA	19277	Rank Transform	Sex+Age+Area[a]	mmol/L
Tissue Plasminogen Activator (tPA)	912	NA	1004	NA	NA	Rank Transform	Sex+Age	ng/mL
Triglycerides	928	2674	2006	2093	NA	ln	Sex+Age+Age*Age	mg/dL
Urea	927	NA	1006	2093	19293	Rank Transform	Sex+Age	mg/dL
Uric acid1	948	2674	2003	2093	NA	Rank Transform	Sex+Age	mg/dL
Uric acid2	940	2464	1715	2089	NA	Rank Transform	Sex+Age+BMI+Alc_gday	mg/dL
Von Willebrand Factor (vWF)	922	NA	1003	NA	NA	No	Sex+Age	IU/dL
Cardiometabolic traits								
Diastolic Blood Pressure[b]	945	2551	1933	2093	19429	No	Sex+Age+Age*Age+BMI	mmHg
Pulse Pressure	945	2551	1933	2091	19429	Rank Transform	Sex+Age+Age*Age+BMI	mmHg
Systolic Blood Pressure[b]	945	2552	1935	2091	19430	Rank Transform	Sex+Age+Age*Age+BMI	mmHg

Trait	Vis	Korčula	Orkney	Shetland	GS	Normalisation	Covariates	Units
Heart Rate	NA	1498	NA	2090	19798	Rank Transform	Sex+Age+Age*Age+BMI	bpm
Eye Traits								
Axial Length1	550	853	1194	1825	NA	Rank Transform	Sex+Age	mm
Axial Length2	548	841	1156	1816	NA	Rank Transform	Sex+Age+Height	mm
Central Corneal Thickness	561	858	1112	1888	NA	Z-Transform	Sex+Age	μm
Intra-ocular Pressure (IOP)	NA	NA	1111	1987	NA	No	Sex+Age	mmHg
Lens Thickness	533	853	NA	NA	NA	Z-Transform	Sex+Age	mm
Spherical Equivalent Refraction	527	835	1165	1927	NA	Rank Transform	Sex+Age	diopter
Pulmonary function traits								
Forced Expiratory Flow	NA	NA	NA	NA	15782	Rank Transform	Sex+Age+Age*Age+Height+ever_smoked[s]	L/s
Forced Expiratory Volume in 1 second (FEV1)	925	2397	1828	1688	15847	Rank Transform	Sex+Age+Age*Age+Height+ever_smoked[s]	L
FEV1/FVC	925	2397	1828	1688	15847	Rank Transform	Sex+Age+Age*Age+Height+ever_smoked[s]	NA
Forced Vital Capacity (FVC)	925	2397	1828	1703	15867	Rank Transform	Sex+Age+Age*Age+Height+ever_smoked[s]	L
Sociodemographic traits								
Alcohol Consumption	946	2491	1720	2089	18213	Rank Transform	Sex+Age+Age*Age+Weight	g/day
Educational Attainment	897	2632	1860	2081	18917	No	Sex+YOB+YOB^2+YOB^3	NA

2.2.1 Significance Thresholds and Independent Traits

The canonical genome-wide significance (GWS) *p-value* thresholds used in the literature are 5×10^{-8} for GWAS [73] and 4.9×10^{-5} for linkage (equivalent to a decimal logarithm of the likelihood ratio, LOD score, of ~3.3) [74]. For RH, a threshold of 4.13×10^{-6} is used (0.05/12101 regions). From simulation studies (discussed in Chapter 6), I obtained an empirical linkage significance threshold of LOD=3.41, which I will use in this thesis.

As more than one phenotype was analysed within this thesis, these thresholds should be more stringent to account for additional multiple testing. The phenotypes used here are not all independent, so in order to determine the number of independent traits (and adjust significance thresholds accordingly), principal component analysis was carried out in each cohort, on all analysed phenotypes (pre-adjusted for covariates and kinship), using all phenotyped individuals, using the "prcomp" function in R. Table 2 shows the number of input phenotypes and the number of principal components (PCs) that cumulatively explain 99% of the variance.

When reporting results in individual cohorts, the GWS threshold is adjusted for the number of PCs explaining 99% variance in that cohort. Table 2 shows the adjusted genome-wide significance thresholds for GWAS, linkage analysis and regional heritability. When reporting meta-analysis results, the genome-wide significance threshold is adjusted by 32, reflecting the largest number of PCs explaining 99% variance in any one cohort. Results that exceed the unadjusted GWS threshold of but do not reach the adjusted GWS threshold are considered strongly suggestive hits.

Table 2 - Trait principal components and significance thresholds

The genome-wide significance thresholds for GWAS, RH and linkage, adjusted for the number of PCs within each cohort, as well as in the meta-analysis, are shown.

Cohort	**Input traits**	**PCs explaining 99% variance**	**GWAS threshold** $-\log_{10}(p)$	**RH threshold** $-\log_{10}(p)$	**Linkage threshold** $-\log_{10}(p)$	**Linkage threshold** (LOD)
Vis	39	32	8.80	6.89	5.93	4.84
Korčula	31	24	8.68	6.76	5.81	4.73
Orkney	39	32	8.80	6.89	5.93	4.84
Shetland	36	29	8.76	6.85	5.89	4.80
GS	23	19	8.58	6.66	5.71	4.63
Meta	-	32	8.80	6.89	5.93	4.84

2.3 Genetic Data

DNA for genotyping was extracted by colleagues at the University of Edinburgh, the University of Zagreb and the University of Split Medical School, and subsequently genotyped as shown in Table 3.

2.3.1 Genotyping and Quality Control

Each cohort was genotyped on a commercially-available genotyping platform (Table 3) and genotypes were called using the BeadStudio software. I performed quality control checks in each cohort prior to downstream analysis, as detailed below.

The final number of autosomal SNPs that were analysed in each cohort is reported in Table 3, as is the final number of individuals available after quality control. Quality control (QC) was carried out using PLINK [75, 76]. These SNPs were filtered using the following quality control procedures:

- SNPs that were not in Hardy-Weinberg Equilibrium (HWE test $P<10^{-6}$) were excluded
- Mendelian errors (MEs) were detected and zeroed using the '`--mendel --mendel-duos --mendel-multigen --set-me-missing`' flags in PLINK, after merging the genotyped individuals with a file containing only ungenotyped individuals in order to preserve the full social pedigree for accurate ME detection
- Monomorphic SNPs were removed
- SNPs that did not lift over to Genome Reference Consortium build 37 were removed (the UCSC liftOver tool was used to perform the SNP position conversion)
- SNPs with call rates lower than 98% were excluded
- Individuals with call rates lower than 97% were excluded

Within this thesis, only autosomal SNPs are analysed, but Y chromosome and mitochondrial DNA were used to assist with pedigree assembly (discussed in the next section). Additionally, X and Y chromosome marker counts were used to identify sex discrepancies that could indicate potential sample swaps. This check was performed in PLINK, using the '`--check-sex ycount`' option, which calculates F statistics (X chromosome homozygosity) and counts the number of Y chromosome markers in each individual. Individuals who are recorded as females in the input file but have F statistics higher than 0.8 and/or carry Y chromosome markers are flagged as male, while males that do not carry any Y chromosome markers are also flagged. If such inconsistencies could not be resolved, the individuals in question were removed from the data prior to downstream analysis.

In Orkney and Korčula, individuals were genotyped using one of several different genotyping platforms, and analyses within this thesis were performed on the overlapping markers from these platforms, hence the lower number of genotyped markers in these cohorts. The histograms in Figure 2 show the allele frequency distributions of the genotyped variants in each cohort and show that there is an enrichment of variants with allele frequencies below 5% in GS and Shetland, where denser genotyping chips were used. In contrast, most variants in the other cohorts have minor allele frequencies larger than 5%. There is a roughly even distribution of SNPs across the rest of the allele frequency spectrum in each cohort.

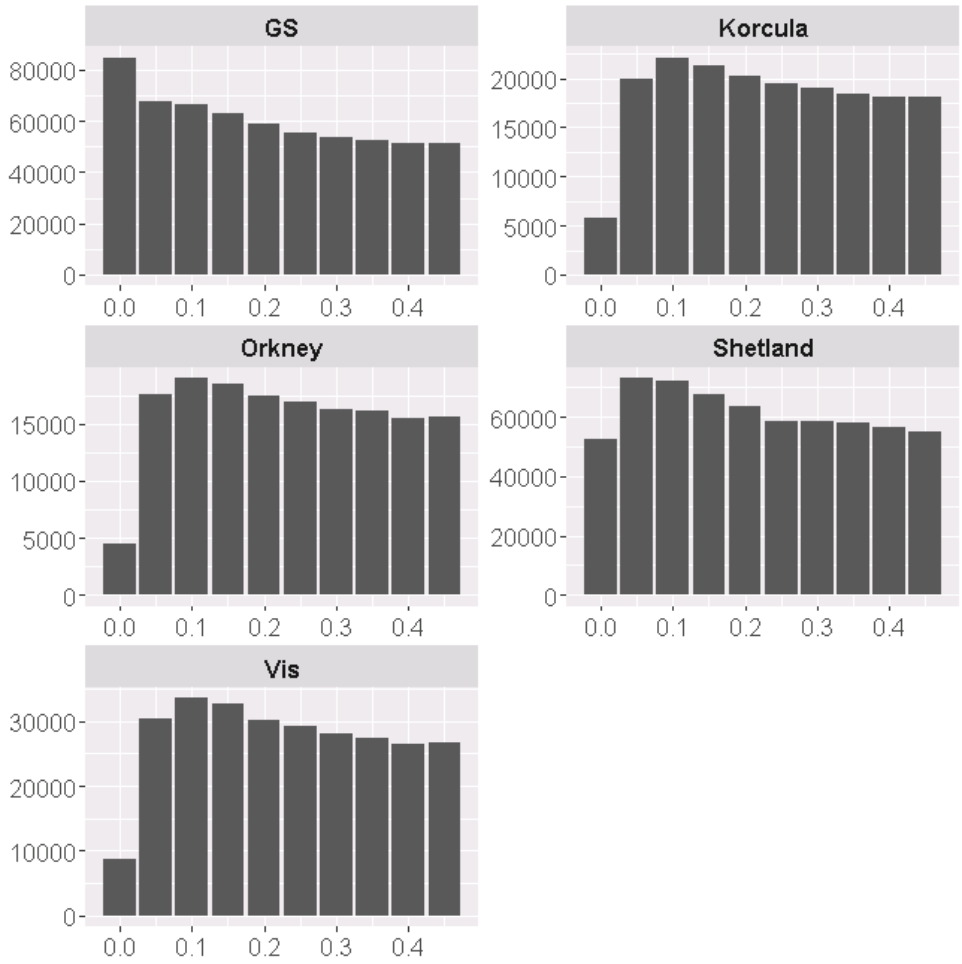
Table 3 - Genotype Data Overview

This table lists the genotyping platforms that were used to type individuals in each cohort. It also shows the number of genotyped SNPs that pass quality control, as well as the number of imputed SNPs that are not monomorphic and have imputation quality scores > 0.4.

Cohort	Genotyping platform and facility	Autosomal SNPs after QC	Imputed SNPs for GWAS	IDs after QC
Vis	**Illumina HumanHap300v1** (Clinical Research Facility, University of Edinburgh)	273,645	12,468,939	960
Korčula	**Illumina HumanHap370CNV DUO/QUAD Phase 1** (Helmholz Zentrum München, Germany) **HumanOmniExpressExome8v1-2_A** (Clinical Research Facility, University of Edinburgh)	182,809	12,382,834	2701
Orkney	**Illumina HumanHap 300v2 Phase 1** (Helmholz Zentrum München, Germany) **Illumina HumanHap370CNV DUO** (Integragen, Paris, France)	157,552	12,696,745	2027
Shetland	**HumanOmniExpressExome8v1-2_A** (Clinical Research Facility, University of Edinburgh)	616,374	14,267,501	2182
GS	**HumanOmniExpressExome8v1-2_A** **HumanOmniExpressExome8v1_A** (Clinical Research Facility, University of Edinburgh)	604,858	24,161,581	20,032

Figure 2 - Histograms of allele frequency distributions of genotyped SNPs in each cohort

The number of variants within each 0.05 MAF interval is plotted. Note that the Y axis scales differ between cohorts.



2.3.2 Imputation

In order to enable meta-analysis of GWAS results, genotyped SNPs were imputed to the Haplotype Reference Consortium (HRC) panel v1.1 [77] using the Sanger Imputation Service. The same genotype quality control steps were used to prepare the files for imputation as those mentioned above, with the exception that where several genotyping platforms were used in a cohort (Orkney and Korčula), platform-specific call rate filters were used prior to merging in order to retain high quality variants even if they were only typed on one platform.

Prior to imputation, autosomal SNPs were phased with Shapeit2 v2r837 [78, 79], using the 'duohmm option 11' flag that takes advantage of the family-based nature of the data [80]. Imputed variants with low imputation quality scores (INFO < 0.4) were removed prior to downstream analysis, as were monomorphic variants. Table 3 presents the number of SNPs that were available for GWAS in each cohort. Genotype imputations were performed by Dr. Thibaud Boutin (MRC Human Genetics Unit).

2.4 Pedigree Assembly, Correction and Summaries

Linkage analysis traditionally requires social pedigrees alongside genetic data. Since all the studies used within this thesis are family-based, they provide the optimal setting for conducting linkage studies. Since the linkage analysis results depend on the accuracy and completeness of these social pedigrees, they were carefully checked and amended in order to remove incorrectly recorded relationships and, where possible, to add new connections.

2.4.1 Vis

The Vis social pedigree had been carefully sought out from church and census records dating back from the 1830s and had already been checked and corrected prior to the start of this study [81], so was used as-is in the linkage analyses.

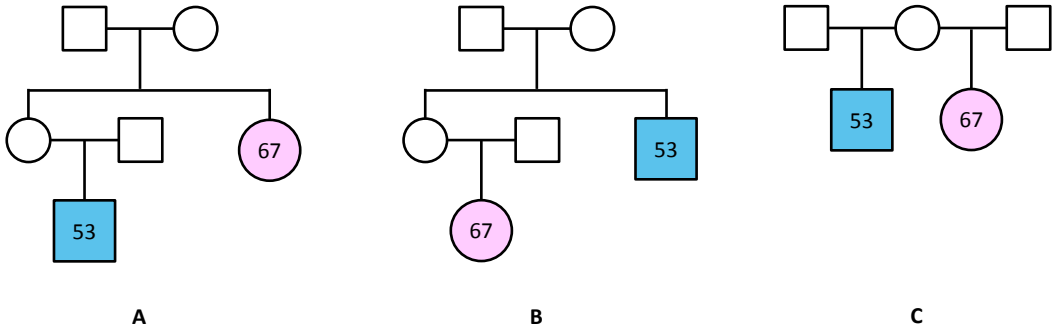
2.4.2 Korčula

No social pedigree was available for Korčula, but since most linkage analyses require a pedigree, a pedigree was manually assembled based on the genetic kinship of the study participants. This was done by first calculating the pairwise genome-wide identity-by-descent (IBD) estimates for all pairs of individuals, using all autosomal SNPs with minor allele frequencies above 1%, using the program PLINK version 1.90b2c [75, 76]. This IBD matrix was passed onto the program PRIMUS [82] which creates networks of related individuals (families), where each person in a network has to be related to at least one other person in that network. The relatedness threshold was set to 0.1, or 10% of markers shared IBD, in the first instance. This grouped 1227 people into the same network, so PRIMUS was re-run in this network using a relatedness threshold of 0.15, in order to group these people into several smaller families where relationships could be reconstructed manually.

Once these networks were created, PRIMUS attempts to reconstruct all possible pedigrees for each network, based on additional data such as age, sex and mitochondrial and Y chromosome matching information. This process works well with nuclear families, but often there can be more than one possible pedigree, so the decision of which pedigree is the most likely one lies with the researcher (Figure 3).

Figure 3 - Pedigrees reconstructed by PRIMUS.

The coloured shapes correspond to two genotyped individuals assigned to the same network. They are estimated to share 25% of their alleles IBD, their mitochondria match, and their ages are shown. They are not related to any other genotyped individuals. Squares represent males; Circles represent females. PRIMUS reconstructs three equally-likely pedigrees (A, B, C) based on these data. Pedigree B looks less likely, due to the younger person being assigned to the older generation, but due to the 14-year age difference, it is not obvious whether the individuals belong to different generations (A) or the same generation (C). In such cases, one pedigree was arbitrarily selected.

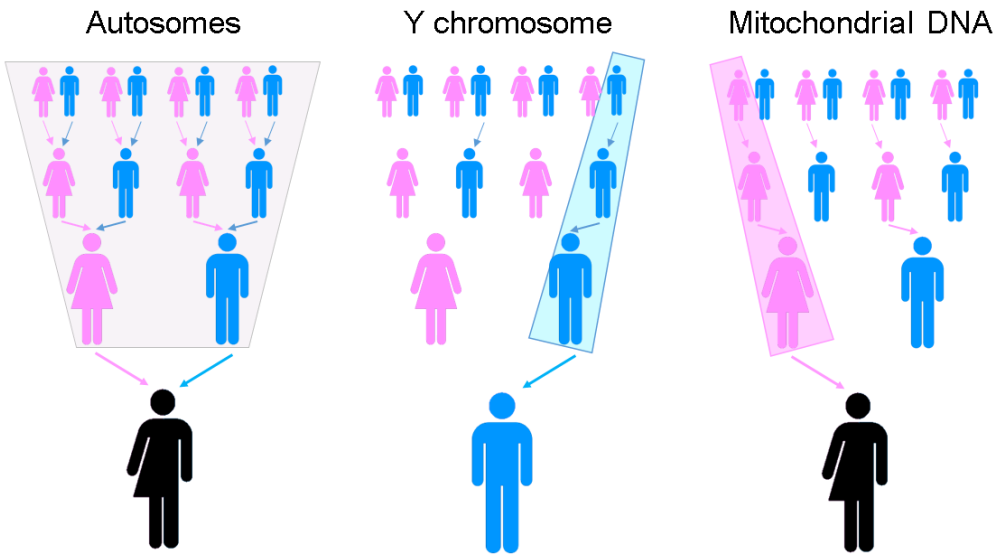


Furthermore, PRIMUS was unable to reconstruct any pedigrees in many networks that contained more than 5 individuals, so pedigree reconstruction in these networks was done manually, making use of Y chromosome data to check paternal lineages and mitochondrial DNA data to check maternal lineages (Figure 4).

In ambiguous cases where two individuals were equally likely to be related in several different ways, one was arbitrarily selected. This is not a problem for linkage analyses, as it only looks at the genetic sharing between relatives, but it might be an issue for other types of analyses, for example parent-of-origin studies.

Figure 4 - Y chromosome and Mitochondrial DNA sharing

Males and females (black shapes) inherit autosomal DNA from both the maternal (pink) and paternal (blue) lines (left). Males inherit the Y chromosome along their paternal line, while females do not have a Y chromosome (centre). Males and females inherit mitochondrial DNA along their maternal line. Fathers do not pass their mitochondrial DNA on to their children (right).



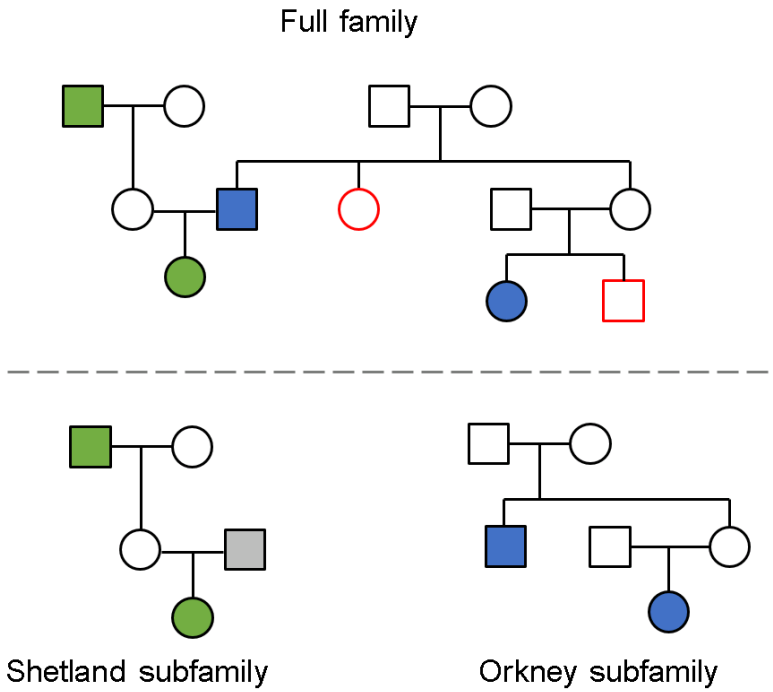
2.4.3 Orkney and Shetland

Both islands are part of Scotland and records of the births, deaths and marriages are all kept at the Edinburgh Register House. Emily Weiss, a PhD candidate in Prof. James F. Wilson's group at the University of Edinburgh, used these records along with relationship information obtained from study participants and genealogies available online to assemble the social pedigree for Orkney and Shetland. This pedigree was corrected to reflect the genetic kinship between individuals, using the merged Orkney and Shetland genotype data. Y chromosome haplogroups (provided by Prof. James F. Wilson) were used to resolve some ambiguities.

The joint Orkney-Shetland master pedigree file was then subdivided into Orkney-specific and Shetland-specific pedigrees. These pedigrees were pruned, keeping genotyped individuals and trimming uninformative individuals (defined as individuals with no genotype data in a cohort, who also do not link two individuals with genotype data), using the program PedStats [83] (Figure 5).

Figure 5 - Pedigree subdivision and trimming

This is a fictitious family tree that is used to illustrate the points of pedigree subdivision and trimming in Orkney and Shetland. The combined Orkney – Shetland family (top) is subdivided into Shetland-specific and Orkney-specific *informative* subfamilies (bottom left and right, respectively). Green shapes indicate individuals with genotype data in Shetland, blue shapes indicate individuals with genotype data in Orkney. Individuals outlined in red indicate entirely non-informative individuals that were removed from the sub-pedigrees and the trimmed Orkney – Shetland pedigree. Note that the genotyped Orkney male (blue square in the top panel) is still informative in the Shetland subfamily (because he links genotyped family members), but will have no genotype information in this dataset (indicated by the grey colour).



The complete social pedigree dates back to ~900 A.D. and is very detailed, consisting of 42000 individuals. After trimming, 22442 individuals were left in the combined Orkney – Shetland pedigree, 21599 of which belonged to one very large 36-generation family linking even very distant relatives. In the Orkney and Shetland-specific pedigrees, 14417 out of 14436 and 8899 out of 9740 individuals, respectively, belonged to this large family. The IBD estimation software Loki [84] is not able to cope with such large family sizes, so the program PedCut [85] was used to generate families of bitsizes no larger than 50, where bitsize is defined as

$$2 * [number\ of\ nonfounders] - [number\ of\ founders]$$

This was done to maximize the number of informative (genotyped/phenotyped) individuals in each family, and it resulted in some genotyped individuals getting dropped if they were unrelated to other individuals, and some individuals becoming duplicated as they were informative in several families. In order to avoid this affecting the phenotype distribution, phenotypes were first adjusted for covariates and transformed, where appropriate, without counting duplicated individuals twice, and the same residuals were assigned to each copy of a duplicated individual.

2.4.4 Generation Scotland

Social pedigrees were corrected as described in [62]. Briefly, pedigrees constructed from self-reported data were checked using genetic relatedness (calculated as described for Korčula above, with the modification that only autosomal SNPs with minor allele frequencies above 5% were used), breaking or modifying links to first and second-degree relatives where the expected and calculated IBD differed by more than 25%. The pedigree was updated to account for previously unrecorded first and second degree relationships (expected IBD = 0, calculated IBD $\geq$ 25%). After these corrections, the expected versus observed IBD sharing was plotted and the pairs were coloured by their pedigree kinship, which helped identify and correct further discrepancies (Figure 6). Mitochondrial and Y chromosome markers were used to resolve lineage-related ambiguities, where possible. Table 4 summarises the number and type of relationships that were changed as a consequence of these correction.

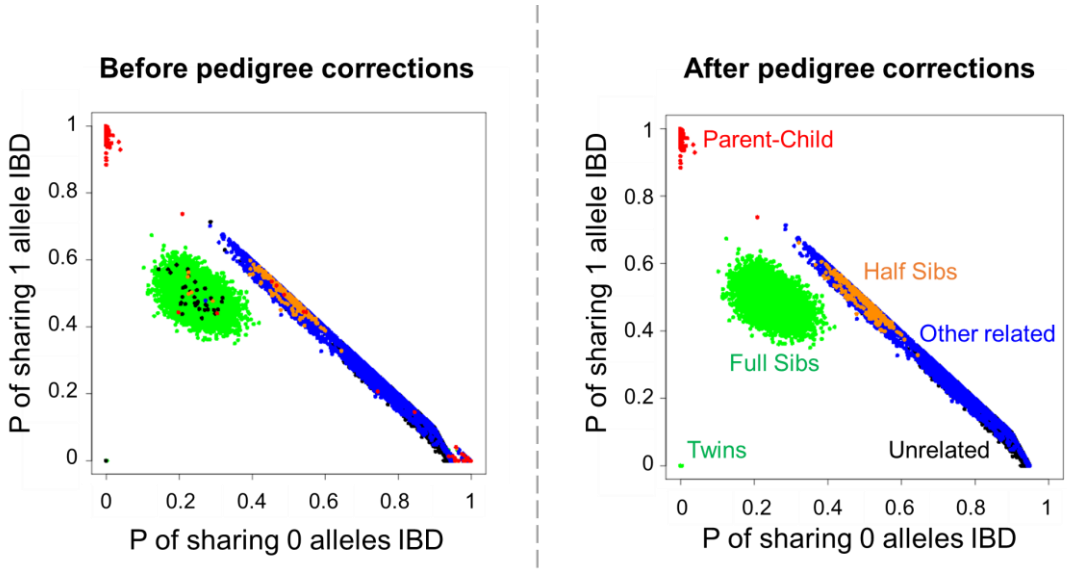
Table 4 - Number and type of relationships that were changed following pedigree QC in Generation Scotland

The first element of each pair reports the initial relationship between a pair of individuals, while the second element reports the corrected relationship, as ascertained from genetic data. FS, full sibling. PO, parent-offspring. OT – other related. UN – unrelated. HS – half sibling. 2^{nd} – second degree relative (can be either grandparent-grandchild, avuncular, or half-sibling).

FS - 2nd	95	PO - FS	2	UN - PO	31	HS - FS	4
FS - UN	13	PO - OT	5	UN - FS	8	HS - UN	3
FS - PO	1	PO - UN	12	UN - 2nd	42	OT - PO	8

Figure 6 - Social pedigree vs Genotype Sharing

This plot shows the average probability of two pairs of individuals sharing 0 alleles IBD (X axis) or 1 allele IBD (Y axis) across the genome, between every pair of individuals in Generation Scotland. Pairs are coloured based on their relationship in the social pedigree – red for parent-child pairs, green for full siblings, orange for half siblings, blue for other family relationships and black for unrelated pairs. The panel on the right represents the calculated genotype sharing coloured using the uncorrected pedigree while the panel on the left represents the same genotype sharing but coloured using the corrected pedigree. There is no unexpected clustering of different types of relationships, indicating that the pedigrees have been adequately corrected.



2.4.5 Pedigree Summaries

The number of people in each pedigree, as well as information about family size and generation number are provided in Table 5. The number of genotyped people in each linkage study, as well as the number of different types of family relationships between genotyped individuals is presented in Table 6.

Table 5 - Pedigree summaries

The pedigree size refers to all informative individuals in the pedigree, regardless of genotyping status. Ork – Orkney, Shet – Shetland.

Population	Pedigree size	Families	Family size (all)		Family size (genotyped)		Generations	
			Average	Maximum	Average	Maximum	Average	Maximum
Croatia - Vis	1843	559	3	351	1.8	127	1.39	6
Croatia - Korčula	4932	1226	4	184	2.2	83	1.62	6
Orkney	14423	14	1030	14404	144.8	2012	3.57	35
Orkney - bitsize 50	7296	170	43	74	11.6	32	6.35	9
Shetland	9740	273	35	8899	7.9	1740	1.67	34
Shetland - bitsize 50	5822	326	17	74	5.8	29	4.07	10
Orkney + Shetland	22442	277	81	21599	15.2	3762	1.66	35
Ork + Shet - bitsize 50	12642	479	26	74	8.1	33	4.74	9
Generation Scotland	30286	6762	6	66	3	35	2.34	5

Table 6 - Summary of genotyped pairs used in linkage analysis

The avuncular relationship type refers to aunt/uncle-niece/nephew relationships. Ork – Orkney, Shet – Shetland.

Population	Number genotyped	Males	Females	Parent-child	Full sibs	Half sibs	Cousins	Avuncular	Grandparent-Grandchild
Croatia - Vis	960	402	558	202	126	11	119	117	33
Croatia - Korčula	2701	987	1714	772	480	55	615	607	124
Orkney	2027	803	1224	796	719	51	1918	1082	132
Orkney - bitsize 50	2081 (1968 unique)	818	1263	781	628	30	997	556	79
Shetland	2182	864	1318	699	612	39	1952	1112	79
Shetland - bitsize 50	1923 (1903 unique)	760	1143	692	512	33	1079	703	70
Orkney + Shetland	4207	1665	2542	1541	1367	84	4010	2272	219
Ork + Shet - bitsize 50	4015	1543	2343	1515	1151	58	2160	1325	159
Generation Scotland	19745	8085	11660	9853	8495	381	2443	6599	848

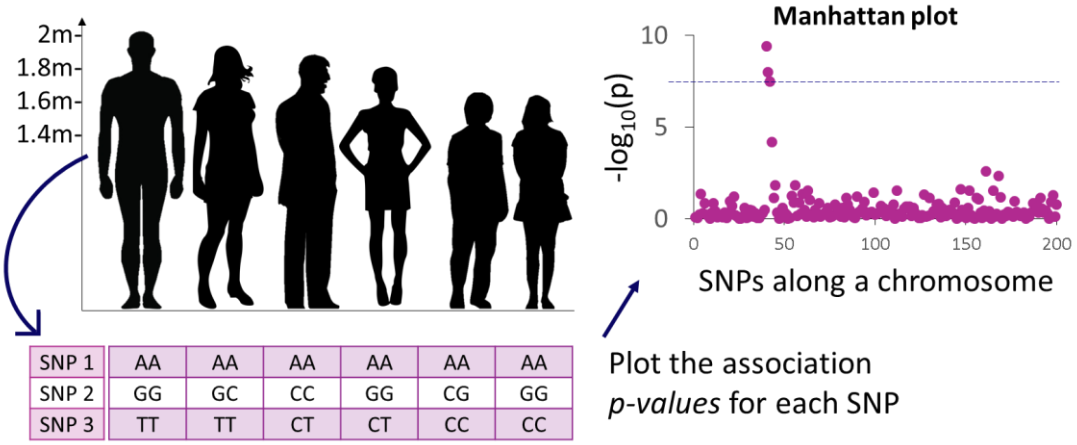
Chapter 3 Genome-Wide Association Studies

3.1 Introduction

Genome-wide association studies (GWAS) use statistical approaches to scan many genetic markers across the genome in order to identify genetic variation that is associated with a trait or disease (Figure 7). These studies leverage the presence of linkage disequilibrium (LD), that is the non-random association, of a genetic locus that exerts an effect on a trait (also referred to as a quantitative trait locus or QTL) and nearby genotyped variants that tag the effect of the QTL.

Figure 7 - Outline of a GWAS

A complex trait, here height, is measured in a population of individuals and the alleles carried by each individual at each SNP is ascertained through genotyping or sequencing. Then, trait values are regressed onto the genotypes in order to ascertain the strength of the association between the phenotype and each studied SNP. The more often a particular allele of a SNP is seen in individuals on one extreme of the trait distribution, the stronger the association signal between that SNP and the trait under study. For example, here, SNP 1 is monomorphic, which means that it is the same in all individuals regardless of their trait value, so it does not contribute to the variation of height in this population. The two alleles of SNP 2 seem to be present randomly in taller and shorter individuals, while the number of T alleles carried by each individual is correlated with their height, with this allele not appearing in the shortest people, so SNP 3 is said to associate with the trait. The results of this analysis on many SNPs across the genome can be visualised in a Manhattan plot, shown on the right of this figure.



SNP 1	AA	AA	AA	AA	AA	AA
SNP 2	GG	GC	CC	GG	CG	GG
SNP 3	TT	TT	CT	CT	CC	CC

GWAS have initially been conducted using datasets consisting of unrelated individuals, in order to avoid false positives that may arise if the genotypes of study participants are not independent of each other [50]. Additionally, false positives or false negatives may also occur if population stratification is present in the studied dataset, as then an association signal may simply be due to different SNP allele frequencies in different populations [86]. While there is no substantial population stratification within the cohorts used within this thesis, and quality control checks have been performed to exclude individuals who are ancestry outliers, there is a large amount of relatedness between individuals owing to the family-based nature of these cohorts. Historically, GWAS have excluded related individuals from the study sample prior to analysis, but excluding related individuals from the cohorts used in this thesis would lead to a huge loss in statistical power. Instead, linear mixed models (LMMs) can be used as these account for both population stratification and the presence of related individuals [87].

Over the past decade, GWAS have been the main tool for discovering loci contributing to complex trait variation [88]. GWAS associations reported in research papers are continuously being aggregated into the NHGRI GWAS Catalog [16], providing a simple means of checking whether variants discovered with a GWAS have previously been reported. It is also possible that a SNP reported in one study is not present on the genotyping panel used to conduct GWAS in a different study. In such cases, if a significant hit is detected at the same locus, LD between these two SNPs can be calculated in order to assess whether they are tagging the same signal or whether they are due to independent signals. Within this thesis, this LD lookup is performed with the help of HaploReg [89], which provides pre-calculated LD statistics between SNPs present in the 1000 Genomes data [90].

In this chapter, I present the statistical framework of conducting GWAS in datasets consisting of related individuals, and apply these to analyse complex human traits measured in five population cohorts of European ancestry. In addition to cohort-specific GWAS, I also perform a meta-analysis using the results of the GWAS that have been performed on the SNP genotyped imputed to the haplotype reference consortium (HRC) panel [77]. The results presented here are used a baseline against which the results obtained with linkage analysis and regional heritability are compared in later chapters.

3.2 Methods

3.2.1 Genetic Relationship Matrix

In contrast to identity-by-descent matrices, presented later in section 4.2, “traditional” genetic relationship matrices (GRM) are calculated without making use of LD between SNPs and they

do not take SNP position into account. The matrix is symmetrical with elements corresponding to estimates of kinship coefficient between pairs of individuals that are calculated using the following formula, as implemented in the GenABEL, DISSECT and GCTA programs.

$$A_{ij} = \frac{1}{N}\sum_{k=1}^{N}\frac{(s_{ik} - 2p_k)(s_{jk} - 2p_k)}{2p_k(1-p_k)}$$

where s_{ik} is the number of copies of the reference allele for SNP k of individual i and p_k is the frequency of the reference allele for this SNP. N is the total number of SNPs used in the calculation [18]. Genome-wide GRMs were used to correct for relatedness in genome-wide association studies (GWAS) and regional heritability (RH) analyses within this thesis.

For use in GWAS, the GRMs were calculated with the 'ibs' function of the GenABEL R package [72], which uses genotype data to estimate the realized kinship coefficient between all pairs of individuals. The weight="freq" option was used, which weighs SNPs based on their minor allele frequency. No minor allele frequency cut-off was used when calculating these GRMs, and SNPs were not pruned for LD prior to GRM calculation.

3.2.2 Linear mixed models

Under a linear mixed model, a phenotype (y) can be expressed using the following equation:

$$y = \mathrm{X}\beta + \mathrm{Z}u + e,$$

where y is a vector of measured values, β is the vector of fixed effects with design matrix X (relating observations to fixed effects), u is the vector of random genetic effects with design matrix Z (relating observations to random effects), and e is the vector of residual error. Values of u and e follow normal distributions with means of 0 and variances var(u)=Gσ_a^2 (where G is the relationship matrix and σ_a^2 is the variance of additive genetic effects) and var(e)= $I\sigma_e^2$ (where I is the identity matrix). A maximum likelihood method is used to estimate parameters (effect of fixed effects and the variance components), based on the multivariate normal feature of the data. This allows the estimation of the narrow sense heritability of the traits as $h^2 = \frac{\sigma_a^2}{(\sigma_a^2+\sigma_e^2)}$.

3.2.3 GWAS

GWAS were performed separately on genotyped SNPs and imputed data. In both cases, phenotypes were adjusted for covariates as well as for relatedness using a genome-wide kinship matrix, using the 'polygenic' function within GenABEL which implements the linear mixed model described above. While fitting one SNP as an additional fixed effect to test

association with a trait is possible, fitting each available marker across the whole genome this way would take too long computationally, so various solutions have been implemented to shorten the analysis time [91]. In GenABEL, GWAS is performed using a two-step process. First, the polygenic effects and the environmental residuals are estimated with the 'polygenic' function, and these estimates are then used to test association with every SNP in the analysis using the 'mmscore' function within the same package, which performs the GWAS under an additive model, using the score test devised by Chen and Abecasis [92]. The 'polygenic' function also produces residuals that are free of correlations (due to polygenic or environmental effects) that can be extracted and used as the phenotype inputs in external GWAS pipelines that do not account for relatedness. As such, the pgresidualY values were used to perform the imputed data GWAS, using the RegScan v0.2 software [93]. After this analysis was completed, the resulting *p-values*, effect sizes and effect size standard errors were corrected to account for relatedness using the GRAMMAR-Gamma correction factors provided by the 'polygenic' function [94].

When using the imputed dataset, imputed SNP dosages, rather than hard-called genotype values, were used to perform the GWAS. Only SNPs with imputation quality scores greater than 0.4, and minor allele counts greater than 20 were used. Note that this minor allele count varies depending on phenotype, as it is calculated using only the individuals with non-missing phenotypes and covariates. Also, because the number of individuals varies between cohorts, the minor allele frequency that a minor allele count of 20 corresponds to will be different between cohorts, ranging from 1% in Vis to in 0.05% in GS. SNPs with Hardy-Weinberg equilibrium *p-values* $< 10^{-6}$ were also removed. For a distribution of allele frequencies in each cohort, see Figure 2.

When reporting the results of GWAS, if a region contained one or more SNPs that passed the genome-wide significance (GWS) threshold, the SNP with the lowest *p-value* in this region is reported as the lead SNP. The results also record the number of additional SNPs that pass the GWS threshold within a 500kb window up- and downstream of the lead SNP.

Additionally, if the region within 100kb up- or downstream of a lead SNP contained an association reported in the NHGRI GWAS catalog [16], this is noted, regardless of the LD status between the two SNPs. To refine these results, the top SNPs reported here have also been analysed with the help of HaploReg v4.1 [89], which extracts all SNPs that are in strong LD with a query SNP ($r^2 > 0.8$) in the 1000 Genomes European dataset, and indicates whether any of these SNPs have been reported in any published GWAS.

In some cases, a SNP reported here is within 100kb of one or more SNPs reported in the GWAS catalog [16], but the two SNPs are not in strong LD according to HaploReg. This can occur due to four reasons. First, the two SNPs might really be independent variants. Second, one or both SNPs could be absent from the 1000 Genomes dataset. Third, novel associations have not yet been incorporated into HaploReg. Fourth, the two SNPs may have very different allele frequencies, which leads to a low r^2 between them. Because HaploReg only filters based on r^2, but not D', two SNPs might appear not to be in LD if their allele frequencies are different. In order to discern whether the SNPs identified here flag QTLs independent to the ones already reported in the literature, or whether the two SNPs are in LD within the cohort where the signal was detected, cohort-specific LD calculations were carried out between these SNPs, using the --ld option in PLINK v1.9 [75], which reports both r^2 and D'.

3.2.4 Meta-Analysis

Each cohort was processed using the same quality control protocols, phenotypes were in the same units and genotypes were imputed and analysed using the same imputation and GWAS pipelines. This means that the cohort-level GWAS results are particularly well-suited for meta-analysis. In addition to providing additional evidence to the presence of a signal detected in one cohort, meta-analysis enables the detection of variants that have modest effect sizes in several cohorts, but do not exceed the genome-wide significance threshold in any one cohort.

Meta-analyses of GWAS results were carried out using the METAL software [95], version 2011-03-25. This software uses GWAS summary statistics to calculate an aggregate *p-value* for each SNP, and, if the phenotypes were in the same units across all analysed cohorts, calculates the meta-analysis level effect sizes and their standard errors as well.

3.3 Results

3.3.1 Cohort-Specific GWAS

The GWAS results using imputed genotypes contain all of the hits identified with GWAS that only used genotyped SNPs, as well additional hits that are not present when only the genotyped SNPs are analysed. These additional associations are primarily with rarer SNPs that were not present on the genotyping chip. Table 7 reports the GWS loci identified in the GWAS of imputed data, and also indicates the position and test statistic of the genotyped SNP with the lowest *p-value* within 500kb of the reported SNP. Supplementary Table 1 reports all loci that exceeded the suggestive but not the genome-wide significance threshold ($-\log_{10}(p\text{-}value)>7.3$).

In total, 8 new loci are identified, 4 in the smaller cohorts, and an additional 4 in GS. These hits are distinct and driven by low-frequency variants. There are no new loci identified in Vis,

where hits for von Willebrand factor, serum uric acid levels and CRP have already been described, respectively, at the well-documented *ABO*, *SLC2A9* and *CRP* gene regions. All of the genome-wide associated SNPs are in high LD with SNPs previously associated with these traits. In Orkney, one new association is discovered with serum uric acid levels.

Most of the other hits reported in Table 7 recapitulate loci that have been previously identified with other GWAS, as they either correspond to, or are in LD with, GWAS hits reported in the literature. In some cases however, the top SNP reported here is within 100kb of one or more reported associations, but HaploReg does not report that they are in LD. In such cases, the imputed data were used to calculate cohort-specific LD statistics between the SNP reported here and all SNPs within 100kb for which associations with the relevant trait have been reported in the literature. This was done to determine whether several independent QTLs segregate at a locus, or whether both GWAS have identified the same signal. These LD statistics are reported in Table 8 and show that, in most cases, the same signal is being detected by the GWAS presented here and the GWAS reported in the literature.

In some cases, LD between such SNPs is stronger within a cohort than in the 1000 Genomes data. For example, cohort-specific LD calculation reveals that in Shetland, rs6540217 (the top SNP that associates with central corneal thickness (CCT) in this cohort) is in strong LD (r^2=0.96, D'=0.98) with rs12447690, the top SNP reported by our group [96] in a CCT GWAS that included the Korčula, Vis and Orkney studies as well as a Croatian metropolitan population from the city of Split cohort. In the 1000 Genomes data, the r^2 between these two SNPs is 0.73. Different allele frequencies can cause two SNPs to have low r^2 values. For example, rs138326449, the top hit for HDL on chromosome 11 in GS, has a MAF of 0.32% in this cohort, while the SNPs reported in the literature have allele frequencies ranging from 8 to 16%. R^2 between the two SNPs is 0 in GS, but D' is 1, which indicates that they are in LD. Sometimes, new hits reported in the GWAS catalog have not yet been imported to HaploReg. This is the case with 5 different associations in GS. For example, two different studies (in addition to our GS study) report associations between rs116843064 and HDL and triglycerides [97, 98], but HaploReg shows no GWAS catalog associations for this SNP.

One exception is the chromosome 15 association with height in GS. The top SNP reported here (rs16942323) is within the *ACAN* gene, but it is not in LD with the two SNPs in this gene for which hits are reported in the literature (rs2238300 [22], rs4932217 [99]).

Table 7 - GWAS genome-wide significant loci in each cohort following HRC imputation

This table summarizes the hits that passed the GWS threshold in the GWAS using imputed genotypes, providing their -$\log_{10}$(*p-value*), effect size and its standard error (Beta and Beta_SE columns), the allele for which the effect size is reported (EA column) as well as the cohort and trait-specific frequency of this allele (EAF column). The nhits column indicates the number of SNPs within 500kb of the reported SNP that also exceeded the GWS threshold in the GWAS using imputed genotypes. Within this 1Mb interval, the position and -$\log_{10}$ (*p-value*) of the SNP with the most significant test statistic in the GWAS using genotyped SNPs is shown (Pos_G and logP_G columns). The name of, and distance to, the gene closest to the reported SNP is indicated – the distance is 0 if the SNP lies within the gene itself. The final column indicates whether other GWAS have identified this hit before. The first value indicates whether any SNPs in the 1000 Genomes data that are in strong LD with the reported SNP (R^2 and D' > 0.8) have been identified with other GWAS, while the second value looks at all SNPs within 100kb of the reported SNP, regardless of LD.

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Orkney														
HDL	16	56991363	rs183130	18.91	0.308	0.0339	T	0.3801	42	56993324	18.19	*CETP*	4470	1|1
LDL	19	45412079	rs7412	19.55	-0.571	0.0619	T	0.0765	15	45395619	5.61	*APOE*	0	1|1
Uric acid1	4	9993838	rs7663032	22.81	0.396	0.0396	T	0.7808	473	10001861	22.02	*SLC2A9*	0	1|1
Uric acid1	11	46015152	rs149931947	9.66	-1.07	0.1686	C	0.0101	4	45975130	1.83	*PHF21A*	0	0|0
Uric acid1	11	47995865	rs148724450	9.09	-1.027	0.1672	T	0.0104	4	48085189	3.02	*PTPRJ*	6243	0|0
Uric acid1	11	50611953	rs182791637	9.38	-0.999	0.1600	C	0.0109	2	50114708	0.67	*LOC646813*	232150	0|0
Uric acid1	11	51531051	rs145325964	9.67	-0.961	0.1514	G	0.0136	1	51440610	0.59	*OR4C46*	14839	0|0
Uric acid1	11	54900697	rs571831626	9.77	-0.9	0.1410	A	0.0153	9	55091268	1.07	*TRIM48*	108961	0|0
Uric acid1	11	63807869	rs183897166	10.86	-1.1	0.1627	A	0.0117	3	64150370	3.8	*MACROD1*	0	0|0
Uric acid1	11	64773463	rs143417571	14.04	-1.164	0.1502	A	0.0127	19	65197393	3.32	*ARL2-SNX15*	8120	0|0
Uric acid1	11	65965632	rs148723727	12.75	-0.982	0.1334	A	0.0161	10	66058082	2.56	*PACS1*	0	0|0

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Uric acid1	11	71120029	rs117991852	9.06	-0.909	0.1483	T	0.0138	1	70919484	2.63	*LOC339902*	0	0|0
Uric acid1	11	72900306	rs79750124	9.1	-0.873	0.1421	T	0.0158	2	73340690	2.49	*P2RY2*	29036	0|0
Uric acid2	4	9951819	rs11723439	22.66	-0.461	0.0463	T	0.1825	521	10001861	20.54	*SLC2A9*	0	1|1
Uric acid2	11	51531051	rs145325964	9.59	-1.055	0.1668	G	0.0128	1	51440610	0.42	*OR4C46*	14839	0|0
Uric acid2	11	54900697	rs571831626	9.6	-0.999	0.1578	A	0.0141	8	55339652	1.22	*TRIM48*	108961	0|0
Uric acid2	11	63807869	rs183897166	10.19	-1.113	0.1703	A	0.0121	1	63327186	3.12	*MACROD1*	0	0|0
Uric acid2	11	64773463	rs143417571	12.8	-1.189	0.1612	A	0.0127	12	64305452	3.07	*ARL2-SNX15*	8120	0|0
Uric acid2	11	66156532	rs193078128	11.52	-1.137	0.1630	C	0.0129	5	66058082	1.71	*SLC29A2*	17240	0|0
Uric acid2	11	66944662	rs150613065	10.05	-0.816	0.1259	A	0.0226	1	67423892	2.35	*KDM2A*	0	0|0
vWF	9	136142203	rs514659	35.19	22.477	1.7964	C	0.3423	193	136139265	25.07	*ABO*	0	1|1
Vis														
CRP	1	159689388	rs2027471	9.16	-0.336	0.0544	A	0.3479	12	159302033	3.02	*CRP*	5008	1|1
Uric acid1	4	9928017	rs13137069	9.64	0.326	0.0514	C	0.7097	94	9611013	9.55	*SLC2A9*	0	1|1
Uric acid2	4	9928017	rs13137069	8.89	0.313	0.0516	C	0.7083	108	9611013	8.86	*SLC2A9*	0	1|1
vWF	9	136137106	rs687289	22.38	18.768	1.8959	A	0.4143	95	136331174	2.14	*ABO*	0	1|1
Shetland														
Central Corneal Thickness	16	88310910	rs6540217	14.44	0.28	0.0357	G	0.6677	58	88298124	12.73	*ZNF469*	182967	0|1
Glucose	11	92708710	rs10830963	13.97	0.284	0.0368	G	0.26	18	92708710	13.68	*MTNR1B*	0	1|1
Glucose_nodiab	11	92708710	rs10830963	16.34	0.311	0.0371	G	0.26	22	92708710	15.96	*MTNR1B*	0	1|1
HDL	16	56997233	rs1864163	15.27	-0.308	0.0380	A	0.2395	36	56997233	15.1	*CETP*	0	1|1
HDL	18	46578242	rs74489351	10.48	0.84	0.1266	A	0.0175	13	47017820	4.37	*DYM*	0	0|0
LDL	19	45412079	rs7412	14.69	-0.532	0.0670	T	0.0575	16	45412079	12.29	*APOE*	0	1|1
Total Cholesterol	19	45422846	rs56131196	10.81	0.3	0.0444	A	0.1738	11	45410002	7.86	*APOC1*	239	1|1

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Triglycerides	11	116559553	rs11824135	9.82	0.218	0.0341	T	0.0681	65	116648917	9.4	*BUD13*	59331	0|1
Uric acid1	4	9997112	rs4529048	22.08	0.409	0.0416	A	0.8125	221	10001861	21.03	*SLC2A9*	0	1|1
Uric acid2	4	9997112	rs4529048	21.56	0.403	0.0416	A	0.8121	297	9998493	20.82	*SLC2A9*	0	1|1
Korčula														
fev1perfvc	4	72034716	rs62302428	8.93	1.041	0.1712	C	0.0123	1	71724706	1.97	*SLC4A4*	18285	0|0
fev1perfvc	7	12963872	rs34079	8.92	0.3	0.0493	A	0.2046	1	12583080	3.56	*ARL4A*	233313	0|0
HDL	8	19912370	rs115849089	10.22	0.292	0.0446	A	0.1216	56	19890612	3.03	*LPL*	87599	1|1
HDL	16	56993324	rs3764261	14.03	0.233	0.0300	A	0.3099	28	56993324	13.95	*CETP*	2509	1|1
LDL	19	45412079	rs7412	18.07	-0.545	0.0616	T	0.0577	14	45333834	2.32	*APOE*	0	1|1
Total Cholesterol	19	45412079	rs7412	10.61	-0.409	0.0612	T	0.0579	2	45585167	2.34	*APOE*	0	1|1
Triglycerides	11	116648917	rs964184	9.4	-0.107	0.0171	C	0.8079	29	116621963	7.33	*ZNF259*	357	1|1
Uric acid1	4	9984541	rs9994216	20.97	0.314	0.0328	T	0.7759	218	10001861	18.78	*SLC2A9*	0	0|1
Uric acid2	4	9984541	rs9994216	19.62	0.315	0.0341	T	0.7763	201	10001861	17.23	*SLC2A9*	0	0|1
GS														
Alcohol consumption	4	100239319	rs1229984	9.21	0.282	0.0456	C	0.9826	1	100256984	2.8	*ADH1B*	0	1|1
BMI	2	630075	rs73139123	8.87	0.088	0.0145	T	0.817	183	653195	8.47	*TMEM18*	37896	1|1
BMI	16	53809123	rs55872725	20.24	0.104	0.0111	T	0.3952	102	53800954	19.63	*FTO*	0	1|1
Body fat	16	53809123	rs55872725	15.26	0.643	0.0794	T	0.3949	100	53800954	14.52	*FTO*	0	1|1
Creatinine	5	176813404	rs3812036	9.95	0.093	0.0144	T	0.2304	17	176801131	9.59	*SLC34A1*	0	0|1
Educational Attainment	5	44719845	rs150675176	9.35	1.68	0.2694	T	0.0056	1	44646453	3.36	*MRPS30*	89180	0|0
Forced Expiratory Flow	4	106819053	rs34712979	8.92	-0.088	0.0144	A	0.2675	2	106698892	4.75	*NPNT*	0	0|1

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
fev1perfvc	15	93685164	rs72749974	9.37	-0.118	0.0189	G	0.1223	2	93681470	4.11	*RGMA*	52720	0|0
Forced Vital Capacity	6	7807702	rs1225986	8.66	0.098	0.0165	T	0.8306	17	7789943	7.52	*BMP6*	0	0|1
Glucose	2	169763148	rs560887	67.22	0.224	0.0129	C	0.7085	152	169763148	66.42	*G6PC2*	0	1|1
Glucose	3	170733076	rs9873618	11.01	-0.089	0.0131	A	0.2869	23	170715830	10.73	*SLC2A2*	0	0|1
Glucose	7	44245853	rs917793	23.6	0.156	0.0153	T	0.1829	35	44240407	23.03	*YKT6*	0	1|1
Glucose	8	118184783	rs13266634	10.44	-0.083	0.0126	T	0.316	7	118184783	10.32	*SLC30A8*	0	1|1
Glucose_nodiab	2	169763148	rs560887	74.68	0.244	0.0133	C	0.7087	163	169763148	73.61	*G6PC2*	0	1|1
Glucose_nodiab	3	170724883	rs8192675	10.08	-0.087	0.0134	C	0.284	18	170715830	9.71	*SLC2A2*	0	0|1
Glucose_nodiab	7	44245853	rs917793	27.84	0.175	0.0158	T	0.182	37	44240407	27.08	*YKT6*	0	1|1
Glucose_nodiab	8	118185733	rs11558471	12.33	-0.093	0.0129	G	0.3235	8	118185733	12.28	*SLC30A8*	0	1|1
Glucose_nodiab	13	28499962	rs7981781	9.3	0.089	0.0143	A	0.2342	31	28491198	8.91	*PDX1*	0	1|1
HDL	8	19824667	rs15285	17.94	0.109	0.0124	T	0.2671	264	19824492	17.55	*LPL*	0	1|1
HDL	11	116701354	rs138326449	19.54	0.981	0.1064	A	0.0032	7	116660686	6.95	*APOC3*	0	0|1
HDL	15	58678720	rs261290	24.56	-0.122	0.0117	C	0.6556	124	58679668	23.39	*LIPC*	45453	1|1
HDL	16	56993324	rs3764261	112.85	0.266	0.0117	A	0.3263	165	56993324	111.02	*CETP*	2509	1|1
HDL	17	41926126	rs72836561	10.81	-0.236	0.0351	T	0.0297	2	41978756	2.31	*CD300LG*	0	0|1
HDL	19	8429323	rs116843064	9.25	0.248	0.0399	A	0.0231	1	8458960	4.7	*ANGPTL4*	0	0|1
HDL	19	45412079	rs7412	13.23	0.156	0.0208	T	0.0777	12	45410002	9.83	*APOE*	0	1|1
Heart Rate	14	23861811	rs365990	9.39	0.784	0.1254	G	0.3635	6	23861811	9.33	*MYH6*	0	1|1
Height	3	141133450	rs1991431	12.28	0.005	0.0007	A	0.4337	47	141102833	12.04	*ZBTB38*	0	1|1
Height	6	26184102	rs7766641	12.48	-0.006	0.0008	A	0.255	13	26233387	11.61	*HIST1H2BE*	0	1|1
Height	6	34219698	rs57026767	10.35	-0.006	0.0009	T	0.8448	36	34194866	10.09	*C6orf1*	2793	1|1
Height	6	126851160	rs1490384	9.15	0.004	0.0007	T	0.4851	153	126744087	8.97	*CENPW*	181405	1|1
Height	6	142745883	rs7753012	13.12	-0.005	0.0007	G	0.3072	115	142767633	12.34	*GPR126*	0	1|1

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Height	10	130682872	rs144225905	8.67	-0.092	0.0154	T	0.001	1	130588415	3.2	*MGMT*	582580	0|0
Height	11	67024534	rs7952436	11.72	-0.009	0.0012	T	0.0896	5	67472145	3.75	*KDM2A*	0	0|0
Height	15	89383764	rs16942323	10.96	-0.013	0.0019	C	0.0345	12	89408081	6.34	*ACAN*	0	0|1
Height	18	20713215	rs8096254	11.36	0.005	0.0008	A	0.7404	21	20716805	11.16	*CABLES1*	1311	0|1
Height	20	34005240	rs6060402	12.55	-0.005	0.0007	C	0.6415	136	34001250	11.79	*UQCC*	5294	1|1
Total Cholesterol	1	55505647	rs11591147	16.74	-0.361	0.0424	T	0.0168	1	55505647	16.68	*PCSK9*	0	1|1
Total Cholesterol	1	62957030	rs10889333	9.67	-0.072	0.0113	A	0.3599	218	62905893	9.53	*DOCK7*	0	1|1
Total Cholesterol	1	109817590	rs12740374	21.09	-0.124	0.0129	T	0.2294	20	109817192	21.44	*CELSR2*	0	1|1
Total Cholesterol	2	21319016	rs672889	15.42	0.13	0.0159	G	0.8644	289	21288321	15.07	*APOB*	52070	1|1
Total Cholesterol	2	44069772	rs75331444	10.73	-0.142	0.0211	A	0.0723	22	44065090	10.16	*ABCG8*	0	0|1
Total Cholesterol	5	74656539	rs12916	10.2	0.073	0.0112	C	0.3973	51	74656539	10.31	*HMGCR*	0	1|1
Total Cholesterol	8	9183664	rs4841133	8.6	0.112	0.0187	G	0.9074	7	9177732	6.42	*LOC157273*	0	1|1
Total Cholesterol	19	11193949	rs10412048	24.3	-0.181	0.0175	G	0.1084	69	11202306	23.64	*LDLR*	6087	1|1
Total Cholesterol	19	45412079	rs7412	93.28	-0.42	0.0204	T	0.0776	155	45415640	51.48	*APOE*	0	1|1
Urea	3	187687840	rs16862780	9.52	-0.093	0.0147	A	0.1572	23	187687890	9.5	*LOC339929*	181152	0|1
Waist	6	127454893	rs72959041	9.99	0.161	0.0249	A	0.0567	3	127519234	2.02	*RSPO3*	0	0|1
Waist Hip Ratio	6	127454893	rs72959041	13.59	0.188	0.0246	A	0.0567	4	127481154	5.04	*RSPO3*	0	0|1

Table 8 - LD statistics between SNPs reported in this study and GWAS Catalog (GWC) SNPs

Cohort	Trait	Chr	SNP	GWC SNP	R^2	D'
Shetland	Central Corneal Thickness	16	rs6540217	rs12447690	0.96	0.98
				rs9938149	0.88	0.98
Shetland	Triglycerides	11	rs11824135	rs12272004	0.5	0.72
				rs1558861	0.008	0.09
				rs28927680	0.47	0.73
				rs12286037	0.45	0.72
				rs964184	0.31	0.8
				rs603446	0.06	0.91
				rs7350481	0.01	0.13
				rs4938303	0.18	0.94
				rs2160669	0.008	0.09
				rs6589566	0.008	0.09
Korčula	Uric Acid	4	rs9994216	rs12498742	0.65	0.92
				rs734553	0.55	0.78
				rs16890979	0.61	0.87
				rs7442295	0.71	1
				rs3775948	0.82	0.91
				rs11722228	0.2	1
				rs13129697	0.83	0.98
				rs6855911	0.58	0.81
				rs6449213	0.66	0.99
				rs737267	0.58	0.81
				rs6832439	0.61	0.87
GS	Creatinine	5	**rs3812036**	**rs3812036**	1	1
GS	Forced Expiratory Flow	4	**rs34712979**	**rs34712979**	1	1
GS	Forced Vital Capacity	6	rs1225986	rs6923462	0.74	0.88
GS	Glucose	3	rs9873618	rs11920090	0.32	1
GS	Glucose_nodiab	3	rs8192675	rs11920090	0.32	1
GS	HDL	11	rs138326449	rs662799	0.0001	1
				rs964184	0.0003	0.96
				rs651821	0.0001	1
				rs11216126	0.0004	1
GS	HDL	17	rs72836561	rs77697917	0.72	0.95
GS	HDL	19	**rs116843064**	**rs116843064**	1	1
GS	Height	15	rs16942323	rs2238300	0.0001	0.06
				rs4932217	0.001	0.19
				rs2351491	0.04	0.82
				rs16942341	0.69	0.84

Cohort	Trait	Chr	SNP	GWC SNP	R^2	D'
				rs2280470	NA	NA
				rs3817428	0.03	0.6
				rs8041863	0.007	0.5
				rs8042988	0.05	0.88
GS	Height	18	rs8096254	rs4800148	0.71	0.95
				rs4369779	0.73	0.97
				rs4800452	0.72	0.95
				rs8098316	0.05	0.37
				rs11082304	0.35	0.95
GS	Total Cholesterol	2	rs75331444	rs6756629	0.93	0.98
				rs4299376	0.04	0.99
				rs76866386	0.99	1
GS	Urea	3	rs16862780	rs10937329	0.34	0.93
GS	Waist	6	**rs72959041**	**rs72959041**	1	1
GS	Waist Hip Ratio	6	**rs72959041**	**rs72959041**	1	1

3.3.2 GWAS Meta-analysis

The loci that exceeded the GWS threshold in the meta-analysis of GWAS using the HRC imputed genotypes are summarised graphically in Figure 8, and described in detail in Table 9. Supplementary Table 2 lists all loci that exceeded the suggestive but not the genome-wide significance threshold ($-\log_{10}(p\text{-}value)>7.3$).

The meta-analysis identifies 109 GWS loci, 37 of which were not GWS in any cohort-specific GWAS. There are 7 novel loci where no GWAS hits are reported in the literature within 100kb of the lead SNP, and 5 of these had no GWS association in any cohort-specific GWAS. Forest plots showing the per-cohort and meta-analysis effect size and direction for the 5 novel loci that had signals in more than one cohort are shown in Figure 9.

Figure 8 - GWAS meta-analysis results overview

This figure shows a plot of the -$\log_{10}$(*p-value*) of every hit that passed the GWS threshold in the GWAS meta-analysis, plotted by chromosome position and coloured by trait. In the legend on the right, the highest -$\log_{10}$(*p-value*) was taken from each trait, and was used to rank the traits in decreasing order – this rank number precedes each trait name. Each trait is allocated a colour that is used to represent all points associated with that trait. In order to help distinguish colours with similar hues, on each chromosome, the trait number is plotted immediately to the right of the SNP that yielded the highest -$\log_{10}$(*p-value*) in that trait.

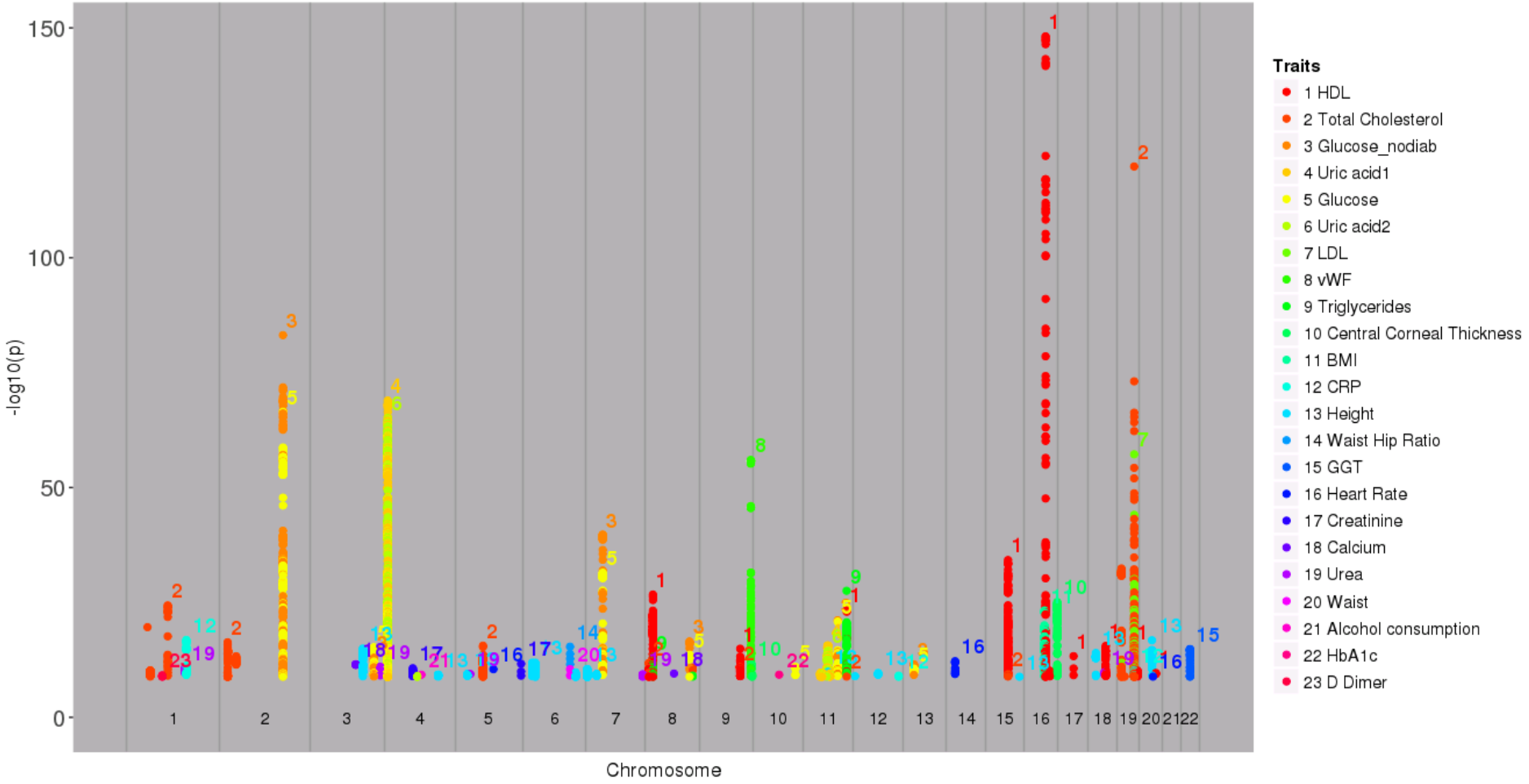


Table 9 - Genome-wide significant hits in the GWAS meta-analysis

This table summarizes the hits that passed the GWS threshold in the GWAS meta-analysis, providing their -$\log_{10}$(*p-value*), effect size (Beta) of the effect allele (EA) and its standard error (Beta_SE) as well as the direction of the effect in each cohort (O=Orkney, V=Vis, K=Korčula, S=Shetland, G=GS; - or + values indicate effect size direction, ? indicates that this SNP was not available in the cohort, x indicates that the trait was not analysed in the cohort). The nhits column indicates the number of SNPs within 500kb of the reported SNP that also exceeded the GWS threshold in the meta-analysis. The name of, and distance to, the gene closest to the reported SNP is indicated – the distance is 0 if the SNP lies within the gene itself. The final column indicates whether other GWAS have identified this hit before. The first column indicates whether any SNPs in the 1000 Genomes data that are in strong LD with the reported SNP (R^2 and D' > 0.8) have been identified with other GWAS, while the second value looks at all SNPs within 100kb of the reported SNP, regardless of LD.

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
BMI	16	53809123	rs55872725	0.4019	23.32	T	0.0974	0.0096	113	27488	+	+	+	+	+	*FTO*	0	1|1
Calcium	3	122013465	rs73186030	0.1476	11.46	T	0.2009	0.0289	5	5045	+	+	x	+	x	*CASR*	8120	1|1
Central Corneal Thickness	9	137437183	rs943423	0.3341	11.58	A	0.1691	0.0242	19	4416	+	+	+	+	x	*COL5A1*	96467	0|1
Central Corneal Thickness	9	139862633	rs57024841	0.4074	9.46	A	-0.1467	0.0234	15	4416	-	-	-	-	x	*PTGDS*	9321	0|1
Central Corneal Thickness	16	88314452	rs11117401	0.3485	25.14	A	0.2434	0.0231	113	4416	+	+	+	+	x	*ZNF469*	179425	1|1
Creatinine	4	77358987	rs10023335	0.4111	10.72	T	-0.0679	0.0101	56	23950	-	-	-	-	-	*SHROOM3*	0	1|1
Creatinine	5	176801131	rs10866705	0.2522	11.73	A	-0.0815	0.0116	13	23950	-	-	-	-	-	*RGS14*	1531	0|1
CRP	1	159668984	rs11265257	0.3866	17.16	T	-0.2142	0.0249	49	4970	-	-	x	-	x	*CRP*	13093	1|1
CRP	12	121420260	rs7979473	0.3858	8.92	A	-0.1522	0.025	26	4970	-	-	x	-	x	*HNF1A*	0	1|1
CRP	19	45411941	rs429358	0.1562	10.55	T	0.2361	0.0355	24	4970	+	+	x	+	x	*PVRL2*	0	0|1
D Dimer	1	95052032	rs2022030	0.3127	9.12	A	-0.2188	0.0356	19	1926	-	-	x	x	x	*F3*	44618	1|1

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
GGT	22	24996582	rs2330795	0.345	14.88	A	0.2259	0.0283	34	3075	+	x	x	+	x	*GGT1*	0	1|1
Glucose	2	169763148	rs560887	0.2892	66.07	T	-0.1847	0.0107	144	23631	-	-	-	-	-	*G6PC2*	0	1|1
Glucose	3	170709193	rs1604038	0.2903	15.3	T	-0.0869	0.0107	35	23631	-	-	-	-	-	*SLC2A2*	4942	0|1
Glucose	7	44243438	rs2971668	0.1877	31.66	C	0.1498	0.0126	46	23631	+	+	+	+	+	*YKT6*	0	1|1
Glucose	8	118191475	rs35859536	0.3197	13.56	T	-0.0797	0.0105	8	23631	-	+	-	-	-	*SLC30A8*	2521	1|1
Glucose	10	114758349	rs7903146	0.2791	11.14	T	0.0752	0.011	26	23631	+	+	+	+	+	*TCF7L2*	0	1|1
Glucose	11	92708710	rs10830963	0.2709	20.77	C	-0.1049	0.011	45	23631	-	-	-	-	-	*MTNR1B*	0	1|1
Glucose	13	28498265	rs60353775	0.2332	10.59	C	-0.0772	0.0116	32	23631	-	-	-	-	-	*PDX1*	0	1|1
Glucose_nodiab	2	169763148	rs560887	0.2891	82.82	T	-0.2138	0.011	169	22080	-	-	-	-	-	*G6PC2*	0	1|1
Glucose_nodiab	3	170709193	rs1604038	0.291	12.93	T	-0.0822	0.0111	26	22080	-	-	-	-	-	*SLC2A2*	4942	0|1
Glucose_nodiab	7	44248828	rs2908282	0.1873	39.61	A	0.1737	0.0131	51	22080	+	+	+	+	+	*YKT6*	0	1|1
Glucose_nodiab	8	118185733	rs11558471	0.3216	16.56	A	0.0909	0.0107	8	22080	+	-	+	+	+	*SLC30A8*	0	1|1
Glucose_nodiab	11	92708710	rs10830963	0.2688	21.22	C	-0.11	0.0114	42	22080	-	-	-	-	-	*MTNR1B*	0	1|1
Glucose_nodiab	13	28498265	rs60353775	0.2329	11.53	C	-0.0835	0.012	32	22080	-	-	-	-	-	*PDX1*	0	1|1
HbA1c	10	71094504	rs17476364	0.1117	10.34	T	0.1972	0.0299	7	6892	+	+	+	+	x	*HK1*	0	0|1
HDL	8	9183596	rs4841132	0.0904	10.6	A	-0.1109	0.0166	12	26920	-	-	-	-	-	*LOC157273*	0	1|1
HDL	8	19824667	rs15285	0.2762	26.91	T	0.1152	0.0106	384	26920	+	+	+	+	+	*LPL*	0	1|1
HDL	9	107661742	rs2740488	0.2738	15.21	A	0.0866	0.0107	43	26920	+	+	+	+	+	*ABCA1*	0	1|1
HDL	11	116701354	rs138326449	0.0038	23.52	A	0.9351	0.092	29	23902	+	?	+	?	+	*APOC3*	0	0|1
HDL	15	58680178	rs261291	0.3423	34.42	T	-0.1248	0.0101	128	26920	-	-	-	-	-	*LIPC*	43995	1|1
HDL	16	56993324	rs3764261	0.3336	148.14	A	0.26	0.01	228	26920	+	+	+	+	+	*CETP*	2509	1|1
HDL	16	67327250	rs530515679	0.0032	10.36	A	-0.9837	0.1493	75	21229	-	?	?	?	-	*LRRC29*	0	0|1
HDL	17	41926126	rs72836561	0.0372	13.13	T	-0.2084	0.0279	3	26920	-	-	-	-	-	*CD300LG*	0	0|0
HDL	18	47106028	rs149615216	0.0124	16.34	T	0.3808	0.0453	40	25992	+	?	+	+	+	*LIPG*	0	0|1
HDL	19	8429323	rs116843064	0.0236	12.03	A	0.2451	0.0343	1	26920	+	+	+	+	+	*ANGPTL4*	0	0|1

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
HDL	19	45412079	rs7412	0.0751	15.63	T	0.1506	0.0184	14	26920	+	+	+	+	+	*APOE*	0	1|1
HDL	19	54837635	rs17634081	0.3224	10.22	T	0.0744	0.0114	49	26920	+	+	+	+	+	*LILRA4*	7055	0|0
HDL	20	44547970	rs2868346	0.2318	9.67	T	0.0714	0.0112	4	26920	+	+	+	+	+	*PLTP*	6966	1|1
Heart Rate	5	102601697	.	0.0049	10.48	T	-15.7929	2.3818	1	2087	x	x	?	-	?	*C5orf30*	0	0|0
Heart Rate	14	23861811	rs365990	0.3614	12.46	A	-0.8306	0.1142	24	23383	x	x	-	-	-	*MYH6*	0	1|1
Heart Rate	20	36841914	rs3746471	0.4574	8.93	A	-0.6672	0.1097	25	23383	x	x	-	-	-	*KIAA1755*	0	1|1
Height	3	141133450	rs1991431	0.4217	14.9	A	0.0047	0.0006	67	27555	+	+	+	+	+	*ZBTB38*	0	1|1
Height	4	145566848	rs13125694	0.4247	9.71	T	0.0038	0.0006	34	27555	+	+	+	+	+	*LOC646576*	42363	0|1
Height	5	32773275	rs72742734	0.0527	9.43	A	-0.0082	0.0013	10	27555	-	-	-	-	-	*NPR3*	0	0|1
Height	6	26186200	rs9379832	0.2763	11.33	A	0.0046	0.0007	21	27555	+	+	+	+	+	*HIST1H2BE*	1741	0|1
Height	6	34219698	rs57026767	0.1557	11.78	T	-0.0057	0.0008	35	27555	-	+	-	-	-	*C6orf1*	2793	1|1
Height	6	142745883	rs7753012	0.3092	9.14	T	0.0039	0.0006	52	27555	+	+	-	+	+	*GPR126*	0	1|1
Height	7	2836848	rs2533884	0.2892	10.75	T	0.0044	0.0006	108	27555	+	+	+	+	+	*GNA12*	0	1|1
Height	7	28185091	rs849141	0.2908	9.6	A	0.0041	0.0006	14	27555	+	+	+	+	+	*JAZF1*	0	1|1
Height	11	67024534	rs7952436	0.0881	9.88	T	-0.0072	0.0011	5	27555	-	+	-	-	-	*KDM2A*	0	0|0
Height	12	4384844	rs76895963	0.0336	9.15	T	-0.0116	0.0019	1	27555	-	-	-	-	-	*CCND2*	0	0|0
Height	12	66376091	rs7306710	0.4818	9.8	T	0.0038	0.0006	12	27555	+	+	+	+	+	*HMGA2*	16019	1|1
Height	15	89383764	rs16942323	0.0316	9.41	T	0.0111	0.0018	12	27555	+	-	+	+	+	*ACAN*	0	0|1
Height	18	20724328	rs4800148	0.2163	13.78	A	0.0055	0.0007	30	27555	+	+	+	+	+	*CABLES1*	0	1|1
Height	20	34025756	rs143384	0.405	16.81	A	-0.0052	0.0006	139	27555	-	-	-	-	-	*GDF5*	0	1|1
LDL	19	11190534	rs142158911	0.0935	12.05	A	-0.2101	0.0294	36	7663	-	-	-	-	x	*LDLR*	9502	1|1
LDL	19	45412079	rs7412	0.0688	57.47	T	-0.5432	0.0338	54	7663	-	-	-	-	x	*APOE*	0	1|1
Total Cholesterol	1	55505647	rs11591147	0.0182	20.07	T	-0.3373	0.0361	2	26960	-	-	-	-	-	*PCSK9*	0	1|1

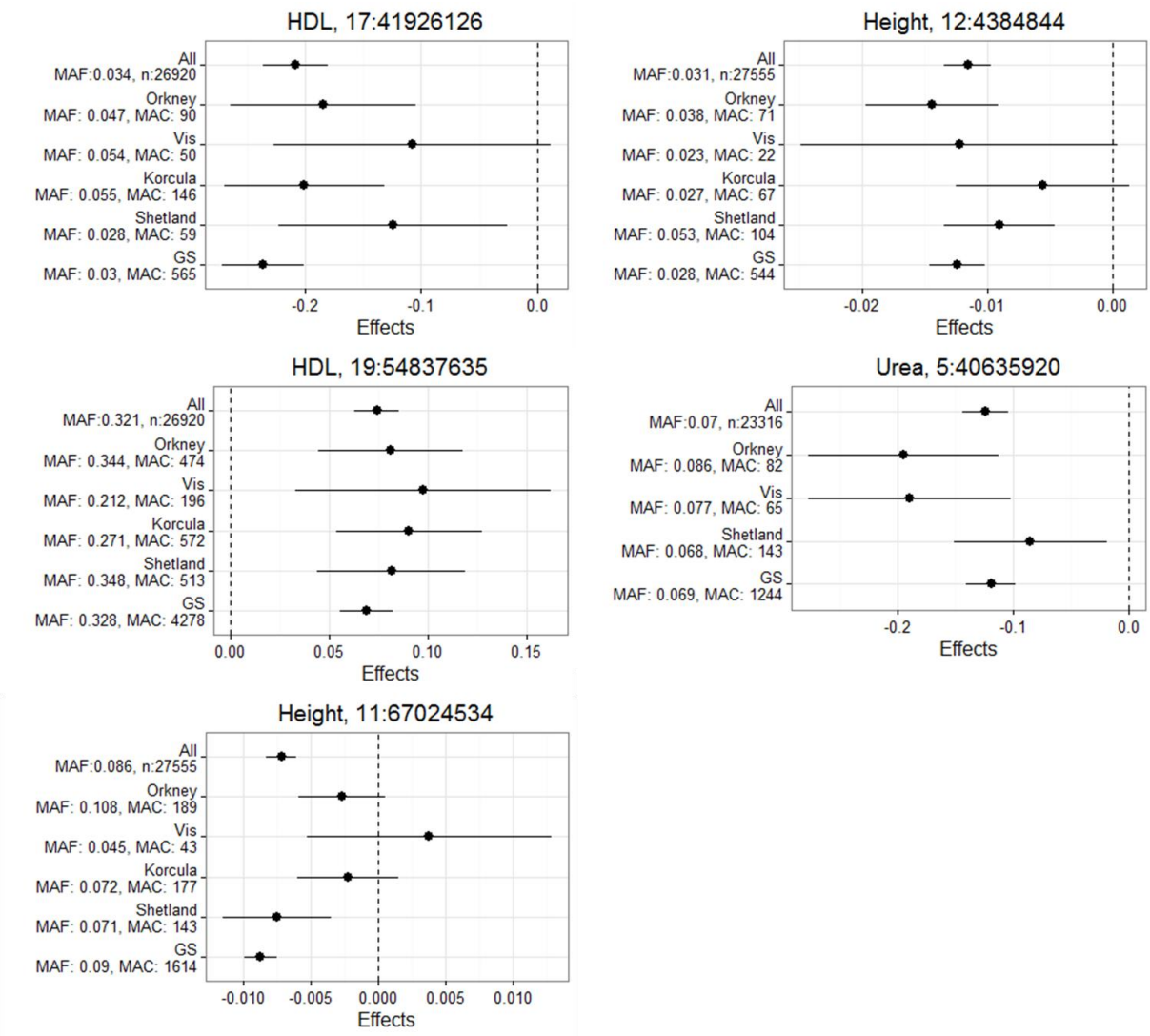
Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
Total Cholesterol	1	62940097	rs1979722	0.3427	10.24	A	-0.064	0.0098	218	26960	-	+	-	-	-	*DOCK7*	0	1|1
Total Cholesterol	1	109817590	rs12740374	0.2288	24.24	T	-0.1132	0.011	24	26960	-	-	-	-	-	*CELSR2*	0	1|1
Total Cholesterol	2	21271323	rs1713222	0.1513	16.18	A	-0.1088	0.013	293	26960	-	-	-	-	-	*APOB*	4377	0|1
Total Cholesterol	2	44069772	rs75331444	0.0699	13.46	A	-0.1387	0.0183	28	26960	-	-	-	-	-	*ABCG8*	0	1|1
Total Cholesterol	5	74656539	rs12916	0.396	15.71	T	-0.0779	0.0095	249	26960	-	-	-	-	-	*HMGCR*	0	1|1
Total Cholesterol	6	160997118	rs74617384	0.0802	9.02	A	-0.1117	0.0183	5	26960	-	-	+	-	-	*LPA*	0	0|1
Total Cholesterol	8	9183664	rs4841133	0.0909	11.81	A	-0.1141	0.0161	20	26960	-	-	-	-	-	*LOC157273*	0	1|1
Total Cholesterol	8	126506694	rs112875651	0.4082	10.99	A	-0.0654	0.0096	21	26960	-	-	-	-	-	*TRIB1*	56049	0|1
Total Cholesterol	9	107661742	rs2740488	0.2739	10.96	A	0.0707	0.0104	21	26960	+	+	+	+	+	*ABCA1*	0	1|1
Total Cholesterol	11	116651115	rs11604424	0.2113	8.82	T	-0.0695	0.0115	3	26960	-	-	-	-	-	*ZNF259*	0	1|1
Total Cholesterol	15	58723426	rs1077835	0.225	9.42	A	-0.0703	0.0112	12	26960	-	-	-	-	-	*LIPC*	747	1|1
Total Cholesterol	19	11193949	rs10412048	0.1049	32.65	A	0.1827	0.0152	105	26960	+	+	+	+	+	*LDLR*	6087	1|1
Total Cholesterol	19	19379549	rs58542926	0.0637	8.91	T	-0.1152	0.019	4	26960	-	-	-	-	-	*TM6SF2*	0	1|1
Total Cholesterol	19	45412079	rs7412	0.0752	120.07	T	-0.4146	0.0177	176	26960	-	-	-	-	-	*APOE*	0	1|1
Triglycerides	7	73001021	rs35173225	0.1383	8.82	T	-0.074	0.0122	38	7698	-	-	-	-	x	*MLXIPL*	6501	0|1
Triglycerides	8	19933089	rs79445051	0.0826	13.23	C	0.1173	0.0156	172	7698	+	+	+	+	x	*SLC18A1*	69275	0|1

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
Triglycerides	8	126495818	rs10808546	0.4511	9.17	T	-0.0528	0.0086	8	7698	-	-	-	-	x	*TRIB1*	45173	1|1
Triglycerides	11	116648917	rs964184	0.165	26.82	C	-0.1261	0.0116	107	7698	-	-	-	-	x	*ZNF259*	357	1|1
Urea	1	155178782	rs760077	0.4201	10.75	A	0.0682	0.0101	5	23316	+	+	x	+	+	*MTX1*	0	0|1
Urea	3	187719348	rs9880162	0.3183	11.06	A	0.0727	0.0106	25	23316	+	+	x	+	+	*LOC339929*	149644	1|1
Urea	5	40635920	rs112647987	0.07	9.58	T	-0.1239	0.0196	24	23316	-	-	x	-	-	*PTGER4*	44110	0|0
Urea	7	151413194	rs10224210	0.2806	9.36	T	-0.0718	0.0115	15	23316	-	-	x	-	-	*PRKAG2*	0	1|1
Urea	18	43252053	rs12963357	0.0268	10.07	T	0.2194	0.0338	2	23316	+	+	x	+	+	*SLC14A2*	0	0|1
Uric acid1	4	9995240	rs3775947	0.2297	68.07	T	0.3476	0.0198	1132	7715	+	+	+	+	x	*SLC2A9*	0	1|1
Uric acid1	11	44713083	rs536771513	0.01	9.33	C	-1.132	0.1817	2	2003	-	?	?	?	x	*TSPAN18*	34932	0|0
Uric acid1	11	46015152	rs149931947	0.0101	9.66	T	1.0701	0.1686	2	2003	+	?	?	?	x	*PHF21A*	0	0|0
Uric acid1	11	47081384	rs757626652	0.0087	9.12	A	-1.1432	0.1858	1	2003	-	?	?	?	x	*C11orf49*	0	0|0
Uric acid1	11	49008491	rs534640316	0.0091	9.74	T	-1.1318	0.1775	10	2003	-	?	?	?	x	*TRIM51CP*	41245	0|0
Uric acid1	11	50740928	rs117993891	0.0081	8.83	A	-1.1264	0.1862	3	2003	-	?	?	?	x	*LOC646813*	361125	0|0
Uric acid1	11	50751235	.	0.0081	8.83	A	-1.1264	0.1862	3	2003	-	?	?	?	x	*OR4A5*	660143	0|0
Uric acid1	11	51531051	rs145325964	0.0136	9.67	A	0.9614	0.1514	2	2003	+	?	?	?	x	*OR4C46*	14839	0|0
Uric acid1	11	55186768	rs142280180	0.0138	9.07	A	0.9075	0.1479	4	2003	+	?	?	?	x	*OR4A15*	50373	0|0
Uric acid1	11	57113738	.	0.0088	9.43	A	-1.1348	0.1811	3	2003	-	?	?	?	x	*P2RX3*	0	0|0
Uric acid1	11	65943116	rs185509475	0.0086	15.65	T	-1.4796	0.1802	23	2003	-	?	?	?	x	*PACS1*	0	0|0
Uric acid1	11	67726024	rs193229075	0.008	15.54	A	-1.5199	0.1859	8	2003	-	?	?	?	x	*UNC93B1*	32549	0|0
Uric acid1	11	70619477	rs182623072	0.0065	10.32	T	-1.3599	0.2067	2	2003	-	?	?	?	x	*SHANK2*	0	0|0
Uric acid1	11	72769082	rs544872267	0.0078	10.29	A	-1.3182	0.2007	15	2003	-	?	?	?	x	*FCHSD2*	0	0|0
Uric acid2	4	9993838	rs7663032	0.2301	64.22	T	0.3472	0.0204	1122	7205	+	+	+	+	x	*SLC2A9*	0	1|1
Uric acid2	4	89039082	rs1481012	0.0934	9.05	A	-0.1843	0.0301	9	7205	-	-	-	-	x	*ABCG2*	0	1|1
Uric acid2	11	46015152	rs149931947	0.0096	9.36	T	1.156	0.1853	3	1715	+	?	?	?	x	*PHF21A*	0	0|0
Uric acid2	11	48999881	rs141399760	0.0081	9.2	T	1.2284	0.1987	10	1715	+	?	?	?	x	*LOC120824*	0	0|0

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
Uric acid2	11	51531051	rs145325964	0.0128	9.58	A	1.0548	0.1669	2	1715	+	?	?	?	x	*OR4C46*	14839	0|0
Uric acid2	11	55186768	rs142280180	0.0126	9.07	A	1.0166	0.1657	4	1715	+	?	?	?	x	*OR4A15*	50373	0|0
Uric acid2	11	66208468	rs553549722	0.0092	14.34	C	1.4905	0.1902	20	1715	+	?	?	?	x	*MRPL11*	0	0|0
Uric acid2	11	67846680	rs370311822	0.0082	14.49	T	-1.5557	0.1974	8	1715	-	?	?	?	x	*CHKA*	0	0|0
Uric acid2	11	70619477	rs182623072	0.0067	8.94	T	-1.3306	0.2186	2	1715	-	?	?	?	x	*SHANK2*	0	0|0
Uric acid2	11	72769082	rs544872267	0.0082	10.23	A	-1.3834	0.2114	6	1715	-	?	?	?	x	*FCHSD2*	0	0|0
vWF	9	136137106	rs687289	0.3765	56.06	A	20.7077	1.304	245	1925	+	+	x	x	x	*ABO*	0	1|1
Waist	6	127454893	rs72959041	0.0563	10.69	A	0.1454	0.0217	4	27212	+	-	-	+	+	*RSPO3*	0	0|1
Waist Hip Ratio	6	127454893	rs72959041	0.0563	15.78	A	0.1755	0.0213	4	27184	+	-	-	+	+	*RSPO3*	0	0|1

Figure 9 - Forest plots of novel GWAS meta-analysis hits

The trait name and top SNP position are shown in the plot titles. The rows of the plot show the effect size and direction of this top SNP (bars represent the standard error). The first row ("All") in each plot shows the meta-analysis results, and indicates the aggregate minor allele frequency as well as the number of individuals that were included in the meta-analysis. Subsequent rows show the results in each analysed cohort, indicating the minor allele frequency in that cohort, as well as the minor allele count.



3.4 Discussion

Most GWAS hits identified in cohort-specific GWAS, as well as the GWAS meta-analysis, replicate previously-identified GWAS loci, by either corresponding to, or being in LD with, SNPs reported in the literature. Yet, a few additional loci were found, and most of these have evidence of consistency of effect across populations, which provides a promising source of preliminary replication.

Within this thesis, the function of genes nearest the GWS hits is discussed, but it is acknowledged that the target gene of an association signal is not necessarily the closest gene. For example, intronic variants within the *FTO* gene consistently associate with obesity phenotypes but no functional connection between this gene and obesity has been established. Instead, it was found that this locus directly interacts with the promoter of the *IRX3* gene 500kb downstream of *FTO*, and *irx3*-deficient mice were shown to have markedly reduced body weight as a consequence of increased metabolic rate and loss of fat mass [100]. Nonetheless, the relevance of the known function of the closest gene to the associated phenotype is always checked as potentially revealing.

In Shetland, there is one novel association with HDL cholesterol, on chromosome 18. The top SNP has a 1.5% MAF and lies within *DYM*, a gene implicated in skeletal development and brain function [101, 102]. There is no known link between this gene and cholesterol regulation, however. This SNP is 3Mb upstream of the *LIPG* (Lipase G) gene, located at 49Mb, which has reported associations with HDL, but it is not in LD with any of the reported associated SNPs.

In Korčula, two novel associations are detected with a lung function trait – fev1/fvc, or the ratio of forced expiratory volume in 1 second (fev1) and forced vital capacity (fvc).

One association is on chromosome 4 and the top SNP has a 1.2% MAF and lies upstream of the *SLC4A4* gene, which encodes a sodium bicarbonate transporter. The lungs are the primary site of the bicarbonate-carbonic acid buffer system, and the rate of breathing can change to compensate for changes in the blood concentration of CO_2 [103], so altering bicarbonate transport can have an effect on lung function in this way.

The second association is on chromosome 7 and the top SNP is common, with a 20% MAF in Orkney. The lead SNP, rs34079, is 230kb downstream of the nearest gene, *ARL4A*, which encodes a GTP-binding protein. This SNP is also within the RBMX2P4 pseudogene, which encodes a long non-coding RNA.

The novel hits identified with imputed data in GS have been reported in [62], with the exception of those for lung function traits and educational attainment, which are reported here. The educational attainment hit, flagged by a low MAF SNP on chromosome 5, is near the MRPS30 gene, and while no SNPs within 100kb of the SNP reported here are in the NHGRI GWAS catalog, common SNPs within the MRPS30 gene associated with educational attainment and have been reported in a large GWAS meta-analysis [104]. The novel lung function hit for fev1/fvc ratio, is driven by a common SNP in an intergenic region, 52kb away

from the *RGMA* gene, which is primarily expressed in neurons and encodes an axon guidance protein.

In Orkney, results of the GWAS of serum uric acid levels shows many GWS SNPs along a ~30 Mb stretch of the centromeric region of chromosome 11. When the trait is only adjusted for age and sex (referred to as 'Uric acid 1'), there are 60 GWS SNPs in the 46-73Mb region, all with MAFs lower than 2%. When the trait is additionally adjusted for BMI and grams of alcohol consumed per day, there are 28 GWS SNPs in this region (Figure 11). The SNPs in this region that are reported in Table 7 are not in LD with each other in the 1000 Genomes European set of individuals, but they are in LD with each other within Orkney. Re-running the GWAS after conditioning on the SNP with the lowest *p-value* in this region causes the signals of all other SNPs to disappear, indicating that these markers are all flagging the same QTL. This region contains the *SLC22A11* and *SLC22A12* genes at the 64.3-64.5 position, and these genes code for known urate transporters [105, 106] that have been flagged in a large uric acid meta-analysis conducted by the Global Urate Genetics Consortium (GUGC) [107]. Interestingly, the 64-68 Mb interval harbouring these genes has the smallest *p-values* in the Orkney GWAS, but the GUGC meta-analysis does not have GWS signals in the broader region flagged in the Orkney GWAS (Figure 10).

Figure 10 - Orkney and GUGC uric acid GWAS results

This figure shows a plot of the $-\log_{10}$(*p-value*) of every SNP on chromosome 11 in the GUGC uric acid meta-analysis in black, and the Orkney uric acid GWAS using imputed data in red. The blue arrow indicates the position of the top GUGC hit that is within the *SLC22A11* gene, and adjacent to the *SLC22A12* gene.

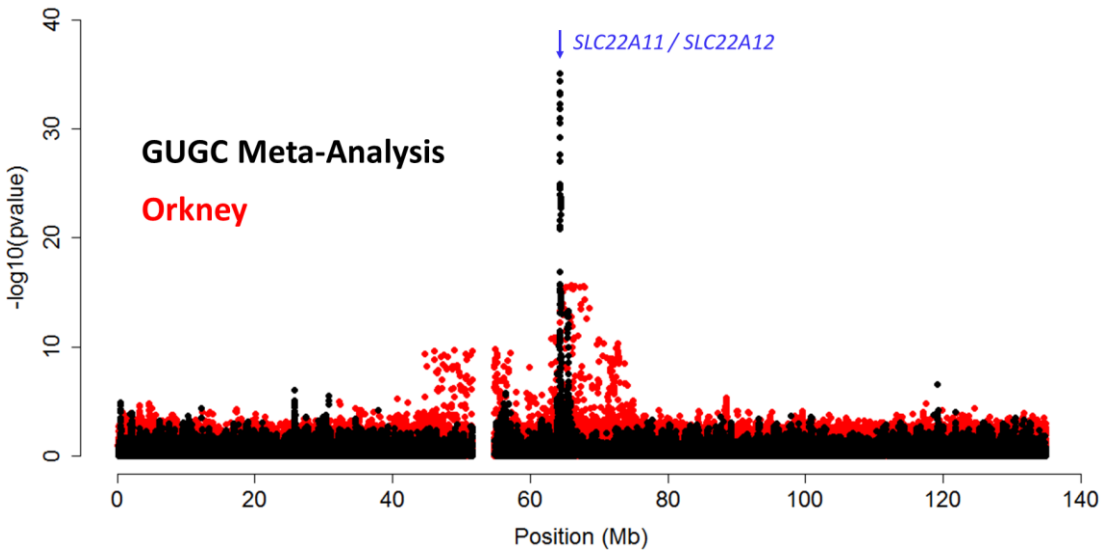
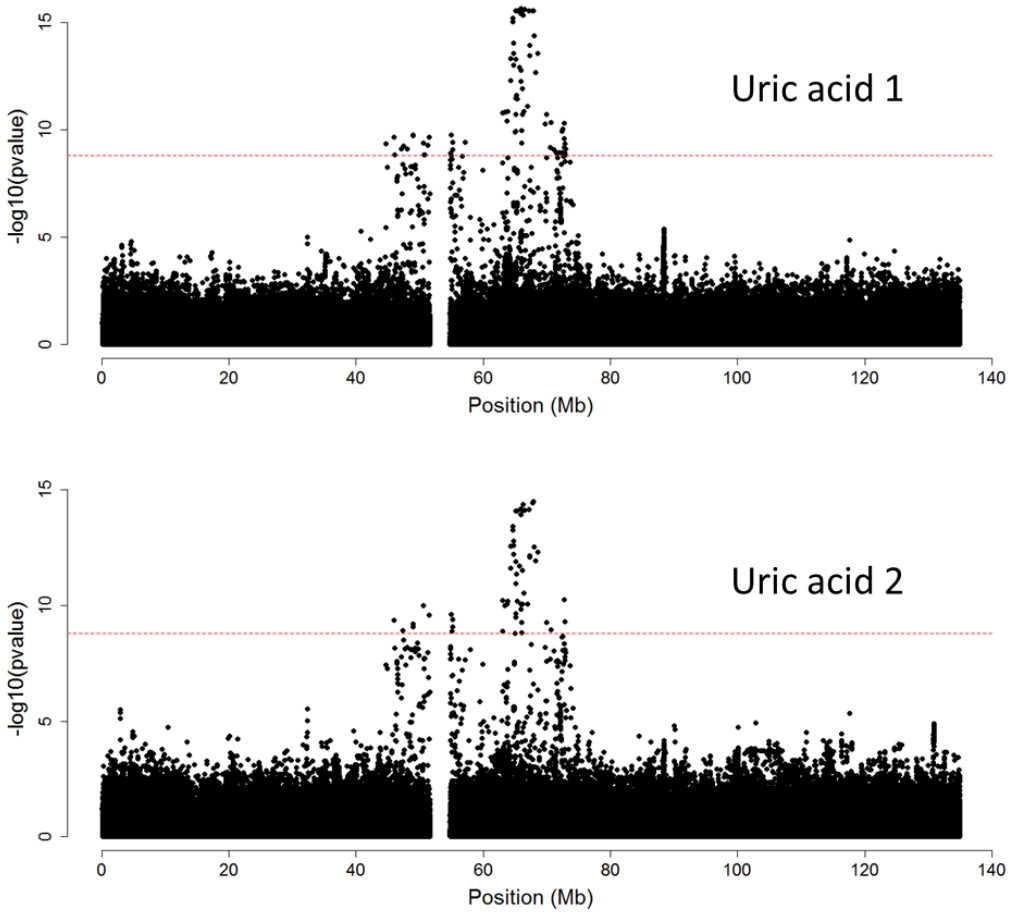


Figure 11 - Manhattan plot of chromosome 11 results for uric acid GWAS

These Manhattan plots depict every imputed SNP with minor allele count > 20 and imputation quality score > 0.4 on chromosome 11 in Orkney. Uric acid 1 was adjusted for age and sex, while Uric acid 2 was additionally also adjusted for BMI and grams of alcohol consumed per day. The red dashed line is the GWS threshold.



Look-up of the effect of these associated variants across the different studies should reveal whether these signals are study specific or show some replication. This was examined as part of a general meta-analysis of the datasets, and some of the meta-analysis results are discussed below.

As was seen in the results of GWAS of serum uric acid levels in Orkney, several SNPs along chromosome 11 also appear in the GWAS meta-analysis of this trait. The meta-analysis signal originates from Orkney only, as these SNPs were filtered out in the other cohorts due to low allele frequency, while the haplotype carrying them drifted to a higher frequency in Orkney.

SNPs in, and near, *CD300LG* and *LILRA4*, associate with HDL cholesterol levels, and the top SNPs show a consistent direction of effect in all 5 cohorts. Both genes encode immunoglobulin-like cell surface glycoproteins and are implicated in immune system pathways. High levels of cholesterol lead to cholesterol accumulation in immune cells, leading to increased inflammation through Toll-like receptor signalling. This, in turn, leads to decreased cholesterol efflux from these cells, amplifying the inflammatory response [108], so it is feasible that there could be an interaction between inflammation genes and cholesterol levels.

Meta-analysis of urea levels yields an association with a locus that is 44kb upstream of the *PTGER4* gene, which encodes a prostaglandin E2 receptor. This trait was not measured in Korčula, and the other 4 cohorts show a consistent direction of effect at the top SNP. Prostaglandin E2 has been shown to act as a second messenger that inhibits the effects of arginine vasopressin on water permeability and water, sodium and urea transport in the kidney inner medullary collecting duct [109]. Mutations that may affect the expression of its receptor can therefore have an indirect effect on urea levels by modulating prostaglandin activity.

In general, the direction of the effect at the lead SNP is consistent across every analysed cohort. Exceptions to this may occur when one cohort does not contribute to the signal. An example is the fasting glucose hit on chromosome 8, where the lead SNP in Vis has an opposite direction of effect compared to other 4 cohorts (rs35859536 has a 0.37 *p-value* in Vis). Vis also has this inconsistency at three loci that show associations with height (p=0.68 at rs7952436, p=0.61 at rs57026767 and p=0.26 at rs16942323) and one of the chromosome 1 associations with total cholesterol (p=0.93 at rs1979722). The lack of association signals at this SNP in Vis leads to effect sizes that are close to 0 and happen to have an opposite sign in Vis compared to the other cohorts.

The top association with waist circumference and waist to hip ratio, on chromosome 6 at rs72959041, shows a negative direction of effect in both Vis and Korčula but a positive direction of effect in the three Scottish cohorts. The association signals originate from the Scottish cohorts only (*p-values* are 0.84 in Vis, 0.79 in Korčula, 4.3×10^{-3} in Orkney, 2.8×10^{-5} in Shetland and 2.8×10^{-14} in GS), and the effect sizes in the Croatian cohorts are again close to 0 (beta values are -0.02 in Vis, -0.01 in Korčula, 0.21 in Orkney, 0.30 in Shetland and 0.18 in GS).

These results also highlight the value of using the genotypes imputed to the HRC reference panel [77] as this permits access to many low and rare frequency variants (0.5-5% MAF) that are often not present on genotyping arrays and imputed poorly with previous reference panels

[110]. The datasets used within this thesis have been also been imputed to the 1000 Genomes reference panel version 3 [90], which yielded approximately 9.5 million SNPs with high imputation quality in each cohort. In comparison, the HRC imputation panel, which consists of sequence data from 38000 individuals including the 1000 Genomes samples, yielded 12.3-24.1 million high imputation quality SNPs in these cohorts.

Rare variants can be imputed with high accuracy especially when the genotyped samples originate from the same population as some of the individuals sequenced as part of the reference panel. In fact, the HRC reference panel contains 400 low-depth (4x) whole genome sequences of individuals from the ORCADES study, so it should be particularly valuable for the imputation into the remaining individuals within this cohort.

Indeed, the unusual GWAS signals originating from the serum uric acid GWAS in Orkney are due to associations with rare variants with high imputation qualities that were excluded from analysis in the other cohorts because they did not have a minor allele count of at least 20. This indicates that these variants drifted to a slightly higher frequency in Orkney. Interestingly, the pedigree-free linkage, which is discussed in Chapter 4, also flags a 2.5 cM interval within this region (at 58-61Mb) where IBD sharing explains some of the serum uric acid trait variance in Orkney and this region is 3Mb away from the urate transporter genes mentioned above.

One possible explanation for this signal is that a small number of (distantly) related carriers of the rare variants share a large haplotype IBD which has not recombined due to chance recombination events. Due to the association pattern spanning the centromere, it is possible that this shared haplotype is tagging a pericentric inversion in some individuals within Orkney. Such an inversion has previously been documented on this chromosome [111], and pericentric inversions on other human chromosomes are not uncommon in the general population [112, 113]. Because the centromere regions are usually comprised of heterochromatin, these types of balanced inversions do not generally affect phenotypes noticeably.

To confirm the presence of such an inversion, experimental validation would be required, as SNP genotype data are not sufficient for determining the presence of a balanced inversion. Experimental validation could be done using cytogenetic methods or targeted long-read sequencing, as genotyping data or short sequencing reads that do not span the entire inversion site are not able to detect large balanced inversions.

This locus also harbours two urate transporter genes, and associations with common variants have been reported at this locus [114]. The fact that no signal is detected in any of the other cohorts hint at the presence of a signal that is distinct from the one tagged by common variants

at this locus. It is known that both common and rare variants within a gene can have an effect on a phenotype, and one of the best-known examples of this is the *PCSK9* gene where it has been shown that coding and non-coding variants across the whole spectrum of allele frequencies (0.2-34%) modulate the levels of LDL cholesterol [115]. The rare variant associations within this gene have been replicated in other cohorts, indicating that it is likely to be a true positive signal.

The association with the rare variants described above could be a testament to the value of imputation by using sequences from the same cohort as a reference, as it could help detect haplotypes carrying these rare variants. However, one cannot exclude the possibility that, due to the low number of carriers, this association is detected by chance as all the carriers just happen to have similar phenotype values, or that these genotypes were incorrectly imputed. The latter could be checked by sequencing the candidate locus in order to assess the presence of the variants obtained from imputation. In Generation Scotland, exome sequence data were available for 864 individuals, but these were not used in the imputation, so they can provide an important source of imputed variant validation. 20 of the GWS SNPs reported in [62] were available in the exome sequencing data, 2 of which were rare (rs142101835 (in *IRS1*) and rs138326449 (in *APOC3*)), and each of which had at least a 97% concordance with the imputed genotypes.

The findings reported here serve as the baseline results against which the results of variance component linkage analysis and regional heritability mapping are compared in the following chapters, in order to assess whether these methods are able to uncover these and additional loci.

Chapter 4 Identity by Descent and Linkage Analysis

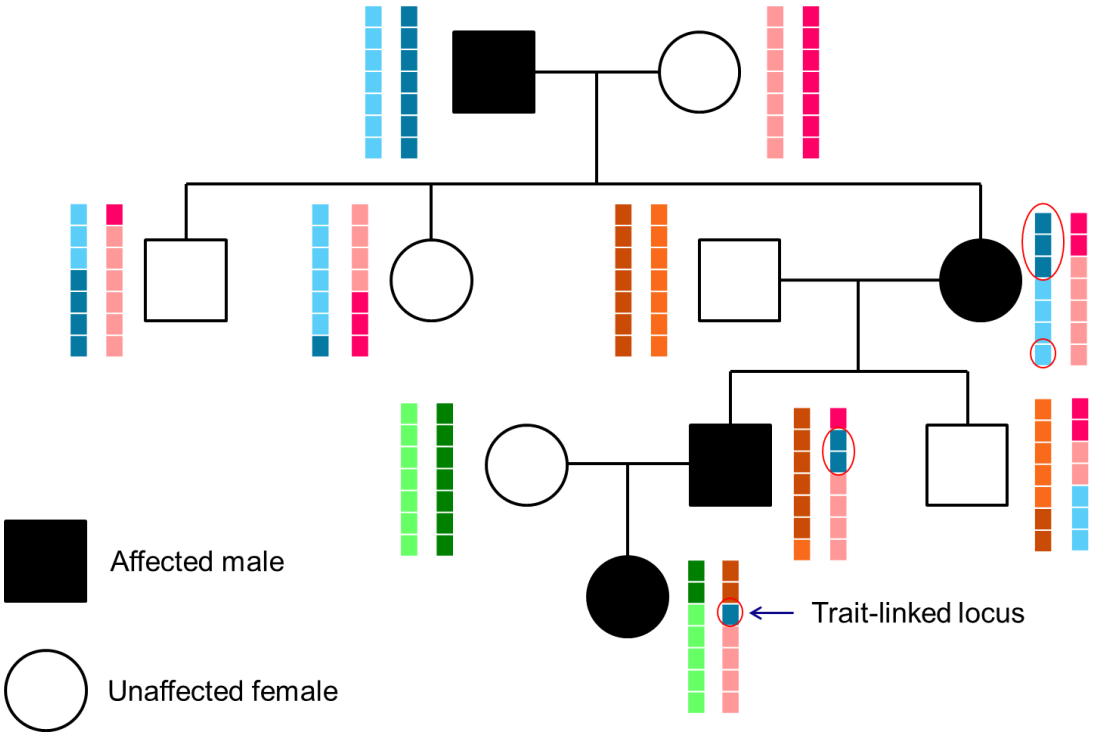
4.1 Introduction

Before the advent of dense genotype data and computational methods capable of analysing these data, linkage analysis was the predominant genetic mapping approach. This approach relies on the fact that loci that are physically close to each other along a chromosome are linked – that is, the closer they are, the less likely it is that they will be broken up by recombination during meiosis. The long-term consequence of recombination occurring in past generations can be observed through the presence of linkage disequilibrium (LD) – that is, the non-random occurrence of specific alleles at different loci [116]. Linkage should not be confused with linkage disequilibrium - two loci that are in LD will also generally be linked, but the reverse it not true – genetic linkage can extend beyond the range of LD [117]. By tracking the recombination rate between marker loci along a genome, a genetic map can be created. It should be noted that recombination rates are not even along the genome, as there are 'hotspots' where recombination happens more often than expected by chance [118].

In classical linkage analysis, microsatellite markers were typed and used to track segregating haplotypes within families in order to identify markers inherited by affected, but not by unaffected, individuals. That is, attempts were made to map major disease loci based on the fact that the inheritance pattern of a marker physically close to a disease-causing variant tracks the inheritance pattern of a disease, and recombination narrows down the potential location of this variant (Figure 12) [116]. When a pair of people share copies of a chromosomal segment inherited from a common ancestor without any recombination, this locus is said to be identical by descent (IBD), and it is the increased sharing of these IBD segments amongst individuals with similar phenotypes, compared to what is expected under random allocation of inherited material, that is the focal point of linkage analysis.

Figure 12 - Classical linkage analysis

This figure is a schematic representation of a family segregating a disease-affecting locus. Four founders exist within this family and their two haplotypes are indicated by light and dark shades of blue, pink, orange or green. The two haplotypes segregating within each individual are shown, with colour-coding corresponding to haplotype segments inherited by each individual. The haplotype segments inherited only by affected descendants of the blue founder are circled. The more individuals in a family, the higher the number of observed meioses, which allows researchers to narrow down the position of the disease-linked locus that is inherited from one common ancestor in all affected individuals.



The first step to performing a linkage analysis is therefore to identify the presence of segments that are IBD between pairs of individuals in a family. When both the number of individuals as well as the number of marker loci are small, an exact algorithm can be used to calculate IBD coefficients at each locus [119], which enumerates all possible inheritance patterns for each pair of individuals. As the complexity of the data increases, applying these methods becomes unfeasible, so approximate algorithms relying on Markov Chain Monte Carlo methods (as implemented in e.g. Loki [84]) or Hidden Markov Models (as implemented in e.g. IBDLD [120]) are used, which calculate the most likely IBD sharing coefficient for each pair of individuals by sampling different realisations of possible inheritance patterns, and yielding IBD sharing posterior probabilities.

As linkage analysis relies on using genetic data from related individuals belonging to a small number of generations, this also means that there will be a limited number of observable meioses. A consequence of this is that regions identified by linkage analysis are broad, encompassing many genes, and often, secondary analyses are required to fine-map the region and pinpoint the causal variant(s). A classic example of this is the *CFTR* gene locus, responsible for cystic fibrosis. In the steps taken to identify this locus, first a large region on the long arm of chromosome 7 was identified through linkage analysis [33], but this region had to be fine-mapped with molecular genetic approaches such as chromosome walking and positional cloning [121] in order to home in on the causal gene.

While linkage analyses have been successful in identifying loci underlying Mendelian diseases of simple genetic architecture, they have had less success with mapping complex trait loci, because the effect of any one locus influencing a complex trait is often too small to be detected unless data with very large sample sizes are available [116]. Yet if some rare variants of strong effect contribute to the genetic architecture of complex traits, those would cluster amongst related individuals and linkage analysis methods would be suited to flag regions harbouring those co-segregating with high or low trait values. Within this chapter, I study complex traits in large populations of related individuals, including population isolates, and explore an analytical framework for pedigree-free linkage analysis, which aims to maximise the number of informative individual pairs that share regions that are IBD by including relationships that are not indicated in a social pedigree.

4.1.1 Pedigree-free Linkage Analysis

Linkage analysis, by definition, requires the use of individuals who shared a recent common ancestor and it cannot be performed on individuals who are completely unrelated. Linkage studies traditionally use social pedigrees to track this relatedness between individuals. While the cohorts studied within this thesis consist of related individuals, detailed genealogy information was not always available. Social pedigrees can be stitched together with the help of genetic data, as is discussed in section 2.4, but this is very time-consuming and it also becomes difficult to ascertain the correct way in which distantly related individuals are connected.

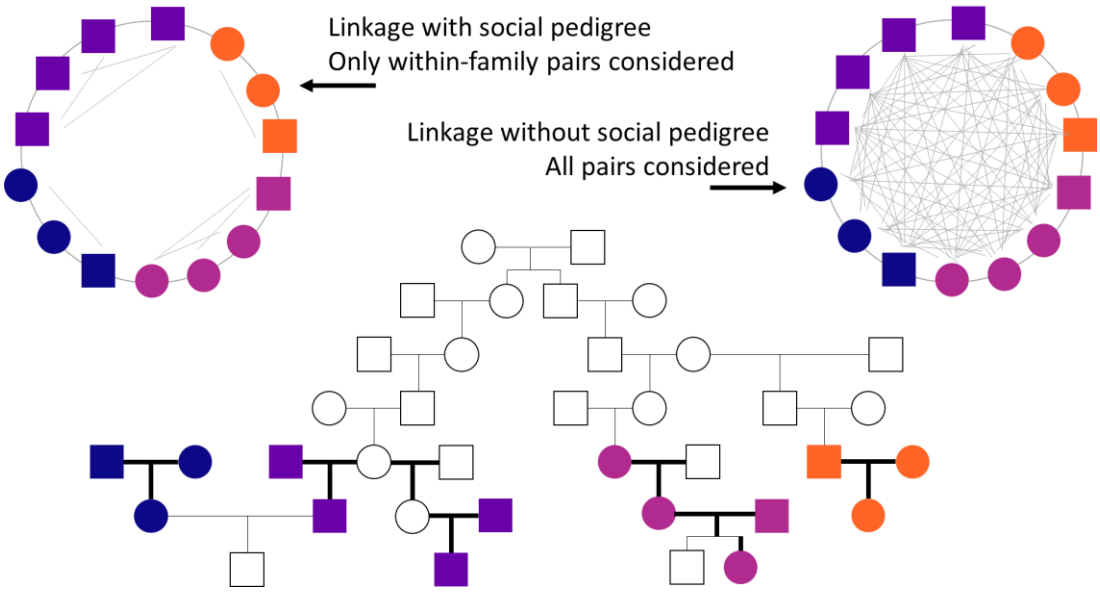
Another drawback of using known pedigrees to perform linkage analysis is the loss in power due to the inability to use more distant relatives who are not connected in the social pedigree, but still share some DNA segments IBD. These individuals might not be linked in the pedigree simply because of insufficient genealogical information or the need to "clip" large pedigrees to a manageable size so that each family falls under a certain bit-size threshold to accommodate

the computational requirements of software used in traditional linkage analysis (e.g. Loki). This problem is illustrated in Figure 13. With the availability of novel IBD coefficient estimation algorithms, there is no longer a need to clip pedigrees and, in theory, social pedigrees do not need to be used at all. Nonetheless, limitations remain in the accuracy of the IBD coefficient estimations and some DNA segments that are IBD between (distant) relatives are likely to go undetected.

Within this chapter, I also discuss the framework for a pedigree-free linkage analysis methodology that aims to circumvent the need for a social pedigree, by relying entirely on genotype-derived kinship.

Figure 13 - Gain in sample size from pedigree-free linkage analysis

This is a schematic representation of pedigrees. Genotyped individuals are represented by coloured shapes, and the thick lines represent relationships as they exist in the pedigree files. White symbols represent individuals who are not genotyped, and thin lines represent genealogical connections that are not represented in the pedigree, either because they are unknown or because pedigrees had to be clipped. The circle schematic on the left represents the pairs that are considered in traditional linkage analysis, while the circle schematic on the right represents the pairs considered using the pedigree-free linkage analysis.



4.2 Methods

The basic principle of linkage analysis is that within a genetic region harbouring a QTL that affects a trait, relatives who have similar trait values will share more alleles IBD within this region compared to relatives with dissimilar trait values.

In this section, the statistical basis of variance component linkage analysis in complex traits is introduced, describing its implementation in the program SOLAR [122]. Also described are the statistical methods used to estimate IBD coefficients using genetic data and, optionally, social pedigrees, as implemented in different software packages. Finally, the linkage meta-analysis methodology is described.

4.2.1 IBD Coefficient Calculations with Loki

The datasets used within this thesis consist of thousands of people and hundreds of thousands of marker loci, and some large pedigrees have been reconstituted. As such, it is impossible to calculate IBD coefficients between individuals in these pedigrees using exact enumerations of all inheritance vectors as implemented in algorithms such as the Elston-Stewart or Lander-Green algorithms. In contrast, approximate methods using Markov Chain Monte Carlo (MCMC) algorithms can handle large pedigrees, multiple families (note that complex families may still be a limiting factor, as noted above) and a larger number of markers because they consider underlying segregation patterns in relation to their likelihood of occurring. The program Loki [84] version 2.4.5 was used to estimate IBD coefficients within 2.5 cM intervals along each chromosome, between pairs of related individuals, resulting in 1462 regions across the whole genome. Using the social pedigree, Loki uses an MCMC algorithm to estimate IBD coefficients between related individuals by drawing segregation patterns over 10000 sampling iterations and calculating the identity coefficients associated with the segregation pattern drawn. Each iteration tries to update every parameter in the segregation model. For the purposes of IBD coefficient estimation, these parameters consist of ordered genotypes and allele frequencies.

Loki does not take LD between markers into account, and because the inheritance probability of one marker is conditional on the inheritance probability of a neighbouring marker, this could result in distorted inheritance probabilities if markers are in LD. To avoid this problem, Loki needs to be supplied with a pruned genotype set to avoid incorrect estimation of IBD coefficients. Pairwise pruning was done using PLINK's `--indep-pairwise` command, using windows of 100 SNPs, shifting the window by 25 SNPs at each step and using an r^2 threshold of 0.2 to remove one of a pair of SNPs when their LD is greater than this value. This left the following number of SNPs in each cohort: 127770 (Generation Scotland), 71881 (Korčula), 81041 (Vis), 134099 (Shetland), 66818 (Orkney).

The input marker files were updated to include positions in cM (using Kosambi's map function [123]), using the map positions calculated from the 1000 Genomes Phase 3 dataset (available for download from http://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3/, using the files

called genetic_map_chr*_combined_b37.txt). For SNPs that were not in this database, cM position was calculated from surrounding SNPs, using the following formula:

$$cM_n = \frac{(cM_{n+1} - cM_{n-1}) * (bp_{n+1} - bp_n)}{bp_{n+1} - bp_{n-1}}$$

Where cM_n is the cM position of SNP *n* and bp_n is the basepair position of SNP *n*. Care was taken to assign a unique cM position to each SNP, as Loki is unable to deal with duplicated cM positions at multiple SNPs.

Loki can output results that can be used directly by the variance component linkage analysis software SOLAR, recoding IDs into SOLAR's internal IDs (IBDIDs), if the pedindex.out and pedindex.cde files (obtained by loading the social pedigree into SOLAR) are provided prior to performing the IBD coefficient estimation. These output files consist of pairs of individuals, twice their kinship coefficient at the given position (phi2, calculated using the following formula: $2\Phi = 2 * (\Delta_1 + \frac{1}{2} * (\Delta_3 + \Delta_5 + \Delta_7) + \frac{1}{4} * (\Delta_8)$, where Δ_n corresponds to the Jacquard's condensed identity state probabilities [124]), and their inbreeding coefficient (delta7), equivalent to Jacquard's $\Delta_7 + \Delta_1$ (the sum of the probabilities of both alleles between, and within individuals are shared IBD). Note that SOLAR ignores the inbreeding coefficient when performing linkage analyses without a dominance model.

4.2.2 IBD Coefficient Calculations with IBDLD

The program IBDLD [120] v3.32 was used to estimate IBD coefficients between pairs of people at SNPs along each chromosome, as well as to estimate whole-genome kinship. Given unphased genotype data, this program uses Hidden Markov Models (HMMs) to estimate multipoint IBD coefficients at each locus, taking into account the genotypes of all markers on the chromosome, and can also utilise the LD between SNPs to accurately infer haplotypes, so genotypes do not need to be LD pruned. It has options to estimate IBD coefficients with (LD-RR method) or without (GIBDLD method) social pedigree information.

IBDLD is robust to missing genotypes, accurately inferring the presence of IBD segments even when 20% per-individual, or 5% per-SNP missingness is present [39]. As the genotype data used in this thesis have been through stringent quality control, there will be, at most, 3% per-individual and 2% per-SNP missingness. Finally, IBDLD is robust to pedigree errors (in the LD-RR method, where pedigrees are used), as it can correctly infer IBD coefficients even when relationships are grossly misspecified – this again should not be the case in the data that were used in this thesis, as the pedigrees have been screened prior to analysis and gross pedigree errors (that did not match the genetic kinship) were fixed. While the GIBDLD method

(that estimates IBD coefficients without using a social pedigree) was not part of the initial release of IBDLD, this robustness to pedigree errors is the reason it was developed and implemented into later version of the program.

IBDLD estimates IBD coefficients through two separate computational steps. In step 1, the background LD parameters are estimated, while step 2 estimates the IBD coefficients at each SNP, between pairs of individuals.

Within this thesis, the LD-RR and GIBDLD methods were used. Both methods calculate LD between markers prior to inferring IBD coefficients, but GIBDLD does not require social pedigree information and was used to set up the pedigree-free linkage analysis methodology introduced in section 4.1.1 and described in more detail in section 4.2.5 below. Both methods use ridge regression to calculate LD patterns, jointly conditioning on the genotypes of two sets of markers: the genotypes of *n* previous SNPs (here set to 10, the default option) as well as the genotypes of all previous SNPs within *k* cM (here set to 2.5 cM to align with the input parameters used with Loki).

cM positions were assigned to each SNP as described in section 4.2.1 above and, while all SNPs along the chromosome are used to infer IBD at each SNP, output was only requested at 0.1 cM intervals prior to performing pedigree-based linkage analysis, resulting in IBD coefficients at ~33,000 SNPs. This was done to reduce the computational time of the linkage analysis that was performed with SOLAR, and because, due to the LD pattern between SNPs, linkage peaks tend to be broad, encompassing several cM, and denser sampling does not narrow the width of the associated "peak" region (although can more accurately pinpoint the position of the apex), as shown in Figure 15, where pedigree-based linkage analysis results using IBD coefficients calculated by Loki and the two IBDLD methods are plotted together.

IBDLD can output probabilities for the 9 Jacquard's condensed identity coefficients [124] at each locus as well as for each chromosome (by averaging the identity coefficients from all markers on the chromosome). It also outputs the IBD sharing (the kinship coefficient) at each locus and each chromosome (calculated by averaging the kinship coefficients from all markers on the chromosome), as well as the genome-wide kinship coefficient. The genome-wide kinship coefficient is calculated using the following formula:

$$\hat{\pi}_{ij} = \frac{1}{\sum_{h=1}^{nChr} W_h} \sum_{h=1}^{nChr} W_h \hat{\pi}_h$$

Where W_h is the number of SNPs on each chromosome and $\hat{\pi}_h$ is the hth chromosome's chromosome-wide kinship coefficient between individuals i and j. $nChr$ is the total number of chromosomes (22 in humans).

As IBDLD does not automatically output SOLAR-formatted files, I have written R scripts that convert IBDLD output files into the SOLAR input files, creating IBD files in the same format as described in section 4.2.1 above, by adding $\Delta_7 + \Delta_1$ to obtain the "delta7" value used by SOLAR, and multiplying the kinship coefficient by 2 to obtain 2Φ.

When pedigree data were used (LD-RR method), per-SNP IBD output was requested for every genotyped pair *within* each family. In Orkney and Vis, where the pedigree-free IBD estimation (using IBDLD's GIBDLD method) and linkage analysis was trialled, output was requested for *all* genotyped ID pairs in the dataset (rather than only pairs of individuals within families), at each genotyped SNP. Analysing a region in SOLAR with pedigree-free linkage analysis takes approximately 8 minutes in Orkney, so instead of analysing SNPs at 0.1 cM intervals as described above, in order to reduce the analysis time, the genome was split into 2.5 cM regions and the kinship coefficients of all SNPs within each region were averaged to obtain a regional kinship coefficient. This resulted in 1446 non-empty regions (that is, regions that contain at least one SNP). The median number of SNPs in each region was 107 SNPs in Orkney (180 SNPs in Vis), and the interquartile range was 47 SNPs in Orkney (93 SNPs in Vis). Figure 14 shows the distribution of the number of SNPs allocated to each region, within each cohort.

Figure 14 - Distribution of SNPs into 2.5 cM regions in Orkney and Vis

The X axis shows the number of SNPs in each 2.5 cM region across the autosomal genome, grouped into 8 bins. The Y axis shows the number of regions belonging to each bin, in each cohort.

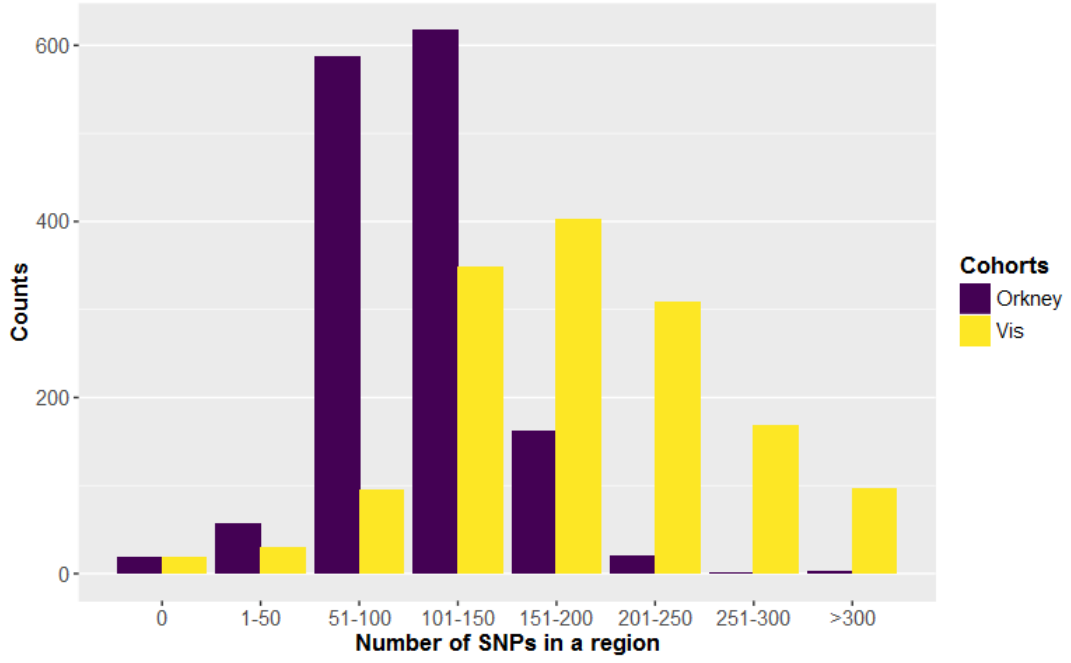
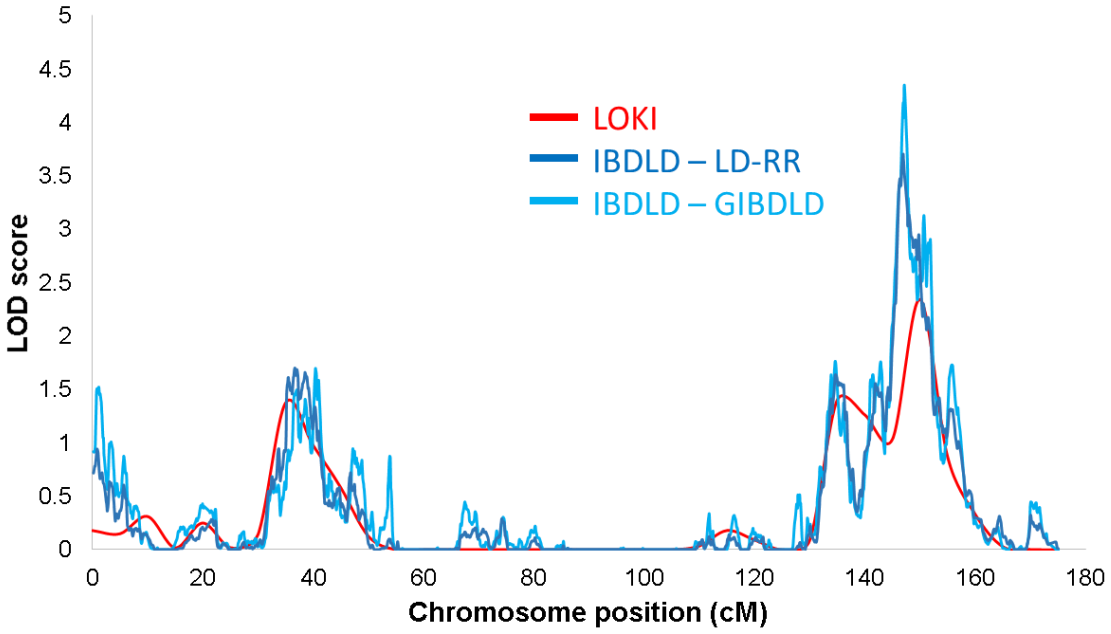


Figure 15 - Results of linkage analysis using SOLAR, with IBD coefficients calculated by Loki, LD-RR or GIBDLD.

This figure shows the educational attainment linkage analysis results on chromosome 12 in Orkney. The SOLAR analysis is identical in all three cases (it uses the same social pedigree to estimate genome-wide kinship), only the regional IBD estimations differ. In red, IBD coefficients calculated at 2.5 cM intervals (for a total of 64 regions on this chromosome) by Loki, using pruned SNP data. In dark blue, IBD coefficients calculated using the LD-RR method in IBDLD (uses pedigree information). In light blue, IBDs calculated using the GIBDLD method in IBDLD (does not use pedigree information). IBD coefficients at 1540 SNPs along chromosome 12 were obtained from both IBDLD methods. Note that while GIBDLD does not use pedigree information to calculate IBD coefficients, SOLAR was still provided with a pedigree to estimate linkage (in order to make the results comparable), so this figure does not show results obtained with the pedigree-free method since only relative pairs who are connected in the social pedigree are used.



4.2.3 Variance Component Linkage Analysis in Complex Traits

In variance component linkage analysis, phenotypic variance is partitioned into components derived from fixed effects and random effects consisting of additive polygenic variance, additive QTL variance and residual variance. The polygenic term Zu described in section 3.2.2 is therefore split into two components, hence there is an additional variance

component, Wv, that needs to be estimated. In variance component linkage analysis, the trait *y* can therefore be expressed as

$$y = X\beta + Zu + Wv + e$$

where Z and W are the design matrices for random effects, u is the whole-genome additive effect with variance $var(u) = \Phi\sigma_u^2$, v is the regional additive effect with variance $var(v) = \widehat{\Pi}\sigma_v^2$, e is the residual effect with variance $var(e) = I\sigma_e^2$. Matrices Φ, $\widehat{\Pi}$ and I are the whole-genome kinship coefficient, regional IBD coefficient and the identity matrix.

This generalises to pedigrees of arbitrary complexity, such as those used within this thesis, and the covariance matrix for such data can be written as $\Omega = \sum_{i=1}^{n} \widehat{\Pi}_i \sigma_{\gamma i}^i + 2\Phi\sigma_g^2 + I\sigma_e^2$ where $\widehat{\Pi}_i$ is a matrix with elements π_{ikj} that provide the proportion of alleles that are IBD between individuals *k* and *j* at the *i*th QTL, Φ is the matrix of kinship coefficients and I is the identity matrix.

Note that the kinship coefficient can be derived in two ways. It can be estimated from a social pedigree, in which case the pairwise kinship coefficients are calculated based on the average expected amount of genetic sharing given the number of meioses separating the two individuals through their most recent common ancestor or ancestors (where common ancestor is defined as an individual whose descendants are not also a common ancestor of the two individuals), using the formula $\Phi_{jk} = \sum_{x=1}^{n}(1 + F_x)\frac{1}{2^{M_x+1}}$, where Φ_{jk} is the kinship coefficient between individuals j and k, F_x is the inbreeding coefficient of their most recent common ancestor x, and M_x is the number of meioses separating them through ancestor x. The pedigree-based kinship is always the same fixed value for pairs with the same degree of relatedness. For example, in the absence of inbreeding, a parent-child pair or a pair of full siblings will always have an estimated kinship coefficient of 0.25 ($2\Phi = 0.5$). While a child inherits exactly half of a parent's genome, realised kinship for full siblings can vary greatly around this average value due to the randomness of inherited chromosomes at each meiosis.

Whole-genome kinship coefficients and regional IBD coefficients can also be derived solely from genetic data, as described in the sections above, and these values reflect the true amount of genetic sharing between a pair of individuals.

Linkage analysis was performed using the program SOLAR [122], which carries out linkage analyses using variance component methods, considering the likelihood of the entire pedigree (which may consist of multiple, complex families) jointly. SOLAR can calculate multipoint IBD coefficients internally but also accepts externally-computed IBD coefficients as long as

the files containing these are provided in the appropriate format. Note that while SOLAR allows for the estimation of dominance and epistasis effects, within this thesis, only additive effects are assessed.

SOLAR uses maximum likelihood functions to test for the presence of a QTL in a given region by comparing the full model (QTL effect + polygenic effect) with a null model that does not include the QTL effect. The null hypothesis is (H_0) is the absence of regional contribution to the variance, while the presence of regional contribution constitutes the alternative hypothesis (H_1). Within each region, these hypotheses will have likelihoods L_0 and L_1, respectively and a likelihood ratio test (LRT) statistic can be calculated from these with $LRT=-2\ln(L_0/L_1)$. The LRT follows a distribution that is a 50:50 mixture of a point mass at 0 and a χ^2 distribution with 1 degree of freedom [74]. To correctly account for the fact that this is a one-sided test, the *p-value* for the LRT needs to be divided by 2. SOLAR expresses the test statistic as a log of odds (LOD) score, to be in line with previous linkage studies, which is calculated by dividing the LRT by $2\ln(10)$.

Within this thesis, the "multipoint" command of SOLAR was used to run linkage analyses using IBD coefficients estimated at 2.5 cM intervals along each chromosome by the program Loki. In contrast, the IBD coefficients calculated by IBDLD were analysed using the "twopoint" command in SOLAR, which calculates linkage at each input SNP. Note that while this is called a twopoint analysis in SOLAR (as the input to SOLAR is one SNP at a time), in reality, this is also a multipoint analysis, as the IBD status at each SNP was calculated using other markers on the chromosome.

4.2.4 Converting LOD Scores to P-values

As stated above, linkage (and also RH) analyses use likelihood ratio tests (LRTs) to test for the presence of a QTL in a given locus. The RH analysis software reports associated *p-values*, while the linkage analysis software converts them to LOD scores, to be in line with the conventional measure of confidence used in linkage studies performed on traits with Mendelian inheritance (parametric linkage analysis). While both LOD scores and *p-values* are used in this thesis, some comparisons are facilitated by all results being reported using the same metric (*p-values*).

In order to convert LOD scores into *p-values*, the method described by Nyholt [74] is used. LOD scores are first converted to LRT statistics with the formula $LRT=LOD\times2\times\ln(10)$, since $LOD_\theta = \log_{10}(\Lambda_\theta)$ and $LRT_\theta=2\times\ln(\Lambda_\theta)$, where θ is the parameter estimated by the maximum likelihood method and Λ is the likelihood function. The distribution of these LRTs is a 50:50

mixture of a chi-square distribution with 1 degree of freedom and a point mass centred at 0 and is converted to a *p-value* using the pchisq() function in R followed by dividing these *p-values* by 2 to reflect the properties of its distribution, using this function: p(LOD)=0.5×pchisq(LOD×2×ln(10), df=1, lower.tail=FALSE).

4.2.5 Pedigree-free Linkage Analysis

Pedigree-free linkage analysis attempts to bypass the need for a social pedigree altogether, using the "GIBDLD" method within the IBDLD3 program [120], as described above. GIBDLD can estimate IBD sharing at given SNPs even in the absence of a social pedigree, which was not possible using the Loki software [84].

As described above, IBD sharing was estimated at each genotyped SNP, between all pairs of individuals in the cohort, the genome was split into 2.5 cM regions and regional IBDs were obtained by averaging the IBDs of all SNPs within each region. This larger interval was used with pedigree-free linkage in order to reduce the analysis time, as analysing each region with pedigree-free linkage analysis takes approximately 8 minutes in Orkney, so analysing SNPs spaced at 0.1 cM intervals would have been unfeasible. Following the IBD estimation step, linkage analysis was carried out in SOLAR as described previously, with the following differences:

1. The true social pedigree was not used by SOLAR. However, SOLAR needs to load a pedigree prior to any analysis. For this purpose, the "pedigree" file containing a list of the genotyped individuals only, without any parental information, is provided.
2. By default, this "pedigree" is used by SOLAR to estimate the genome-wide kinship coefficient (which is going to be 0 between all pairs as no actual relatedness information was recorded in this pedigree). This matrix is overridden by the genome-wide kinship matrix estimated by GIBDLD.
3. By default, SOLAR only calculates linkage using pairs whose kinship coefficient is not 0, as calculated from the social pedigree. To force SOLAR to use all pairs of individuals in the data, the 'option MergeAllPeds 1' command was used, which uses the IBD sharing estimates between all pairs of individuals, regardless of their relatedness according to the social pedigree.

Where a regional LOD score exceeded the multiple testing-corrected GWS threshold, linkage analysis was additionally run on every SNP in the region separately, by using the IBD coefficients estimated at each SNP within that region.

4.2.6 Linkage Analysis Power Calculations in SOLAR

The methods described in this section are used to assess the performance of each cohort for linkage studies. They highlight the importance of large families over a simple increase in sample size in order to increase power to detect a QTL using linkage analysis. The social pedigrees derived in each of the cohorts (see section 2.4) were used in this section.

The power to detect a linkage signal with a LOD score of 2 or 3 was assessed in each population. This was done using the 'power' command in SOLAR. These calculations were done by simulating traits with total heritability ranging from 25% to 80% (increasing in increments of 5%), with one QTL explaining 0 to 100% of the total heritability (increasing in increments of 1%). QTLs with allele frequencies of 2, 10, 25 and 50% were used to assess the power to detect linkage.

SOLAR uses the social pedigree to assign reference or alternate causal alleles to individuals, using a drop-down method. This means that it first randomly assigns alleles to a number of founders based on the minor allele frequency, and then propagates these alleles through the families using chromosomal segregation.

4.2.7 Meta-analysis

Linkage analysis was performed using the same pipeline, and phenotypes quality controlled in the same way, in 5 different cohorts. Meta-analysing the results may reveal common patterns emerging in several cohorts that may not have been obvious when each cohort was analysed on its own.

As described in section 4.2.2, pedigree-based linkage analysis was performed at every 0.1 cM along the chromosome in each cohort, which partitions the genome into 36269 regions. With pedigree-free linkage analysis, linkage analysis was performed at 2.5 cM intervals along each chromosome, which results in 1464 regions across the genome. Some of these regions may contain no genotyped SNPs in some cohorts – in such cases, the LOD score of the nearest non-empty region was used. The LOD scores within the corresponding region in each cohort were meta-analysed using Fisher's combined test, described below.

In 1952, Ronald Fisher described a statistical method that allows the combining of *p-values* into a joint test to determine whether the global null hypothesis can be rejected [125]. To use this method, LOD scores are first converted into *p-values*. This is done by calculating $LOD * (2 * ln10)$, which yields an LRT test statistic with a distribution that is a 50:50 mixture of a point mass at 0 and a χ^2 distribution with 1 degree of freedom [74], from which the *p-value* can be determined using the upper tail of a χ^2 test, using the pchisq() function in R. To account

for the 50:50 mixture of function distributions, the *p-values* are divided by 2. Then, the statistic $-2\sum \ln(p)$, is calculated at each locus, which, following Fisher's combined test, has a χ^2 distribution under the null. The global *p-value*, P_{meta}, for each locus is obtained by taking the upper tail of a χ^2 test with degrees of freedom equal to 2 times the number of studies used.

Regions with $-\log_{10}(P_{meta}) > 4.43$ (corresponding to LOD = 3.41) are considered suggestively significant, while regions with $-\log_{10}(P_{meta}) > 5.93$ (corresponding to LOD = 4.84) are considered genome-wide significant, taking into account the number of traits analysed.

If the linkage signals are generated by rare variants or rare combinations of variants in one population, replication in another population depends on whether these alleles are also segregating within the target populations. Therefore, it could be the case that a QTL segregating in Scotland is not segregating in Croatia (or vice-versa), and therefore cannot be detected in that population. This could dilute the signal of the meta-analysis, so in addition to analysing all the cohorts together (which could reveal common linkage regions across European populations), the cohorts were also subdivided into the following groups: Croatia (Vis and Korčula), Scotland (GS, Shetland, Orkney) and Scottish Islands (Orkney and Shetland) and meta-analysis was performed within these groups as well.

4.3 Results

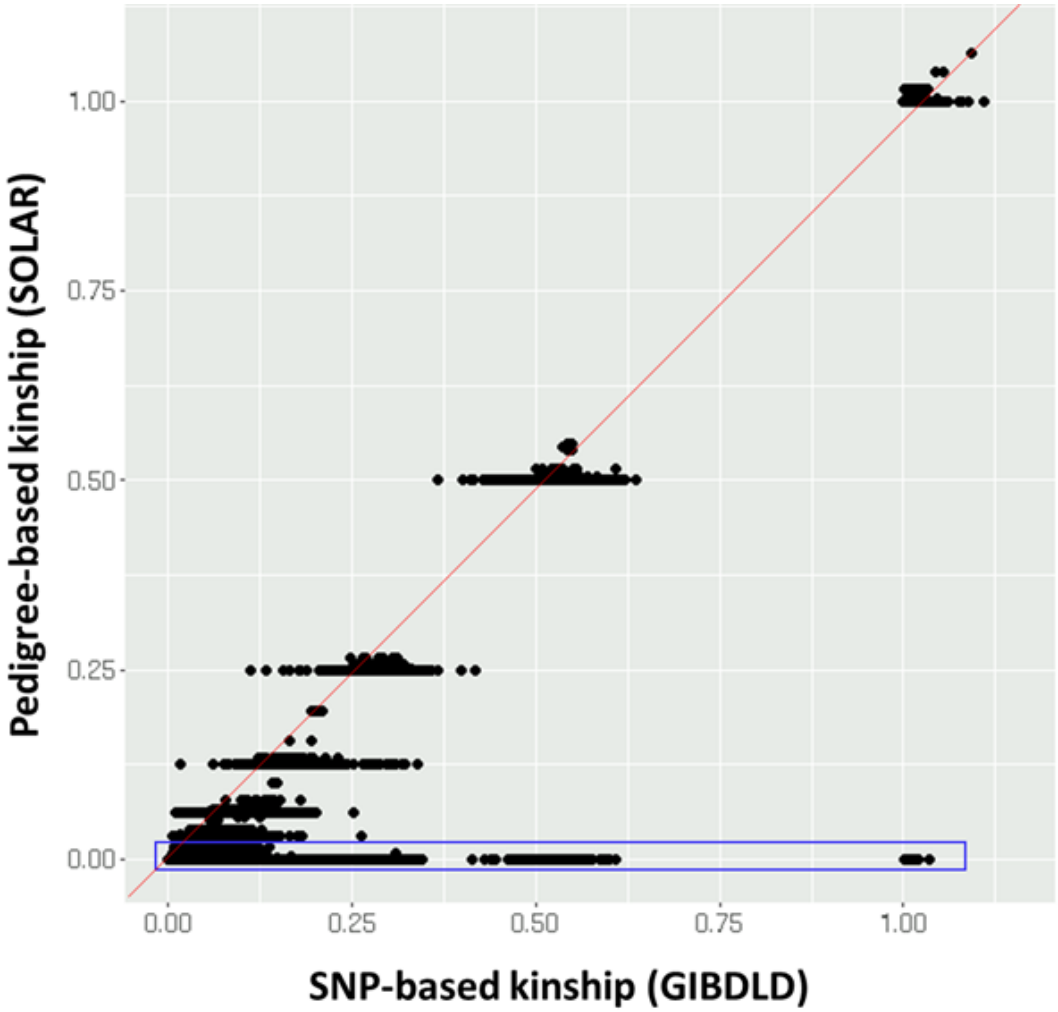
4.3.1 Identity by Descent Estimations

4.3.1.1 Genome-wide Kinship

Figure 16 shows a comparison between the genome-wide pairwise relationship coefficient (twice the kinship coefficient=2Φ) in Orkney, as calculated by SOLAR using the clipped social pedigree and calculated by IBDLD's GIBDLD (pedigree-free) method, which uses solely the genotype data.

Figure 16 - Pedigree-based and SNP-based whole genome kinship

The SNP-based kinship, as calculated by IBDLD's GIBDLD (pedigree-free) method is plotted on the X axis. The Y axis shows the kinship estimated from the clipped social pedigree by SOLAR. All pairs (including self-pairs, in the top right corner) in Orkney are depicted. Pairs in the blue rectangle are pairs who are unrelated in the social pedigree but have non-0 kinship according to IBDLD. Note that here, 2Φ is used (that is, twice the kinship coefficient, which can range from 0 to 2, and takes on a value of 1 in self-pairs in the absence of inbreeding).



The genome-wide kinship, as calculated by GIBDLD, is nearly identical to the genetic relationship matrix calculated using the IBS-based relatedness formula described in section 3.2.1 (for a comparison of the relationship-values calculated by these two methods, see section 7.1.1), but shows marked differences compared to the pedigree-based kinship. For example, 126731 pairs of individuals are unrelated according to the clipped social pedigree but have $2\Phi > 0.03$. 2091 of these are seemingly compatible third degree relatives or closer relatedness ($2\Phi > 0.125$). Other than these pairs of relatives who appear unrelated in the social pedigree, there appear to be no gross misspecifications of relationships in the social pedigree (gross misspecification being defined as a difference of more than 0.25 between the pedigree-based kinship and the genetic relatedness), which is expected, given that such misspecifications were addressed prior to downstream analysis, as described in section 2.4.

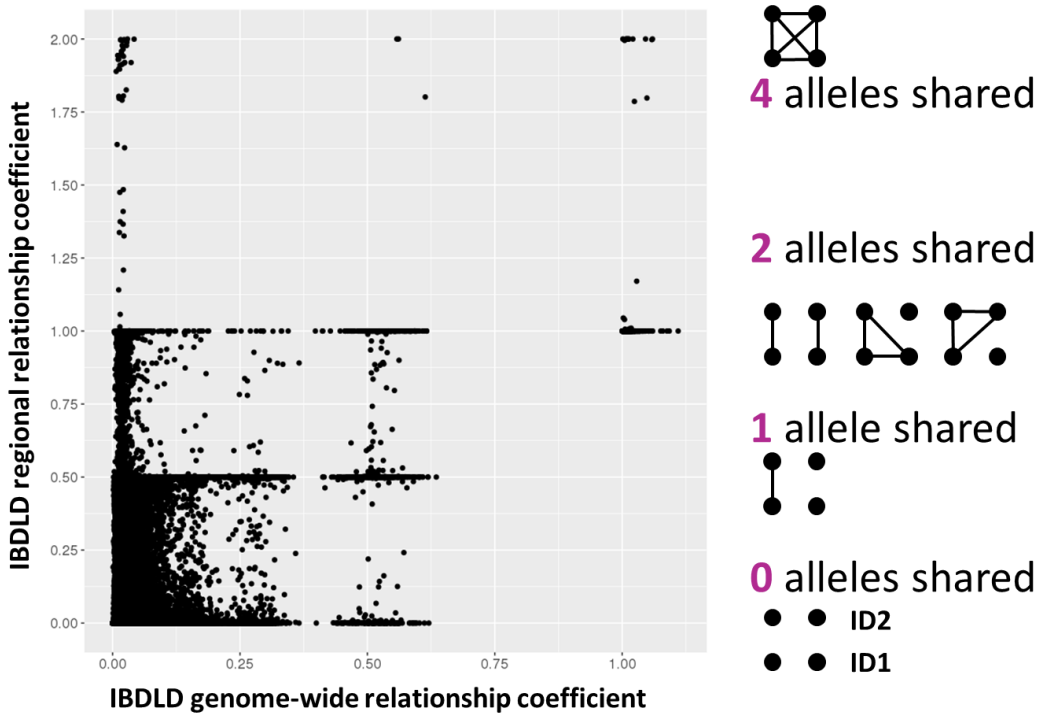
4.3.1.2 Regional Kinship

When two individuals are distantly related, their probability of sharing IBD segments decreases exponentially with each meiosis that separates them, but should they share an IBD segment, the length of it decreases only linearly with each meiosis [126]. The estimation of pairwise kinship coefficients based on the genotype data indicates that many pairs of individuals in Orkney are distantly related, therefore there is a lot of potential in using these pairs in linkage analysis. Additionally, the rationale behind the pedigree-free linkage analysis also makes use of the fact that regions where DNA segments are shared IBD come in large blocks and therefore may be easier to ascertain.

Figure 17 illustrates the difference between the genome-wide kinship coefficient and the regional kinship coefficient, where the latter was obtained by averaging the GIBDLD-IBD estimates output at 28 SNPs in a 0.3 cM region around the *SLC2A9* gene. 48131 pairs who are unrelated in the clipped social pedigree have a kinship-value that indicates the sharing of at least 1 allele IBD throughout this region, and 45866 of these pairs are less than 4th degree relatives ($2\Phi < 0.0625$). Additionally, 34 pairs appear to be homozygous by descent (HBD) in this region (average regional $2\Phi > 1.5$) despite their low overall relatedness (genome-wide $2\Phi < 0.04$).

Figure 17 - Whole genome versus regional kinship at the SLC2A9 locus

The SNP-based whole genome relationship coefficient (2Φ) between all pairs of individuals (including self-pairs) is plotted on the X axis. The regional average relationship coefficient (2Φ) obtained from the 28 SNPs in the 0.3 cM region around the *SLC2A9* gene is plotted on the Y axis. The possible configurations of alleles shared IBD corresponding to regional relationship-values of 0, 1, 2 and 4 are plotted on the right hand side on their respective rows – here, the unordered alleles of two individuals are shown as nodes and the lines connecting these nodes indicate that they are IBD. GRM, genetic relationship matrix.



4.3.1.3 Pedigree-free IBD coefficient Estimation Accuracy

The benefit of pedigree-free IBD estimation is that it enables the use of regions that are shared IBD by distantly-related individuals, increasing the power to detect a QTL with linkage analysis. This way, untyped rare variants may be detected as long as they segregate on a clear haplotype. However, the power gained will critically depend on the accuracy of the IBD coefficients.

In order to evaluate the IBD coefficients estimated by GIBDLD, I chose a relatively rare variant (rs16865292, 2.5% C allele frequency) and explored its segregation within a family in Orkney, as well as in the whole Orkney cohort. Shared haplotypes in the vicinity of this variant, inferred through the IBD coefficients estimated by GIBDLD, were compared to haplotypes

obtained from genotype data that had been phased using the SHAPEIT program version 2.r644, with the duoHMM option that made us of pedigree information (run previously by Dr Peter Joshi for imputation purposes). Because the SHAPEIT output matches well with expected Mendelian inheritance in the families recorded, it is considered as the "true" haplotype. In the phased data, the haplotype region was defined as the region consisting of rs16865292 and 25 flanking SNPs on either side.

Figure 18 shows this family's pedigree, and indicates the rs16865292 alleles carried by specific individuals. Within this family, there is one "major" (more common) haplotype, originating from one of the founders, that carries the C allele. There are three other distinct haplotypes, originating from married-in individuals, which also carry this allele. Table 10 illustrates the four different haplotypes on which the C allele appears in this family.

Using the genotypes of the whole Orkney dataset, GIBDLD accurately detected sharing of rs16865292 C-carrying haplotypes, and correctly distinguished between them in all but one pair within this family. While ORCA1193 and ORCA1607 (outlined in red in Figure 18) carry two different C haplotypes, as well as two different T haplotypes, GIBDLD estimated that they share one allele IBD. GIBDLD correctly distinguished the different haplotypes that carry the T allele, indicating that, for example, ORCA1202 and her sister, ORCA1345, do not share this region IBD, while ORCA1345 and her aunt, ORCA3951, share one of the T haplotypes.

Next, I broadened this comparison to include all pairs of individuals in the cohort. I extracted pairs of individuals where both members of the pair carry at least one C allele at this SNP. There were 5151 such pairs. I then looked at whether they share the haplotype of the C allele (and the T allele, when both carried this allele), as ascertained by the phased data, to evaluate the consistency between the IBD coefficients output by GIBDLD and haplotype sharing. 549 pairs shared at least one C haplotype, and all were reported to share at least one allele by IBD GIBDLD. 170 of these pairs consisted of an individual from the family shown in Figure 18 and an individual who is unrelated to them according to the social pedigree. The median genome-wide kinship between these 549 pairs was 0.017. Three individuals carried homozygous C haplotypes, which was also predicted by GIBDLD.

Conversely, there were 394 pairs who shared no C or T haplotypes, but GIBDLD reported that they share at least one allele. I investigated whether some of these "false positives" could be explained by the way in which GIBDLD constructs the LD patterns – as mentioned in the Methods section of this chapter, at each SNP, IBDLD conditions jointly on the genotypes of the 10 previous SNPs, as well as the genotypes of all previous SNPs within 2.5 cM. Where both the 10 previous SNPs, as well as SNPs within 2.5 cM match between two individuals,

GIBDLD accurately detects IBD sharing (which is the case for all 170 distantly related pairs that share C haplotypes). However, if the 10 previous SNPs match but there is some mismatch in the SNPs within 2.5 cM, IBD sharing accuracy seems to decrease, as detailed below:

There were 30 different haplotypes carrying the C allele in Orkney, based on the phased output. Figure 19 illustrates these haplotypes by showing the 25 SNPs up- and downstream of rs16865292. Looking only at the 10 SNPs upstream of rs16865292, however, these 30 haplotypes can be condensed into 8 different ones (coloured rectangles in Figure 19). Three quarters, or 295 out of the 394, "false positive" pairs (who are reported to share at least 1 allele by GIBDLD but do not share haplotypes according to the phased data) share these reduced haplotypes (and a further 33 share haplotypes on the T allele), but not the 2.5 cM region, which could explain why GIBDLD is not able to accurately distinguish IBD sharing here.

The reason GIBDLD looks at IBD sharing at the previous 10 SNPs as well as all SNPs within 2.5 cM becomes apparent when we look at the pairs that share the smaller, but not the larger region –1075 pairs share these reduced haplotypes (but not the bigger haplotypes), and GIBDLD detects that they do not share rs16865292 IBD – eliminating a lot of potential false positives. It is not clear why GIBDLD determines that 295 pairs share this region IBD despite this not appearing to be the case according to the phased data.

One possible explanation for the remaining 66 pairs that do not share this smaller region's haplotypes but have IBD values that suggest they do could be the fact that GIBDLD was unable to correctly determine the haplotypes segregating in these people (the input to GIBDLD consists of genotype data that have not been phased, and no pedigree information was provided). Alternatively, GIBDLD did accurately estimate sharing and the phasing could be incorrect in these people.

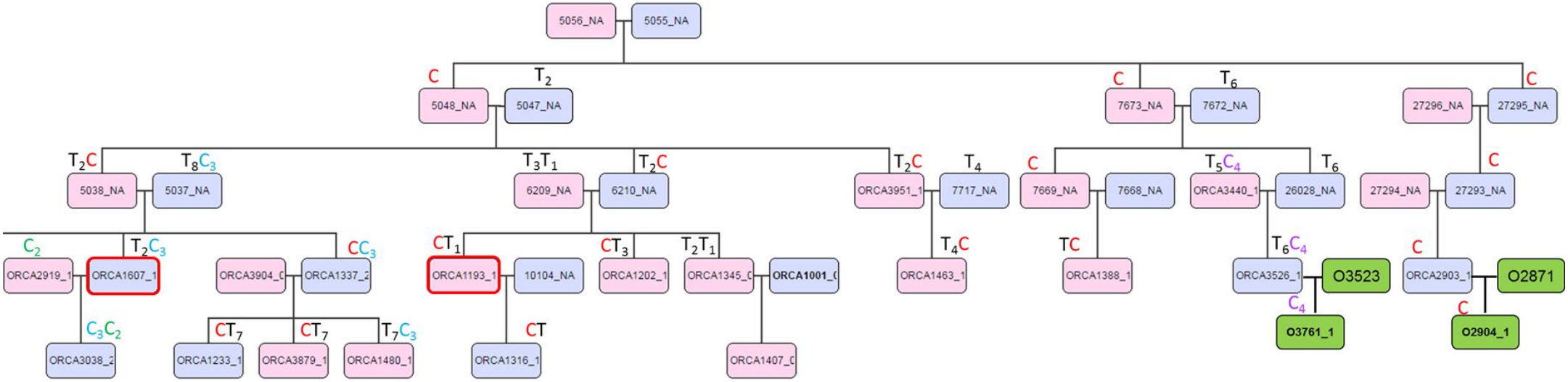


Figure 18 - Tracking rs16865292 in an Orkney family.

This is a schematic representation of the Orkney family. Pink rectangles indicate females, blue rectangles indicate males. The _0, _1, _2 or _NA following each ID indicates the number of C alleles (the minor allele) each individual has at this SNP, as ascertained from the genotype data. I indicate the 4 different haplotypes that carry the C allele with different colours and subscripts, and the "major" haplotype carrying the C allele is shown in red. The polymorphism is C/T and subscripts of the T allele show the haplotypes that carry this allele. These haplotypes were inferred from phased data and are shown in Table 10. The two individuals highlighted by the red rectangles are reported to have 1 allele IBD by IBDLD, even though they have no matching haplotypes in the region. The individuals in green were not part of the clipped social pedigree (they were part of the full pedigree), but IBDLD detected the shared haplotype between them and their parents.

Table 10 - Four distinct haplotypes carrying the C allele of rs16865292 in an Orkney family.

rs16865292 is positioned in the column with thick borders, and the alternate alleles (0) at each SNP have a green background, while the reference alleles (1) have a yellow background.

C	1	1	1	0	0	1	1	1	0	1	1	1	1	1	0	1	0	1	0	1	0	0	1	1	1	1	0	1	1	1	1	1	1	1	0	1	0	0
C2	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1
C3	0	1	1	0	1	1	0	1	0	1	1	1	0	1	0	0	1	1	0	1	1	0	1	1	1	1	1	0	1	0	0	1	1	1	0	1	0	0
C4	1	1	1	0	1	1	1	1	0	0	0	1	1	1	1	1	1	0	1	1	1	0	1	1	1	1	1	0	1	0	1	1	1	0	0	1	0	1

Figure 19 - 30 different haplotypes carrying the C allele at rs16865292.

The genotypes of rs16865292 (at point 0 on the Y axis) and the genotypes of the 25 flanking SNPs are pictured, for each haplotype. Each column represents a haplotype, named along the top and bottom x axes. Each haplotype is distinct when all 51 SNPs are considered, while each differently-coloured box indicates a different haplotype composed of the lead SNP and the 10 previous SNPs only.

4.3.1.4 Regions under Selective Pressure

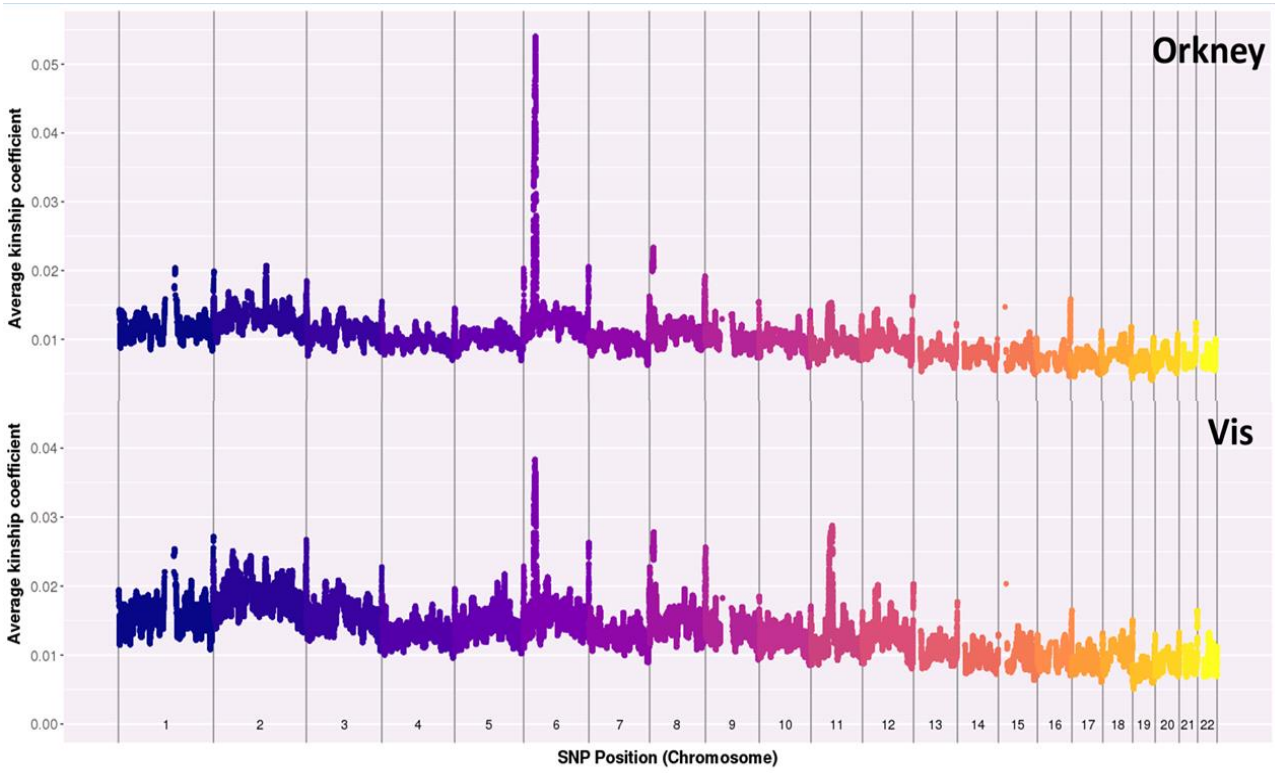
As IBDLD allows for the estimation of IBD sharing between all pairs of people in study samples of arbitrary size, it becomes possible to calculate the average IBD sharing at each genotyped SNP, as well as the average genome-wide IBD. This gives an overview of the population-wide IBD in a cohort, at each SNP, and can help detect regions where the amount of IBD sharing is higher (or lower) than the average IBD sharing.

The average kinship coefficient across the genome is 0.00995 in Orkney and 0.013 in Vis, corresponding to 6th and 5th degrees of relatedness, respectively. These are higher than the sharing observed in datasets of unrelated European individuals, which is around 0.0022 [127]. As both Orkney and Vis are population isolates consisting of many related individuals, this higher average sharing is not unexpected. Previous analyses of the Orkney and the Croatian cohorts show similarly increased levels of IBD sharing [128].

Figure 20 shows the average kinship coefficient at each SNP in Orkney and Vis. One feature that can be observed from these plots is that the average chromosome-wide IBD sharing correlates with chromosome length (correlation coefficient 0.91 in Vis, 0.8 in Orkney), ranging from 0.007 to 0.018 in Vis and 0.006 to 0.015 in Orkney. This is a consequence of crossing over (recombination) during meiosis: homologous chromosomes align during meiosis and chromatids from different chromosomes come into contact at one or more points along their length, resulting in a crossover. To ensure accurate chromosome cohesion and segregation, at least one recombination needs to occur on each chromosome. The recombination rate is higher in shorter chromosomes, so IBD along shorter chromosomes breaks down faster [129]. One crossover inhibits the formation of additional crossovers nearby, in a process known as crossover interference. Crossover interference is higher on shorter chromosomes, which decreases the probability of double crossovers occurring [130]. However, it appears that the reduced rate of double crossovers on shorter chromosomes is not enough to compensate for the higher number of recombination events.

Figure 20 - Average kinship at each SNP across the genome in Orkney and Vis.

The X axis marks the position of SNPs along the genome. The Y axes show the average kinship coefficient (φ) at each SNP, for Orkney (top) and Vis (bottom).



There are also several peaks and trends that can be observed in Figure 20 and these agree with previous reports of genomic features. The peak on chromosome 6 represents the human leukocyte antigen (HLA) region that is well-documented as being under a strong selective pressure, leading to lower haplotype diversity as it contributes to immune function [127, 131]. The telomeric regions show increased IBD sharing followed by a region of decreased IBD sharing. One study ([132]) reported that this is due to a telomere-mediated suppression of recombination at the chromosome extremes, followed by a region of high recombination. This leads to telomeric haplotypes that are unaffected by natural selection and recombination and this can be used for lineage tracing over large time scales. Other studies have found similarly increased sharing at telomeres [133, 134] but others also report reduced sharing [135] at this region. Centromeric regions also show increased sharing, which is likely due to the suppression of recombination here [136]. A region around the centromere of chromosome 11 shows increased sharing, especially in Vis. This region contains clusters of olfactory receptor genes, and other studies have also reported excess sharing across populations in this region [134]. The region of increased sharing between 9 and 11Mb on chromosome 8 has also been previously documented by several studies [134, 137] including the Generation Scotland study [138]. This region is near a cluster of defensin genes, and one study has found that it is also the site of a large polymorphic inversion [139]. The region of increased sharing on chromosome 2 (2q21.3-2q22.1) has also been previously documented [39], and contains 45 genes including the lactase gene, which has strong evidence for being under selective pressure [140].

4.3.2 Power Calculations

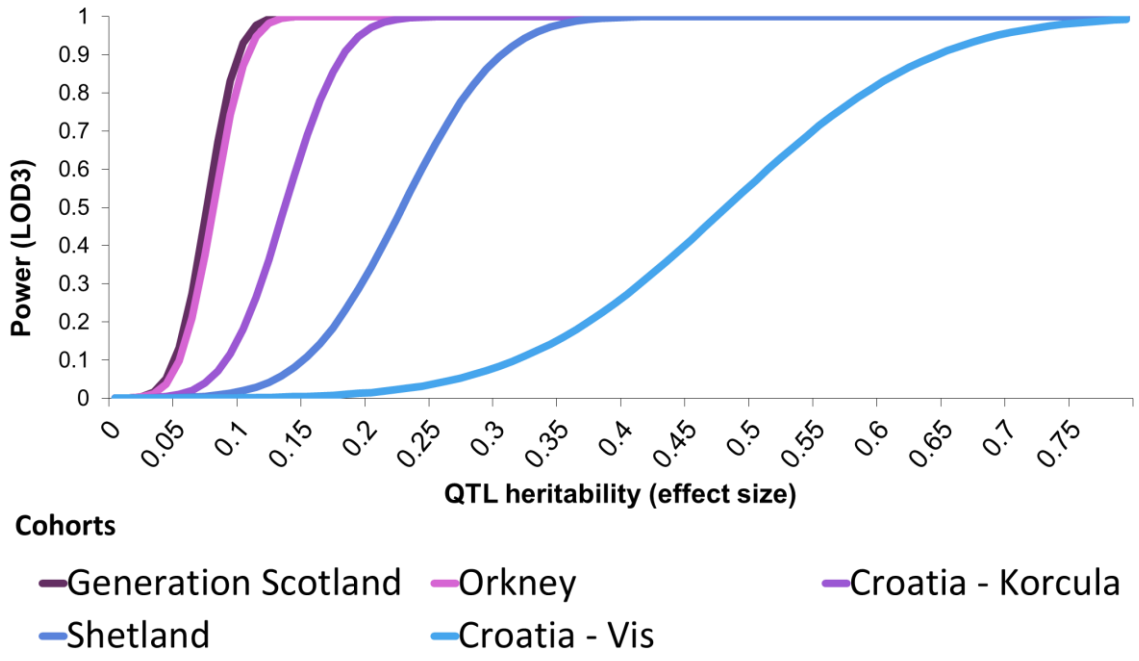
Power calculations done using SOLAR underestimate the power to detect linkage, as they only consider individuals within families. However, we know that there is substantial relatedness between individuals in different families, especially in the Orkney and Shetland populations, where pedigrees had to be clipped in order to accommodate the linkage analysis software. Because of this, individuals who share segments of their genome that are IBD have been assigned into different families and this sharing will not be accounted for in the power calculations, where the simulated alleles are assigned in each family independently.

With this limitation in mind, comparisons of the relative power of linkage analysis in the different cohorts can still be made. As Figure 21 shows, the cohort of Vis is the poorest performer, because in this cohort, a QTL can only be detected if it explains nearly all of the trait heritability. This is not surprising, as Vis has the lowest number of individuals and both it and Korčula have the lowest number of individuals per family – a metric that is very

important in linkage studies [141]. The second important metric in linkage studies is the total number of individuals in a study – this is why Shetland appears to perform worse than Korčula despite having more people per family. Korčula has nearly 1000 more individuals than Shetland, and this 50% larger sample size contributes strongly to the increased power to detect linkage. Yet, the importance of large families over a simple increase in the number of individuals, can be seen when comparing Orkney with Korčula and Generation Scotland. Orkney, like Shetland, has 2000 individuals but outperforms Korčula and performs nearly as well as GS despite GS having a 10-fold higher sample size. Indeed, Table 5 shows that Orkney has the highest number of genotyped individuals per family.

Figure 21 - Power to detect linkage in a trait with 0.8 heritability, using a QTL with 2% MAF across a range of heritabilities

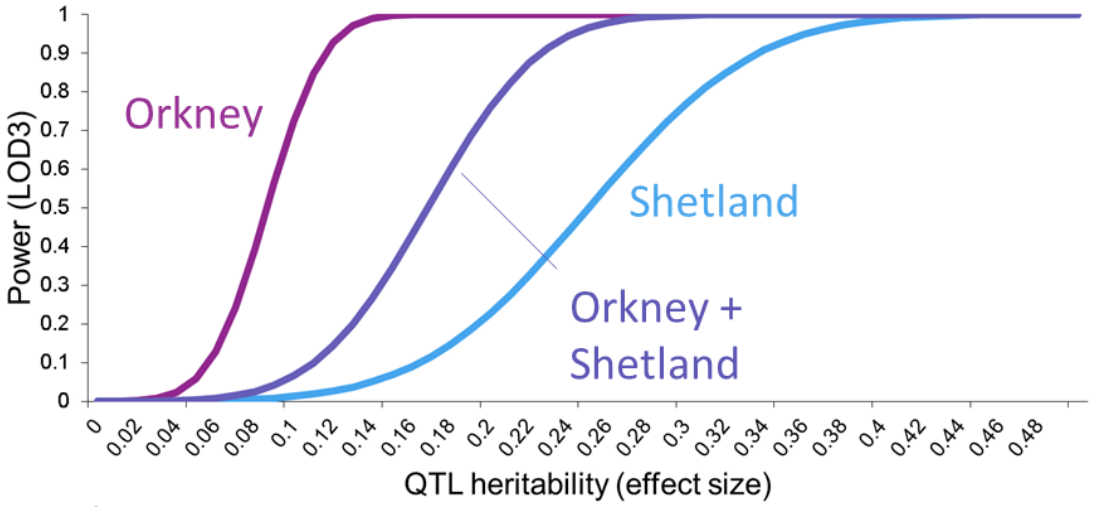
The X axis shows increasing QTL heritability (up to the value of the total trait heritability). The remaining trait heritability (if any) is explained by a simulated polygenic effect.



Interestingly, the combined Orkney + Shetland dataset performs more poorly than Orkney alone (but better than Shetland alone), despite the doubling in sample size (Figure 19). Again, this can be explained by a drop in the number of genotyped people per family once the combined pedigrees are clipped to bitsize 50, compared to Orkney. Also, due to the drop-down QTL simulations, stronger linkage signals are created in deeper families, which are more frequent in Orkney. The families where the simulated QTL does not segregate dilute the signal, which is exactly what is seen when the (smaller) Shetland families are added.

Figure 22 - Power to detect linkage in a trait with 0.5 heritability, using a QTL with 2% MAF

This figure shows the power to detect linkage in Orkney, Shetland and the combined Orkney and Shetland datasets.

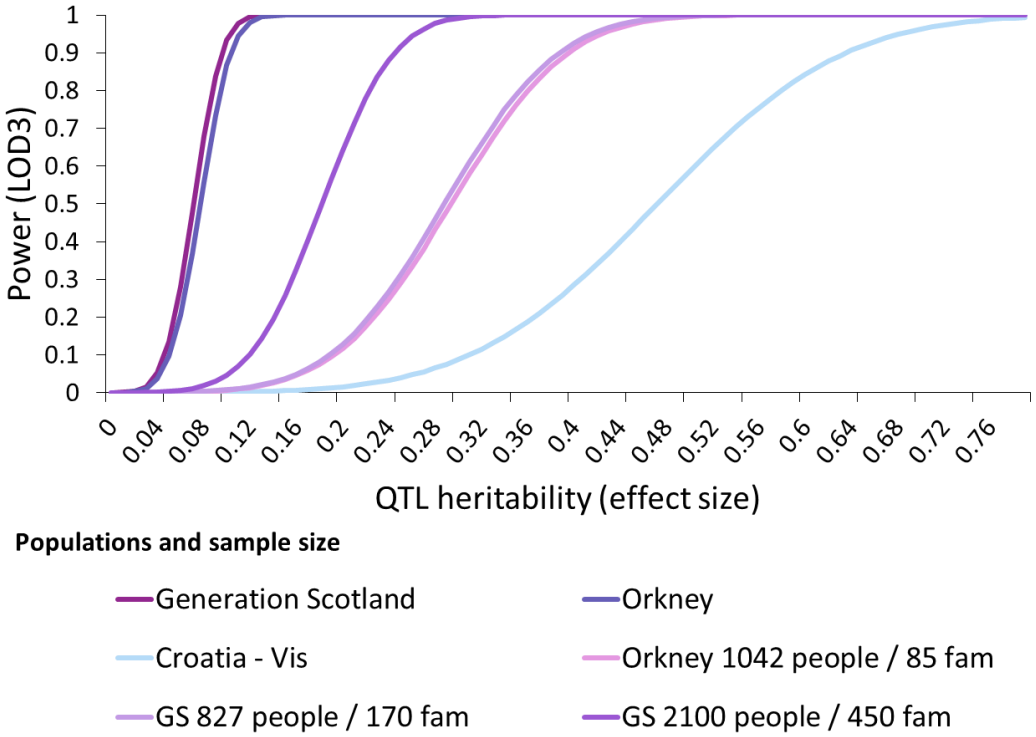


A final validation of the importance of sample size and family size comes from power calculations done on pedigrees where a random subset of genotyped individuals were masked (Figure 23). Reducing the number of individuals in Generation Scotland to 2100 in order to match the number of individuals in Orkney led to a big drop in power as these individuals were spread over 450 families, rather than the 170 families in Orkney. Once the family size in GS was reduced to 170, an even bigger drop in power could be observed, as only 827 individuals were retained. This sample size is comparable to Vis (960 individuals) which still performs worse due to the sparse family structure in this cohort. As families become sparser, power drops significantly and family structure becomes less important compared to sample size. When the sample size in Orkney is halved, its performance matches that of the 827-individual GS sample.

It should, however, be noted that to have an 80% power to detect a QTL with pedigree-based linkage analysis, its effect size needs to be quite large, as it has to explain about 8% of the trait heritability in the cohorts that have the highest power to detect linkage.

Figure 23 - Power to detect linkage in a trait with 0.8 heritability, using a QTL with 1% MAF

Genotyped individuals from Orkney and Generation Scotland were randomly masked from SOLAR when estimating power in order to assess the effect of reducing sample and family size on power to detect linkage.



4.3.3 Linkage Analysis Results

Here, I present the results of linkage analyses performed in individual cohorts, as well as the results of meta-analysis. Estimation of segments shared IBD is an important step in linkage analysis, and since IBD coefficients were calculated using different methods (Loki, IBDLD-LD-RR (pedigree) and IBDLD-GIBDLD (no pedigree)), three different sets of results will be presented. Since regions of IBD sharing can extend over large distances, linkage analysis cannot pinpoint the precise location of a QTL causing a linkage signal, and this is why the interval that is within a 2-LOD drop of the peak is also provided. The suggestive and genome-wide significance thresholds are described in section 2.2.1.

4.3.3.1 Linkage Results Obtained with IBD Coefficients Calculated by Loki

Table 11 presents the pedigree-based linkage analysis results that pass the suggestive significance threshold, and which were obtained by using regional IBD coefficients calculated by Loki. In total, 16 loci pass the suggestive significance threshold. The only locus with a

LOD score exceeding the genome-wide significance threshold is the chromosome 4 hit for fasting plasma glucose levels in Orkney, which is why the suggestive and genome-wide significant results are not presented in separate tables. This locus has a lower, but suggestively significant LOD score (3.44) with the pedigree-based linkage analysis conducted on IBD coefficients calculated by IBDLD.

Table 11 - Linkage analysis loci that reached the uncorrected GWS threshold, using IBD coefficients calculated by Loki

Loki outputs IBD coefficients at every 2.5 cM along the genome. The peak where the highest LOD score is reached is shown, as is the interval where LOD scores that are within a 2-LOD drop of the top hit. The total trait heritability (h^2) and heritability explained by the hit (h^2QTL), as output by SOLAR, are shown, and the chromosome band containing each locus is indicated. The last column shows if the locus has been associated with the relevant trait in the literature.

Trait	Chr	cM	LOD	h^2	h^2QTL	2-LOD drop region (cM)	Chromosome Bands	Known?
Orkney								
BMI	16	80	3.5043	0.42	0.28	60-100	q12.1-q23.2	Yes
FEV1	12	130	3.9894	0.41	0.37	110-135	q22-q24.21	Yes
Glucose	4	30	5.2099	0.35	0.35	25-35	p15.33-p15.31	No
Vis								
GGT	10	35	3.6281	0.58	0.58	20-50	p14-p12.1	Yes
GGT	19	95	3.5588	0.63	0.63	75-105	q13.32-q13.43	Yes
Shetland								
Creatinine	9	35	4.032	0.41	0.3	30-45	p23-p21.3	No
Diastolic BP	1	110	3.5	0.34	0.34	105-120	p31.1-p22.2	Yes
HbA1c	8	105	4.5454	0.45	0.37	100-110	q21.11-q21.3	No
Spherical Equivalent Refraction	22	60	3.8247	0.63	0.38	55-70	q13.31-q13.33	No
Korčula								
Central Corneal Thickness	13	50	3.7958	0.75	0.7	40-60	q13.3-q21.2	No
Glucose_nodiab	10	65	3.6544	0.35	0.35	55-90	p12.1-q21.3	No
GS								
Alcohol consumption	7	60	3.4515	0.29	0.1	50-70	p14.3-p13	Yes
Forced Vital Capacity	9	105	3.5034	0.42	0.13	80-110	q13-q22.2	No
Forced Vital Capacity	21	10	4.0323	0.41	0.13	0-20	p11.2-q21.2	No
Height	7	75	4.1948	0.92	0.08	70-85	p13-q11.22	Yes
Height	17	85	3.8357	0.92	0.08	75-90	q21.33-q24.1	Yes

4.3.3.2 Linkage Results Obtained with IBD Coefficients Calculated by the Pedigree-based method of IBDLD (LD-RR)

In this section, I present the regions that passed the genome-wide LOD significance threshold adjusted for the number of traits analysed, within each cohort (Table 12). For suggestively significant results that exceeded the unadjusted LOD threshold (LOD>3.41) but not the GWS threshold, see Supplementary Table 3.

These results were obtained using regional IBD coefficients calculated by the pedigree-based method of IBDLD (LD-RR), and kinship coefficients were calculated by SOLAR using the social pedigree. Note that the pairs used in this linkage analysis are restricted to those that are connected in the social pedigree rather than all pairs of individuals in the data.

In total, 158 regions pass the suggestive significance threshold (LOD>3.41), and 26 of these regions pass the GWS threshold. Every locus that was suggestively significant in the linkage analysis using the IBD coefficients calculated by Loki also exceeds the suggestive significance threshold in this analysis, with the exception of 4 loci that do not reach this threshold when IBD coefficients calculated with LD-RR are used in the linkage analysis.

Table 12 - Loci that reached the GWS threshold with pedigree-based linkage analysis, using IBD coefficients calculated by IBDLD

The rsID of the SNP at which the highest LOD score was obtained is shown, as is the chromosome (Chr column) and position (Pos column) where this SNP is located. The start and end positions of the interval where LOD scores that are within a 2-LOD drop of the top hit are shown, in Megabases (2-LOD drop column). The total trait heritability (h^2) and heritability explained by the hit (h^2QTL), as output by SOLAR, are shown. The 'Gene' and 'Gene_Dist' columns indicate the gene nearest the top hit, as well as the distance to this gene from the top hit (this distance is 0 when the top hit is within the gene itself). The last column shows if the locus has been associated with the relevant trait in the literature.

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist	Known?
Orkney											
Diastolic BP	2	rs6432025	9758621	6.2592	0.32	0.28	8.63-11.16	p25.1	*YWHAQ*	0	Yes
Vis											
Central Corneal Thickness	1	rs240099	176547669	5.0805	0.84	0.58	175.89-177.02	q25.1-q25.2	*PAPPA2*	0	No
Central Corneal Thickness	12	rs901232	128073805	7.7858	0.79	0.64	127.88-128.15	q24.32	*FLJ37505*	292355	No
Forced Vital Capacity	1	rs1251531	76469334	6.0745	0.31	0.26	75.94-76.57	p31.1	*ST6GALNAC3*	71053	Yes
GGT	1	rs2053675	188967356	6.3055	0.43	0.36	188.62-189.72	q31.1	*FAM5C*	1099439	No
GGT	19	rs217541	54443866	8.4062	0.63	0.36	54.27-54.76	q13.42	*CACNG7*	0	No
Lens Thickness	7	rs4724195	43326197	7.1727	0.47	0.46	43.28-43.57	p14.1-p13	*HECW1*	0	No
Urea	9	rs12001341	86203246	8.0489	0.42	0.35	86.18-86.76	q21.32	*C9orf103*	34785	No
Urea	9	rs12554805	87092773	7.7952	0.43	0.34	86.96-87.41	q21.33	*SLC28A3*	109359	No
Urea	9	rs1885493	38335475	6.931	0.52	0.25	37.73-38.56	p13.2-p13.1	*ALDH1B1*	57225	No
Shetland											
HbA1c	8	rs7012200	81891232	7.2106	0.45	0.36	80.66-82.6	q21.13	*PAG1*	0	No
GS											

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist	Known?
BMI	11	rs7932734	11201684	5.7443	0.55	0.1	11.11-11.3	p15.3	*GALNTL4*	90735	No
Educational Attainment	6	rs2523949	29917591	5.6092	0.49	0.09	29.34-30.77	p22.1-p21.33	*HLA-A*	3929	No
Educational Attainment	9	rs7850169	24152518	6.1377	0.5	0.15	23.52-25.26	p21.3	*ELAVL2*	326454	Yes
Educational Attainment	9	rs10967133	25942793	5.1814	0.5	0.14	25.56-27.34	p21.3-p21.2	*LOC100506422*	123878	Yes
Educational Attainment	9	rs527427	22748466	4.7856	0.5	0.13	20.61-27.34	p21.3-p21.2	*FLJ35282*	0	Yes
Educational Attainment	12	rs987019	63311148	5.539	0.51	0.14	63.16-64.19	q14.2	*PPM1H*	0	No
Educational Attainment	12	rs11175348	64740578	5.028	0.51	0.13	63.05-66.42	q14.1-q14.3	*C12orf56*	0	No
Height	2	rs1371040	148165153	5.145	0.92	0.09	143.93-149.93	q22.2-q23.2	*ACVR2A*	437415	No
Height	7	rs2240090	51096974	5.3466	0.92	0.09	46.71-52.72	p12.3-p12.1	*COBL*	0	Yes
Height	10	rs7089663	71652580	5.0763	0.92	0.09	71.39-73.24	q22.1	*COL13A1*	0	No
Height	11	rs921443	57902114	4.645	0.92	0.07	56.91-59.81	q12.1	*OR9Q1*	0	No
Height	15	rs2439359	66887723	5.2906	0.92	0.09	65.21-67.68	q22.31-q23	*LCTL*	29887	No
Height	15	rs1994534	63226015	5.2442	0.92	0.09	62.86-68.07	q22.2-q23	*TLN2*	89185	No
Sodium	11	rs4936111	130573050	5.187	0.26	0.12	130.22-130.65	q24.3	*SNX19*	172714	No
Total Cholesterol	6	rs2153635	63722285	5.7048	0.3	0.13	57.16-65	p11.2-q12	*LGSN*	263569	No

4.3.3.3 Linkage Results Obtained with IBD Coefficients Calculated by the Pedigree-free Method of IBDLD (GIBDLD)

In this section, I present the regions that passed the suggestive LOD significance threshold (unadjusted for the number of traits analysed) within Orkney and Vis. There are 8 loci that reach the suggestive significance threshold in Orkney and 4 in Vis, and there is some overlap with GWAS and RH results. Only one hit has a LOD score that passes the GWS threshold, so all hits that exceed the suggestive LOD significance threshold are shown in Table 13.

The pedigree-free linkage analysis was performed using IBD coefficients calculated by the pedigree-free method of IBDLD (GIBDLD), between all genotyped pairs at all SNPs. For each pair, IBD coefficients were calculated at every genotyped SNP, and within every 2.5 cM interval across the genome, these IBD coefficients were averaged across all SNPs falling into that region. Whole-genome IBD coefficients between every genotyped pair were also calculated by GIBDLD. These regional and whole-genome IBD matrices were then used by SOLAR to perform linkage analysis.

If a region's LOD score exceeded the suggestive genome-wide significance threshold, linkage analysis was re-run separately using each SNP within that region (that is, without averaging IBD coefficients over SNPs in the 2.5 cM region).

The top hits identified in the pedigree-free linkage analysis do not appear in the pedigree-based linkage analysis results, but some of the peaks identified here overlap with loci flagged by GWAS, indicating that this method may be capturing some population-level IBD sharing. I briefly discuss some of the hits identified with pedigree-free linkage analysis in section 4.4.4 and expand on the Orkney axial length linkage peak in detail in section 4.4.4.1.

Table 13 - Loci that reached the uncorrected GWS threshold with pedigree-free linkage analysis, using IBD coefficients calculated by IBDLD

The regions with the highest LOD scores are shown, with the start and end positions (in bp) of these regions indicated. The total trait heritability (h^2) and heritability explained by the region (h^2Reg), as output by SOLAR, are shown. Within each region, the SNP that had the highest LOD score in individual SNP analysis is shown. 2.5 cM regions were numbered sequentially from the start of each chromosome, and the chromosome and region number are shown in the "Region" column. The last column shows if the locus has been associated with the relevant trait in the literature.

Trait	Chr	Region	Region_Start	Region_End	LOD	h^2	h^2Reg	Band	rsID	Pos	SNP_LOD	Known?
Orkney												
Axial length1	8	8_33	57560788	59785090	3.5578	0.64	0.07	q12.1	rs2939966	59066247	3.5353	No
Fibrinogen	11	11_8	7909715	10372041	3.579	0.28	0.09	p15.4	rs3751050	9091244	3.3997	No
HDL	16	16_29	55535678	56696613	4.5121	0.49	0.05	q12.2	rs7189840	56683626	5.2585	No
Height	3	3_68	153057662	155051740	3.6363	0.7	0.04	q25.2-q25.31	rs1025192	154827787	4.4078	No
Systolic BP	19	19_42	56767367	57747455	3.7334	0.22	0.05	q13.43	rs1860565	57335022	4.0186	No
Uric acid1	4	4_10	8384752	11408142	4.2712	0.4	0.04	p16.1-p15.33	rs7685513	9728599	5.1457	Yes
Uric acid2	4	4_10	8384752	11408142	4.6392	0.41	0.04	p16.1-p15.33	rs7685513	9728599	5.0944	Yes
Uric acid2	11	11_28	57901987	60858840	3.4616	0.41	0.05	q12.1-q12.2	rs7925914	59708244	3.7537	No
vWF	9	9_67	135670467	136511711	10.4832	0.55	0.11	q34.13-q34.2	rs574347	136135659	11.0415	Yes
Vis												
Axial length1	1	1_67	149401422	153179458	4.7575	0.36	0.15	q21.2-q21.3	rs1332506	152724053	5.1656	No
Axial length2	1	1_67	149401422	153179458	4.7332	0.33	0.15	q21.2-q21.3	rs873775	152692472	4.506	No
Pulse Pressure	14	14_34	88727479	90015642	3.4585	0.33	0.09	q31.3-q32.11	rs17714667	89550378	3.62	No
Calcium	14	14_11	32847232	33600856	3.5881	0.13	0.1	q12-q13.1	rs2103781	33258553	3.5635	No
Height	15	15_24	37968377	39155424	3.7772	0.76	0.1	q14	rs11629949	38717660	3.9326	No

4.3.4 Meta-analysis Results

The meta-analysis results for pedigree-based linkage analysis are presented in Table 14, while Table 15 shows the results obtained when grouping cohorts based on geographical location. Only results exceeding the GWS threshold (LOD>4.84 or logP>5.93) are listed in these tables, while Supplementary Table 4 and Supplementary Table 5 list the results that exceed the suggestive but not the genome-wide significance threshold in the joint and geographical location-based meta analyses. Table 16 shows all the pedigree-free linkage meta-analysis results that pass the suggestive significance threshold.

When all cohorts are included in the meta-analysis, 15 regions yield meta-analysis test statistics that exceed the genome-wide significance threshold. While most of these are due to strong signals in one cohort, 4 regions resulted in per-cohort LOD scores of at least 1 in at least 2 different cohorts. Of note, sometimes the position of the point with the highest LOD score in a cohort-level analysis is not the same as the position of the peak in the meta-analysis, and this is discussed in more detail at the end of section 4.4.5.

Meta-analysis restricted to the Croatian or Scottish island populations identifies no loci that contain a signal in both cohorts, as the meta-analysis signals that pass the GWS threshold originate from strong signals in one cohort only. The Scottish cohorts meta-analysis reveals 4 height-linked loci that segregate in GS and one (but not both) Scottish island populations.

Table 14 - Meta-analysis results that exceed the genome-wide significance threshold in pedigree-based linkage analysis

The meta-analysis -$\log_{10}$(*p-value*) is indicated (logP column) for each peak. These peaks represent a 0.1 cM interval that starts at the cM position indicated (cM column). The start and end positions of the region surrounding these peaks where the meta-analysis test statistic continuously exceeded the suggestive significance threshold (logP>4.43) is indicated in cM (Reg_cM) and Mbp (Reg_Mbp), as is the chromosome band where these regions can be found. The per-cohort LOD scores in the 0.1 cM peak region are indicated in the last columns (O = Orkney, S=Shetland, G=Generation Scotland, V=Vis, K=Korčula), and the intensity of the green shading corresponds to the magnitude of the LOD score (darker green = higher LOD score). The last column shows if the locus has been associated with the relevant trait in the literature.

Trait	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G	V	K	Known?
Central Corneal Thickness	12	157.6	7.19	q24.32	157.5-157.7	128.06-128.11	0	0.4	NA	7.79	0	No
Educational Attainment	6	50.8	6.63	p22.1-p21.33	50-51.3	29.34-31.34	1.32	0	3.85	0.12	1.99	No
Forced Vital Capacity	1	105.1	7.03	p31.1	105-105.2	76.1-76.57	0.14	0.6	1.08	6.07	0	Yes
GGT	19	94.3	8.88	q13.42	94.2-94.4	54.41-54.45	0.68	0.02	NA	8.41	NA	No
GGT	19	95	7.31	q13.42	94.9-95.1	54.58-54.68	0.62	0.06	NA	6.73	NA	No
Glucose	4	185	6.06	q34.1	182.1-185.7	173.01-175.79	1.06	4.36	0	0.04	1.34	No
HbA1c	8	103.8	7.03	q21.12-q21.13	102.2-105.2	79.83-83.18	0.3	7.19	NA	0.2	0.01	No
HDL	18	87.5	6.07	q21.33	86.5-87.9	60.82-61.34	0.38	3.47	2.99	0.03	0	Yes
Height	15	89.8	6.03	q22.31	88.4-90.3	65.37-67.04	0.17	0.04	5.29	0.2	1.04	No
Lens Thickness	7	68.7	7.32	p14.1-p13	68.6-68.8	43.28-43.43	NA	NA	NA	7.17	0	No
Lens Thickness	7	69.1	7.20	p13	69-69.2	43.48-43.56	NA	NA	NA	7.05	0	No
Total Cholesterol	6	81.1	6.12	p11.2-q12	80.5-81.3	57.27-64.61	0.39	0	5.7	0.25	0.59	No
Urea	9	101.6	7.55	q21.32-q21.33	101.3-101.7	86.87-87.23	0.5	0	0.08	7.8	NA	No
Urea	9	100.6	7.02	q21.32	100.5-100.7	86.18-86.34	0.05	0	0.01	8.05	NA	No
Urea	9	63.9	6.43	p13.2	63.8-64.2	38.3-38.4	0.07	0	0.23	6.93	NA	No

Table 15 - Meta-analysis results that exceed the genome-wide significance threshold in pedigree-based linkage analysis, with cohorts grouped by geographical location

The meta-analysis $-\log_{10}(p\text{-}value)$ is indicated (logP column) for each peak. These peaks represent a 0.1 cM interval that starts at the cM position indicated (cM column). The start and end positions of the region surrounding these peaks where the meta-analysis test statistic continuously exceeded the suggestive significance threshold (logP>4.43) is indicated in cM (Reg_cM) and Mbp (Reg_Mbp), as is the chromosome band where these regions can be found. The per-cohort LOD scores in the 0.1 cM peak region are indicated in the last columns (O = Orkney, S=Shetland, G=Generation Scotland, V=Vis, K=Korčula), and the intensity of the green shading corresponds to the magnitude of the LOD score (darker green = higher LOD score). The last column shows if the locus has been associated with the relevant trait in the literature.

Trait - Croatia	**Chr**	**cM**	**logP**	**Band**	**Reg_cM**	**Reg_Mbp**	**V**	**K**	**Known?**
Central Corneal Thickness	12	157.6	7.93	q24.32	157.5-157.7	128.06-128.11	7.79	0	No
Forced Vital Capacity	1	105.1	6.25	p31.1	105-105.2	76.1-76.57	6.07	0	Yes
Lens Thickness	7	68.7	7.32	p14.1-p13	68.6-68.8	43.28-43.43	7.17	0	No
Lens Thickness	7	69.1	7.20	p13	69-69.2	43.48-43.56	7.05	0	No

Trait - Scotland	**Chr**	**cM**	**logP**	**Band**	**Reg_cM**	**Reg_Mbp**	**O**	**S**	**G**	**Known?**
Educational Attainment	6	50.4	6.21	p22.1	49.8-50.6	28.26-30.37	0.6	0.05	5.61	No
Forced Vital Capacity	9	101.9	6.16	q21.33	101.6-103.3	87.09-88.1	0.51	0.89	4.54	No
HbA1c	8	104	7.96	q21.12-q21.2	102.2-106.7	79.83-85.64	0.31	7.21	NA	No
HDL	18	87.5	7.09	q21.33	86.4-87.9	60.78-61.34	0.38	3.47	2.99	Yes
Height	2	167.8	6.27	q22.3-q23.2	166.4-169.1	145.3-149.93	1.23	0	5.14	No
Height	5	61.2	6.42	p13.1	60.9-61.5	40.29-41.3	0.28	2.26	3.64	No
Height	15	75.8	6.23	q21.3	75.5-77.5	56.32-58.1	0	2.8	3.41	No
Height	17	86	6.46	q22-q23.2	84.2-88.1	55.85-60.54	1.9	0.01	4.54	No

This table continues on the next page.

Trait - Scottish Isles	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	Known?
Diastolic BP	2	22.7	6.48	p25.1	19.2-25.8	8.64-11.23	6.23	0.01	No
Glucose	4	185	6.05	q34.1	182.1-185.7	173.01-175.79	1.06	4.36	No
Glucose_nodiab	4	185.3	6.25	q34.1-q34.2	182.1-186.6	173.01-176.57	1.2	4.41	No
HbA1c	8	104	7.96	q21.12-q21.2	102.2-106.7	79.83-85.64	0.31	7.21	No

Table 16 - Meta-analysis results of pedigree-free linkage analysis

The meta-analysis $-\log_{10}(p\text{-}value)$ is indicated (logP column) for each peak. These peaks represent a 2.5 cM interval whose start and end positions are indicated in Mbp (Mbp column), and which start at the cM position indicated (cM column). The start and end positions of the region surrounding these peaks where the meta-analysis test statistic continuously exceeded the suggestive significance threshold (logP>4.43) is indicated in cM (Reg_cM) and Mbp (Reg_Mbp), as is the chromosome band where these regions can be found. The per-cohort LOD scores in the 0.1 cM peak region are indicated in the last columns (O = Orkney, V=Vis), and the intensity of the green shading corresponds to the magnitude of the LOD score (darker green = higher LOD score). The last column shows if the locus has been associated with the relevant trait in the literature.

Trait	Chr	cM	Mbp	logP	Band	Reg_cM	Reg_bp	O	V	Known?
Axial length1	1	165	149.4-153.18	5.26	q21.2-q21.3	162.5-167.5	147.64-153.18	0.1	4.76	No
Axial length2	1	165	149.4-153.18	4.99	q21.2-q21.3	162.5-167.5	147.64-153.18	0	4.73	No
Pulse Pressure	14	82.5	88.73-90.02	4.46	q31.3-q32.11	80-85	85.93-90.02	0.45	3.46	No
GGT	22	22.5	24.41-25.93	5.87	q11.22-q12.1	20-25	23.4-25.93	2.21	2.97	Yes
HDL	16	70	55.54-56.7	4.73	q12.2	67.5-72.5	54.53-56.7	4.51	0	No
LDL	19	67.5	43.87-45.23	4.49	q13.2-q13.32	65-70	41.17-45.23	3.26	0.64	Yes
Uric acid1	4	22.5	8.38-11.41	7.11	p16.1-p15.33	20-25	7.67-11.41	4.27	2.15	Yes
Uric acid1	11	77.5	69.79-71.31	4.45	q13.2-q13.4	75-80	68.16-71.31	3.33	0.54	No
Uric acid2	4	22.5	8.38-11.41	7.15	p16.1-p15.33	20-25	7.67-11.41	4.64	1.84	Yes
vWF	9	165	135.67-136.51	13.16	q34.11-q34.2	160-170	133.24-137.1	10.48	2.04	Yes

4.4 Discussion

4.4.1 IBD Estimation

Genome-wide kinship is more accurately estimated with the help of genetic data and IBD sharing matrices estimated this way give a more accurate representation of genetic kinship compared to pedigree-based genome-wide kinship estimates, as shown in Figure 16. This is because pedigree-based kinship estimates will always assume the same fixed kinship-value for a specific relative pair type, e.g. full siblings will always have $2\Phi = 0.5$ (assuming no inbreeding) according to a social pedigree, but the true amount of genetic sharing can vary widely – between 0.36 and 0.63 in Orkney, which is in line with the sib pair empirical genome-wide IBD coefficient that ranged between 0.374 and 0.617 according to one study of 4401 siblings [142].

IBDLD offers many benefits for IBD coefficient estimation compared to Loki. The fact that IBDLD can use compressed PLINK files, as opposed to a highly specific format and uncompressed files, reduces the space needed to store the files, as well as the time required to pre-process them. Its LD-RR method can handle large and complex pedigrees without the need to clip these to a certain bit-size, which offers a substantial increase in power, especially in Orkney and Shetland, which have very detailed pedigrees spanning 35 generations. Because it explicitly models LD, genotype data do not need to be pruned and, in fact, denser marker data allow for IBD coefficients to be estimated more accurately, picking up IBD segments that might otherwise have been missed [39]. Since IBD coefficients can be output at any or all genotyped SNPs, this also offers an increase in resolution for linkage analysis, with the caveat that IBD segments shorter than 3 cM cannot reliably be detected.

One slight drawback is that IBDLD does not offer the option to output files directly into SOLAR format, so some additional downstream processing is required. However, this offers more flexibility in manipulating the IBD files for other purposes. Also, newer versions of SOLAR accepts as input file formats where individual IDs do not need to be coded using SOLAR's internal system, which greatly facilitates file manipulation and accessibility.

One issue with using regional IBD coefficients that only include pairs of individuals that are related according to the social pedigree is that while two individuals may share only a very small proportion of their genome overall, they might still share loci that were inherited from the same common ancestor, as shown in Figure 17. The IBD coefficients calculated by Loki therefore incorrectly indicate pairwise IBD sharing between such pairs is 0 at every locus, which may dilute the linkage signal. IBDLD's GIBDLD (pedigree-free) method is attractive

in principle to circumvent this problem by estimating IBD sharing probabilities along the genome without the need for a social pedigree.

The fact that GIBDLD accurately detects shared haplotypes between distantly related individuals is encouraging and adds to the power to detect a QTL with the help of pedigree-free linkage analysis – these examples of IBD sharing would go unnoticed if pedigree-based linkage analysis was used. However, it could be seen that sometimes, GIBDLD yields IBD coefficients that are incorrect when compared to the phased data. The parameters that GIBDLD uses to estimate LD (instructing it to utilise SNPs in the previous 2.5 cM as well as the previous 10 SNPs) were kept constant within my analyses, but by systematically modifying these parameters, better insight could be gained about the performance of GIBDLD.

While IBDLD circumvents the software limitation for IBD estimation and allows for IBD estimation between all pairs of individuals in the data, the linkage analysis process still experiences a bottleneck with SOLAR. In Orkney, SOLAR takes 8 minutes to calculate linkage statistics in each region, because each region contains IBD coefficients for 2055378 pairs of individuals. This means that calculating linkage statistics at each individual SNP across the entire genome is currently unfeasible. Additionally, SOLAR fails to estimate the reference polygenic model when presented with the Korčula whole-genome kinship matrix (3649051 pairs) and a phenotype that is present in all 2701 individuals, and this is the upstream step that needs to be performed prior to linkage analysis.

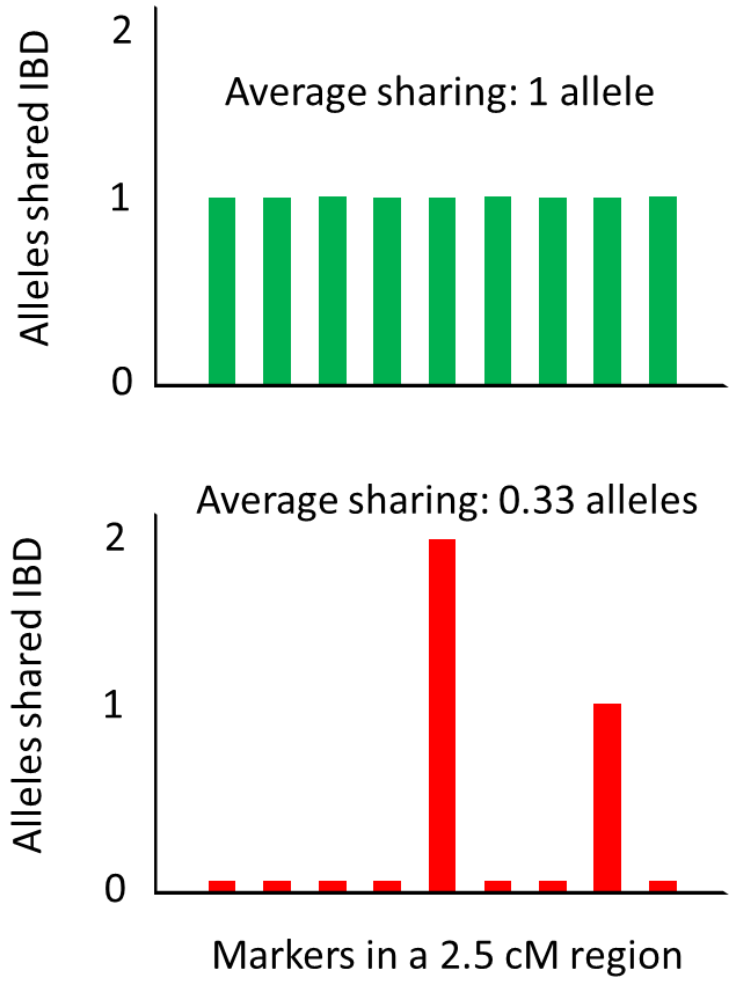
This is one of the reasons for using regional, rather than per-SNP, IBDs to test pedigree-free linkage analysis in Orkney and Vis. Regional IBD sharing is calculated by averaging the IBD sharing at all SNPs within each 2.5 cM region, for every pair of genotyped individuals. Loki also outputs the average IBD sharing at 2.5 cM intervals, but it uses LD-pruned data and does not output per-SNP IBD estimates, making follow-up by fine-mapping impossible.

Using 2.5 cM intervals is also biologically and statistically meaningful: while the probability of IBD sharing decreases exponentially with the number of meioses, the length of shared segments only decreases linearly, so if a segment is shared IBD between distant relatives, this segment tends to be large [126]. It has been shown that IBD segments spanning 3 cM or more can accurately be inferred and give rise to fewer false positives [39, 127]. Additionally, I show that GIBDLD can sometimes give rise to false positive estimates of IBD sharing at specific SNPs (Section 4.3.1.3). Therefore, we have higher confidence that pairs who have high average regional IBD sharing (which is a result of high IBD sharing throughout all markers in the 2.5 cM region) really do share that segment IBD, and averaging the IBD sharing across the 2.5 cM regions also reduces noise due to sporadically high IBD sharing that is not

consistent throughout the region and may be the result of false positive IBD sharing estimates (Figure 24).

Figure 24 - True vs sporadic regional IBD sharing

Where there is true IBD sharing across a 2cM region, the average IBD in that region will be high, due to the IBDs being consistently high at all markers within that region (top). The low average regional IBD sharing reduces the effect of high IBD sharing values that are incorrectly estimated at some markers.



4.4.2 Linkage Analysis

As presented in Chapter 7, the results obtained with pedigree-based linkage analysis do not correlate well with those obtained using the pedigree-free linkage analysis, RH or GWAS. In contrast, the pedigree-free linkage analysis results show higher correlation with the results of RH and GWAS, and some results indicate that this method is capable of identifying signals originating from common variants, as long as these have large effects on a trait. The presence

of loci identified with pedigree-free linkage analysis that do not show a signal with either RH or GWAS, however, suggests that there may be some additional QTLs segregating in the general population that linkage analysis is better suited to discovering.

4.4.3 Pedigree-based Linkage Analysis

With pedigree-based linkage analysis using IBD coefficients calculated by Loki, 16 regions have LOD scores exceeding 3.41, which is the suggestive significance threshold that is not corrected for the number of traits analysed. All but 4 of these regions exceed this LOD threshold in the pedigree-based linkage analysis that uses IBD coefficients calculated by IBDLD's pedigree-based method (LD-RR), showing good agreement between the two methods of local IBD coefficient estimation. The hits that are missed are the refractive error (SER) hit on chromosome 22 in Shetland, the gamma-glutamyl transferase hit on chromosome 10 in Vis, the central corneal thickness hit on chromosome 13 in Korčula and the alcohol consumption hit on chromosome 9 in Generation Scotland. Conversely, when a pedigree-based linkage analysis is done using IBD coefficients calculated by IBDLD (the pedigree-based, LD-RR method), a total of 158 loci pass the suggestive significance threshold, with over half of these originating in Generation Scotland, the cohort that has the highest power to detect linkage signals. This could be a consequence of the fact that IBD coefficients calculated by Loki were output at 2.5 cM intervals, while with IBDLD, they were output at 0.1 cM intervals, allowing the detection of smaller shared segments. With the Loki IBD coefficients, only a single region exceeds the significance threshold corrected for the number of traits analysed, while using the IBDLD IBD coefficients, 27 such regions are identified. Strangely, most of these originate in either Generation Scotland (the highest-powered cohort) or Vis (the lowest-powered cohort), with only a single hit originating in Orkney and Shetland, and no hits originating in Korčula. The reason for this is unclear and the hits in Vis should be treated with caution.

With pedigree-based linkage analysis, the loci with LOD scores that exceed the trait-corrected multiple testing threshold generally explain a high proportion of the trait variance in Vis (25-64%) while in Generation Scotland, they explain between 7 and 15% of the trait variance. This was expected, given the results of the power calculations that show that Vis is very under-powered to detect a QTL effect unless it explains >40% of the trait variance, while Generation Scotland is powered to detect QTL effects explaining >8% of the trait variance. This, and the little overlap between pedigree-based and pedigree-free linkage analyses, casts some doubt on the "true positiveness" of the results. Some replication of signal across studies would lend more confidence in the regions implicated. Sequencing the regions flagged by these analyses,

or performing whole genome sequencing at the cohort level could help pinpoint the location of QTLs giving rise to these signals. Such sequencing efforts are currently being undertaken but were not completed before the due date of this thesis.

In Generation Scotland, many regions across the genome show linkage to height. This is not unexpected, as height is the best-documented, highly-polygenic and highly heritable complex trait. It is interesting that the hits identified with linkage analysis are different to the ones identified by genome-wide association studies on the same cohort [62]. The peak on chromosome 7 is less than 500kb away from GWAS hits reported in the GWAS catalog [16] and the smaller peak on chromosome 15 is ~800kb away from a GWAS hit reported in the GWAS catalog, but in general, most loci identified with linkage analysis do not overlap with those identified by large-scale GWAS, despite the fact that over 180 loci have been found to associate with height [22]. This may suggest the presence of allelic heterogeneity with rare variants that are not tagged by common, genotyped markers and are not well-imputed, or variants segregating within some families, but not at the population level.

In the case of the pedigree-based linkage results that were obtained using regional IBD coefficients calculated by IBDLD, it can often be seen that if the LOD score at the peak is high, the 2-LOD drop region around the peak encompasses only a single gene region. Generally, a clear relationship cannot be established between what is known about the function of these genes and the phenotype they appear to be linked to. However, below are some examples where such a functional relationship has been established experimentally, which lend some confidence to the accuracy of these results and encourages the functional follow-up of (some of) these hits.

In Orkney, there is a strong hit on chromosome 2 for diastolic blood pressure (LOD 6.26). This signal is strongest within the *YWHAQ* (Tyrosine 3-Monooxygenase/Tryptophan 5-Monooxygenase Activation Protein Theta) gene and studies have shown experimentally an interaction between this gene and blood pressure. *YWHAQ* is expressed in the heart and in neurons and has been shown to interact, and be co-expressed, with the *ATP2B* calcium regulator gene family, members of which have been found to associate with blood pressure in European and Asian individuals [143]. SNPs within this gene were associated with changes in heart rate following a 20-week endurance training regime [144] in a GWAS study of 483 individuals belonging to 99 families. Methylation levels at the *YWHAQ* promoter have been shown to affect blood pressure in individuals with preeclampsia [145].

In Vis, the analysis of gamma-glutamyl transferase (GGT) levels yields a signal in the 54,272,420-54,759,571 interval on chromosome 19. This is a gene-rich region that harbours a

cluster of leukocyte and killer cell immunoglobulin-like receptor genes, voltage-dependent calcium channel gene subunits and additionally, the *NDUFA3* (NADH:Ubiquinone Oxidoreductase Subunit A3) gene. The *NDUFA3* gene has been implicated in the progression of non-alcoholic fatty liver disease, which is often associated with metabolic syndrome [146]. GGT levels are used to assess liver function and can aid in the diagnosis of metabolic syndrome, as higher levels of GGT correlate with the prevalence of metabolic syndrome [147].

In Generation Scotland, the 20,610,914-27,344,662 interval on chromosome 9 is linked to educational attainment and its highest point falls within the *ELAVL2* gene. A SNP in this region (rs1360382, at 23,379,721) has been identified by a GWAS of educational attainment performed on 106,000 individuals [148]. The *ELAVL2* gene encodes a neuron-specific RNA binding protein that is highly expressed in the brain and its down-regulation led to a change in the expression and splicing of genes related to autism spectrum disorders, such as *RBFOX1* and *FMRP*, as well as genes involved in neurodevelopment and synaptic function [149]. It has also been shown to interact with transcripts linked to Alzheimer's disease, affecting their splicing [150].

Within the same cohort and the same trait, there was also a peak on chromosome 12, in the 63,048,066-66,422,374 interval. The gene at the apex of this peak, *PPM1H*, is highly expressed in brain and it contains a pseudogene (*GAPDHP44*) that contains SNPs that associate with late-onset Alzheimer's disease [151]. Additionally, other linkage studies have also indicated that this region is linked to late-onset Alzheimer's disease [152, 153].

Also of interest is the sodium peak on chromosome 11 in GS. The *SNX19* (sorting nexin 19) gene in this region affects insulin secretion and homeostasis by preventing insulin-containing vesicles in the pancreatic beta-cells [154]. There is evidence that increased insulin levels lead to sodium retention in healthy individuals [155], increasing sodium reabsorption by modulating the activity of sodium channels in the kidney [156].

4.4.4 Pedigree-free Linkage Analysis

From the results of the pedigree-free linkage analysis, it can be seen that in most cases, the LOD score of the region is lower than the LOD score of a single SNP within the region. This may indicate that IBD sharing at a specific position within that region is most strongly linked to a QTL, and the IBD sharing diminishes throughout the region, diluting this signal. Sometimes, however, the regional LOD score is higher than any of the LOD scores obtained from single SNPs within that region, indicating that such regions may harbour several causal loci as no IBD sharing at any one SNP explains as much variance as the average regional IBD

sharing. Some individuals in the population may carry one QTL, while a different subset may carry the other QTL, giving rise to allelic heterogeneity and the regional IBD sharing better represents this than the IBD sharing at any one SNP.

In general, the amount of trait heritability explained by a GWS region in the pedigree-free linkage analysis, where all pairs of individuals are used, is more modest than the amount of heritability explained by the IBD sharing patterns between closely related individuals in GWS hits reported using the pedigree-based linkage analysis. Where the pedigree-free linkage analysis identifies regions that were also reported in GWAS (such as the *SLC2A9* locus for uric acid and the *ABO* locus for von Willebrand factor), the proportion of heritability explained by these regions matches those quoted in the GWAS literature.

In the Orkney pedigree-free linkage, the chromosome 16 peak for HDL cholesterol is adjacent to the *CETP* (cholesteryl ester transfer protein) gene which has been implicated by association in many HDL GWAS. *CES1* (carboxylesterase1) and *CES5A* (carboxylesterase 5A) are within the region itself and have been shown to participate in cholesterol ester metabolism [157]. This peak also contains a cluster of metallothionein genes which have been implicated in preventing the development of obesity. Specifically, metallothionein-2 null mice that were fed a high-fat diet showed a greater increase in body weight and plasma cholesterol than null mice on a control diet, or mice without the knockout on a high-fat diet [158]. Interestingly, no variants within these two genes have been found to associate with HDL cholesterol in published GWAS.

In Orkney, the chromosome 19 peak for systolic blood pressure falls into a cluster of zinc finger protein genes. Zinc finger proteins have previously been implicated in the modulation of blood pressure [159], and it has been shown that differential methylation of zinc finger genes is associated with changes in blood pressure due to drops in temperature [160]. One of the zinc finger genes in this region, *ZNF667*, is also known as 'myocardial ischemic preconditioning upregulated ortholog 1', and it is named after the same gene discovered in mice, where it is abundantly expressed in the heart and has been implicated in maintaining vascular homeostasis, and is upregulated during myocardial ischemia [161]. Several of the zinc finger proteins in this region are expressed in the nervous system, and are involved in neuronal processes, being implicated in amyotrophic lateral sclerosis and alternating hemiplegia of childhood (a neurological condition that causes temporary paralysis in all or parts of the body). Since blood pressure is regulated by the sympathetic nervous system, this locus presents an attractive candidate for follow-up studies.

The chromosome 11 peak for serum uric acid levels in Orkney lies 3.5 Mb away from the *SLC22A12* gene, which is a urate anion exchanger [105]. The same region has been identified with a variance component linkage study on American Indians [162], and associations with SNPs in the *SLC22A12* region have been reported in GWAS of African [163], Japanese [164] and European [114] individuals. This is also the site of the unusual signal detected with the HRC imputation GWAS in Orkney, where associations were detected with low-frequency SNPs in the 46-72 Mb region. Re-running the linkage analysis by fitting the genotype of rs370311822, the top GWAS SNP, as a covariate causes the trait heritability to drop from 0.41 to 0.375 and the LOD score of this region also drops to 0.83. Interestingly, the SNP that has the overall highest LOD score in this region in the pedigree-free linkage analysis, rs7925914, has a *p-value*=0.99 in the HRC GWAS in Orkney, and the two SNPs are not in LD (R^2=0, D'=0.35) in this cohort. Conditioning on this SNP has no effect on the pedigree-free linkage analysis results.

Pedigree-free linkage analysis of serum uric acid levels in Orkney identifies the *SLC2A9* region that is also identified with GWAS and RH. *SLC2A9* encodes a well-characterised transporter protein affecting serum uric acid levels [69], and common variants explaining 1.7-5.3% of the trait variance have been described in trans-ethnic studies. Conditioning on the top GWAS SNP, rs11723439, causes the heritability to drop from 0.41 to 0.40 and the LOD score to drop from 4.63 to 2.92. This means that the common variant explains 2.5% of the trait heritability (1% of the trait variance) in Orkney, and, because it fails to account for the entire linkage peak signal, suggests that other independent QTL effects are also present in this region.

Similarly, linkage analysis of von Willebrand factor identifies the *ABO* locus, which also yielded strong signals with GWAS and RH, and this is discussed in more detail in section 7.1.4. Briefly, re-running the linkage analysis while conditioning on rs514659, the SNP with the most significant *p-value* in the GWAS, causes the heritability to drop from 0.62 to 0.51. This is in line with the results obtained from the simulation studies in Chapter 6, that show that a QTL needs to explain around 10% of the trait variance to be consistently detected with pedigree-free linkage analysis in Orkney. When conditioning on rs514659, the LOD score drops from 10.48 to 1.24, so this common SNP explains the majority of the pedigree-free linkage signal. The interaction between *ABO* blood types and von Willebrand factor levels has been known for over 40 years [165] and this locus is consistently identified in von Willebrand factor GWAS [166] and linkage analyses [167], explaining around 30% of the heritability of this trait [1].

4.4.4.1 Axial Length in Orkney

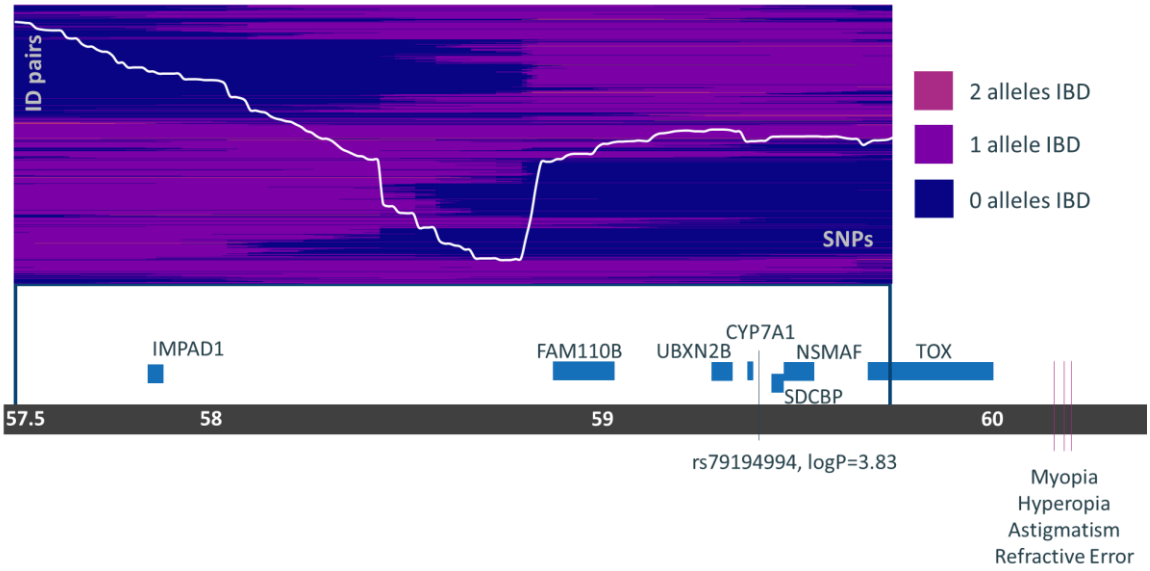
In Orkney, the chromosome 8 peak in the pedigree-free linkage analysis of eye axial length contains 7 genes (Figure 25). In the axial length GWAS using imputed genotypes in Orkney, the SNP with the lowest *p-value* ($-\log_{10}$(*p-value*)=3.83) is rs79194994, located between *CYP7A1* and *SDCBP*. The GWAS catalog reports several associations with refractive error and astigmatism, myopia and hyperopia, ocular disorders that arise because the eye is not the right length in relation to its refractive ability, in the region upstream of *TOX* (just outside the linkage region, also shown in Figure 25) [168–171], but no linkage signal is detected at this locus when analysing refractive error in Orkney. Despite the GWAS catalog hits being upstream of *TOX*, one of these studies [168], which is a large GWAS meta-analysis of eye traits from the CREAM consortium, proposes *SDCBP* (syndecan binding protein, syntenin-1) as the most likely candidate in this region, especially because associations are also detected with SNPs in *CLSTN2* (calsyntenin-2) on chromosome 3, indicating a potential role for syntenins in eye morphology. *SDCBP* shows high expression in all eye tissues and microduplications in the 8q12-8q13 region have been described in individuals who had Duane retraction syndrome, a form of strabismus (cross eye) that is often associated with short eyes [172].

The presence of a cluster where there is an increased amount of IBD sharing could help narrow down the location of the QTL contributing to the linkage signal in this region. IBD coefficients calculated at every SNP in this region by GIBDLD were used to assess this, and Figure 25 shows a plot of the number of alleles shared IBD in this region between pairs of individuals in the dataset. The plot contains 3893 pairs that were selected based on two criteria. First, they had to share at least one allele IBD at 10% (17) of the SNPs within this region. Second, their phenotypes had to be similar, where "similar" is defined as: the phenotype differences between the two members of each pair are calculated and only those pairs are kept where this difference is within half of the first quartile of the distribution of these differences. Overlaid on this plot is the average IBD coefficient at each SNP, calculated by using *all* pairs in the data, as long as both members had the phenotype measured. These results show that this locus has two regions of increased IBD sharing separated by a region of decreased IBD sharing, suggesting that it is the combined signals contributed by both regions that lead to this signal, which would be missed with single-SNP GWAS. This pattern of IBD sharing is not enough to home in on the location of the QTL, however.

Figure 25 - Axial length pedigree-free linkage region in Orkney

The grey horizontal line indicates chromosome position, in Mbp. The region of the linkage peak is indicated by the thick blue vertical lines. Genes are shown as blue rectangles. The location of rs79194994, the SNP reaching the highest -$\log_{10}$(*p-value*) in the GWAS of axial length in Orkney is indicated by the thin vertical blue line. SNPs reported to associate with eye traits in the GWAS catalog are indicated by pink vertical lines.

The plot on top shows the number of alleles shared IBD at each genotyped SNP in this region, between those pairs of individuals who share at least one allele IBD at 10% (17) of the SNPs in this region, and whose phenotypes are similar. Dark blue colour in this plot indicates that a pair shares no alleles IBD at a SNP, purple colour indicates that 1 allele at a SNP is shared IBD by a pair, while pink colour (seen very rarely here) indicates that 2 alleles are shared IBD. The average IBD coefficient at each SNP, across *all* pairs with the phenotype measured, is overlaid on this plot (white line), and ranges from 0.021 to 0.025.



Haplotype analysis was then carried out in order to identify haplotypes that segregate with this trait in Orkney. Using the 1000 Genomes recombination map, this region was divided into haplotype regions based on recombination hotspots – specifically, points where the cM/Mb exceeded 20 were used as the region boundaries, yielding 10 haplotype blocks in this region (Figure 26). Next, every genotyped SNP in each haplotype region was extracted from the phased Orkney genotypes, and was used to construct the segregating haplotypes. For each unique haplotype, every individual was assigned values of 0, 1 or 2 to indicate whether they carried 1, 2 or no copies of this haplotype. The residuals of the normalised phenotype, adjusted

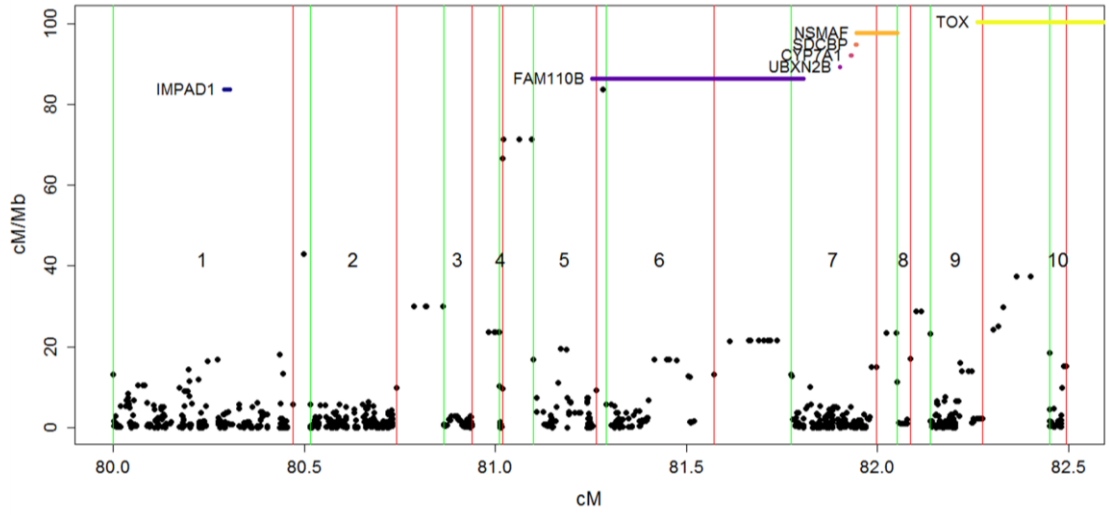
for covariates and genetic kinship, were regressed onto these haplotype counts using a simple linear model.

A haplotype within the regions containing the start of the *TOX* gene showed the strongest association (logP=3.68, 11 instances of this haplotype in Orkney, all in heterozygotes, though the phenotype was measured in only 5 of these). The haplotype analysis was also carried out in Shetland, as this is the cohort most similar to Orkney. While the haplotype identified in Orkney was not present in Shetland, a different haplotype in the same region yielded the strongest association (logP=3.93, 15 heterozygotes) out of all haplotypes across each region.

Fitting this haplotype as a covariate in the linkage analysis caused the LOD score to drop from 3.56 to 3.23, the heritability explained by the region to drop from 0.072 to 0.068 – indicating that this haplotype only explains a small proportion of the signal originating from this locus.

Figure 26 - Haplotypes flanked by recombination hotspots

The cM positions of SNPs present in the 1000Genomes recombination map are plotted along the X axis. Each haplotype is delimited by a point where the cM/Mb recombination rate (plotted on the Y axis) exceeds 20. Haplotype start points are indicated with green lines, while their end points are indicated with red lines, and each haplotype regions is numbered sequentially. Genes in this region are shown with coloured horizontal lines, and the name of each gene is to the left of its corresponding bar.

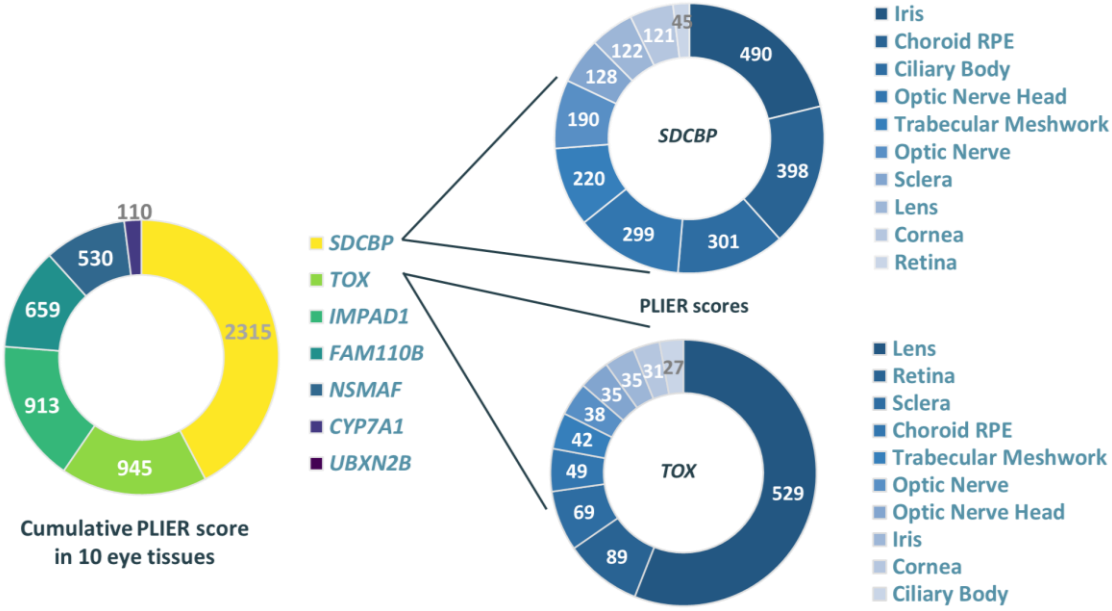


Next, the expression of the 7 genes in this region was assessed using the Ocular Tissue Database [173], which provides gene expression values within 10 different tissues in the eye (Figure 27). *SDCBP* showed the highest levels of overall expression, followed by *TOX*, and the tissue-specific expression patterns of this genes revealed that while *SDCBP* is expressed ubiquitously throughout all eye tissues, *TOX* is predominantly expressed in the lens, which

implies that *TOX* may have an effect on refractive error (as the lens is one of the refractive media in the eye), which is why hits upstream of it were detected in GWAS meta analyses of refractive error.

Figure 27 - Gene expression in the Ocular Tissue Database

This figure shows the expression levels of the 7 genes within the linkage region, normalised to account for the use of different arrays and expressed in PLIER (Probe Logarithmic Intensity Error) scores. On the left is an aggregate PLIER score for each gene, obtained by summing the PLIER scores reported for 10 different eye tissues. On the right, eye tissue-specific PLIER scores for *SDCBP* and *TOX* are shown.



Finally, the function of *SDCBP* and *TOX* was assessed. *TOX* (Thymocyte Selection Associated High Mobility Group Box) encodes a DNA-binding protein that plays a role T-cell development [174], but there are no studies of its function in relation to eye development. *SDCBP* encodes a syndecan binding protein and it is primarily localised to the membrane-associated adherens junctions and focal adhesions, interacting with many transmembrane proteins. Functional enrichment of *SDCBP* and interacting genes shows an enrichment in cytoskeletal functions and ephrin receptor signalling, as well as axon guidance, phototransduction, retinoid metabolism, cell adhesion and extracellular matrix development. Ephrins affect neuronal, vascular and epithelial development and are especially involved in axon guidance [175]. Interestingly, neuronal functions are intimately linked with eye morphology. Axial length is not fixed, increasing until humans each adulthood, and decreasing

with old age, but it also shows circadian changes. The eye needs to adjust to this change in length in order to ensure that the images are focused on the focal plane of the retina and it is suggested that this happens through an active emmetropisation mechanism by altering the extracellular matrix composition of the sclera [176]. The fact that *SDCBP* and its interacting gene partners are enriched in functions pertaining to extracellular matrix development and cell adhesion is an encouraging sign that this gene could really effect axial length.

4.4.5 Meta-analysis

The results of the meta-analysis reveal regions where several cohorts contribute weak signals that would have gone unnoticed if the results were only analysed within each cohort separately. In some cases, these signals appear in cohorts that are geographically distant, suggesting the presence of common variants present in all these cohorts, or the presence of cohort-specific QTLs segregating in the same region.

For example, Orkney, GS and Korčula all contribute to the educational attainment peak on chromosome 6. This peak is within the major histocompatibility complex (MHC) region and one case-control study fine-mapped a QTL in this region using individuals with extremely high intelligence in an attempt to identify QTLs influencing IQ, and while one variant passed their study-wide significance threshold, it failed to explain a large proportion of trait variance [177]. A large meta-analysis of GWAS of educational attainment using SNPs imputed to the 1000 Genomes reference panel identified several associated common variants 3-10Mb away from this region [104]. The peak region contains three genes, *MUC21*, *STFA2* and *DPCR1*, none of which have obvious links to cognitive functions or intelligence.

Meta-analysis of fasting glucose levels identifies a peak at the 4q34.1 locus, with signals contributed by the Orkney, Shetland and Korčula studies. This locus has been linked to type 2 diabetes with obesity in a study of Korean individuals, where the region was identified with whole-genome linkage analysis and this region was then fine-mapped with association analysis [178]. The 0.1 cM peak encompasses a single gene, *HPGD* (hydroxyprostaglandin Dehydrogenase 15-(NAD)). One study suppressed diabetes in mice with the help of *Isaria sinclarii*, a fungus cultured on silkworm, and studied the subsequent changes in gene expression [179]. *Isaria sinclarii* produces myriocin, which modulates the sphingosine-1-phosphate receptor, and sphinghosine-1-phosphate is an upstream regulator of prostaglandin production [180]. The researchers found that the drop in glucose levels correlated with the drop in *HPGD* expression, suggesting a possible role for the *HPGD* gene product in regulating glucose levels.

Conversely, some signals are only present in cohorts from the same geographical region, suggesting the presence of region-specific QTLs that may be absent in geographically distant cohorts.

For example, meta-analysis of HDL cholesterol levels reveals a region at the 18q21.33 locus. The three Scottish cohorts contribute to this signal (LOD = 0.38, 3.47 and 2.99 in Orkney, Shetland and GS, respectively), while it is absent in Vis and Korčula. The peak contains a single gene, *VPS4B* (vacuolar protein sorting 4 homolog B), but no functional associations between this gene and cholesterol are reported in the literature.

The 15q21.3 peak is due to signals from GS and Shetland, and this locus contains the *TCF12* gene and part of the *ZNF280D* gene. *TCF12* (Transcription Factor 12) is a transcription factor that is expressed in skeletal muscle and cardiac muscle and GWAS have identified associations between this gene and coronary artery disease (CAD). Its targets, identified with chromatin immunoprecipitation sequencing (ChIP-Seq) were enriched for growth-factor binding and matrix interaction functions, and TCF12 target genes are over-represented among genes associated with height. Short stature has repeatedly been linked to increased risk of CAD [181] with a Mendelian randomisation study establishing a causal link between height-affecting SNPs and CAD risk [182], hinting at a common pathway affecting both traits.

Generally, the pedigree-free linkage meta-analysis results highlight regions that are primarily driven by the results obtained in one cohort, while a signal is either absent or very modest in the other cohort. Although limited to two cohorts only, this suggests the presence of cohort-specific QTLs that are only segregating within Croatian or Scottish populations, but not both.

One exception to this is the hit on chromosome 22 in gamma-glutamyl transferase levels. Both Orkney and Vis pedigree-free linkage results contribute roughly equally (but individually not significantly) to this peak. This region contains the *GGT1* and *GGT5* genes, giving support for this result to be a true positive as the genes are paralogs and give rise to gamma-glutamyl transferase enzymes. No SNPs (in the case of GWAS) or regions (in the case of RH) at this locus reach genome-wide significance in either cohort. However, GWAS using imputed genotypes in this region reveals an association with rs2330795 (MAF 36% in Orkney, 41% in Vis) that has a $-\log_{10}(p\text{-}value)$ of 7.52 in both cohorts. Additionally, in Shetland, RH yields a sharp peak at this locus, further validating this region as a QTL for circulating GGT enzyme levels in a third cohort.

The two regions that yield the strongest meta-analysis signals in the pedigree-free linkage analysis are the *ABO* region for von Willebrand factor and the *SLC2A9* region for serum uric

acid levels, which are both well-characterised loci that affect these traits, having previously been identified by GWAS [69, 166]. The stronger signal in these regions is detected in Orkney, but some signal is also present in Vis, as shown in Table 16.

An important consideration is the size of the regions used in meta-analysis. For pedigree-based linkage analysis, IBD sharing was output at every 0.1 cM interval along each chromosome. The per-cohort results can easily be aggregated for meta-analysis regardless of what SNPs are within these intervals. Problems may arise when several cohorts have high LOD scores in close, but non-overlapping regions.

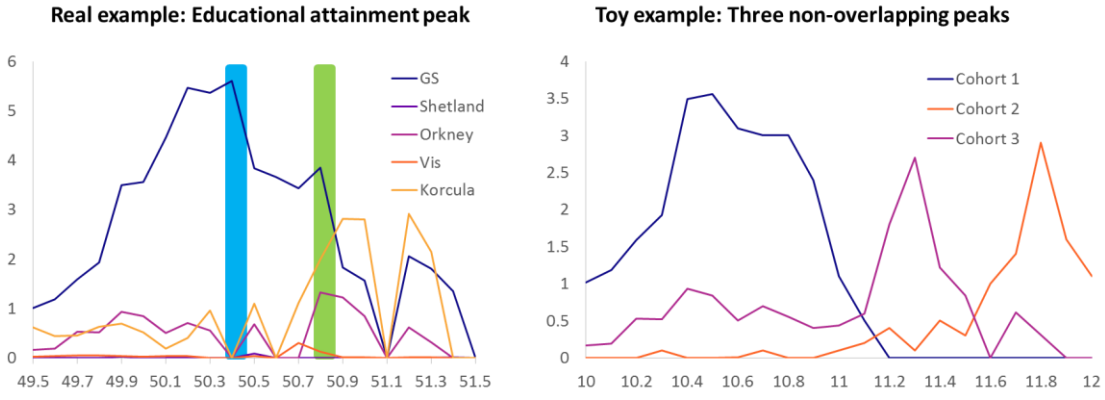
The educational attainment meta-analysis hit (in the region spanning 50-51.3 cM) on chromosome 6 exemplifies this: the 2-LOD drop region around the peak (at 50.8 cM) contains the HLA region. In GS, IBD sharing at a 50.4 cM in this region yielded a LOD score of 5.61 but no signal was detected in the other cohorts. However, at 50.8 cM, the GS LOD score is smaller but not insignificant (3.85), and Orkney and Korčula also have LOD scores above 1 here, which is why this is detected in the meta-analysis (Figure 28, left panel). Another scenario, however, could be that there are three distinct, non-overlapping peaks within a 2 cM region. If results are aggregated at every 0.1 cM, this region would be missed in the meta-analysis. If, however, the maximum LOD score is selected out of all regions within a larger interval (for example, every 2 cM along each chromosome), then this region would result in a meta-analysis signal that all analysed cohorts contribute to (Figure 28, right panel). This would come at the expense of resolution, as these regions would necessarily be larger and may span several genes.

Figure 28 - Challenges with defining meta-analysis regions

LOD scores, obtained at every 0.1 cM position within 2 cM interval, are plotted for each cohort.

Right panel: Region that yielded a peak for educational attainment on chromosome 6. The meta-analysis peak is indicated by the green rectangle while the peak in GS is indicated by the blue rectangle.

Left panel: Toy example showing three cohorts with non-overlapping peaks in a 2 cM region.



Chapter 5 Regional Heritability

5.1 Introduction

Yang *et al.* [18] proposed a method to address the missing heritability problem that arose when it was found that the genome-wide significant loci from GWAS only explained a small amount of trait heritability [17] and implemented it in a program called GCTA. They demonstrated that most of the heritability is "hidden" rather than missing, because collectively, genotyped variants do explain a large proportion of trait heritability, but the effect sizes of individual variants are often too small to be detected by GWAS. Regional heritability (RH) mapping [45] was developed following on from this rationale – if one can calculate the total proportion of trait variance captured by all genotyped SNPs across the genome, then the genome could be partitioned into smaller regions, and the variance captured by all genotyped SNPs within each region can also be calculated.

Using simulated data, Riggio and Pong-Wong showed that RH mapping outperformed association and linkage analyses in terms of power to detect a QTL, as it identified more true QTLs than these other methods [183]. RH mapping has been used to identify a novel relationship between major depressive disorder and a transcription factor binding site within the *TOX2* gene, in GS [184]. Variants within this locus modulate the expression of *TOX2* and RP1-269M15.3, a long non-coding RNA, in the brain. A comparison of GWAS and RH mapping of blood lipid traits (LDL cholesterol, HDL cholesterol, triglycerides and total cholesterol) in Vis, Korčula and a metropolitan cohort from the city of Split had revealed that RH can identify additional significant signals compared to GWAS performed in the same cohorts [48], all corresponding to loci reported in large lipid GWAS meta-analyses ([97, 185]). The same work also showed that a locus that was flagged by GWAS in our cohorts, but not the large GWAS meta-analysis, did not yield a significant test statistic with RH, indicating that RH may be less sensitive to false positive signals.

Within this chapter, I use RH mapping to systematically analyse the quantitative traits listed in Section 2.2, in each cohort separately, followed by a meta-analysis. In order to enable the meta-analysis, I use a modified version of RH mapping that is based on 'haplotype' blocks defined by cohort-independent recombination maps, rather than regions defined by a sliding window consisting of a fixed number of SNPs. This has the twofold benefit of reducing LD block complexity within each region and ensuring that regions are comparable between different cohorts, regardless of the number of SNPs genotyped in each cohort.

5.2 Methods

Regional heritability analyses were carried out using the program REACTA [186] and its successor, DISSECT [187], developed by the same team at the University of Edinburgh. These programs are based on the statistical principles employed by GCTA [18] but extend its functionality by implementing and automating regional analysis.

5.2.1 Linear mixed models

These programs use linear mixed models (LMMs) to compute the trait variance explained by all, or a subset of, genotyped SNPs, using the restricted maximum likelihood (REML) method. As was the case for variance component linkage analysis, presented in section 4.2.3, the regional heritability approach, as described in [45], adjusts the LMM equation presented in section 3.2.2 to account for both whole-genome and regional effects:

$$y = X\beta + Zu + Wv + e$$

where Z and W are the design matrices for random effects, u is the whole-genome additive effect with variance $var(u) = G\sigma_u^2$, v is the regional additive effect with variance $var(v) = Q\sigma_v^2$, e is the residual effect with variance $var(e) = I\sigma_e^2$. Matrices G, Q and I are the whole-genome GRM, regional GRM and the identity matrix. The phenotypic variance σ_p^2 can be expressed as $\sigma_u^2 + \sigma_v^2 + \sigma_e^2$ and the whole genome heritability h_u^2 is $\frac{\sigma_u^2}{\sigma_p^2}$ while the regional heritability h_v^2 is $\frac{\sigma_v^2}{\sigma_p^2}$.

In RH analyses, as shown above, two random effects are fitted: a whole-genome GRM (calculated using all the available genotyped SNPs) to account for background relatedness and polygenic effect, and a regional GRM (calculated using a subset of adjacent SNPs) that accounts for local effects. These GRMs are calculated as described in section 3.2.1, using either all autosomal SNPs across the genome in the case of whole-genome GRMs, or a subset of SNPs in the case of the regional GRM.

The LMM described in [45] is the same that is used by DISSECT, but there is one difference in the whole-genome GRM used: while only one whole-genome GRM is computed for each cohort, the regional GRM is subtracted from the whole-genome GRM prior to each regional analysis. This ensures that a regional effect is not "counted twice", which might reduce the likelihood of detecting a regional effect.

5.2.2 Defining a Region

The genome can be broken down into regions consisting of arbitrary combinations of SNPs, such as regions containing all SNPs within a gene, or regions flanked by recombination hotspots. The original implementation of RH mapping in REACTA uses sliding windows of *n* SNPs, sliding the window *m* SNPs forward after each analysis. While this keeps the number of SNPs in each region constant, the number of resulting windows depends on the number of genotyped SNPs available.

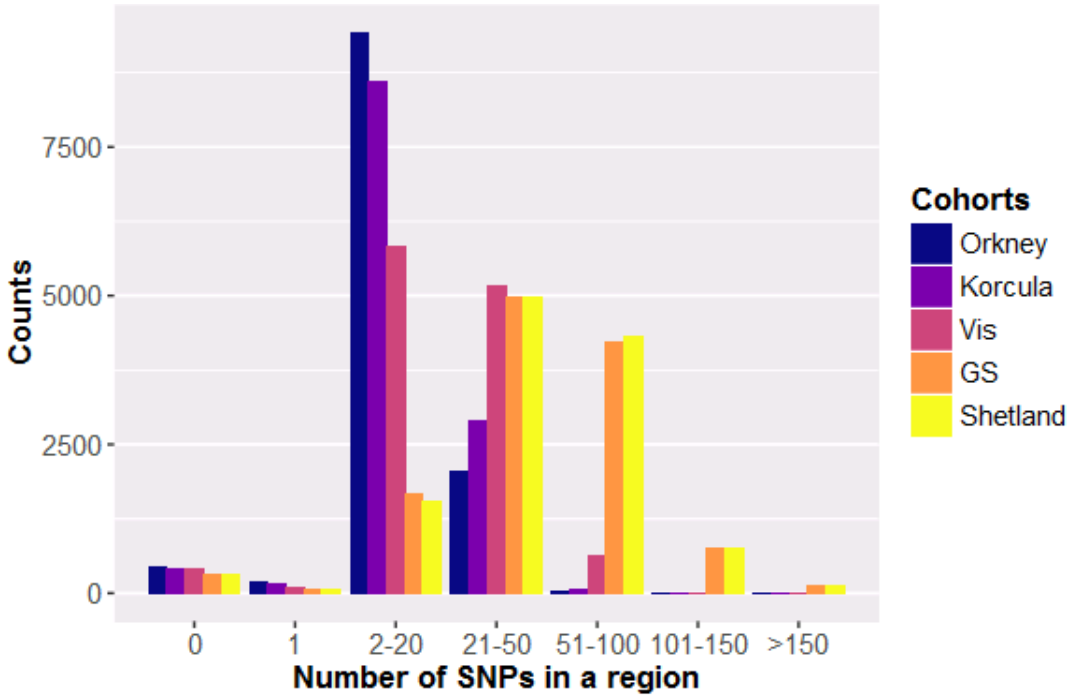
In order to keep the window positions consistent across different cohorts, regardless of the number of genotyped SNPs available in a cohort, I created 0.3 cM windows spanning the whole genome. This gave rise to 12101 non-overlapping windows, some of which may contain no SNPs in some cohorts. This definition of a window enables the meta-analysis of results, as the regions will be located at the same positions regardless of cohort. Windows of this size were chosen in order to ensure that most regions have >1 SNPs in all cohorts. Figure 29 shows the distribution of the number of SNPs allocated to each region, within each cohort.

REACTA only implements on-the-fly GRM calculations if the sliding window approach described above is used, so is not particularly well-suited for analysing arbitrarily-defined regions. This means that for each region, the regional GRM has to be pre-calculated and stored, the whole-genome GRM has to be adjusted by the regional GRM with the help of external programs, then the regional and adjusted whole-genome GRMs have to be read into REACTA to perform the analysis, which adds significant time and storage overheads.

In contrast, DISSECT implements a grouping approach where an input file may assign genotyped SNP to one (or more) groups (a SNP may also be assigned to no group). The whole-genome GRM and whole-genome genotype files are read into the program once, and then, for each group, the regional GRM is calculated using all its constituent SNPs on-the-fly, subtracted from the whole-genome GRM and the REML is performed on the full and reduced models, outputting the variances, heritabilities and logLikelihoods attributed to these models, as well as the LRT and its associated *p-value*. Within this thesis, I have initially used REACTA to test the window-overlap approach, but the results I present here were obtained with DISSECT.

Figure 29 - Distribution of the number of SNPs in 0.3 cM regions across the genome

The X axis shows the number of SNPs in each 0.3 cM region across the autosomal genome, grouped into 7 bins. The Y axis shows the number of regions belonging to each bin, in each cohort.



5.3 Results

5.3.1 Regional Heritability Results by Cohort

RH analyses have been performed in DISSECT for each trait available in each cohort. The genome was partitioned into 0.3 cM segments, and all SNPs falling into each segment were used to calculate the regional GRM. In total, 97 hits yielded RH test statistics that exceed the suggestive significance threshold ($-\log_{10}(0.05/12101)=5.38$), and Table 17 lists the 70 hits that exceed the genome-wide significance threshold that is adjusted for the number of traits analysed in each cohort. The hits that exceed the suggestive but not the genome-wide suggestive threshold are presented in Supplementary Table 6. The majority (49) of the GWS hits originate in GS, which has the largest sample size out of all the cohorts analysed.

All of the hits reported in Table 17, and most of the hits reported in Supplementary Table 6, replicate the findings of GWAS reported in the GWAS catalog [16] or within this thesis, and

persist in the meta-analysis. 18 of the per-cohort hits are not found in the meta-analysis, either because the trait was only available in one cohort, or because the result is no longer significant in the meta-analysis, these are indicated with asterisks in the table and they all contain genes for which associations have been reported in the GWAS catalog.

The body fat percentage phenotype was only available in GS, so this trait was not meta-analysed, but the region identified in GS corresponds to the well-known *FTO* locus, which was shown to associate with BMI and obesity for the first time in 2007, in several independent studies [188–190]. The same locus has a strong signal in RH analyses of BMI in GS, and some signal is also present in Vis and Korčula, but not Orkney and Shetland.

Table 17 - Regional heritability results that exceeded the GWS threshold in individual cohorts

The RH $-\log_{10}(p\text{-}value)$ of each region is displayed in the logP column. The chromosome and start and end positions of each region (in Mbp) are provided. Additionally, the start of each 0.3 cM region is also indicated in cM. The total trait heritability (h^2) and heritability explained by the region (h^2reg), are shown, as well as the chromosome band (Band) where the region is located. Rows with asterisks (*) indicate hits that are not present in the RH meta-analysis. GWAS in the literature have reported associations between the relevant trait and a SNP within every region reported here.

Trait	Chr	cM	Mbp	logP	h^2	h^2reg	Band
Orkney							
HDL	16	72.9	56.95-57.04	13.32	0.53	0.051	q13
LDL	19	70.5	45.34-45.43	7.77	0.32	0.051	q13.32
Uric acid1	4	23.4	9.91-10.54	21.01	0.44	0.057	p16.1
Uric acid2	4	23.4	9.91-10.54	20.53	0.43	0.057	p16.1
vWF	9	166.2	136.13-136.37	32.81	0.66	0.275	q34.2
Vis							
vWF	9	166.2	136.13-136.37	22.65	0.51	0.134	q34.2
Shetland							
Central Corneal Thickness	16	127.5	88.24-88.38	10.75	0.8	0.017	q24.2
CRP	1	175.2	159.53-159.76	6.85	0.29	0.025	q23.2
GGT	22	22.8	24.98-25.14	8.22	0.38	0.029	q11.23
Glucose	11	99.9	92.63-92.8	9.45	0.3	0.03	q14.3
Glucose_nodiab	11	99.9	92.63-92.8	10.95	0.29	0.033	q14.3
HDL	16	72.9	56.95-57.04	15.34	0.44	0.031	q13
LDL	19	70.5	45.34-45.43	11.86	0.13	0.038	q13.32
Total Cholesterol	19	70.5	45.34-45.43	7.29	0.16	0.025	q13.32
Triglycerides	11	126	116.55-117.1	7.19	0.31	0.026	q23.3

Trait	Chr	cM	Mbp	logP	h^2	h^2reg	Band
Uric acid1	4	23.4	9.91-10.54	18.49	0.46	0.037	p16.1
Uric acid2	4	23.4	9.91-10.54	19.80	0.46	0.042	p16.1
Korčula							
HDL	16	72.9	56.95-57.04	13.24	0.42	0.044	q13
Heart Rate*	2	246.3	228.29-228.52	7.60	0.15	0.033	q36.3
Uric acid1	4	23.4	9.91-10.54	16.11	0.35	0.029	p16.1
Uric acid2	4	23.4	9.91-10.54	15.92	0.32	0.032	p16.1
GS							
BMI	16	64.5	53.77-53.85	16.18	0.47	0.002	q12.2
Body fat*	16	64.5	53.77-53.85	12.08	0.44	0.001	q12.2
Creatinine*	5	200.1	176.78-176.96	7.45	0.44	0.002	q35.3
Forced Vital Capacity*	6	20.4	7.68-7.92	6.83	0.35	0.004	p24.3
Glucose	2	185.4	169.65-169.82	71.38	0.22	0.019	q24.3-q31.1
Glucose*	3	179.7	170.38-170.77	8.73	0.23	0.004	q26.2
Glucose	7	69.3	43.89-44.27	24.01	0.24	0.013	p13
Glucose	8	133.2	118.18-118.3	8.20	0.23	0.004	q24.11
Glucose_nodiab	2	185.4	169.65-169.82	77.85	0.25	0.023	q24.3-q31.1
Glucose_nodiab*	3	179.7	170.38-170.77	8.17	0.26	0.005	q26.2
Glucose_nodiab	7	69.3	43.89-44.27	28.90	0.27	0.015	p13
Glucose_nodiab	8	133.2	118.18-118.3	9.54	0.26	0.004	q24.11
Glucose_nodiab*	13	20.7	28.45-28.59	7.75	0.26	0.003	q12.2
HDL	8	47.1	19.71-19.94	28.38	0.5	0.012	p21.3
HDL	9	124.2	107.63-107.68	11.11	0.5	0.006	q31.1
HDL*	11	126	116.55-117.1	8.86	0.5	0.008	q23.3
HDL	15	78.3	58.63-58.71	26.02	0.5	0.005	q21.3
HDL	16	72.9	56.95-57.04	165.85	0.5	0.042	q12.2-q13
HDL	19	70.5	45.34-45.43	12.08	0.5	0.006	q13.32
Heart Rate	14	11.7	23.76-23.98	10.31	0.25	0.005	q11.2
Height*	1	150.6	118.84-119.31	6.99	0.82	0.002	p12
Height*	1	33.9	17.19-17.5	6.89	0.82	0.002	p36.13
Height	2	83.1	55.9-56.24	10.35	0.82	0.003	p16.1
Height	3	153.6	141.01-141.34	13.73	0.82	0.002	q23
Height	4	153.6	145.2-146.17	8.17	0.82	0.004	q31.21
Height*	4	33.6	17.74-18.27	7.42	0.82	0.001	p15.32-p15.31
Height	5	51	32.7-32.89	8.09	0.82	0.006	p13.3
Height	6	151.5	142.63-143.06	11.11	0.82	0.002	q24.1-q24.2
Height*	6	49.2	25.69-26.67	9.16	0.82	0.006	p22.2
Height*	6	54.3	34.04-34.24	7.99	0.82	0.003	p21.31

Trait	Chr	cM	Mbp	logP	h^2	h^2reg	Band
Height*	6	132.6	126.44-127.53	7.53	0.82	0.002	q22.32-q22.33
Height	7	108.3	92.06-92.49	6.92	0.82	0.001	q21.2
Height*	12	79.2	66.3-66.42	7.31	0.82	0.001	q14.3
Height	15	115.2	89.37-89.45	9.44	0.82	0.004	q26.1
Height	18	42.9	20.68-21.07	10.68	0.82	0.002	q11.2
Height	20	53.7	32.9-34.72	19.41	0.82	0.007	q11.22-q11.23
Total Cholesterol	1	138.6	109.73-110.12	14.48	0.27	0.005	p13.3
Total Cholesterol	1	80.7	55.47-55.51	12.20	0.27	0.005	p32.3
Total Cholesterol*	1	92.1	62.83-63.38	8.02	0.27	0.001	p31.3
Total Cholesterol	2	42	21.25-21.54	16.43	0.27	0.003	p24.1
Total Cholesterol	2	68.4	43.91-44.1	9.39	0.28	0.008	p21
Total Cholesterol	5	85.8	74.24-74.95	8.25	0.27	0.002	q13.3
Total Cholesterol*	8	24.9	9.1-9.2	6.71	0.27	0.003	p23.1
Total Cholesterol	9	124.2	107.63-107.68	6.99	0.27	0.003	q31.1
Total Cholesterol*	16	87.3	72.1-72.93	7.25	0.27	0.004	q22.2-q22.3
Total Cholesterol	19	70.5	45.34-45.43	90.24	0.32	0.083	q13.31-q13.32
Total Cholesterol	19	31.8	11.01-11.28	25.92	0.28	0.017	p13.2
Urea	3	206.7	187.63-187.8	11.89	0.22	0.005	q27.3
Urea	18	63	43.13-43.3	6.96	0.22	0.004	q12.3

5.3.2 Meta-analysis Results

Meta-analysis was conducted by applying Fisher's combined test, as described in section 4.2.7, to the RH results obtained in each cohort. Regions with meta $-\log_{10}(P_{meta}) > 5.38$ are considered suggestively significant, while regions with meta $-\log_{10}(P_{meta}) > 6.89$ are considered genome-wide significant (GWS), taking into account the number of traits analysed.

After meta-analysis, 45 regions yield GWS meta-analysis results, and these are presented in Table 18. In total, 69 regions pass the suggestive significance threshold, and the suggestively significant results are presented in Supplementary Table 7. Most of the meta-analysis hits are predominantly led by strong signals originating in GS, and published GWAS [16] have reported hits within, or near, every GWS RH meta-analysis peak that I present here.

Table 18 - RH meta-analysis results that pass the GWS threshold

The meta-analysis -$\log_{10}$(*p-value*) is indicated (logP column) for each peak. These peaks represent 0.3 cM regions that start at the cM position indicated (cM column). The start and end positions of these regions are also shown in Mbp, as is the chromosome band (Band) where these regions can be found. The per-cohort -$\log_{10}$(*p-values*) at the peak region are shown (O = Orkney, S=Shetland, G=GS, V=Vis, K=Korčula). The final column shows the genes in each region (or on the same chromosome band) that have been implicated in GWAS of the corresponding trait reported in the literature.

Trait	Chr	cM	Mbp	logP	Band	O	S	G	V	K	GWAS
BMI	16	64.5	53.77-53.85	17.79	q12.2	0.38	0.30	16.18	1.54	4.96	*FTO*
Central Corneal Thickness	16	127.5	88.24-88.38	14.73	q24.2	0.61	10.75	NA	2.32	5.22	*ZNF469*
CRP	1	175.2	159.53-159.76	14.96	q23.2	5.38	6.85	NA	5.68	NA	*CRP*
GGT	22	22.8	24.98-25.14	8.64	q11.23	0.36	8.22	NA	2.61	NA	*GGT1*
Glucose	2	185.4	169.65-169.82	70.27	q24.3-q31.1	3.69	1.45	71.38	0.67	0.73	*G6PC2*
Glucose	7	69.3	43.89-44.27	23.52	p13	0.30	3.45	24.01	0.30	1.43	*GCK*
Glucose	11	99.9	92.63-92.8	14.29	q14.3	2.92	9.45	2.82	1.87	2.51	*MTNR1B*
Glucose	8	133.2	118.18-118.3	9.26	q24.11	0.42	4.75	8.20	0.30	0.30	*SLC30A8*
Glucose	3	179.7	170.38-170.77	7.19	q26.2	0.30	0.32	8.73	0.45	1.78	*SLC2A2*
Glucose_nodiab	2	185.4	169.65-169.82	79.83	q24.3-q31.1	3.91	2.37	77.85	0.96	2.60	*G6PC2*
Glucose_nodiab	7	69.3	43.89-44.27	31.93	p13	0.69	4.38	28.90	1.16	3.23	*GCK*
Glucose_nodiab	11	99.9	92.63-92.8	13.71	q14.3	2.79	10.95	1.41	0.63	3.13	*MNTR1B*
Glucose_nodiab	8	133.2	118.18-118.3	10.47	q24.11	0.42	4.67	9.54	0.30	0.40	*SLC30A8*
HDL	16	72.9	56.95-57.04	201.08	q12.2-q13	13.32	15.34	165.85	2.70	13.24	*FTO*
HDL	15	78.3	58.63-58.71	32.15	q21.3	4.20	2.74	26.02	3.17	2.44	*LIPC*
HDL	8	47.1	19.71-19.94	30.61	p21.3	0.90	2.03	28.38	0.34	5.32	*LPL*
HDL	9	124.2	107.63-107.68	15.43	q31.1	0.30	2.95	11.11	3.88	2.56	*ABCA1*

Trait	Chr	cM	Mbp	logP	Band	O	S	G	V	K	GWAS
HDL	19	70.5	45.34-45.43	8.93	q13.32	0.30	0.30	12.08	0.49	0.42	*APOE*
HDL	18	68.7	46.56-47.18	8.18	q21.1	0.30	4.27	6.07	0.76	1.33	*LIPG*
Heart Rate	14	11.7	23.76-23.98	9.41	q11.2	NA	1.11	10.31	NA	0.61	*MYH6*
Height	20	53.7	32.9-34.72	19.37	q11.21-q11.23	0.93	2.46	19.41	0.89	1.37	*GDF5*
Height	3	153.6	141.01-141.34	15.36	q23	0.30	5.23	13.73	NA	0.30	*ZBTB38*
Height	15	115.2	89.37-89.45	9.32	q26.1	2.38	0.30	9.44	NA	0.89	*ACAN*
Height	4	153.6	145.2-146.17	9.12	q31.21	0.55	2.06	8.17	1.21	1.82	*HHIP*
Height	6	151.5	142.63-143.06	9.12	q24.1-q24.2	1.06	0.30	11.11	NA	0.32	*GPR126*
Height	18	42.9	20.68-21.07	9.01	q11.2	0.70	0.30	10.68	0.71	1.30	*CABLES1*
Height	2	83.1	55.9-56.24	8.56	p16.1	0.30	1.21	10.35	NA	0.30	*EFEMP1*
Height	7	3.9	2.69-2.92	7.79	p22.3-p22.2	3.74	1.68	5.33	1.15	0.37	*GNA12, AMZ1*
Height	7	108.3	92.06-92.49	7.38	q21.2	0.95	2.16	6.92	1.24	0.53	*CDK6*
Height	5	51	32.7-32.89	6.95	p13.3	0.30	1.31	8.09	0.46	1.13	*NPR3*
LDL	19	70.5	45.34-45.43	19.97	q13.32	7.77	11.86	NA	1.00	3.84	*APOE*
Total Cholesterol	19	70.5	45.34-45.43	94.92	q13.31-q13.32	3.51	7.29	90.24	0.87	1.12	*APOC1, APOE*
Total Cholesterol	19	31.8	11.01-11.28	24.91	p13.2	1.22	0.30	25.92	2.37	1.16	*LDLR*
Total Cholesterol	2	42	21.25-21.54	12.94	p24.1	0.66	0.35	16.43	0.30	0.34	*APOB*
Total Cholesterol	1	138.6	109.73-110.12	12.72	p13.3	0.96	1.67	14.48	0.34	0.37	*SORT1, CELSR2*
Total Cholesterol	5	85.8	74.24-74.95	9.24	q13.3	0.93	1.81	8.25	1.85	1.09	*HMGCR*
Total Cholesterol	1	80.7	55.47-55.51	8.76	p32.3	0.30	0.30	12.20	0.30	0.30	*PCSK9*
Total Cholesterol	2	68.4	43.91-44.1	8.64	p21	0.71	0.56	9.39	0.59	2.01	*ABCG8, ABCG5*
Total Cholesterol	9	124.2	107.63-107.68	7.25	q31.1	0.30	1.13	6.99	1.32	1.92	*ABCA1*

Trait	Chr	cM	Mbp	logP	Band	O	S	G	V	K	GWAS
Triglycerides	11	126	116.55-117.1	12.28	q23.3	2.80	7.19	NA	1.10	5.17	*APOC3, APOA1, APOA4, APOA5*
Urea	3	206.7	187.63-187.8	11.28	q27.3	0.99	0.41	11.89	1.88	NA	*LPP, BCL6*
Urea	18	63	43.13-43.3	7.24	q12.3	2.42	1.01	6.96	0.30	NA	*SLC14A1, SLC14A2*
Uric acid1	4	23.4	9.91-10.54	55.92	p16.1	21.01	18.49	NA	5.99	16.11	*SLC2A9*
Uric acid2	4	23.4	9.91-10.54	56.1	p16.1	20.53	19.80	NA	5.55	15.92	*SLC2A9*
vWF	9	166.2	136.13-136.37	53.36	q34.2	32.81	NA	NA	22.65	NA	*ABO*

5.4 Discussion

Every peak that is GWS in the RH meta-analysis contains SNPs that associated with the relevant trait in the GWAS literature. The benefit of RH over GWAS is that here, these effects are detected in a sample size of 23000 European individuals, while some of the GWAS hits reported in the GWAS catalog [16] required over 100000 individuals to be detected. This could either be because of the lower multiple testing penalty applied with RH mapping, or it could also be due the presence of multiple independent signals in a region. If the RH signal disappears when conditioning on the top GWAS SNP, this would be evidence for the presence of a single causal variant in the region, while if some signal remains, this could suggest the presence of multiple independent signals. In section 7.1.4, both scenarios are demonstrated to occur in adjacent regions at the *ABO* locus. Some of the GWS and suggestively significant meta-analysis hits are discussed in more detail below, and illustrate the ability of RH to identify 'true positive' hits using sample sizes that are smaller than those used in large GWAS meta-analyses.

Meta-analysis of gamma-glutamyl transferase levels flags the *GGT1* gene locus as GWS. This locus was identified with the pedigree-free linkage meta-analysis and also yields a GWS signal in Shetland RH analysis. In the RH meta-analysis, Vis contributes a weak signal (Vis logP=2.61) (note, GS does not have this phenotype measured). It is interesting that this locus has a strong signal in Shetland while a signal is completely absent in Orkney.

The GWS peak on chromosome 3 for urea levels falls into an intergenic region. A GWAS meta-analysis of blood urea nitrogen (BUN) levels in east Asian populations reports a hit in the same intergenic region [191]. The flanking genes (*BCL6* and *LPP*) have not been implicated in kidney function. In contrast, the peak on chromosome 18 for the same trait contains the *SLC14A1* and *SLC14A2* genes, both of which produce urea transporter proteins. This locus is also flagged by the same study, but no studies of European individuals have linked it to this trait before.

The suggestively significant heart rate signal on chromosome 2 originates in Korčula, with neither GS nor Shetland contributing signal. A heart rate GWAS has identified a signal in the *COL4A3* gene 155kb away [192], but the RH region only contains the *AGFG1* and *C2orf83* genes, neither of which have functions associated with heart rate. This does not preclude that they may have regulatory functions that affect heart rate, however.

The suggestively significant signal on chromosome 5 for serum creatinine levels contains the *SLC34A1* gene, with GS contributing most strongly to the signal (GS logP=7.45). This signal

is also identified in the GS GWAS using genotyped SNPs, while a GWAS meta-analysis of 5 European population isolates, including Orkney and Vis, failed to identify this signal [193], as did a larger meta-analysis of ~24000 European individuals [194]. This locus has, however, been associated with chronic kidney disease and glomerular filtration rate of creatinine in a meta-analysis of 67000 individuals [195] as well as a meta-analysis of 133000 individuals [196].

Some of the genes implicated in this meta-analysis encode proteins whose functions can clearly be linked to the measured trait with which they associate. For example, *SLC2A9* is a urate transporter, *SLC14A1* and *SLC14A2* are urea transporters, *SLC2A2* is a glucose transporter, the apolipoproteins and lipases are involved in the regulation of cholesterols, glucokinase mediates glucose uptake while glucose-6-phosphatase is involved in glucose homeostasis, *GGT1* and *CRP* produce the proteins that are detected in GGT tests and CRP test, respectively, the calcium-sensing receptor encoded by *CASR* is responsible for calcium homeostasis, ZNF469 regulates the organisation of collagen fibres in the cornea, *MYH6* produces cardiac muscle myosins and aggrecan is a component of cartilage that withstands compression. Such results lend confidence to the validity of the signals detected with RH mapping.

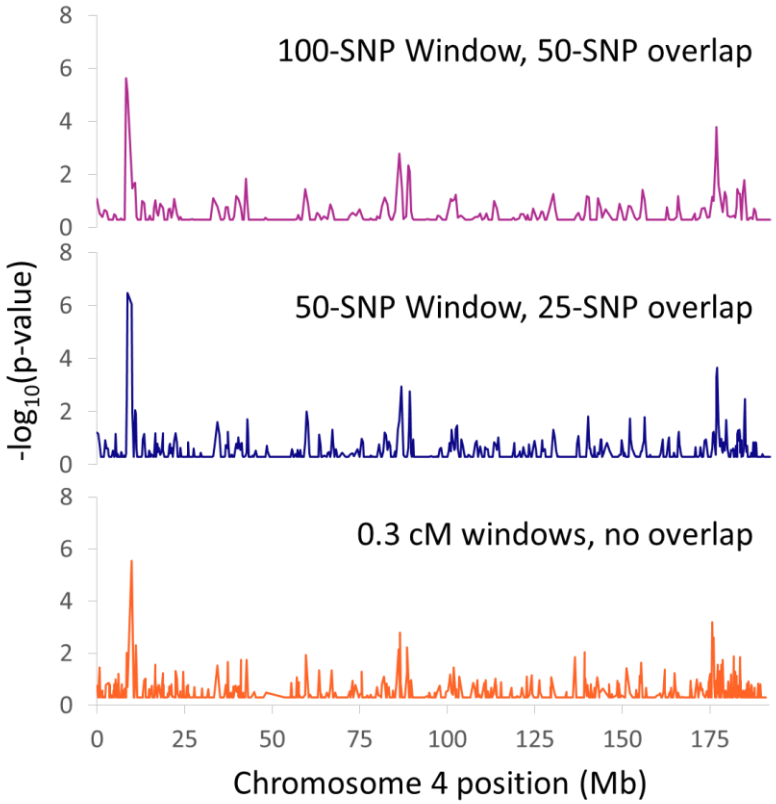
One possible concern with using RH is that the effect of a QTL may be 'diluted' by the effects of other SNPs included in the same region. Nagamine *et al.*[45] show that this is not the case, because the magnitude of the regional effect does not change substantially in a region containing a QTL, regardless of whether 10, 20 or 100-SNP windows are used. This also means that there is no requirement to keep the number of SNPs in each region constant, which is reassuring because within this thesis, the genome was divided into 0.3 cM regions, and SNPs were allocated to each region based on their cM position. Because each cohort has a different number of genotyped SNPs, this results in a different number of SNPs allocated to the same region in different cohorts. In Figure 30, I show a comparison of the RH mapping results obtained from regions defined in 3 different ways: the first and second analyses both used the sliding window approach, with windows consisting of 100 or 50 SNPs, with a 50 or 25 SNPs overlap between regions, respectively. The third analysis corresponds to 0.3 cM windows that contain between 1 and 77 SNPs in each window, with 23 SNPs per window on average. It can be seen that the results of all three analyses are similar, corroborating the statement that the effect of a QTL is not substantially diluted by the number of SNPs in the window.

Nagamine *et al.*[45] also demonstrate that the average GWAS test statistic obtained from all SNPs in a region harbouring a QTL is substantially worse than the RH test statistic. This is

why effects that are detected by GWAS are also detected by RH, but additional regions are also discovered with RH. In part, this is due to QTL effects originating from single SNPs with test statistics that fell below the multiple-testing significance threshold with GWAS – with RH, this threshold is lower, because the number of regions analysed is smaller than the number of genotyped SNPs. It may also be because of the presence of regions containing several QTLs with effects that would be too small to be detected with single-SNP GWAS. RH also provides an advantage here, as it accounts for the joint heritability explained by all SNPs within each region. Chapter 7 provides a systematic comparison of RH, GWAS and linkage analysis results obtained in each cohort, and demonstrates that RH detects all but two of the 56 GWS hits identified by GWAS but also identifies 26 additional hits that were missed by GWAS using genotyped SNPs only.

Figure 30 - Regional heritability results obtained from regions defined in three different ways

These three plots show the results on chromosome 4 of regional heritability mapping performed in Vis using serum uric acid as the phenotype. Chromosome positions are plotted along the x-axis in Mb, while the $-\log_{10}(p\text{-}value)$ is plotted on the Y axis. The region definitions are shown alongside each plot.



Chapter 6 Simulations

The methods used within this thesis identify novel loci that have not been reported in the literature. These could be due to real signals that are specific to these cohorts, but they could also be false positives. Additionally, these analyses miss some known QTLs, either because they are underpowered to detect the QTL effect, or because the QTL does not segregate in these populations.

To shed light on these issues, GWAS, linkage and RH analyses can be performed on simulated traits, where the true underlying QTL is known and its effect size can be specified prior to analysis. This enables the assessment of true positive as well as false positive rates and also enables the comparison of the power to detect QTLs with these methods. Traits over a range of heritabilities were generated using combinations of a stronger genetic effect originating from a "sentinel" SNP, a polygenic background consisting of 1000 SNPs that individually have a small effect on the simulated trait as well as residual environmental noise.

While the canonical threshold for classical linkage analyses is a LOD score of 3.3 [74], there is no well-established threshold for variance component linkage methods. These simulations will also serve to establish a genome-wide significance threshold for variance component linkage analysis.

6.1 Methods

6.1.1 Phenotype Simulations

The program DISSECT was used to simulate quantitative traits based on genotype data in Orkney. DISSECT simulates quantitative traits based on the additive effects of an arbitrary number of causal loci: $y_i = g_i + e_i$ where y_i is the phenotype for individual *i* and consists of the genetic effect $g_i = \sum_{k=1}^{N} w_{ik} u_k$ and environmental noise e_i. e_i is a random normally distributed variable with a mean of 0 and a variance determined by the total trait heritability and the variance of g_i. The absolute genetic effect of each SNP, w_{ik}, is calculated using $\frac{(s_{ik}-2p_k)}{\sqrt{2p_k(1-p_k)}}$, where s_{ik} is the number of copies of the reference allele at SNP *k* in individual *i* and p_k is the frequency of this allele. This genetic effect is multiplied by u_k, which is the weight (effect size) assigned to that SNP. This effect size can be specified separately for each SNP in the data and can be set to 0, in which case that SNP does not contribute to the simulated phenotype.

6.1.1.1 Initial Models

The initial phenotype generation process is outlined in Figure 31, as described below:

39 sentinel SNPs were selected to have an effect on the phenotype, emphasizing SNPs with minor allele frequencies < 5%. A summary of these SNPs can be found in Table 19. When presenting summaries, these SNPs will be broken down into three groups – the low MAF group (MAF < 5%), the medium MAF group (10% > MAF > 5%) and the high MAF group (MAF > 10%).

For each simulated phenotype, one of these SNPs was given an effect size that explained either 4, 10 or 40% of the trait heritability. Environmental noise was added by varying the trait heritability, so traits with heritabilities ranging from 10% to 100% were generated.

A polygenic background was generated using 1000 randomly selected SNPs (the same 1000 SNPs were used in all simulations), each of these SNPs was assigned the same effect size, which was kept constant throughout all simulations. Individual components of the polygenic variance never had an effect greater than the sentinel SNP, with any one polygene explaining at most 0.096% of the trait heritability.

Figure 31 - Phenotype simulation process

Each phenotype is composed of the effects of three components – the sentinel SNP effect, the effects of 1000 polygenes and environmental noise. The effect size of the sentinel SNP was varied so it explained 40, 10 or 4% of the heritability. Maintaining this sentinel SNP effect:polygene effect ratio, environmental noise was added to vary the trait heritability between 10 and 100%.

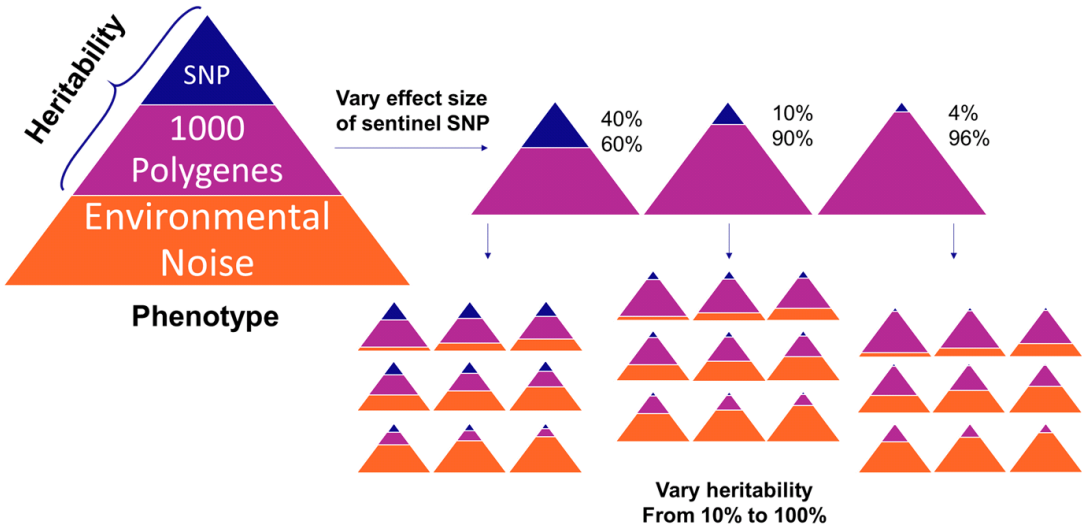


Table 19 - Summary of sentinel SNPs used in the simulation

The minor allele frequencies (MAF) are shown for each SNP and the table is sorted by MAF. On the first set of 4 columns, SNPs with MAF < 5% are shown, while on the second set of 4 columns, SNPs with MAF > 5% are shown. The SNPs under the thick line in this set of columns have MAF > 10%. The columns labelled nAlt refer to the number of alternate alleles present in Orkney, while FamAlt indicates the highest number of alternate alleles carried in one family (here, family is defined by the social pedigree).

rsID	MAF	nAlt	FamAlt	rsID	MAF	nAlt	FamAlt
Low MAF				**Medium MAF**			
rs4977114	0.016	66	5	rs6750185	0.050	197	8
rs9516949	0.018	71	11	rs7307400	0.054	211	8
rs7914943	0.019	76	12	rs1341598	0.063	240	12
rs2200674	0.020	77	6	rs12718123	0.067	259	13
rs3784635	0.023	77	15	rs739999	0.071	270	10
rs16865292	0.026	100	18	rs10494476	0.074	267	9
rs4568351	0.026	101	6	rs867191	0.087	350	18
rs17670378	0.027	105	8	rs2958431	0.089	345	13
rs1673443	0.031	117	9	rs13029379	0.090	327	16
rs6030760	0.038	146	4	rs3113176	0.092	363	14
rs12610125	0.039	151	11	**High MAF**			
rs10144225	0.041	162	8	rs10938462	0.110	432	11
rs4129315	0.043	170	5	rs6985783	0.116	451	16
rs5749011	0.044	174	5	rs1014290	0.221	864	18
rs12636173	0.045	177	5	rs2665739	0.273	1062	27
rs4393596	0.045	178	5	rs6481838	0.349	1357	35
rs1405040	0.047	184	5	rs657152	0.361	1408	30
rs9659165	0.047	185	10	rs3764261	0.380	1397	34
rs10499266	0.049	190	5	rs2291334	0.388	1531	30
				rs1360738	0.404	1587	28
				rs7204262	0.473	1854	37

6.1.1.2 Follow-up Modelling

Following the results of the initial simulation analyses, a smaller follow-up was conducted to assess the gain in power obtained by using pedigree-free linkage analysis, using phenotypes that resemble the observed traits more closely. The reason the full-scale analysis described above was not conducted here was due to the long time required to perform pedigree-free linkage analysis. For this follow-up, the same overall phenotype simulation process was used, with the following modifications: The trait heritability was set to either 0.4 (the mean

heritability of traits analysed within this thesis is 0.37) or 0.7 (to represent a trait with high heritability). The sentinel SNP's effect was set such that it explained either 4 or 10% of the overall trait variance, as shown in Table 20. Varying the trait heritability but keeping the proportion of total trait variance that the sentinel SNP explains constant means we can gain some insight into how this affects the analysis.

Table 20 - Trait heritabilities and sentinel SNP effects in the follow-up simulation

Trait Heritability	Proportion of heritability due to sentinel SNP	Proportion of total trait variance due to sentinel SNP
0.4	0.1	0.04
0.7	0.057	0.04
0.4	0.25	0.1
0.7	0.14	0.1

6.1.2 Analysis Process

The first simulated phenotypes were analysed using linkage analysis (using a pedigree, and IBD sharing as output by Loki) and GWAS, as described previously. The phenotypes in the follow-up simulation were additionally analysed by regional heritability (RH) and pedigree-free linkage analysis.

One aim of the simulation study was to see whether the sentinel SNP could be detected if it is not directly genotyped (to mimic the common situation where incomplete genotyping information is available). In the case of GWAS, all SNPs are analysed independently of each other, so after obtaining its association *p-value*, the sentinel SNP was removed from the summary statistics prior to downstream analysis of the results. In the case of RH, the sentinel SNP was removed from the genotype file prior to the analysis of the region it belonged to. As linkage analysis uses IBD sharing between pairs of individuals instead of genotype data, and because IBD segments extend over several SNPs, the sentinel SNPs were not removed when performing linkage analyses.

6.2 Results

6.2.1 Initial Model Analysis

After performing linkage analysis and GWAS, I extracted the loci with the highest test statistic (LOD score or $-\log_{10}(p\text{-}value)$) from the chromosome that harbours the sentinel SNP. With GWAS, the test statistic of the sentinel SNP exceeded the genome-wide significance threshold

50% of the time when this SNP explained 1.5% of the trait variance, and always exceeded the significance threshold as long as this SNP explained at least 3% of the trait variance. This is an indicator of the power of GWAS to detect a QTL effect in Orkney. After ascertaining this, the sentinel SNP was removed prior to downstream analysis to assess whether other SNPs were able to 'tag' this signal. I also extracted the loci with the highest test statistic from the rest of the genome in order to assess the prevalence and magnitude of false positives.

Figure 32 shows whether the highest test statistic on the target chromosome exceeded the canonical genome-wide significance threshold (LOD 3.3 for linkage analysis, and $-\log_{10}(p\text{-}value)=7.3$ for GWAS) in each of the 39 independent loci tested under each variation of the genetic architecture. This figure shows that in Orkney, pedigree-based linkage analysis is only reliably able to detect the signal originating from the sentinel SNP if this explains at least 30% of the trait variance, but detection is not dependent on the allele frequency of the sentinel SNP. In contrast, GWAS can detect the signal originating from a sentinel SNP that explains at least 2% of the trait variance, even when the sentinel SNP itself is removed, as long as the sentinel SNP is common (MAF > 20%). With a few exceptions, when the sentinel SNP is rare (MAF < 5%), it needs to explain at least 12% of the trait variance for 'tagging' SNPs to yield GWS signals with GWAS.

Table 21 and Table 22 show the summaries for the pedigree-based linkage analysis and GWAS, respectively. Within each set of 39 simulations, I have averaged the highest test statistic on the target chromosome, the highest test statistic in the rest of the genome as well as the number of non-target chromosomes that had test statistics exceeding the canonical genome-wide significance threshold. I have also calculated these averages within each SNP allele frequency class.

In Table 21, the summary for pedigree-based linkage reinforces what could be observed in Figure 32 – that detecting the QTL with this method is largely independent of the allele frequency of the sentinel SNP, and it also shows that the magnitude of the LOD score depends on the proportion of trait variance explained by the sentinel SNP. Because the LOD scores shown in this table are averaged across all 39 sentinel SNPs, it appears that pedigree-based linkage analysis is well-powered to detect a signal as long as these explain at least 24% of the trait variance, even though when studied individually (as is shown in Figure 32), it can be seen that over half of the simulated QTLs do not yield signals exceeding a LOD of 3.3 when the sentinel SNP explains 24% of the trait variance. This apparent discrepancy is due to stronger signals originating from a few simulated traits rather than all simulated traits having a consistently detectable signal. Table 21 also shows that the occurrence and magnitude of false

positive hits with pedigree-based linkage analysis is positively correlated with the variance explained by the sentinel SNP. Allele frequency does not in general appear to be a factor in the prevalence of false positives, except in the most extreme example where the sentinel SNP explains 40% of the trait variance in a trait that is 100% heritable, where rarer sentinel SNPs tend to lead to more false positives.

In Table 22, in addition to the magnitude of the association signal increasing as the proportion of trait variance explained by the sentinel SNP increases, a clear positive correlation can also be seen between the magnitudes of the association signals and the MAF of the sentinel SNP. As stated previously, the summaries presented in this table were generated after removing the sentinel SNP, which yielded GWS signals in all analyses as long as it explained at least 2% of the trait variance. Variants that tag the sentinel SNP yield GWS signals when the sentinel SNP is rare (MAF < 5%) and explains at least 8% of the trait variance. Variants tagging sentinel SNPs with intermediate allele frequencies (5% < MAF < 10%) yield GWS signals when the sentinel SNP explains at least 6% of the trait variance, while variants tagging common sentinel SNPs (MAF > 10%) yield GWS signals when the sentinel SNP explains at least 3% of the trait variance. GWAS produce many fewer false positive signals than pedigree-based linkage analysis.

Figure 32 - Highest test statistic on the target chromosome for each simulated phenotype

The loci with the highest test statistic have been extracted from the chromosome harbouring the sentinel SNP. In the case of GWAS, the sentinel SNP was removed prior to this step. If the highest test statistic on the target chromosome exceeded the canonical significance threshold (LOD > 3.3 for linkage analysis, $-\log_{10}(p\text{-}value) > 7.3$ for GWAS), its box is coloured green. The Y axis is sorted by the allele frequency of the lead SNP and the thicker horizontal bars indicate the limits for low (MAF < 5%), medium (10% > MAF > 5%) and high (MAF > 10%) minor allele frequencies. The top X axis shows the trait heritability while the bottom X axis shows the proportion of variance explained by the lead SNP.

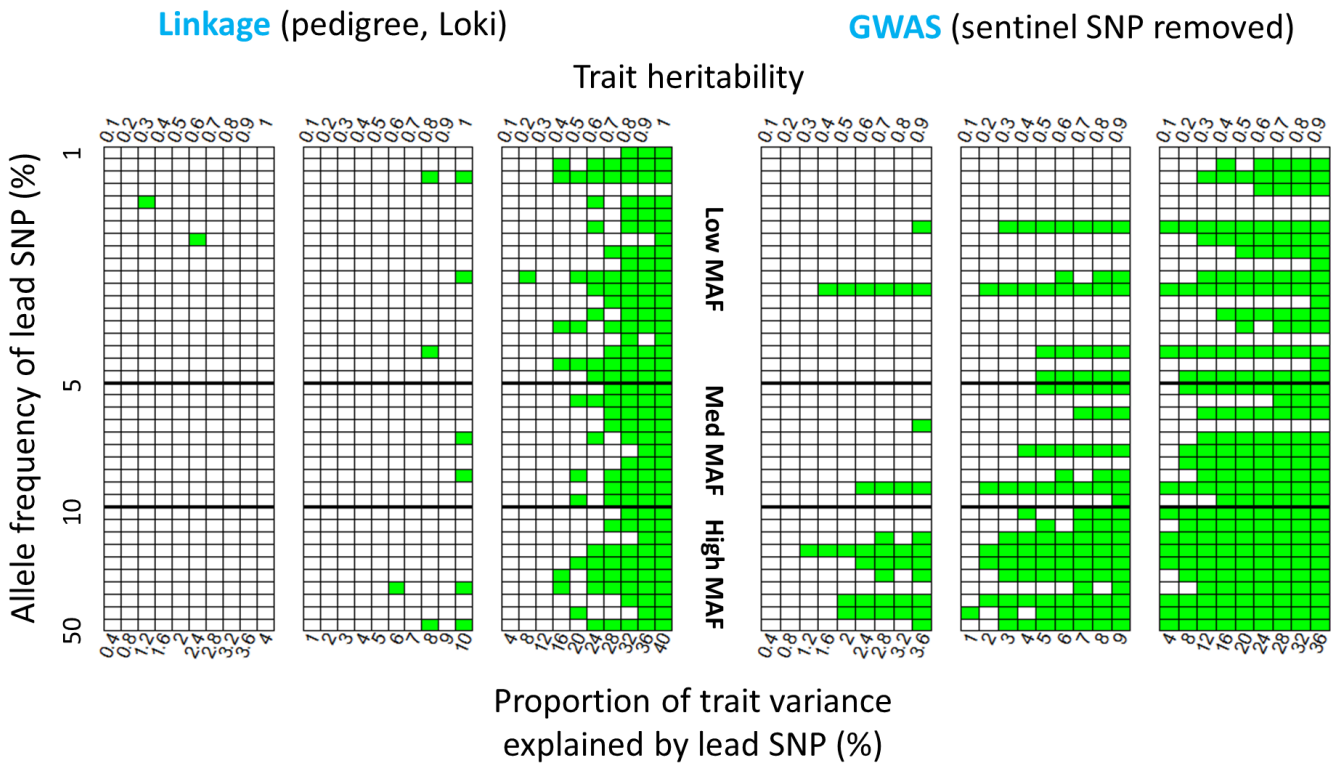


Table 21 - Summary of pedigree-based linkage analysis of simulated phenotypes

The results of 39 simulated traits have been averaged and broken down by sentinel SNP allele frequency into Low, Medium and High MAF groups. The maximum LOD score on the chromosome containing the lead causal SNP (target chromosome, A) or the maximum LOD score in the rest of the genome (false positives, B) is shown. The average number of non-target chromosomes that had LOD scores exceeding 3.3 is also shown (C). In each case, more intense shading denotes a higher number, and values in bold and surrounded by a frame are LOD scores exceeding 3.3 (A and B) or non-zero values in the case of C.

Heritability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Trait variance explained by lead SNP (%)	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	4	1	2	3	4	5	6	7	8	9	10	4	8	12	16	20	24	28	32	36	40
	A	Maximum LOD score on target chromosome (mean)																												
All MAF	0.5	0.8	1.1	1.1	0.9	1.2	0.9	1.3	1.4	1.4	0.4	0.9	0.8	1.2	0.9	1.1	1.2	1.4	1.4	2.0	0.8	1.1	1.3	1.9	2.4	**3.3**	**4.4**	**6.2**	**8.0**	**11.9**
Low MAF	0.5	0.9	1.0	0.9	0.9	1.4	0.8	1.1	1.2	1.3	0.4	1.1	0.7	1.0	0.8	1.1	1.1	1.4	1.2	1.9	0.7	1.1	1.2	2.1	2.4	**3.8**	**4.6**	**6.1**	**7.5**	**11.5**
Medium MAF	0.4	1.1	1.1	1.3	0.9	1.1	0.8	1.8	1.4	1.5	0.4	0.8	0.6	1.2	1.3	1.1	1.3	1.4	1.4	2.4	0.8	1.2	1.3	1.5	2.4	2.1	**4.3**	**6.2**	**8.5**	**12.5**
High MAF	0.7	0.4	1.3	1.1	0.9	1.2	1.3	1.1	1.7	1.5	0.5	0.7	1.2	1.5	0.8	1.2	1.1	1.3	1.7	2.1	0.8	0.8	1.5	1.9	2.5	**3.8**	**4.2**	**6.5**	**8.5**	**12.1**
	B	Maximum LOD score in rest of genome (mean)																												
All MAF	1.2	1.9	2.2	2.1	2.3	2.5	2.2	2.7	2.8	2.4	1.1	2.1	2.3	2.2	2.1	2.2	2.4	2.4	2.8	2.4	1.6	2.0	2.0	2.2	2.2	2.4	2.5	2.7	2.8	**3.3**
Low MAF	1.2	1.9	2.4	2.0	2.5	2.4	2.2	2.6	2.9	2.4	1.3	2.0	2.5	2.1	2.3	2.3	2.3	2.4	2.9	2.4	1.6	1.9	2.0	2.3	2.2	2.4	2.6	2.7	3.2	**3.8**
Medium MAF	1.3	1.9	2.1	2.4	2.0	2.5	2.3	3.0	2.8	2.4	0.9	2.4	2.2	2.2	1.8	1.9	2.3	2.3	2.8	2.3	1.7	1.8	2.0	1.9	2.3	2.6	2.3	2.7	2.4	3.0
High MAF	1.3	1.9	1.9	2.2	2.2	2.5	2.1	2.6	2.8	2.4	1.0	2.0	2.1	2.4	2.0	2.5	2.7	2.3	2.8	2.5	1.6	2.3	2.2	2.3	2.1	2.2	2.4	2.5	2.7	2.8
	C	Number of non-target chromosomes with GWS hits (mean)																												
All MAF	0.0	0.0	**0.1**	**0.2**	**0.2**	**0.2**	0.0	**0.4**	**0.6**	0.0	0.0	**0.1**	**0.2**	**0.1**	**0.2**	**0.1**	**0.2**	**0.2**	**0.5**	**0.1**	0.0	**0.1**	**0.1**	**0.1**	**0.1**	**0.2**	**0.2**	**0.3**	**0.4**	**0.8**
Low MAF	0.0	0.0	**0.2**	**0.1**	**0.3**	**0.2**	0.0	**0.3**	**0.6**	0.0	0.0	**0.1**	**0.3**	**0.1**	**0.3**	**0.1**	**0.1**	**0.3**	**0.6**	**0.1**	0.0	**0.1**	**0.1**	**0.1**	**0.1**	**0.1**	**0.2**	**0.3**	**0.6**	**1.3**
Medium MAF	0.0	0.0	0.0	**0.3**	0.0	**0.2**	0.0	**0.7**	**0.5**	0.0	0.0	**0.2**	**0.1**	0.0	0.0	0.0	**0.2**	0.0	**0.5**	0.0	0.0	0.0	**0.1**	**0.1**	**0.2**	**0.3**	**0.2**	**0.5**	**0.3**	**0.6**
High MAF	0.0	0.0	0.0	**0.1**	0.0	**0.1**	0.0	**0.3**	**0.6**	0.0	0.0	0.0	0.0	**0.1**	**0.1**	**0.2**	**0.4**	0.0	**0.5**	**0.1**	0.0	**0.1**	**0.1**	**0.1**	0.0	**0.2**	**0.1**	**0.1**	**0.2**	**0.3**

Table 22 - GWAS simulation summaries

The results of 39 simulated traits have been averaged and broken down by sentinel SNP allele frequency into Low, Medium and High MAF groups. The maximum -$\log_{10}$(*p-value*) on the chromosome containing the lead causal SNP (target chromosome, A) or the maximum -$\log_{10}$(*p-value*) in the rest of the genome (false positives, B) is shown. The average number of non-target chromosomes that had -$\log_{10}$(*p-values*) exceeding 7.3 is also shown (C). In each case, more intense shading denotes a higher number, and values in bold and surrounded by a frame are -$\log_{10}$(*p-values*) exceeding 7.3 (A and B) or non-zero values in the case of C. Note that in the case of table A, the sentinel SNP was removed prior to generating these summaries.

Heritability	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
Trait variance explained by lead SNP (%)	0.4	0.8	1.2	1.6	2	2.4	2.8	3.2	3.6	1	2	3	4	5	6	7	8	9	4	8	12	16	20	24	28	32	36
	A									Maximum -log10(pvalue) on target chromosome (mean)																	
All MAF	3.9	4.2	4.1	4.6	4.8	5.2	5.6	5.6	6.3	4.3	4.9	5.8	6.6	**7.6**	**8.4**	**9.7**	**10.8**	**11.8**	6.9	**10.3**	**14.2**	**18.6**	**22.3**	**27.2**	**31.3**	**35.7**	**42.0**
Low MAF	3.8	4.2	3.9	4.4	4.3	4.9	4.6	5.0	5.2	4.1	4.3	5.0	5.2	6.3	6.6	7.0	**7.7**	**8.2**	5.6	**7.7**	**10.0**	**11.9**	**14.4**	**17.3**	**20.4**	**21.8**	**25.2**
Medium MAF	4.0	4.3	4.0	4.2	4.3	5.0	4.9	4.9	6.1	4.6	4.4	4.9	6.3	6.7	**7.4**	**8.5**	**10.1**	**10.7**	5.4	**8.2**	**12.2**	**16.8**	**20.2**	**25.1**	**28.8**	**33.5**	**39.8**
High MAF	4.0	3.9	4.7	5.2	6.1	6.1	**8.1**	**7.3**	**8.8**	4.5	6.6	**8.2**	**9.5**	**10.9**	**13.0**	**15.9**	**17.3**	**19.8**	**10.8**	**17.5**	**24.0**	**32.9**	**39.5**	**48.1**	**54.6**	**64.5**	**76.2**
	B									Maximum -log10(pvalue) in rest of genome (mean)																	
All MAF	5.2	5.2	5.6	5.0	5.6	5.8	5.6	5.1	5.7	5.5	5.6	5.5	5.3	6.0	5.5	5.3	5.5	5.6	5.3	5.2	5.4	5.4	5.4	5.3	5.5	5.5	5.4
Low MAF	5.2	5.1	5.5	5.1	5.6	6.2	5.8	5.1	6.0	5.6	5.7	5.6	5.3	6.3	5.3	5.1	5.6	5.7	5.4	5.3	5.1	5.4	5.5	5.4	5.4	5.5	5.3
Medium MAF	5.4	5.4	5.2	5.0	5.5	5.2	5.4	5.1	5.5	5.2	5.6	5.4	5.3	5.8	6.0	5.4	5.4	5.7	5.2	5.1	5.5	5.5	5.2	5.2	5.7	5.7	5.5
High MAF	4.9	5.3	6.2	4.8	5.7	5.4	5.3	5.1	5.4	5.5	5.2	5.4	5.3	5.8	5.6	5.7	5.6	5.2	5.1	5.0	5.8	5.2	5.5	5.2	5.4	5.3	5.7
	C									Number of non-target chromosomes with GWS hits (mean)																	
All MAF	0.0	0.0	**0.3**	0.0	0.0	**0.2**	**0.1**	0.0	**0.1**	**0.1**	**0.0**	0.0	**0.3**	0.0	**0.1**	**0.1**	**0.1**	**0.1**	0.0	**0.0**	**0.1**	**0.0**	**0.1**	**0.1**	**0.0**	**0.1**	**0.1**
Low MAF	0.0	0.0	**0.2**	0.0	0.0	**0.4**	**0.2**	0.0	**0.3**	**0.2**	**0.1**	0.0	**0.5**	0.0	0.0	0.0	**0.1**	**0.1**	0.0	**0.1**	0.0	**0.1**	**0.1**	**0.1**	0.0	**0.1**	0.0
Medium MAF	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	**0.1**	0.0	**0.2**	0.0	0.0	0.0	0.0	0.0	**0.2**	0.0	0.0	0.0	**0.1**	**0.1**	**0.2**
High MAF	0.0	0.0	**0.6**	0.0	0.0	**0.1**	0.0	0.0	0.0	0.0	0.0	0.0	**0.1**	0.0	0.0	**0.2**	0.0	0.0	0.0	0.0	0.0	0.0	**0.1**	0.0	0.0	0.0	**0.2**

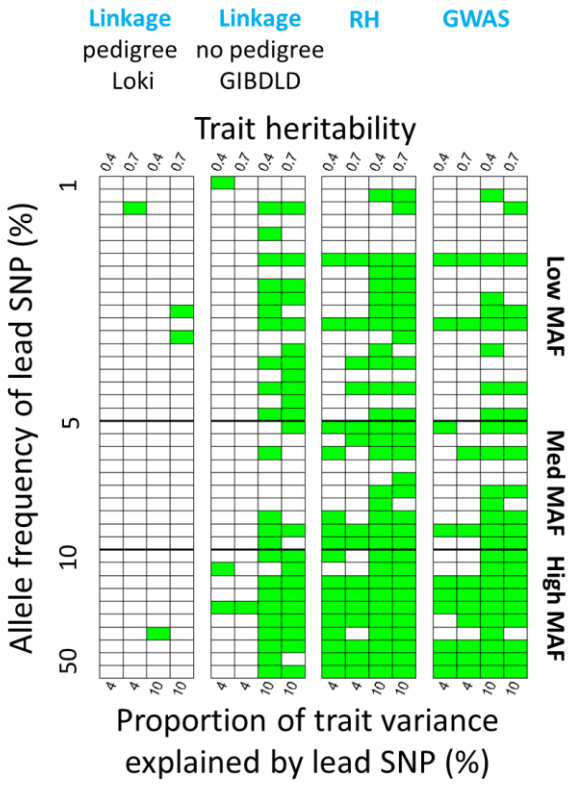
6.2.2 Follow-up Analysis

I performed linkage analysis (using IBD coefficient calculated by Loki in the case of pedigree-based linkage analysis, or IBD coefficients calculated by GIBDLD in the case of pedigree-free linkage analysis), regional heritability (RH) analysis and GWAS on the follow-up simulated phenotypes. As described in section 6.2.1, I extracted the highest test statistic (LOD score or $-\log_{10}(p\text{-}value)$) from the chromosome that harbours the sentinel SNP. I also extracted the highest test statistic from the rest of the genome in order to assess the prevalence of false positives.

Figure 33 shows whether the highest test statistic on the target chromosome in a simulation exceeded the canonical genome-wide significance threshold (LOD 3.3 for both linkage analyses, $-\log_{10}(p\text{-}value)$=7.3 for GWAS, $-\log_{10}(p\text{-}value)$=5.38 for RH).

Figure 33 - Highest test statistic on the target chromosome for each simulated follow-up phenotype

If the highest test statistic on the target chromosome exceeded the canonical significance threshold (LOD > 3.3 for linkage analysis, $-\log_{10}(p\text{-}value)$ > 7.3 for GWAS, $-\log_{10}(p\text{-}value)$ > 5.3 for RH), its box is coloured green. The Y axis is sorted by the allele frequency of the lead SNP and the thicker horizontal bars indicate the limits for low (MAF < 5%), medium (10% > MAF > 5%) and high (MAF > 10%) minor allele frequencies. The top X axis shows the trait heritability while the bottom X axis shows the proportion of variance explained by the lead SNP. RH, regional heritability.



From this figure, it can be seen that pedigree-based linkage is the poorest performer as it is unable to detect most QTLs. This is not surprising given that the sentinel SNPs here only explain up to 10% of the trait variance, and I have shown with the help of the initial models that pedigree-based linkage does not have the power to detect QTLs with this magnitude of effect in Orkney. Pedigree-free linkage shows some gain in power compared to pedigree-based linkage, as it is able to detect about half of the loci harbouring the sentinel SNP when this SNP explains 10% of the trait variance. This depends slightly on the allele frequency of the sentinel SNPs as signals originating from common (MAF > 10%) sentinel SNPs are more reliably detected. When the sentinel SNP explains only 4% of the trait variance, neither linkage method is able to detect the loci harbouring these SNPs. GWAS perform worse than pedigree-free linkage when detecting associations with variants tagging low MAF sentinel SNPs, but performs better when detecting associations with variants tagging medium MAF sentinel SNPs. The two methods perform similarly when detecting signals originating from common sentinel SNPs that explain 10% of the trait variance, but GWAS can also detect associations with variants tagging common SNPs that explain 4% of the trait variance. RH performs best across the board, being able to most reliably pinpoint the region where a sentinel SNP was, even when this SNP was removed prior to the RH analysis. Out of the four methods compared here, RH also appears to have the highest power to detect loci harbouring sentinel SNPs that explain 4% of the trait variance. As was the case with GWAS, signal detection with RH appears to be dependent on the allele frequency of the sentinel SNP.

I summarise these results in Table 23, where within each set of 39 simulations, I have averaged the highest test statistic on the target chromosome, the highest test statistic in the rest of the genome as well as the number of non-target chromosomes that had test statistics exceeding the canonical genome-wide significance threshold. I have also calculated these averages within each SNP allele frequency class. The results presented here show similar trends to what could be gleaned from Figure 33: the allele frequency of the sentinel SNP seems to affect detection rate with every method except pedigree-based linkage. In contrast, allele frequency does not have a strong effect on the prevalence of false positives in any of the methods tested here. GWAS and RH have a lower prevalence of false positives than the two linkage methods, with pedigree-free linkage showing the overall highest prevalence of false positive hits.

Table 23 - Follow-up simulation summaries

The simulation results have been averaged by analysis type across all MAFs and also broken down by SNP allele frequency into Low, Medium and High MAF groups. The maximum test statistic (LOD score or -$\log_{10}$(*p-value*)) on the chromosome containing the lead SNP (target chromosome, A) or the maximum test statistic in the rest of the genome (false positives, B) is shown. The average number of non-target chromosomes that had test statistics exceeding their respective significance thresholds is also shown (C). In each case, more intense shading denotes a higher number, and values in bold and surrounded by a frame exceed their respective significance thresholds (A and B) or are non-0 values in the case of C (might appear as 0.0 due to rounding). Note that since the significance thresholds between methods are different, the numbers presented here are not directly comparable between different analyses.

	Linkage-Pedigree				Linkage-No Pedigree				RH				GWAS			
Heritability	0.4	0.7	0.4	0.7	0.4	0.7	0.4	0.7	0.4	0.7	0.4	0.7	0.4	0.7	0.4	0.7
Trait variance explained by lead SNP (%)	4	4	10	10	4	4	10	10	4	4	10	10	4	4	10	10
	A	Maximum test statistic on target chromosome (mean)														
All MAF	1.0	1.2	1.4	1.4	1.7	1.5	**4.0**	**4.0**	**6.1**	**6.2**	**14.6**	**16.1**	6.6	6.7	**12.8**	**12.8**
Low MAF	1.0	1.2	1.3	1.5	1.7	1.6	3.2	**3.5**	4.4	4.8	**9.5**	**11.4**	5.5	5.6	**8.5**	**9.3**
Medium MAF	1.1	1.2	1.1	1.3	1.3	1.1	3.2	3.0	**5.9**	**5.5**	**14.3**	**15.1**	5.9	6.0	**11.9**	**10.4**
High MAF	1.0	1.2	1.7	1.4	2.4	1.9	**6.4**	**6.1**	**10.0**	**10.2**	**25.8**	**27.2**	**9.8**	**10.0**	**23.2**	**23.0**
	B	Maximum test statistic in rest of genome (mean)														
All MAF	2.2	2.5	2.0	2.1	2.5	2.5	2.4	2.2	4.0	4.2	4.2	4.1	5.5	5.4	5.5	5.3
Low MAF	2.3	2.5	1.9	2.2	2.3	2.4	2.5	2.2	4.0	4.1	4.2	4.2	5.6	5.4	5.5	5.3
Medium MAF	2.4	2.7	2.1	1.8	2.7	2.6	2.4	2.2	4.0	4.2	4.3	4.1	5.5	5.4	5.5	5.3
High MAF	2.0	2.1	2.1	2.0	2.6	2.6	2.4	2.4	4.0	4.5	4.1	3.9	5.5	5.4	5.5	5.3
	C	Number of non-target chromosomes with GWS hits (means)														
All MAF	**0.0**	**0.2**	0.0	**0.1**	**0.1**	**0.1**	**0.1**	**0.0**	**0.0**	0.0	**0.0**	0.0	0.0	0.0	**0.1**	**0.0**
Low MAF	0.0	**0.2**	0.0	**0.1**	**0.1**	**0.1**	**0.1**	0.0	0.0	0.0	**0.1**	0.0	0.0	0.0	**0.1**	0.0
Medium MAF	**0.1**	**0.2**	0.0	0.0	**0.2**	0.0	**0.1**	0.0	**0.1**	0.0	0.0	0.0	0.0	0.0	0.0	0.0
High MAF	0.0	**0.1**	0.0	0.0	**0.1**	0.0	0.0	**0.1**	0.0	0.0	0.0	0.0	0.0	0.0	**0.2**	**0.1**

6.3 Discussion

The simulation scenario presented here simplifies complex trait architecture for tractability, but highlights some of the features of the methods used within this thesis.

Pedigree-based linkage analysis had very low power to detect the simulated SNP effect in the Orkney dataset used for the simulation, only being able to reliably detect the locus if the SNP explains over 30% of the total trait variance. Effect sizes of this magnitude are expected for Mendelian traits.

As long as there is a segregating allele, its allele frequency does not appear to affect the detection accuracy with linkage analysis. This makes sense because linkage analysis records

the segregation of large chunks of DNA carrying a QTL in individuals with similar trait values (linkage). In contrast, GWAS relies on linkage disequilibrium (LD) between markers such that if a causal variant is not in strong LD with any genotyped markers, it will go undetected. Indeed, these simulations show that if the actual casual variant is removed, the power of GWAS to detect it via other variants declines as the allele frequency of the causal variant diminishes. This is compounded by the fact that low allele frequency variants are poorly represented on genotyping arrays.

Compared to the pedigree-based linkage analysis, there is a large gain in power when the pedigree-free linkage analysis is used, as it is can identify the causal locus when the sentinel SNP explains 10% of the total trait variance. It is acknowledged that this is still quite a large effect size originating from a single locus, however. This gain in power compared to pedigree-based linkage originates from the ability to use individuals who share IBD the DNA segment carrying the sentinel SNP, but who are not recorded as related in the social pedigree.

There is a higher incidence of false positives with linkage analysis than with GWAS and RH. This could be because of imperfect IBD estimation, or because of excess IBD sharing at other loci across the genome, since if two individuals share segments that are IBD around the sentinel SNP, they are also likely to share other IBD segments in their genome. The former cause may be improved with the use of denser genotyping in the future, while the latter is a more intrinsic problem that might be circumvented by obtaining a null distribution of IBD sharing probabilities in a cohort.

While it is affected by allele frequency, RH appears to perform best across the board, detecting the highest number of true hits out of all the methods tested, and yielding the lowest false positive rates. This additional power over GWAS was also reported by Nagamine *et al.* [45] where RH results correlate with those obtained by GWAS but also uncover new loci, in simulated and real traits.

Due to the significant time required to prepare files for, and perform, linkage analysis, GWAS appear to be a more straightforward means to carry out genetic analyses to detect single-SNP effects. Due to its lack of reliance on LD, linkage analyses may however hold an advantage when the allele frequency of a causal SNP is low and all genotypes are not available. Since only the Orkney dataset was used in these simulations, it cannot be ruled out that linkage analysis may have more power to detect a segregating QTL in a population with a larger sample size or a different relationship structure. Additionally, in theory, linkage analysis should be able to better detect loci harbouring several independent causal variants (allelic heterogeneity), but this more complicated scenario was not implemented in this set of

simulations. The ability of RH to detect such loci has been demonstrated in a simulation study by Uemoto *et al.* [46].

6.4 Linkage Analysis Significance Threshold

While the LOD score value corresponding to a 5% significance threshold is easily determined pointwise (due to the distribution of scores relating to a chi-squared distribution), that corresponding to a 5% significance threshold genome-wide is not obvious. In a seminal paper, Lander and Kruglyak set this threshold in human genetics to 3.3 for the classical parametric linkage studies used for Mendelian traits, and to between 3.3 and 3.8 (depending on the degree of relatedness of the affected individuals compared) for allele sharing methods, the popular methods used for complex trait analysis at the time [197]. Variance components linkage analyses in the literature often use 3.3 or 3.4 (e.g. [198]) in reference to this paper, but it seems to be arbitrary and values up to 3.8 may be just as appropriate.

Lander and Kruglyak used the mathematical theory of large deviations to determine the genome-wide threshold, which allowed them to derive the number of regions that would exceed this threshold by chance based on a Poisson distribution. The mean of this Poisson distribution critically depends on the number of chromosomes, the genome length and a measure ρ of how rapidly the statistic fluctuates across the genome, which reflects the total crossing over rate between genotypes being compared (Box 1 in [197]). It is clear that this measure ρ varies depending on the degree of relatedness of the individuals compared in a pair, and is more easily calculated when only one type of relative is used to build the allele sharing statistics (as in the methods described by Lander and Kruglyak). Hence the genome-wide significant LOD threshold varies from 3.3 when grandparent-grandchild pairs are used in allele sharing methods to 3.8 when second cousins are used. This makes sense, because compared to closely related individuals, in more distant relatives, more cross-over have occurred so the regions assessed along the genome are more likely to be independent. As a consequence, there is a higher number of independent test performed, resulting in a higher multiple testing penalty. I would argue that the LOD score corresponding to the IBD sharing variance component used within this thesis follows the same principle, but it is difficult to derive a significance threshold using this calculation due to the presence of varying degrees of relatedness between, and within each population used here. For an upper limit of the score corresponding to a genome-wide significant threshold of 5%, a Bonferroni correction could be applied based on the number of regions being tested (as was done for the RH method). This is very conservative given the fact that the regions are not independent: the threshold calculated this way will be 0.05/33000 (33000 SNPs at 0.1 cM intervals are tested in the pedigree-based linkage analysis), which

corresponds to a LOD score of 4.7. Therefore simulations, for every population under study, should ideally be carried out to estimate this threshold. Another way would be to assign random phenotype values drawn from a normal distribution to each genotyped individual, and perform linkage analysis many times to obtain a distribution of the LOD score statistics.

The simulations I had performed in the Orkney study in order to evaluate the power of the different methods to detect a major QTL can be used to this end. Since the simulated phenotypes aimed to mimic a complex trait caused by one SNP of large effect and 1000 polygenes of small effect, if the chromosome containing the sentinel SNP is excluded, the rest of the genome should provide an adequate approximation for the null hypothesis of no major SNP effects. The prevalence of false positives on these chromosomes can be assessed, and this can be used to calculate an empirical significance threshold. Therefore, I obtained the most extreme LOD score from the pedigree-free linkage analysis results (4×39=156 samples) after removing the chromosome containing the sentinel SNP. I ordered these values and isolated the highest 5% of scores. The lowest among these was the LOD score of 3.41. This value is in line with the LOD significance threshold used in most linkage studies, and it is used as the genome-wide significance threshold within this thesis.

Chapter 7 Method Comparisons and Conclusions

7.1 Method Comparisons

The simulations discussed in Chapter 6 test the performance of the methods used throughout this thesis using simulated traits. The simulation only considered one type of scenario – a single segregating major causal variant and a polygenic background – while in reality, complex traits are the result of many combinations of genetic and environmental effects – with additional complexity added by dominance or epistatic effects, gene by gene interactions or gene by environment interactions. While in chapters 3-5 I applied different analytical methods to the same set of real data, the results from each kind of analysis were presented in isolation from the other methods. In this chapter, I aim to systematically compare the results obtained with GWAS, RH and linkage analysis in order to identify trends obtained by analysing real data. I also compare the relationship coefficients estimated using IBD and IBS-based methods in order to dissect the reason why loci flagged by linkage analysis are often not picked up by GWAS or RH, and vice versa. These trends can then be compared to the trends observed when simulated data were used.

7.1.1 Genetic Relatedness Calculated using IBS or IBD methods

The GRMs calculated by GCTA-based methods utilise identity-by-state between SNPs in pairs of individuals to estimate relationships. In contrast, the kinship matrices calculated by IBDLD utilise identity-by-descent. While all markers that are identical-by-descent are also identical-by-state, the opposite is not true. Here, I use Orkney as an example to demonstrate that while on the whole-genome level, IBD and IBS-based relatedness is comparable, this does not hold at the regional level.

I calculated whole-genome kinship coefficients between all pairs of individuals in Orkney using the GIBDLD (pedigree-free) method of IBDLD, as described in section 4.2.2, as well as the whole-genome GRMs using GCTA [199], as described in section 3.2.1. Figure 34 compares these relationship coefficients and shows that both IBD and IBS-based methods result in similar whole-genome relatedness values. It also reveals that GIBDLD gives slightly higher estimates of genetic kinship than GCTA.

Additionally, I calculated regional kinship coefficients and GRMs using the same 28 genotyped SNPs in a 0.3 cM region around the *SLC2A9* gene, and plotted them against each other in Figure 35. The regional kinship coefficient is the same as that which was also used in section 4.3.1.2, which was calculated by averaging the IBD sharing output at each of the 28 SNPs.

Figure 34 - IBS vs IBD-based whole-genome relatedness

Genetic relatedness between all pairs of genotyped individuals in Orkney was estimated by IBD-based methods (using GIBDLD, IBDLD's pedigree-free method, on the Y axis) and IBS-based methods (using GCTA, on the X axis). All pairs (including self-pairs, in the top right corner) in are depicted. The red line depicts x=y. Note that here, 2Φ is used (that is, twice the kinship coefficient, which can range from 0 to 2, and takes on a value of 1 in self-pairs in the absence of inbreeding).

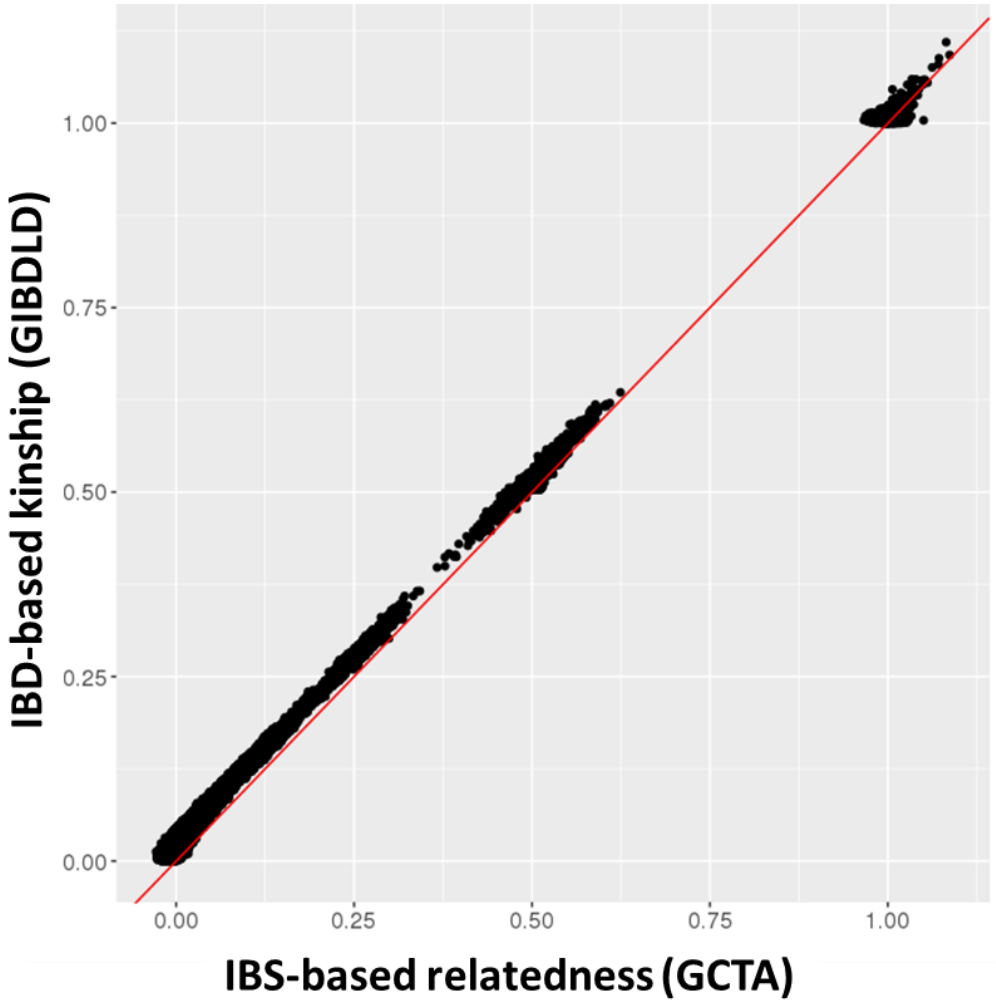
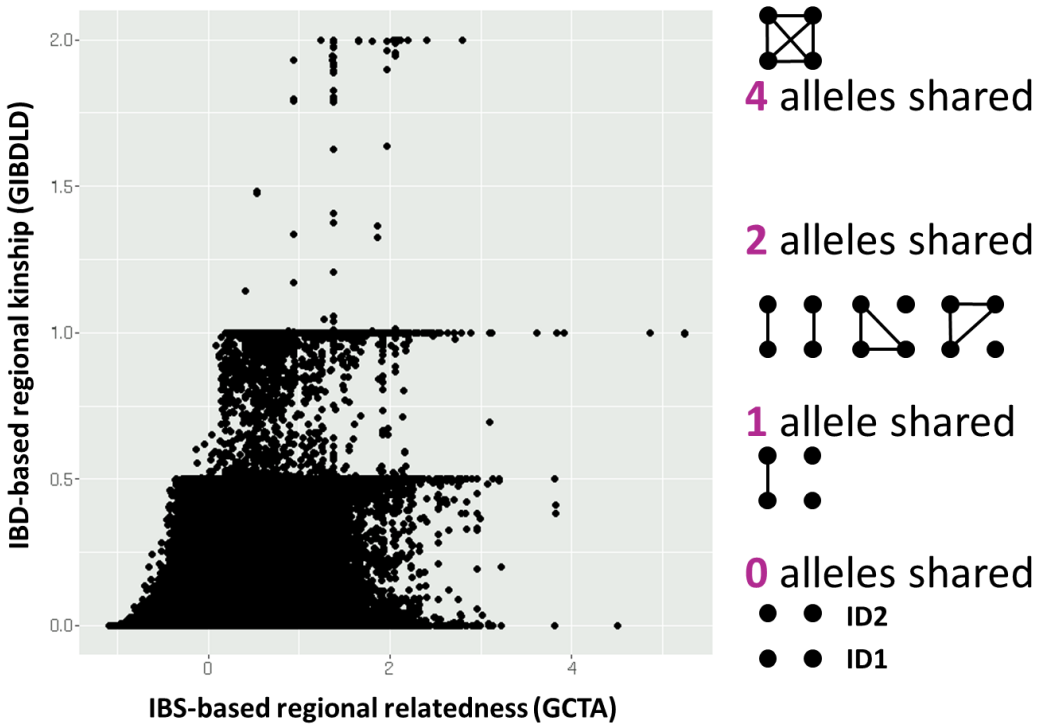


Figure 35 - IBS vs IBD-based regional relatedness at the *SLC2A9* locus

The same 28 SNPs in the 0.3 cM region around the *SLC2A9* gene were used to calculate both IBD (Y axis) and IBS-based (X axis) relatedness. For IBD-based kinship, the IBD sharing probabilities obtained at all 28 SNPs were averaged for each pair of individuals. On the right side, a diagrammatic representation shows the expected IBD values given a number of alleles shared IBD. Here, the unordered alleles of two individuals are shown as nodes and the lines connecting these nodes indicate that they are IBD.



Several things are revealed by Figure 35: First, that some pairs of individuals share all of their alleles IBD across this region. As established from Figure 16, these individuals have very low genome-wide kinship-values. While these pairs have the highest amount of possible IBD sharing, they do not have the highest regional relatedness values according to the IBS-based regional relatedness matrix. Second, while at the whole-genome level, IBD and IBS-based relatedness yields similar values, this does not hold at the regional level. The IBS-based relatedness values are not constrained to the 0-2 interval as they take on negative values and values that are higher than 2. This is a consequence of the way the kinship is calculated, as the equation described in section 3.2.1 is only dependent on the frequency of the reference allele of a SNP, and the number of these alleles carried by the two individuals being compared. Across the genome, these SNP-wise relatedness values average out, but at the regional level

they can yield very high or very low values. For example, using the allele frequencies of these 28 SNPs in Orkney (6 of which have MAFs below 10%), if two individuals both carry two copies of the reference allele at each of these SNPs, their regional relatedness value will be 11.32, while if one individual carries two copies of the reference allele while the other carries 0 copies of the reference allele at each of these SNPs, their regional relatedness value is equal to -2, indicating a high level of genetic dissimilarity in this region.

In IBS-based relatedness, calculations are done based on the allele frequency of each SNP, and LD between SNPs is often ignored. This means that each SNP is assumed to explain the same proportion of trait variance, so rarer SNPs will have larger effect sizes, and clusters of rare SNPs in high LD with each other can skew these estimates. In addition to leading to kinship coefficients that are different to those obtained with IBD-based methods, this can also cause trait heritabilities to be incorrectly estimated, and it is suggested to weight SNPs based on their allelic correlations to circumvent this [200]. The software LDAK was developed specifically to address this problem, and it weights SNPs based on local LD between SNPs before calculating trait heritabilities [201].

To summarise, when whole-genome level SNP information is used, the resulting GRM provides an unbiased estimate of the genome-wide kinship coefficient [202] that is comparable to the whole-genome kinship calculated using IBD-based methods such as IBDLD [120]. On the other hand, this method does not result in a good approximation of regional kinship coefficients as it only models identity-by-state, but not identity-by-descent. This is also apparent in the fact that regional GRM values are not constrained between the values of 0 and 1, and can even take on negative values. By definition, kinship coefficients do not exist outside of this interval – a kinship coefficient of 0 means a pair of individuals is completely unrelated and shares no regions IBD, while a kinship coefficient of 1 means that all four alleles in this pair are identical-by-descent. This difference in the regional relationship-values is likely to be the reason for the non-overlapping results obtained with RH and linkage analysis.

7.1.2 Overlap of Results Obtained with GWAS, Linkage Analysis and RH

Within this section, regional results of two methods are compared at a time, broken down by cohort. Note that the definition of 'region' differs based on which two methods are being compared (Figure 36): the pedigree-free linkage analysis (Orkney and Vis only) has the largest regions (encompassing 2.5 cM), RH regions encompass 0.3 cM, regions used in the pedigree-based linkage analysis encompass 0.1 cM, and GWAS results were obtained at each genotyped SNP. To make the results of different methods comparable, all results from the method using the coarser scale were retained and the regions defined by this method were used to subdivide the results of the method with the finer scale. The SNP or region with the highest $-\log_{10}(p\text{-}value)$ was selected from each subdivided region to represent the result of the finer scale method. It should be noted that here, only the GWAS results obtained using genotyped SNPs are used.

Table 24 presents the number of independent loci detected with each method, as well as the number of overlapping loci across methods. This table is broken down by cohort and groups methods pairwise, indicating the number of GWS loci obtained with each method, the number of regions where both methods obtained GWS results, as well as the Pearson correlation coefficient between the results reported by the two methods.

The plots presented in Figure 37 to Figure 42 depict the $-\log_{10}(p\text{-}value)$ of every analysed region, in each trait, and are a graphical representation of the results shown in Table 24. The figures presented here were chosen as exemplars to illustrate the similarities and differences between results obtained with different methods, so figures are not presented for each pair of methods in each cohort. In these figures, points can fall into one of four different areas depending on whether the region reached GWS with both methods (white background), with one method (light grey background) or with neither method (dark grey background). The GWS threshold for a given method corresponds to the inside border of these areas. Pearson's correlation coefficient between the results obtained with these methods is shown under each plot title.

Figure 37 compares GWAS and RH results in the GS study and Figure 38 zooms in on these results to better show the results with $-\log_{10}(p\text{-}values)$ of up to 15, emphasizing the hits that were GWS with RH only. Figure 39 compares GWAS and linkage analysis results in the Korčula study while Figure 40 compares pedigree-free linkage analysis and GWAS results in the Orkney study. Figure 41 compares RH results to the linkage analysis results in the Korčula study and Figure 42 compares pedigree-free linkage analysis and RH results in the Orkney study.

In general, the results of RH and GWAS agree with each other well (mean correlation 0.7), and all GWS regions identified by GWAS are also identified with RH except in Vis (the smallest cohort), where 2 out of 3 GWS GWAS hits fall below the RH GWS threshold. These two hits correspond to the well-established *SLC2A9* locus in serum uric acid levels [69], corrected ($-\log_{10}(p\text{-}value)$=5.54 in RH) or uncorrected for BMI and alcohol consumption ($-\log_{10}(p\text{-}value)$=5.99 in RH).

In contrast, RH identifies several regions that yield no GWS SNPs with GWAS performed on genotyped SNPs – 2 regions in Orkney, 3 in Shetland, 1 region in Korčula and 21 regions in GS. In most cases, this gain in power is due to the lower GWS threshold with RH – indeed, in all of these regions, the highest single-SNP $-\log_{10}(p\text{-}value)$ obtained by GWAS always exceeds 4. Therefore, there are no examples of regions where a RH signal is present in the complete absence of a GWAS signal, but these signals would not normally be flagged in the course of a standard GWAS. Additionally, RH may be better suited to capturing additional variance contributed by several causal alleles at a locus, as has been demonstrated in a simulation study [46].

The results of pedigree-based linkage analysis do not correlate well with those obtained with GWAS and RH. This is not unexpected, since here, only the effects of variants segregating within families can be detected and these effects might not generalise to the whole population. This observation also works the other way around – variants with relatively large effects on traits such as HDL, glucose or serum uric acid concentration yield strong signals with RH and GWAS but signals at these loci are generally absent in linkage results. This is likely due to these variants segregating on many haplotypes that are not necessarily IBD, so no one haplotype segregates strongly enough within families for linkage analysis to detect it.

Linkage signals also tend to be quite broad, spanning a large region, which is a consequence of large stretches of IBD sharing. This is why, in some cohorts, apparently many linkage signals are identified but these signals are not independent of each other, similar to large 'towers' observed in GWAS Manhattan plots, where many SNPs that in LD with each other and the causal variant all pick up the signal. For the purposes of the summary presented in Table 24, such signals have been collapsed into one region.

The pedigree-free linkage analysis aims to increase the power to detect a QTL with linkage analysis by looking at IBD sharing between all pairs of individuals in the data, not just pairs that are within the same family according to a social pedigree. Its results are mildly correlated with pedigree-based linkage analysis, as well as with GWAS and RH results (average correlation coefficients calculated from Orkney and Vis, 0.14, 0.20, 0.27, respectively),

indicating that it can detect effects segregating within close families and effects segregating at the cohort level (due to the increased power provided by IBD segments shared by distantly-related individuals). Despite this correlation, with the exception of the *ABO* locus, which is GWS in GWAS, RH and pedigree-free linkage of von Willebrand Factor levels in Orkney, there is no overlap between the GWS results obtained with linkage analysis (pedigree-based and pedigree-free) and either GWAS or RH.

Table 24 - Summary of GWS hits obtained with different methods

These tables are broken down by cohort and show pairwise comparison of GWS hits obtained with GWAS performed on genotyped SNPs, RH, pedigree-based linkage analysis and pedigree-free linkage analysis (NoPed, Orkney and Vis only), as well as indicating the number of overlapping regions (where both methods obtained a GWS test statistic). The number of traits that were analysed in each cohort is indicated. Pearson's correlation coefficient is shown (Corr column) and it was calculated by comparing the test statistics yielded by the two methods in corresponding regions within the same trait, across all traits and regions.

Orkney (39 traits)	GWAS hits	RH hits	Linkage hits	NoPed hits	Overlap	Corr
RH - GWAS	4	6	-	-	4	0.71
Linkage - GWAS	4	-	1	-	0	0.03
RH - Linkage	-	6	1	-	0	0.04
NoPed - GWAS	4	-	-	1	1	0.21
NoPed - Linkage	-	-	1	1	0	0.15
NoPed - RH	-	6	-	1	1	0.26

Vis (39 traits)	GWAS hits	RH hits	Linkage hits	NoPed hits	Overlap	Corr
RH - GWAS	3	1	-	-	1	0.69
Linkage - GWAS	3	-	9	-	0	0.03
RH - Linkage	-	1	9	-	0	0.04
NoPed - GWAS	3	-	-	0	0	0.18
NoPed - Linkage	-	-	9	0	0	0.13
NoPed - RH	-	1	-	0	0	0.28

Korčula (31 traits)	GWAS hits	RH hits	Linkage hits	Overlap	Corr
RH - GWAS	3	4	-	3	0.72
Linkage - GWAS	3	-	1	0	0.01
RH - Linkage	-	4	1	0	0.03

This table continues on the next page.

GS (23 traits)	GWAS hits	RH hits	Linkage hits	Overlap	Corr
RH - GWAS	38	59	-	48	0.71
Linkage - GWAS	38	-	15	0	0.02
RH - Linkage	-	59	15	0	0.02

Shetland (36 traits)	GWAS hits	RH hits	Linkage hits	Overlap	Corr
RH - GWAS	8	11	-	8	0.66
Linkage - GWAS	8	-	1	0	0.02
RH - Linkage	-	11	1	0	0.03

Figure 36 - Region definitions used when comparing results of different methods

When discussing individual methods, all results are used (first row). When methods are compared, region boundaries are defined by the method with the coarser scale (dashed lines). Within each region, only the most significant result obtained with the finer scale method is kept. Coloured blocks indicate the results that are kept in each case.

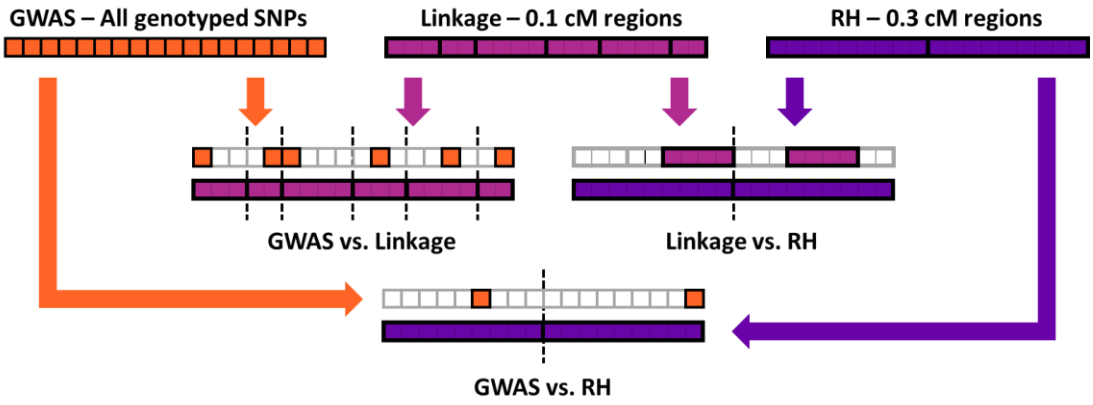


Figure 37 - GWAS vs RH in GS

For RH, the -$\log_{10}$(*p-value*) of each region is plotted. For GWAS, the SNP with the highest -$\log_{10}$(*p-value*) within each RH region is plotted. Traits are ordered based on the highest -$\log_{10}$(*p-value*) in each trait. Trait numbers are assigned to points that exceeded the GWS threshold in one (light grey background) or both analyses (white background). Points in the dark grey background did not reach GWS in either analysis.

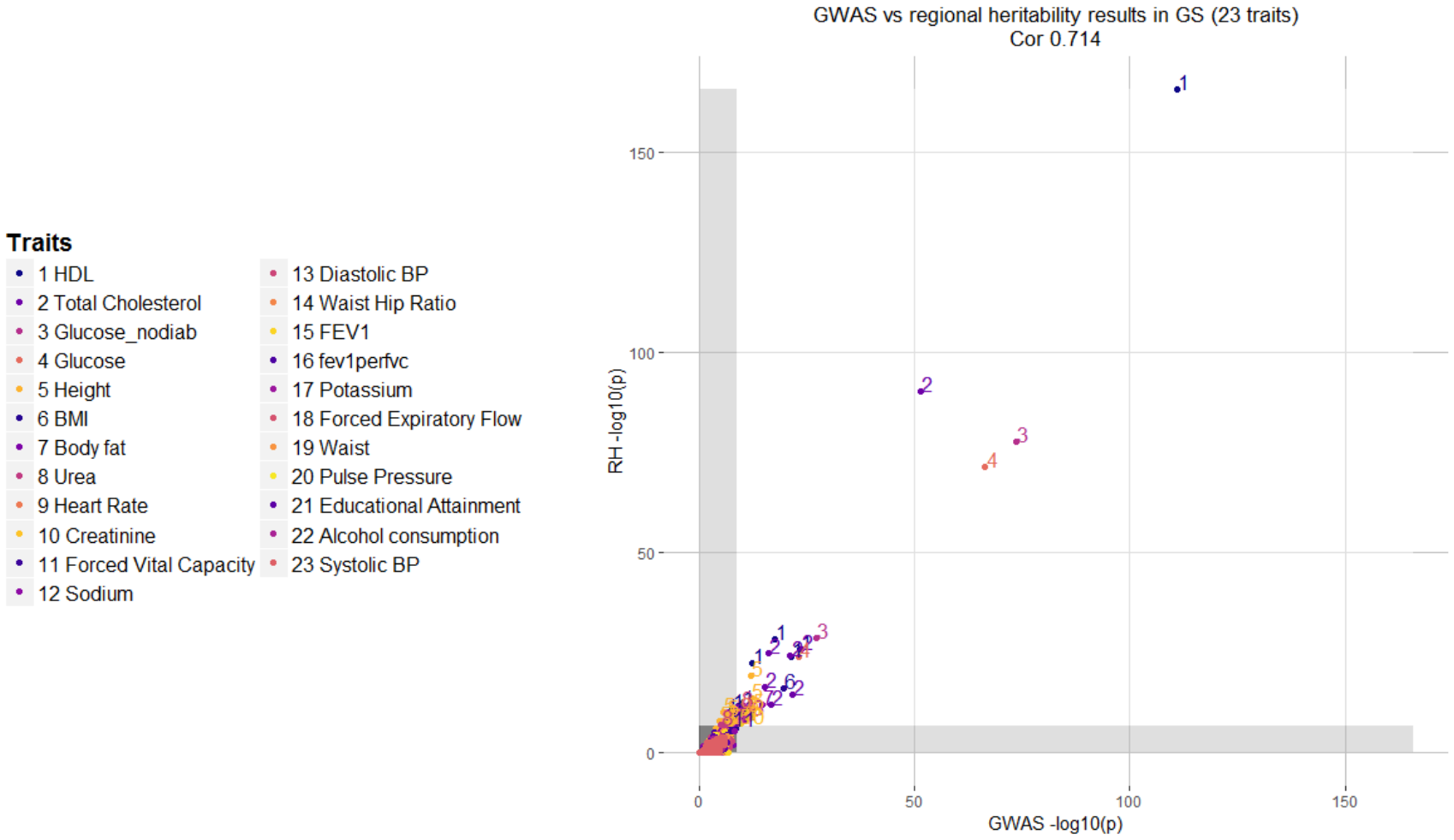


Figure 38 - GWAS vs RH results in GS, zoomed

This image zooms in on Figure 37, having removed all results with -$\log_{10}$(*p-value*) > 15. Traits are ordered based on the highest -$\log_{10}$(*p-value*) in each trait. Trait numbers are assigned to points that exceeded the GWS threshold in one (light grey background) or both analyses (white background). Points in the dark grey background did not reach GWS in either analysis.

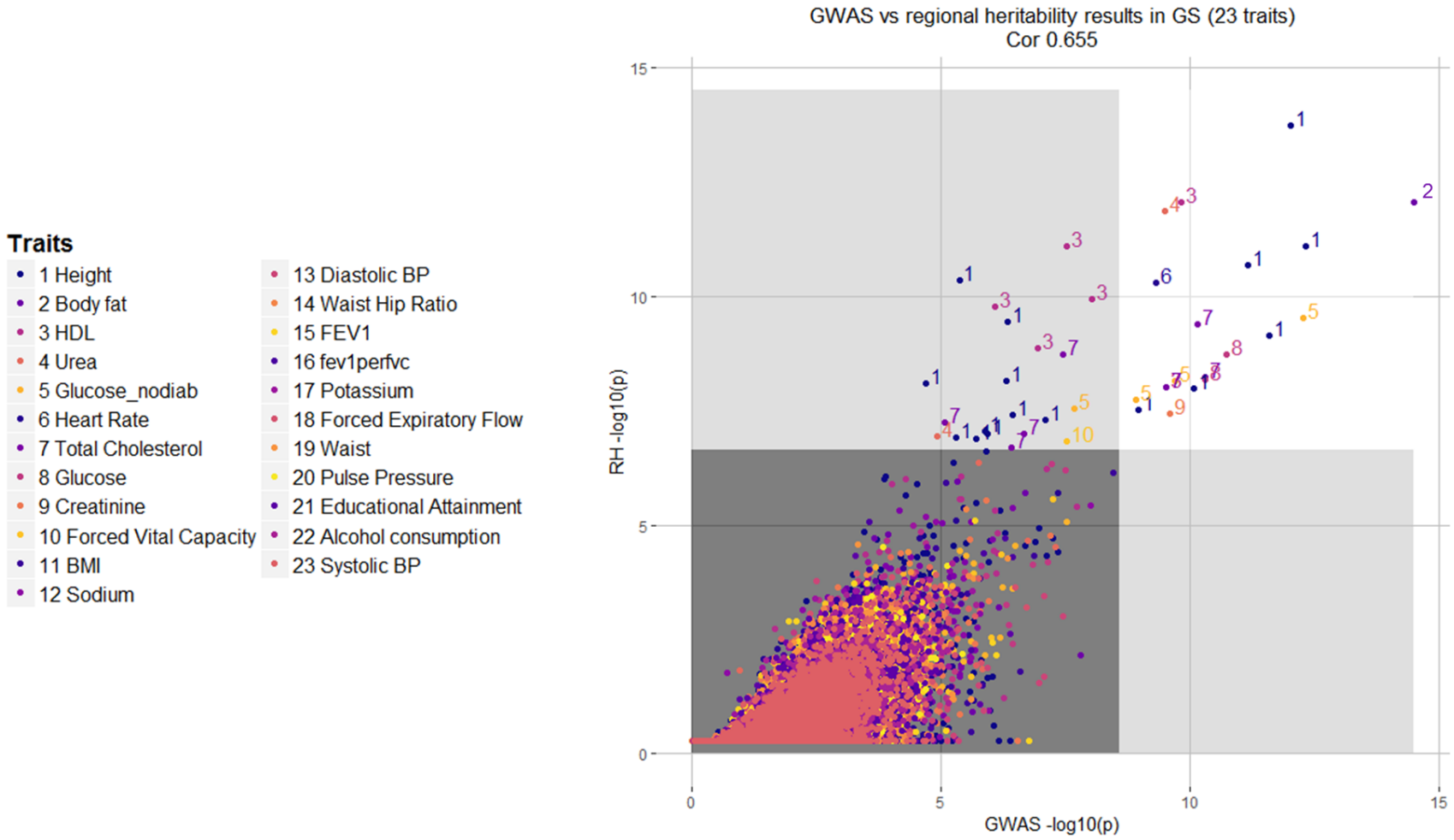


Figure 39 - GWAS vs Linkage analysis results in Korčula

For linkage analysis, the -$\log_{10}$(*p-value*) of each region is plotted. For GWAS, the SNP with the highest -$\log_{10}$(*p-value*) within each linkage region is plotted. Traits are ordered based on the highest -$\log_{10}$(*p-value*) in each trait. Trait numbers are assigned to points that exceeded the GWS threshold in one (light grey background) or both analyses (white background). Points in the dark grey background did not reach GWS in either analysis.

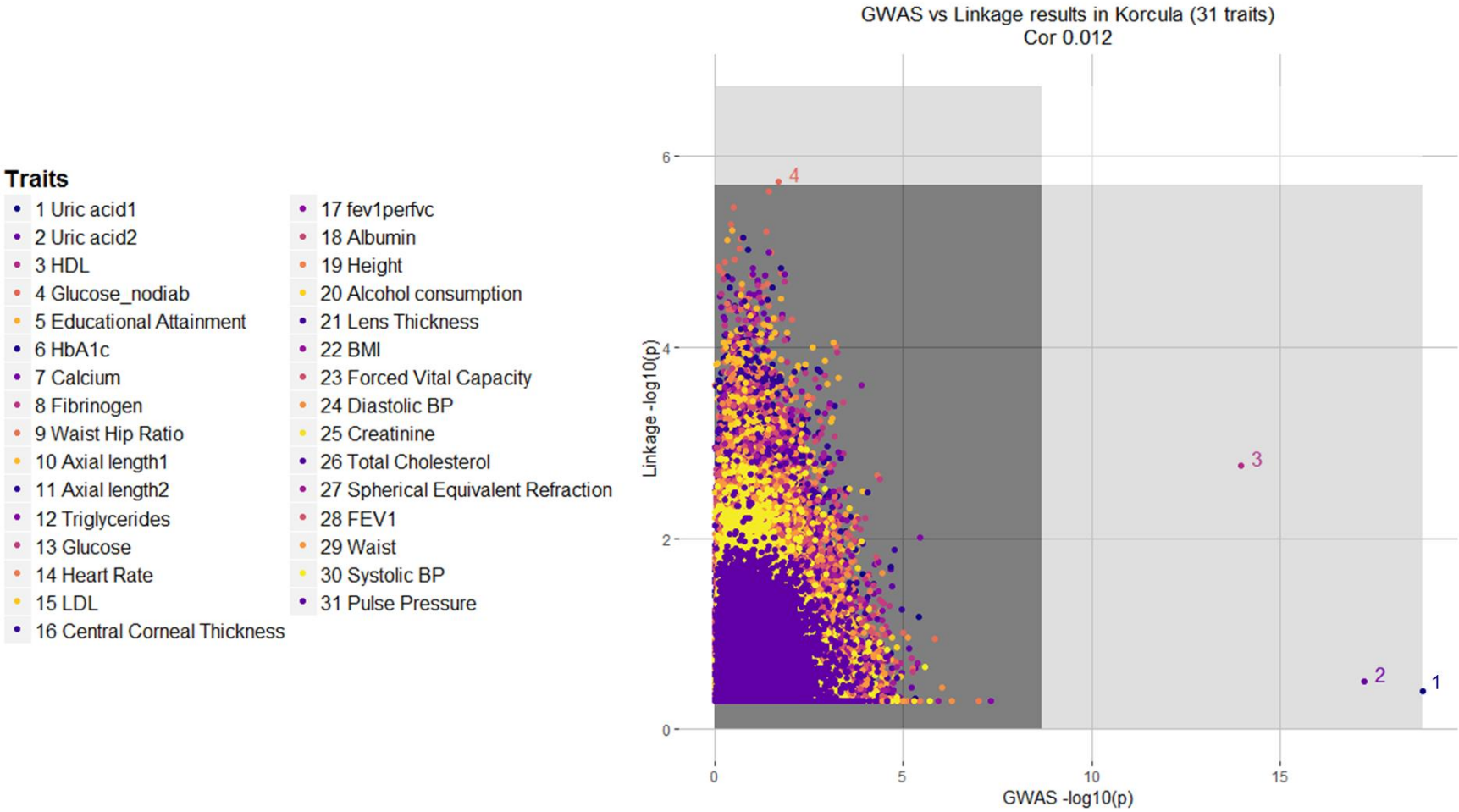


Figure 40 - GWAS vs Pedigree-free linkage analysis in Orkney

For linkage analysis, the -$\log_{10}$(*p-value*) of each region is plotted. For GWAS, the SNP with the highest -$\log_{10}$(*p-value*) within each linkage region is plotted. Traits are ordered based on the highest -$\log_{10}$(*p-value*) in each trait. Trait numbers are assigned to points that exceeded the GWS threshold in one (light grey background) or both analyses (white background). Points in the dark grey background did not reach GWS in either analysis.

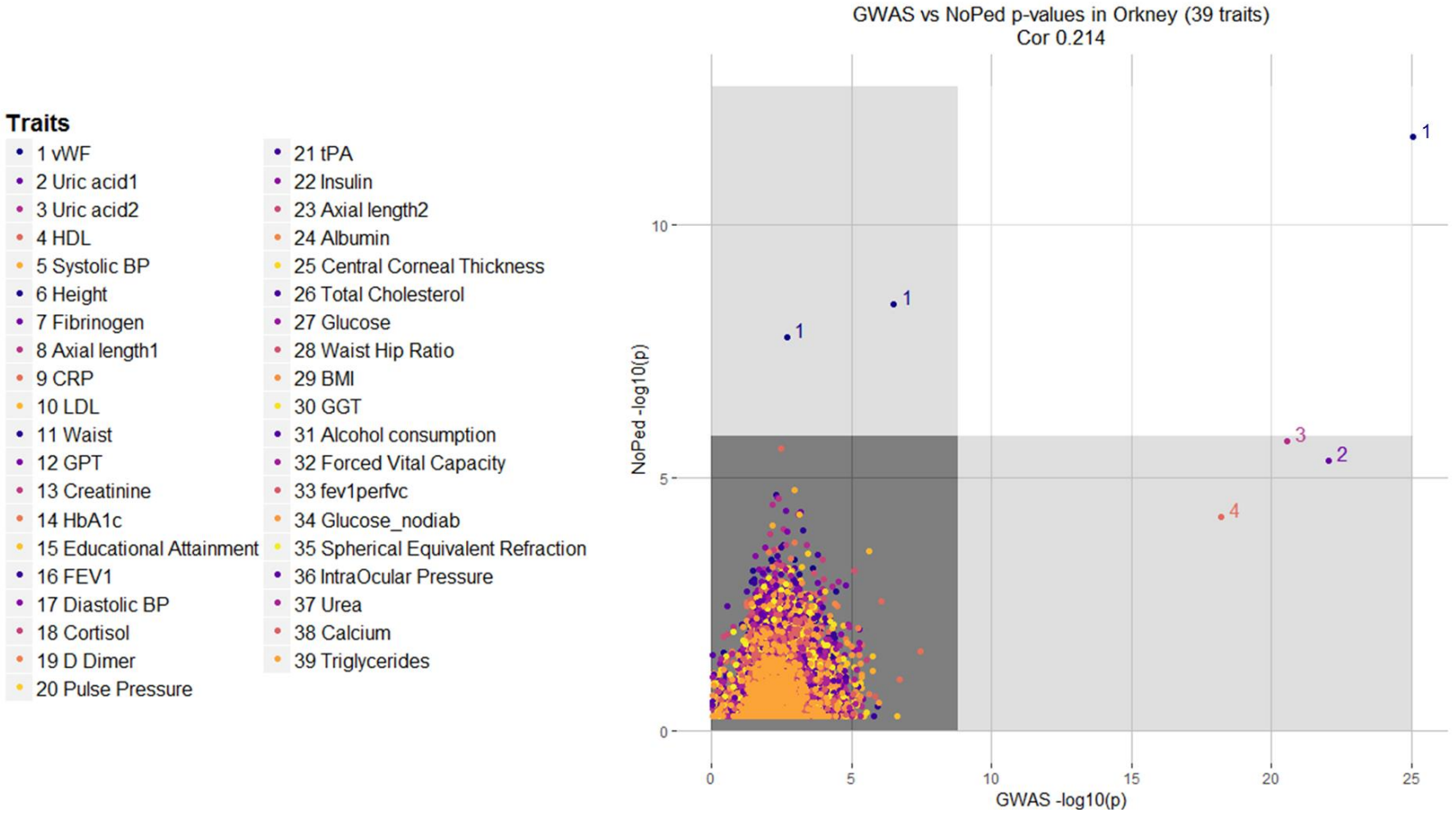


Figure 41 - RH vs Linkage analysis results in Korčula

For linkage analysis, the -log$_{10}$(*p-value*) of each region is plotted. For RH, the region with the highest -log$_{10}$(*p-value*) within each linkage region is plotted. Traits are ordered based on the highest -log$_{10}$(*p-value*) in each trait. Trait numbers are assigned to points that exceeded the GWS threshold in one (light grey background) or both analyses (white background). Points in the dark grey background did not reach GWS in either analysis.

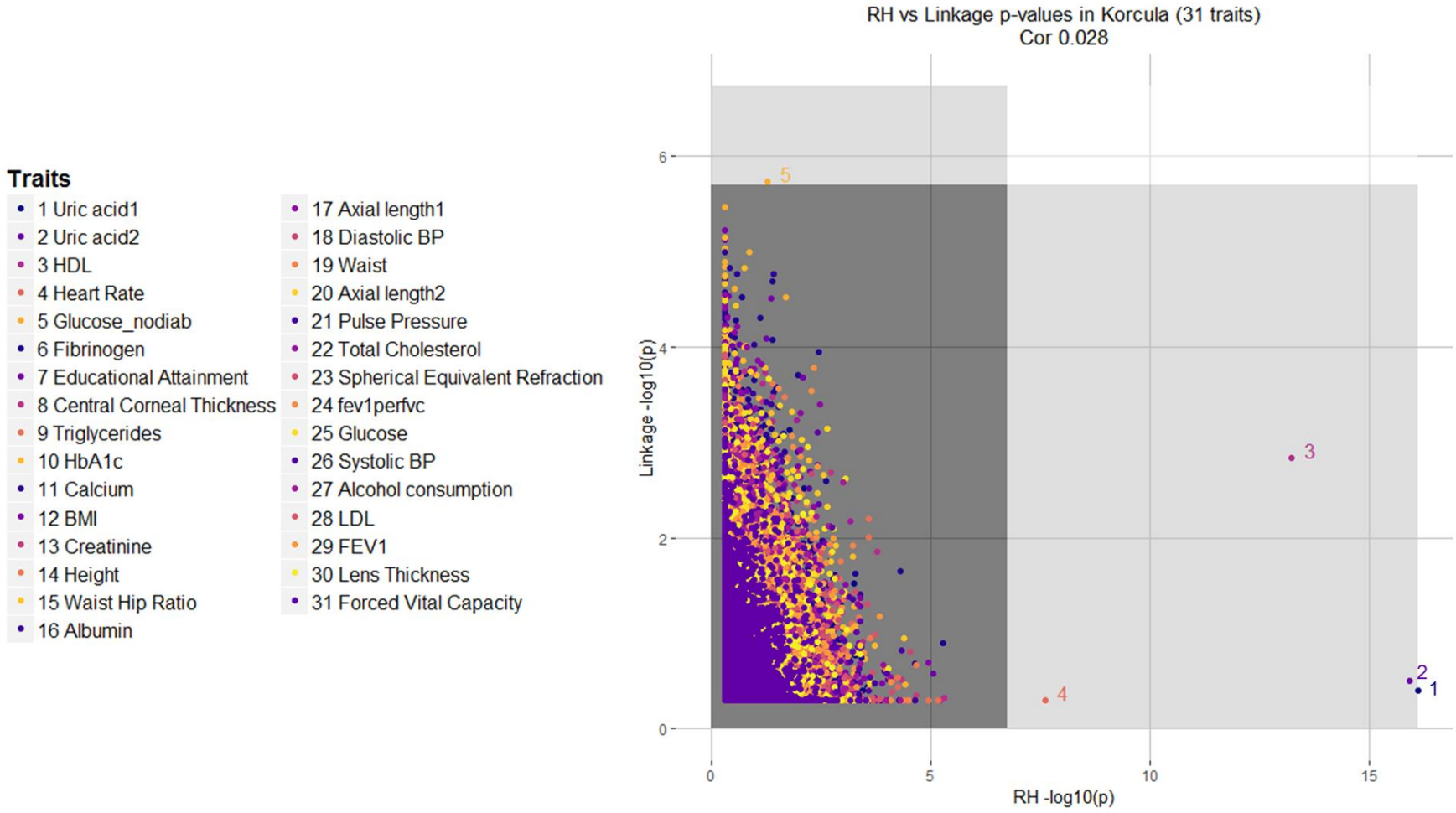
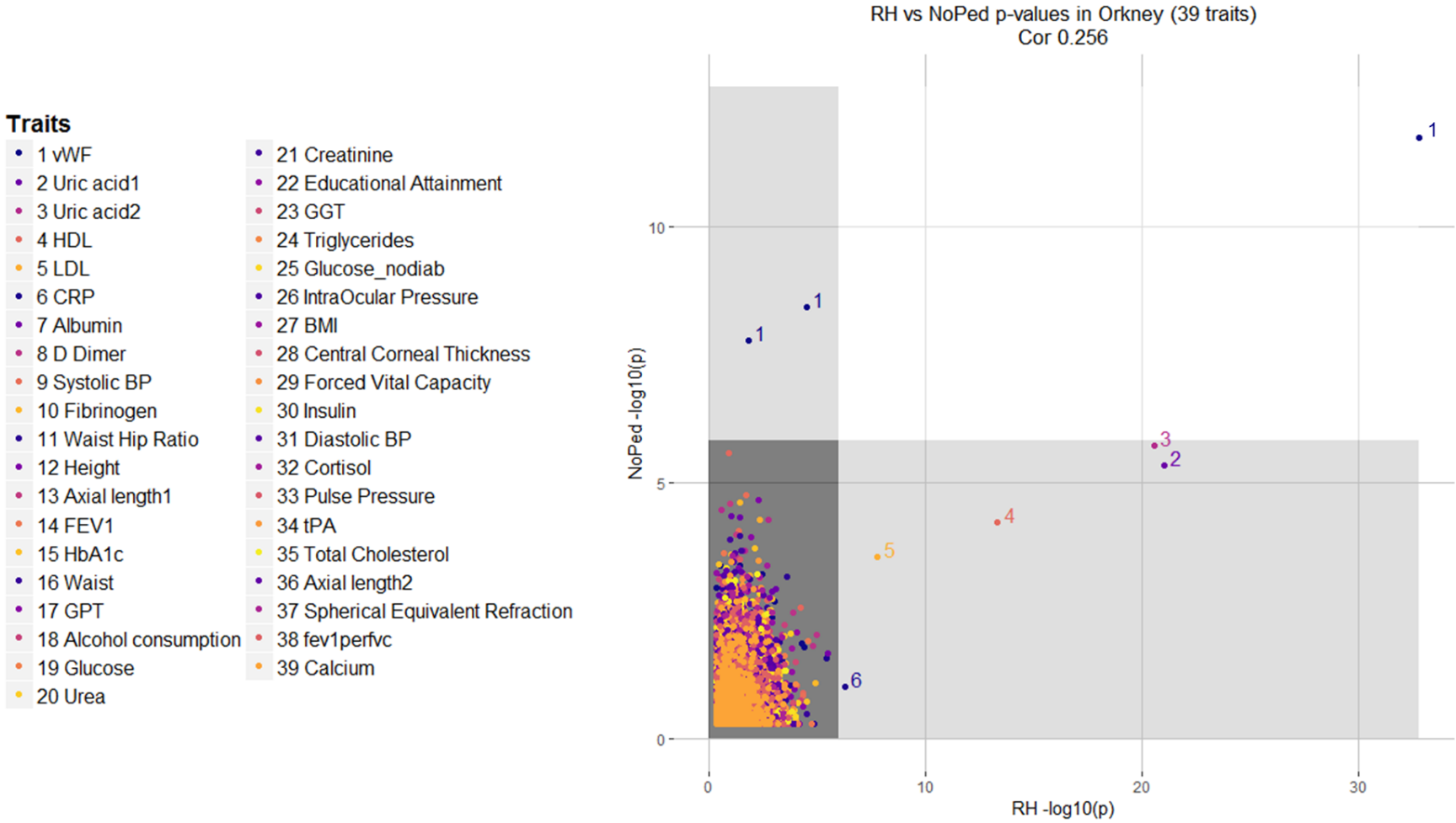


Figure 42 - RH vs Pedigree-free linkage analysis in Orkney

For linkage analysis, the -$\log_{10}$(*p-value*) of each region is plotted. For RH, the region with the highest -$\log_{10}$(*p-value*) within each linkage region is plotted. Traits are ordered based on the highest -$\log_{10}$(*p-value*) in each trait. Trait numbers are assigned to points that exceeded the GWS threshold in one (light grey background) or both analyses (white background). Points in the dark grey background did not reach GWS in either analysis.



7.1.3 Trait Heritabilities

The narrow-sense heritability (that is, the heritability attributable to additive genetic effects) was calculated for each trait (after applying normalisation and adjusting for covariates), within each cohort. To calculate heritabilities, three different genome-wide relationship matrices were used, to reflect those used in pedigree-based linkage analysis, GWAS/RH (which use genetic relationship matrices estimated using the same method) and pedigree-free linkage analysis:

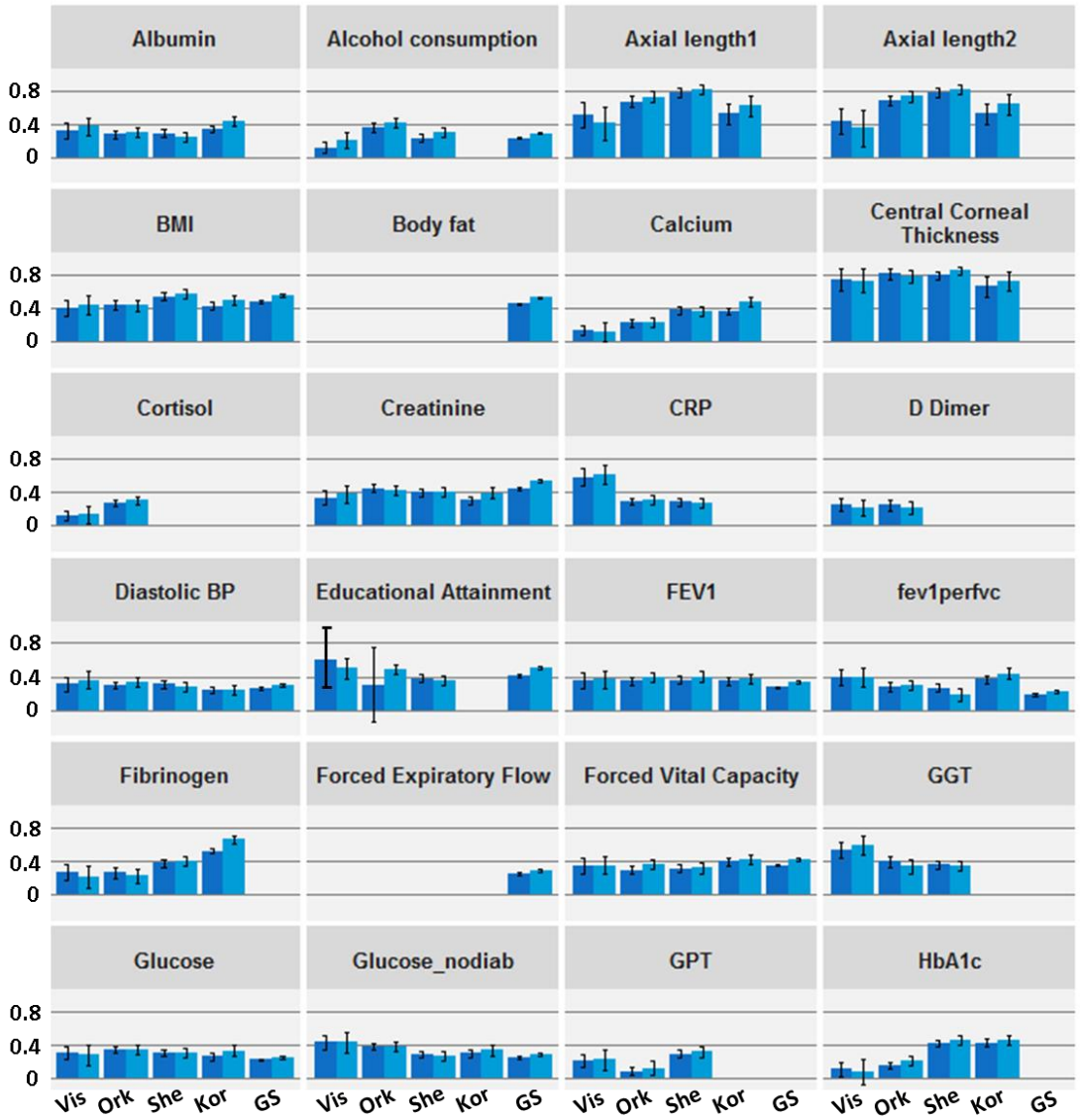
1. Kinship coefficients estimated from relationships indicated in the social pedigree. This kinship matrix is not based on marker genotype information and depends purely on the relationships recorded in the social pedigree, giving the average expected kinship-value for each type of relationship (e.g. full siblings will always have a 0.5 kinship coefficient regardless of the actual proportion of their genomes they share IBD). This social pedigree-based heritability was calculated by the polygenic command within SOLAR [203].
2. A genetic relationship matrix (GRM) calculated from the genotype data using the ibs() function in the GenABEL R package [72]. This GRM is based on SNPs that are identical by state between pairs of individuals and is described in more detail in section 3.2.1. The heritability was calculated using the polygenic() function in GenABEL. This function uses maximum likelihood to estimate the variance explained by the SNPs, as described in [18].
3. In Orkney and Vis, an IBD-based kinship matrix was calculated from the genotype data using the GIBDLD program within IBDLD3 (described in section 4.2.2). Briefly, genetic kinship was calculated between every pair in the data (regardless of their relationship in the social pedigree), and it is based on DNA segments that are identical by descent between pairs of individuals, taking the linkage disequilibrium (LD) between markers into account. The heritability was calculated by the polygenic command within SOLAR after overriding the pedigree-based kinship coefficients with this IBD-based kinship matrix.

Figure 43 shows the social pedigree and GRM-based heritability estimates for each trait, within each cohort. Generally, the social pedigree-based heritabilities are slightly higher than the ones obtained by using the genetic data only. This could be caused by non-additive genetic effects (such as dominance, epistasis or gene-gene interactions) segregating within families [204]. Additionally, as discussed in section 4.3.1.1, the social pedigree-based kinship matrix fails to account for many relationships (mostly between distantly related individuals). This might cause biases in heritability estimates because it only considers closer relatives, who not only share genetic, but also environmental effects, which will cause their phenotypes to be more similar than can be explained by their shared genotypes alone. In Orkney and Vis, genotype-based IBD sharing matrices were calculated by GIBDLD in addition to the social

pedigree to estimate heritability. The heritabilities estimated with these matrices are generally lower than the heritability estimated based on the social pedigree and often more in line with the GRM-based heritability estimates (Figure 44). The fact that these values are still not identical is attributable to the differences in how the GRMs are calculated (either using identity by state (discussed in Section 3.2.1) or identity by descent (discussed in Section 4.2.2)). Overall, however, in Orkney and Vis, the heritabilities estimated using the three methods described here are broadly similar, as indicated by the overlapping error bars in Figure 44

Figure 43 - Trait heritabilities estimated from genetic or pedigree data

This figure shows the heritability of each trait, as estimated from genetic kinship (calculated by GenABEL, dark blue columns) or using social pedigrees only (no genetic data, using SOLAR, light blue columns). These are grouped by cohort (indicated on the bottom of the plot, Ork – Orkney, She – Shetland, Kor – Korčula, GS – Generation Scotland). Note that this figure is continued on the next page.



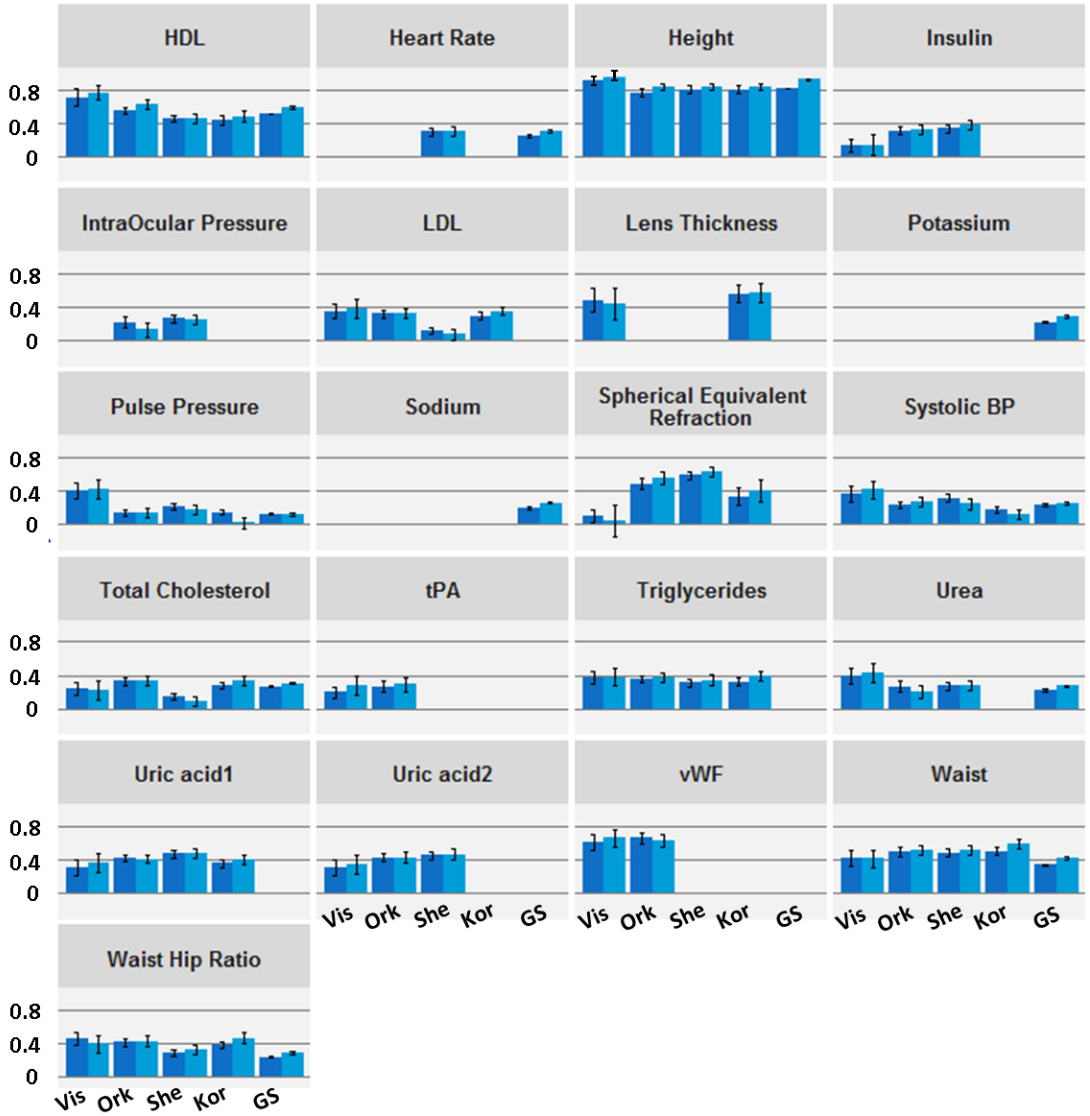
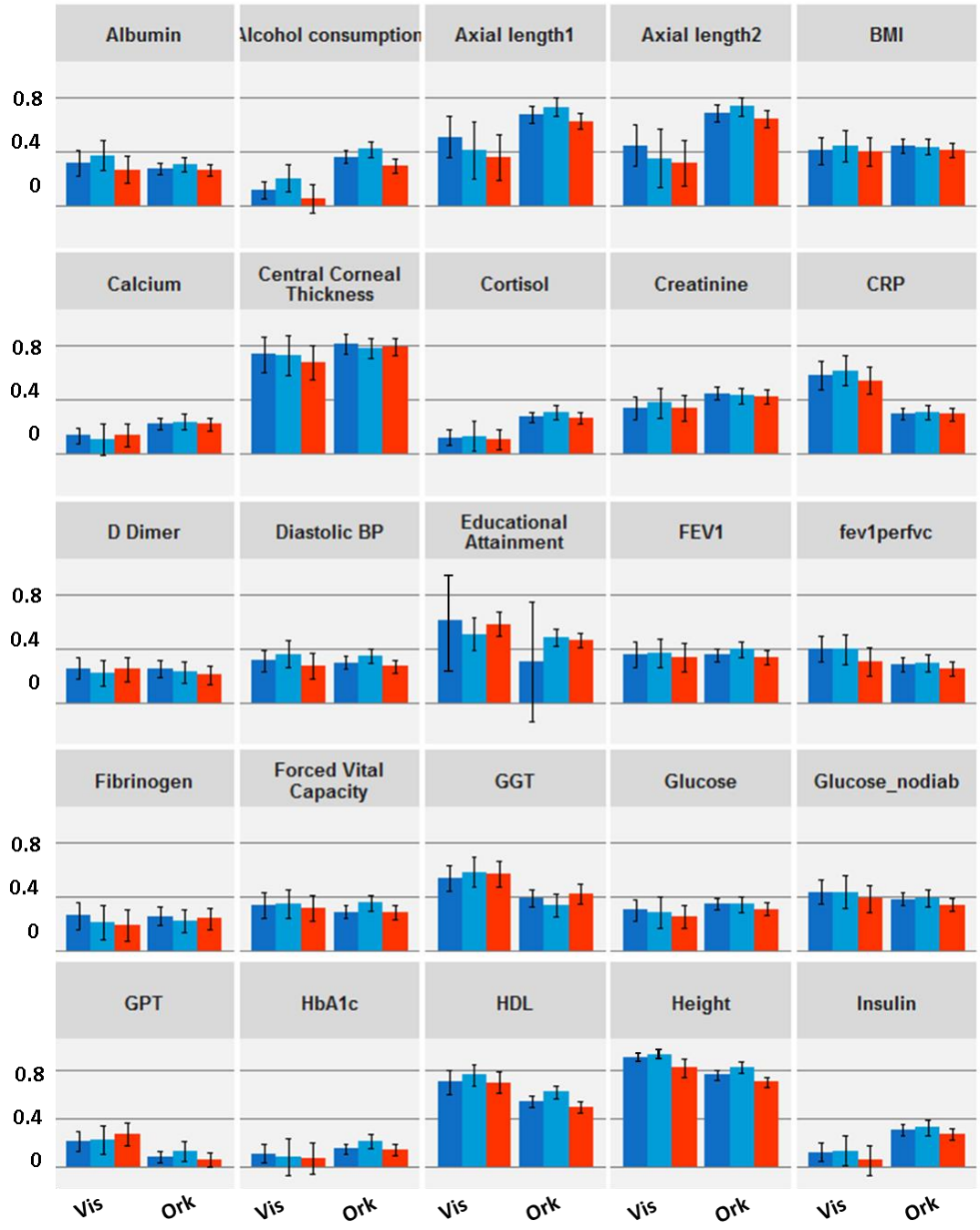
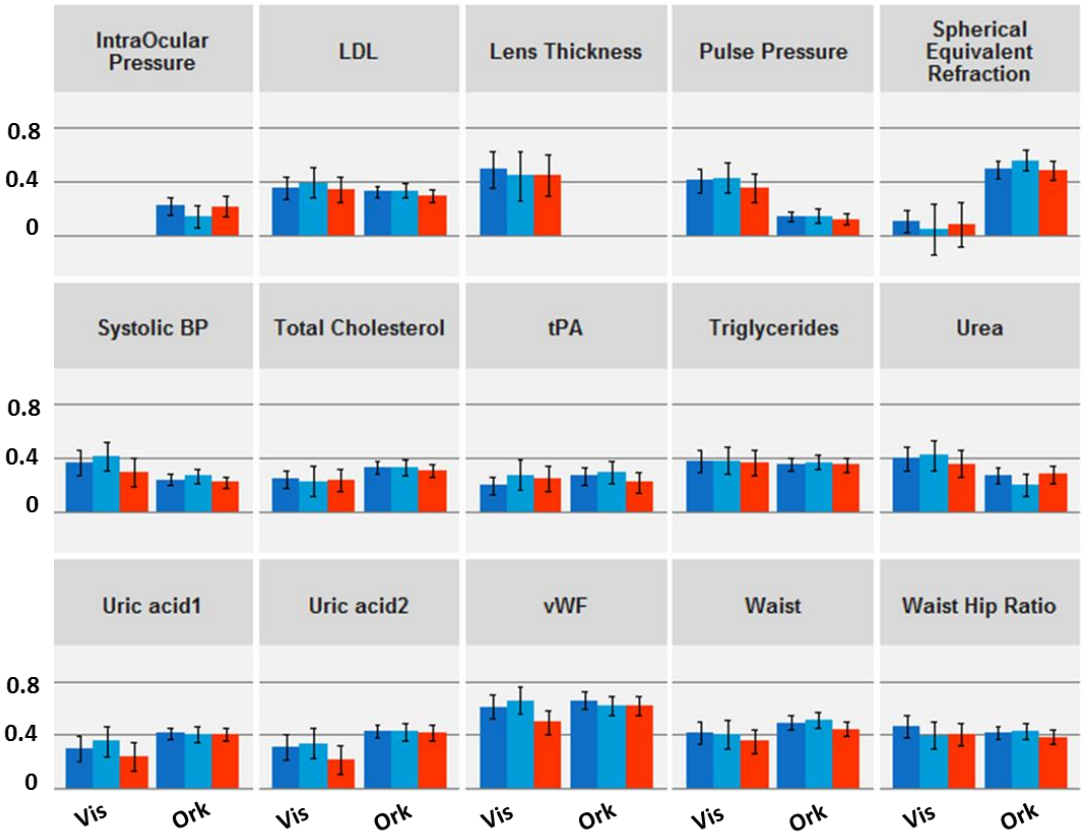

HDL
Heart Rate
Height
Insulin
IntraOcular Pressure
LDL
Lens Thickness
Potassium
Pulse Pressure
Sodium
Spherical Equivalent Refraction
Systolic BP
Total Cholesterol
tPA
Triglycerides
Urea
Uric acid1
Uric acid2
vWF
Waist
Waist Hip Ratio
0.8
0.4
0
Vis
Ork
She
Kor
GS

Figure 44 - Trait heritabilities estimated from pedigree or genetic data (using either marker identity by state or identity by descent)

This figure shows the heritability of each trait, as estimated using kinship matrices obtained from social pedigrees only (no genetic data, using SOLAR, light blue columns) or using GRMs generated from genetic data (dark blue columns - based on identity by state (GenABEL), red columns - based on identity by descent (GIBDLD)). These are grouped by cohort (indicated on the bottom of the plot, Ork – Orkney). Note that this figure is continued on the next page.





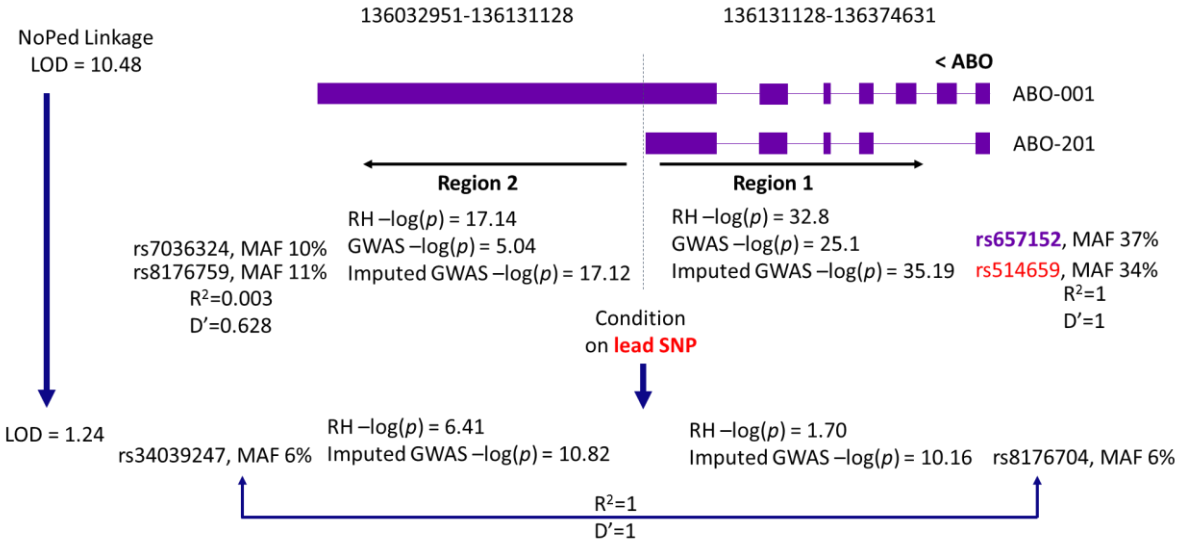
7.1.4 *ABO* Locus

The *ABO* locus contains common variants that strongly associate with von Willebrand Factor (vWF) plasma levels, explaining a large proportion of the heritability in this trait. vWF expresses ABO antigens, and polymorphisms within the *ABO* gene cause the ABO enzyme to attach either N-acetyl glucosamine (A allele) or galactose (B allele) to a precursor antigen, while the O allele results in a non-functional ABO enzyme that leads to an unmodified antigen [205]. Individuals with the O blood group have the highest rates of vWF clearance, individuals with the B blood group have higher vWF levels compared to individuals with the A blood group, while individuals with AB blood groups have the highest vWF levels, which suggests that several polymorphisms at the *ABO* gene may act in concert to affect vWF levels [206].

This locus was selected to evaluate the ability of GWAS (using genotyped or imputed SNPs), RH and pedigree-free linkage analysis to identify major signals originating from common variants, as well as to assess the presence of other independent, potentially rare QTLs of strong effect segregating in this region. A schematic of this region, and a summary of results, is depicted in Figure 45.

Figure 45 - RH and GWAS results at the ABO locus

Two splice variants of the *ABO* gene are depicted. Retained exons are indicated by purple rectangles while introns are indicated by purple lines. The direction of gene transcription is shown. The grey dashed line indicates the border between the two RH regions. RH results and test statistics for the most significant SNPs in the GWAS using genotyped (called GWAS on the plot), or imputed SNPs (called Imputed GWAS on the plot) are shown for each region, and the rsIDs and MAFs of these SNPs are indicated, as are their LD statistics. At the bottom, the pedigree-free linkage, RH and imputed GWAS results of the conditional analyses (conditioning on rs514659) are shown.



In Orkney, GWAS, RH and pedigree-free linkage analysis all yield strong signals at this locus: LOD=10.48 with pedigree-free linkage analysis, $-\log_{10}(p\text{-}value)$=32.8 with RH, and $-\log_{10}(p\text{-}value)$=25.1 with GWAS using genotyped SNPs only, at rs657152 (MAF=37%)). The imputed SNP yielding the strongest signal ($-\log_{10}(p\text{-}value)$=35.19 at rs514659, MAF=34%) is in perfect LD with rs657152, the genotyped SNP that yielded the strongest GWAS signal in this region. This imputed SNP is also in perfect LD with rs687289, the SNP that tagged the O serotype in a vWF GWAS [167]. This RH region contains one splice variant (ABO-201, 937bp) of the *ABO* gene that has a short final exon.

The adjacent RH region contains a part of the *ABO* gene that is only present in a second splice variant (ABO-001, 6341bp) but not the first one (Figure 45), and the RH test statistic in this region far exceeds the GWS threshold ($-\log_{10}(p\text{-}value)$ = 17.14) while no genotyped SNPs in this region reach this threshold with GWAS (highest $-\log_{10}(p\text{-}value)$=5.04 at rs7036324, MAF=10%). This may indicate the presence of causal variants at this locus that are poorly captured by genotyped SNPs. Indeed, the GWAS using imputed SNPs reveals a SNP that

reaches -$\log_{10}$(*p-value*)=17.12 in this region (at rs8176759, MAF=11%), and this SNP is not in LD with rs7036324 (R^2=0.003, D'=0.628). This demonstrates the ability of RH to detect a signal originating from a SNP that is not genotyped.

When the analyses are re-run by conditioning on the genotype of the imputed SNP that yielded the strongest signal at the *ABO* locus, rs514659, the pedigree-free linkage signal is mostly lost (LOD=1.24), indicating that this one variant captures most of the linkage signal. In this conditional analysis, the imputed GWAS still highlights hits with -$\log_{10}$(*p-value*)=10 in both regions (at rs8176704 in the primary region and at rs34039247 in the secondary region, these SNPs are in complete LD with each other and have MAF=6%), indicating that the initial strong GWAS signal was masking a secondary signal originating from these variants. The RH signal is lost in the primary region, but in the secondary region a signal remains with -$\log_{10}$(*p-value*)=6.41, which suggests that the source of the secondary signal may lie within this region. The imputed SNP yielding the most significant *p-value* in this region in the conditional GWAS analysis, rs34039247, is in perfect LD with rs8176704, a SNP that tags the A serotype in the same vWF GWAS mentioned above [167].

Because the pedigree-free linkage analysis region covers the entire locus discussed here, it may be detecting the joint effects of these variants that independently modulate vWF levels. In contrast, the fact that the 'O' and 'A' variants lie in two adjacent RH regions might be the reason why they appear as two separate signals in the RH analysis, while the effect of the strongest ('O') SNP masks the effect of the secondary ('A') SNP in the imputed GWAS.

Chapter 8 Discussion

Within this thesis, I used five family-based datasets to conduct genetic analysis of 45 medically-relevant human complex traits with three distinct statistical methodologies. The aim of this thesis was to provide a systematic assessment of the strengths and weaknesses of GWAS, variance component linkage analysis and regional heritability mapping to detect QTLs under a wide range of genetic architectures; as a consequence of these analyses, novel candidate loci contributing to complex trait variation were also uncovered.

The results presented here reaffirm that human complex traits vary in terms of their underlying genetic architecture. For example, the highly polygenic nature of height, well known from published analyses, is also demonstrated here because in the Generation Scotland study as well as in meta-analyses, all three statistical methods reveal many loci, individually of small effect, influencing this trait. In contrast, some blood biochemistry traits such as the serum levels of cholesterol, von Willebrand Factor or uric acid are influenced by a combination of a few loci of large effect and many additional loci with individually small effects. These findings make it clear that the genetics of complex trait variation cannot be described with any one model (such as the common disease-common variant [6], common disease-rare variant [9] or infinitesimal [11] models described in the Introduction), because complex traits are often the consequence of a combination of large and small effects originating from common as well as rare variants [10].

My results were consistent with GWAS, RH and linkage analysis being capable of uncovering different types of genetic signals. While both GWAS and RH were able to detect signals originating from a single causal variant, as evidenced by the fact that all but a few hits discovered with GWAS also appear in the results of RH, 27 additional hits are detected with RH only. This method might therefore be better powered to detect regions where several independent causal SNPs are present, which may not be detected with single-SNP GWAS. Additionally, most of the loci identified with RH but not GWAS in the cohorts studied here have also been reported in published GWAS meta-analyses, but their detection required sample sizes much larger than the ones used within this thesis. This is a clear advantage of using RH in cohorts where complex traits that are not widely measured are available (for example phenotypes that are costly to measure or require specialist equipment to quantify).

In contrast, pedigree-based linkage analyses often do not detect the loci identified with GWAS or RH, even when these harbour QTLs that have large effects on a trait. This is likely a consequence of the limited power of pedigree-based linkage analysis, as the simulation studies

presented within this thesis suggest that pedigree-based linkage analysis can detect the same loci found with GWAS or RH, but only if the effects of a QTL are sufficiently large. Instead, relying on the assessment of genomic segments that are shared IBD, both pedigree-based and pedigree-free linkage analyses reveal loci that are not detected with GWAS and RH. Such loci could be false positives, but they could also harbour rare causal variants that segregate on specific haplotypes. The simulation studies I performed demonstrated that pedigree-free linkage analysis can sometimes detect signals originating from low frequency (MAF < 5%) variants even when these are missed with GWAS, suggesting that some of the loci flagged when analysing 'real' phenotypes might be true positives.

The fact that linkage analysis results often depend on the family structures (and therefore genetic segregation patterns) present in a study means that while they may reveal novel loci that have not been discovered with GWAS, these loci could be hard (or impossible) to replicate in other cohorts. Indeed, in the literature, novel linkage peaks explaining a large proportion of the trait variance in a specific cohort often lack replication or functional follow-up that would lend more confidence to their validity. For example, in one study [167], linkage analysis of von Willebrand factor levels has resulted in a peak with a LOD score of 2.9 on chromosome 9 at the ABO locus, and this peak explains 24.5% of the trait variance and was also detected with GWAS in the same study as well as with linkage analysis and GWAS within this thesis. However, this study also reports a linkage peak with a higher LOD score of 5.3 on chromosome 2, which explains 19.2% of the trait variance, but this signal is not present in their GWAS and also not present in the GWAS or linkage analysis results reported in this thesis. In the same study, the authors also fail to identify haplotypes in this region that have a strong effect on the trait. While they flag two 1 Mb loci by looking at differences in SNP LD structure, these loci do not harbour any genes that play an obvious role in vWF biology. This emphasizes the difficulty in following up regions identified by linkage analysis but it may also indicate the possibility that this may be a false positive hit.

The advantage of isolate populations over cosmopolitan populations for linkage analysis is evidenced by the fact that, according to the power calculations performed using SOLAR, Orkney and Generation Scotland appear to have the same power to detect a QTL with pedigree-based linkage analysis, despite the tenfold smaller sample size in Orkney. This is a consequence of the more complex family structure and a higher number of individuals in families in Orkney. However, when analysing real phenotypes instead of quantitative traits simulated using a single QTL, more loci are identified with pedigree-based linkage in Generation Scotland than in Orkney, with the caveat that these could again be due to false

positives. This discrepancy may be due to the fact that simulation studies are often unable to account for latent factors that drive complex trait variation so they may provide results that do not hold up when the same method is applied in a 'real' scenario. Similarly, in the simulation studies I performed, phenotypes simulated using DISSECT were not an accurate representation of complex traits as they simulated a single major QTL on a polygenic background, and did not take account of the fact that environmental variance is not distributed randomly within families. This means that they may not be entirely suitable for comparing the performance of the same statistical methods in different populations due to their inability to capture some factors such as cohort-specific environment effects that may lead to cohort-specific trait variation. Nonetheless, such simulation studies still provide useful guidance for interpreting the results obtained when analysing 'real' data.

By setting up the pedigree-free linkage analysis pipeline, I have attempted to provide a boost to the power to detect a QTL compared to pedigree-based linkage analysis by extending linkage analysis to include distantly-related individuals who are not recorded in a social pedigree but who may still share some segments IBD. This should lead to an increase in power particularly in population isolates, due to these having a higher number of relatives and an overall increase in IBD sharing. Interestingly, pedigree-free linkage analysis reveals genetic effects segregating at the population level that have been detected with GWAS but remained undetected with pedigree-based linkage analysis, which is likely due to a gain in power as a larger number of related individuals are used. In theory, linkage analysis can be a useful tool for identifying regions harbouring QTLs missed by single-SNP GWAS, as it can detect the combination of effects of several independent alleles at the same locus or allelic heterogeneity with multiple rare variants. Detecting such loci would be a boon for prioritising regions influencing traits in the age of whole-genome sequencing. However, I acknowledge that linkage analysis overall is not very powerful for studying the genetic basis of complex traits in the cohorts that I have studied here. The power of discovery could be boosted by using larger studies that have a higher number of related individuals. More critically, the use of denser genotyping data (or in the near future, the use of whole genome sequencing) might allow for more accurate estimation of IBD sharing. The inaccurate estimation of IBD sharing may have been a source of false positives in the studies presented here and the accurate determination of segments shared IBD might have been hindered by the fact that fairly sparse genotype data were used, which is a consequence of the use in the same cohort of multiple genotyping platforms that have a limited amount of overlapping SNPs. Another consideration is that pedigree-free linkage analysis is very time-consuming to run even with the pipeline that I have created, because of current software limitations. The two rate-limiting steps are the

conversion of the IBD coefficient files output by IBDLD to SOLAR format (which could easily be circumvented by IBDLD outputting files directly to this format, and its developers have communicated to me that this is something they are already considering), and the linkage analysis step in SOLAR, which is unable to perform the analysis when provided with pairwise IBD sharing between more than 2000 individuals. Instead, the variance component analysis could be performed using DISSECT, the program that was used to conduct RH mapping within this thesis, as it can handle a large number of pairwise relationships, but it is not currently optimised for performing analysis sequentially on many externally-computed matrices.

Within this study, traits that have been extensively studied with GWAS were used as positive controls, and it is reassuring that I find many loci that replicate published findings obtained from large GWAS meta-analyses. These positive controls also lend confidence to the validity of novel loci I identify. Currently, most large GWAS meta-analyses report the results of GWAS performed on genotypes imputed to the 1000 genomes reference panel [90] and many hits reported within this thesis replicate these findings. However, low frequency (MAF in the range of 0.5-5%) variant detection with GWAS has been improved with the help of the HRC imputation panel which allows many such variants to be inferred with a high degree of confidence, and several novel associations with low frequency variants are identified within this thesis, particularly in Generation Scotland as well as the GWAS meta-analysis. It will be interesting to see whether these associations will replicate in independent GWAS performed using genotypes imputed to the HRC panel.

One such dataset is the UK Biobank [207], which consists of genotype and phenotype data for 500,000 participants, who also have had their genotypes imputed to the HRC panel [208]. Due to its sheer size alone, this resource will lead to the discovery of many associations with rare variants, as compared to smaller studies, there will be more instances of a rare variant with the same allele frequency within UK Biobank. One exception to this is if a specific rare variant is enriched within a population isolate due to drift or founder effect for example, in which case there could be a higher number of minor alleles carried by a smaller number of individuals. In addition to its size, another advantage of UK Biobank is the fact that phenotypes were measured the same way across all UK Biobank participants, and consistent genotyping and imputation protocols were used. This should eliminate a lot of noise that would normally be present in a meta-analysis combining the results of many different cohorts. As a consequence, GWAS performed using the UK Biobank data should have an additional boost in power compared to a meta-analysis using the same number of samples. UK Biobank will therefore act as the primary point of reference for replicating the findings of other complex trait mapping

studies. For this purpose, an atlas of genetic associations in UK Biobank has already been created, where GWAS results of over 700 traits are reported [209]. While the searchable database provided by this study is still in its infancy, it (or related tools such as the Global Biobank Engine (http://gbe.stanford.edu/)) will eventually provide simple means of assessing whether a signal detected in a study was also present when analysing the same trait in UK Biobank using GWAS.

A resource of this magnitude provides clear advantages for performing GWAS, especially of well-studied, easily measurable phenotypes. However, one advantage of the smaller cohorts used in this study is that, compared to the UK Biobank dataset which sampled individuals aged 40-69, our populations are richer in 'vertical' family relationships. The 29-yeasr age range results in an underrepresentation of multi-generational family relationships in UK biobank, so studies that make use of, or explicitly model, this kind of relationship will benefit from using cohorts sampled with this in mind. Indeed, in the UK biobank, there are only 1066 trios (two parents and an offspring), 172 families with 5 or more 2nd degree or close relatives and no instances of three-generation families [208]. In contrast, in the Orkney cohort for example, according to the 'clipped' pedigrees, there are 170 families with an average number of 11 genotyped individuals per family, while in Shetland, the clipped pedigrees show 273 families with an average number of 8 people per family. These deep pedigrees are particularly advantageous for studying parent-of-origin effects or pedigree-associated genetic variation [210]. Whether they will be used in linkage analysis will depend on the future assessment and improvement of the pedigree-free linkage analysis methodology. In Orkney, whole-genome sequencing data will soon be available, which should allow for more in-depth follow-up of the novel regions identified within this cohort.

My work has led to a publication presenting the results of the Generation Scotland GWAS using HRC imputed data, of which I am the lead author [62]. By curating the social pedigrees and genetic data in the Generation Scotland study, I also contributed to the work of others studying the genetics of Alzheimer's disease [211], the genetic interactions between longevity and educational attainment [212], the genetic and environmental factors contributing to depression [184, 213], cardiometabolic and anthropometric trait variation [210] and regional differences in obesity-related phenotypes [214]. By performing genetic analyses in the Orkney, Vis and Korčula cohorts, I have contributed to a study assessing the genetic and environmental factors affecting retinal microvasculature [215] as well as a study that aimed to predict complex traits using a combination of SNPs obtained from meta-analysis results and machine learning prioritisation methods [216]. In addition to work that has already led to

publications, I am currently contributing to a project led by Prof. Ruth Jarrett at the University of Glasgow on integrated chromosomally inherited human herpesvirus 6. The pedigree-free linkage pipeline I created and IBD sharing estimation methods are being used to determine the sites of the human genome this virus has integrated into in different families in the Generation Scotland study. Finally, the same methods and pipelines that I have developed can readily be applied to other traits, particularly new phenotypes derived from biological samples such as metabolomics and proteomics that are becoming available in the isolates cohorts but are not currently feasible at scale in large cohorts such as the UK Biobank.

This thesis contributes to the field of complex trait genetics by systematically comparing different statistical study designs, creating robust pipelines and identifying novel genetic loci that affect medically-relevant human traits, which can improve human trait prediction and therefore the diagnosis and management of diseases relevant to these traits.

References

[1] Orstavik KH, Magnus P, Reisner H, Berg K, Graham JB, Nance W (1985) Factor VIII and factor IX in a twin population. Evidence for a major effect of ABO locus on factor VIII level. *Am J Hum Genet* 37(1):89–101.

[2] Lango Allen H, Estrada K, Lettre G, Berndt SI, Weedon MN, et al. (2010) Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 467(7317):832–8.

[3] Iyengar SK, Elston RC (2007) The Genetic Basis of Complex Traits. *Methods in Molecular Biology*, pp 71–84.

[4] Wright AF, Hastie ND (2001) Complex genetic diseases: controversy over the Croesus code. *Genome Biol* 2(8):comment2007.1-comment2007.8.

[5] Schork NJ, Murray SS, Frazer KA, Topol EJ (2009) Common vs. rare allele hypotheses for complex diseases. *Curr Opin Genet Dev* 19(3):212–9.

[6] Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet* 17(9):502–10.

[7] Corder EH, Saunders AM, Strittmatter WJ, Schmechel DE, Gaskell PC, et al. (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 261(5123):921–3.

[8] Corbo RM, Scacchi R (1999) Apolipoprotein E (APOE) allele distribution in the world. Is APOE*4 a "thrifty" allele? *Ann Hum Genet* 63(Pt 4):301–10.

[9] Pritchard JK (2001) Are Rare Variants Responsible for Susceptibility to Complex Diseases? *Am J Hum Genet* 69(1):124–137.

[10] Boyle EA, Li YI, Pritchard JK (2017) An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169(7):1177–1186.

[11] Gibson G (2011) Rare and common variants: twenty arguments. *Nat Rev Genet* 13(2):135–45.

[12] Barton NH, Keightley PD (2002) Multifactorial genetics: Understanding quantitative genetic variation. *Nat Rev Genet* 3(1):11–21.

[13] Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. *Science* 273(5281):1516–7.

[14] Hoffman GE (2013) Correcting for population structure and kinship using the linear mixed model: theory and extensions. *PLoS One* 8(10):e75707.

[15] Kruglyak L (1999) Prospects for whole-genome linkage disequilibrium mapping of common diseasegenes. *Nat Genet* 22(2):139–144.

[16] Welter D, MacArthur J, Morales J, Burdett T, Hall P, et al. (2014) The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42(Database issue):D1001-6.

[17] Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, et al. (2009) Finding the missing heritability of complex diseases. *Nature* 461(7265).

[18] Yang J, Benyamin B, McEvoy BP, Gordon S, Henders AK, et al. (2010) Common SNPs explain a large proportion of the heritability for human height. *Nat Genet* 42(7):565–9.

[19] Yang J, Bakshi A, Zhu Z, Hemani G, Vinkhuyzen AAE, et al. (2015) Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47(10):1114–1120.

[20] Kiezun A, Garimella K, Do R, Stitziel NO, Neale BM, et al. (2012) Exome sequencing and the genetic basis of complex traits. *Nat Genet* 44(6):623–30.

[21] Lee S, Abecasis GR, Boehnke M, Lin X (2014) Rare-variant association analysis: study designs and statistical tests. *Am J Hum Genet* 95(1):5–23.

[22] Wood AR, Esko T, Yang J, Vedantam S, Pers TH, et al. (2014) Defining the role of common variation in the genomic and biological architecture of adult human height. *Nat Genet* 46(11):1173–1186.

[23] Marouli E, Graff M, Medina-Gomez C, Lo KS, Wood AR, et al. (2017) Rare and low-frequency coding variants alter human adult height. *Nature* 542(7640):186–190.

[24] McCarthy MI (2009) Exploring the unknown: assumptions about allelic architecture and strategies for susceptibility variant discovery. *Genome Med* 1(7):66.

[25] McClellan J, King M-C, Bird TD, Bucan M, Abrahams BS, et al. (2010) Genetic Heterogeneity in Human Disease. *Cell* 141(2):210–217.

[26] Stefansson H, Rujescu D, Cichon S, Pietiläinen OPH, Ingason A, et al. (2008) Large recurrent microdeletions associated with schizophrenia. *Nature* 455(7210):232–236.

[27] Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456(7218):18–21.

[28] Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant... or not? *Hum Mol Genet* 11(20):2417–2423.

[29] Romeo S, Pennacchio LA, Fu Y, Boerwinkle E, Tybjaerg-Hansen A, et al. (2007) Population-based resequencing of ANGPTL4 uncovers variations that reduce triglycerides and increase HDL. *Nat Genet* 39(4):513–516.

[30] Ott J, Kamatani Y, Lathrop M (2011) Family-based designs for genome-wide association studies. *Nat Rev Genet* 12(7):465–474.

[31] Beekman M, Blanché H, Perola M, Hervonen A, Bezrukov V, et al. (2013) Genome-wide linkage analysis for human longevity: Genetics of Healthy Aging Study. *Aging Cell* 12(2):184–93.

[32] Gusella JF, Wexler NS, Conneally PM, Naylor SL, Anderson MA, et al. (1983) A polymorphic DNA marker genetically linked to Huntington's disease. *Nature* 306(5940):234–238.

[33] Tsui LC, Buchwald M, Barker D, Braman JC, Knowlton R, et al. (1985) Cystic fibrosis locus defined by a genetically linked polymorphic DNA marker. *Science* 230(4729):1054–7.

[34] Comuzzie AG, Hixson JE, Almasy L, Mitchell BD, Mahaney MC, et al. (1997) A major

quantitative trait locus determining serum leptin levels and fat mass is located on human chromosome 2. *Nat Genet* 15(3):273–276.

[35] Hixson JE, Almasy L, Cole S, Birnbaum S, Mitchell BD, et al. (1999) Normal Variation in Leptin Levels Is Associated with Polymorphisms in the Proopiomelanocortin Gene, POMC. *J Clin Endocrinol Metab* 84(9):3187–3191.

[36] Soria JM, Almasy L, Souto JC, Sabater-Lleal M, Fontcuberta J, Blangero J (2005) The F7 gene and clotting factor VII levels: dissection of a human quantitative trait locus. *Hum Biol* 77(5):561–75.

[37] Sabater-Lleal M, Chillón M, Howard TE, Gil E, Almasy L, et al. (2007) Functional analysis of the genetic variability in the F7 gene promoter. *Atherosclerosis* 195(2):262–268.

[38] Almasy L, Blangero J (2009) Human QTL linkage mapping. *Genetica* 136(2):333–40.

[39] Han L, Abney M (2013) Using identity by descent estimation with dense genotype data to detect positive selection. *Eur J Hum Genet* 21(2):205–11.

[40] Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11(7):499–511.

[41] Krawitz PM, Schweiger MR, Rödelsperger C, Marcelis C, Kölsch U, et al. (2010) Identity-by-descent filtering of exome sequence data identifies PIGV mutations in hyperphosphatasia mental retardation syndrome. *Nat Genet* 42(10):827–9.

[42] Louis-Dit-Picard H, Barc J, Trujillano D, Miserey-Lenkei S, Bouatia-Naji N, et al. (2012) KLHL3 mutations cause familial hyperkalemic hypertension by impairing ion transport in the distal nephron. *Nat Genet* 44(4):456–60, S1-3.

[43] Yan J, Takahashi T, Ohura T, Adachi H, Takahashi I, et al. (2013) Combined linkage analysis and exome sequencing identifies novel genes for familial goiter. *J Hum Genet* 58(6):366–77.

[44] Ott J, Wang J, Leal SM (2015) Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* advance on(5):275–284.

[45] Nagamine Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2012) Localising loci underlying complex trait variation using Regional Genomic Relationship Mapping. *PLoS One* 7(10):e46501.

[46] Uemoto Y, Pong-Wong R, Navarro P, Vitart V, Hayward C, et al. (2013) The power of regional heritability analysis for rare and common variant detection: simulations and application to eye biometrical traits. *Front Genet* 4(November):232.

[47] Riggio V, Matika O, Pong-Wong R, Stear MJ, Bishop SC (2013) Genome-wide association and regional heritability mapping to identify loci underlying variation in nematode resistance and body weight in Scottish Blackface lambs. *Heredity (Edinb)* 110(5):420–429.

[48] Shirali M, Pong-Wong R, Navarro P, Knott S, Hayward C, et al. (2016) Regional heritability mapping method helps explain missing heritability of blood lipid traits in isolated populations. *Heredity (Edinb)* 116(3):333–338.

[49] Peltonen L, Palotie A, Lange K (2000) Use of population isolates for mapping complex traits. *Nat Rev Genet* 1(3):182–90.

[50] Bourgain C, Génin E (2005) Complex trait mapping in isolated populations: Are specific statistical methods required? *Eur J Hum Genet* 13(6):698–706.

[51] Kong A, Masson G, Frigge ML, Gylfason A, Zusmanovich P, et al. (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat Genet* 40(9):1068–75.

[52] Gudbjartsson DF, Helgason H, Gudjonsson SA, Zink F, Oddson A, et al. (2015) Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 47(5):435–444.

[53] Kong A, Steinthorsdottir V, Masson G, Thorleifsson G, Sulem P, et al. (2009) Parental origin of sequence variants associated with complex diseases. *Nature* 462(7275):868–74.

[54] Gudbjartsson DF, Sulem P, Helgason H, Gylfason A, Gudjonsson SA, et al. (2015) Sequence variants from whole genome sequencing a large group of Icelanders. *Sci Data* 2:150011.

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

[56] Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, et al. (2016) Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536(7616):285–291.

[57] Simonson TS, Yang Y, Huff CD, Yun H, Qin GG, et al. (2010) Genetic Evidence for High-Altitude Adaptation in Tibet. *Science (80-)* 329(5987):72–75.

[58] Fumagalli M, Moltke I, Grarup N, Racimo F, Bjerregaard P, et al. (2015) Greenlandic Inuit show genetic signatures of diet and climate adaptation. *Science (80-)* 349(6254):1343–1347.

[59] Zoledziewska M, Sidore C, Chiang CWK, Sanna S, Mulas A, et al. (2015) Height-reducing variants and selection for short stature in Sardinia. *Nat Genet* 47(11):1352–1356.

[60] Newman DL, Abney M, McPeek MS, Ober C, Cox NJ (2001) The importance of genealogy in determining genetic associations with complex traits. *Am J Hum Genet* 69(5):1146–8.

[61] Smith BH, Campbell H, Blackwood D, Connell J, Connor M, et al. (2006) Generation Scotland: the Scottish Family Health Study; a new resource for researching genes and heritability. *BMC Med Genet* 7(1):74.

[62] Nagy R, Boutin TS, Marten J, Huffman JE, Kerr SM, et al. (2017) Exploration of haplotype research consortium imputation for genome-wide association studies in 20,032 Generation Scotland participants. *Genome Med* 9(1):23.

[63] Kerr SM, Campbell A, Marten J, Vitart V, McIntosh AM, et al. (2017) Data Resource Profile: Generation Scotland Electronic Health Record. *bioRxiv*.

[64] Rudan P, Simić D, Smolej-Narancić N, Bennett LA, Janićijević B, et al. (1987) Isolation by distance in Middle Dalmatia-Yugoslavia. *Am J Phys Anthropol* 74(3):417–26.

[65] Vitart V, Biloglav Z, Hayward C, Janicijevic B, Smolej-Narancic N, et al. (2006) 3000

years of solitude: extreme differentiation in the island isolates of Dalmatia, Croatia. *Eur J Hum Genet* 14(4):478–87.

[66] Rudan I, Biloglav Z, Vorko-Jović A, Kujundzić-Tiljak M, Stevanović R, et al. (2006) Effects of inbreeding, endogamy, genetic admixture, and outbreeding on human health: a (1001 Dalmatians) study. *Croat Med J* 47(4):601–10.

[67] Campbell H, Carothers AD, Rudan I, Hayward C, Biloglav Z, et al. (2006) Effects of genome-wide heterozygosity on a range of biomedically relevant human quantitative traits. *Hum Mol Genet* 16(2):233–241.

[68] McQuillan R, Leutenegger A-L, Abdel-Rahman R, Franklin CS, Pericic M, et al. (2008) Runs of homozygosity in European populations. *Am J Hum Genet* 83(3):359–72.

[69] Vitart V, Rudan I, Hayward C, Gray NK, Floyd J, et al. (2008) SLC2A9 is a newly identified urate transporter influencing serum urate concentration, urate excretion and gout. *Nat Genet* 40(4):437–442.

[70] Smith BH, Campbell A, Linksted P, Fitzpatrick B, Jackson C, et al. (2013) Cohort Profile: Generation Scotland: Scottish Family Health Study (GS:SFHS). The study, its participants and their potential for genetic research on health and illness. *Int J Epidemiol* 42(3):689–700.

[71] Pilia G, Chen W-M, Scuteri A, Orrú M, Albai G, et al. (2006) Heritability of Cardiovascular and Personality Traits in 6,148 Sardinians. *PLoS Genet* 2(8):e132.

[72] Aulchenko YS, Ripke S, Isaacs A, van Duijn CM (2007) GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23(10):1294–6.

[73] Pe'er I, Yelensky R, Altshuler D, Daly MJ (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. *Genet Epidemiol* 32(4):381–385.

[74] Nyholt DR (2000) All LODs Are Not Created Equal. *Am J Hum Genet* 67(2):282–288.

[75] Chang CC, Chow CC, Tellier LC, Vattikuti S, Purcell SM, Lee JJ (2015) Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4(1):7.

[76] Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MAR, et al. (2007) PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am J Hum Genet* 81(3):559–575.

[77] McCarthy S, Das S, Kretzschmar W, Delaneau O, Wood AR, et al. (2016) A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 48(10):1279–1283.

[78] Delaneau O, Zagury J-F, Marchini J (2012) Improved whole-chromosome phasing for disease and population genetic studies. *Nat Methods* 10(1):5–6.

[79] Delaneau O, Marchini J, Zagury J-F (2011) A linear complexity phasing method for thousands of genomes. *Nat Methods* 9(2):179–181.

[80] O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, et al. (2014) A general approach for haplotype phasing across the full spectrum of relatedness. *PLoS Genet* 10(4):e1004234.

[81] Vitart V, Bencic G, Hayward C, Herman JŠ, Huffman J, et al. (2010) Heritabilities of

Ocular Biometrical Traits in Two Croatian Isolates with Extended Pedigrees. *Investig Opthalmology Vis Sci* 51(2):737–743.

[82] Staples J, Qiao D, Cho MH, Silverman EK, Nickerson DA, Below JE (2014) PRIMUS: rapid reconstruction of pedigrees from genome-wide estimates of identity by descent. *Am J Hum Genet* 95(5):553–64.

[83] Wigginton JE, Abecasis GR (2005) PEDSTATS: descriptive statistics, graphics and quality assessment for gene mapping data. *Bioinformatics* 21(16):3445–7.

[84] Heath SC (1997) Markov chain Monte Carlo segregation and linkage analysis for oligogenic models. *Am J Hum Genet* 61(3):748–60.

[85] Liu F, Kirichenko A, Axenovich TI, van Duijn CM, Aulchenko YS (2008) An approach for cutting large and complex pedigrees for linkage analysis. *Eur J Hum Genet* 16(7):854–60.

[86] Bouaziz M, Ambroise C, Guedj M (2011) Accounting for Population Stratification in Practice: A Comparison of the Main Strategies Dedicated to Genome-Wide Association Studies. *PLoS One* 6(12):e28845.

[87] Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population stratification in genome-wide association studies. *Nat Rev Genet* 11(7):459–63.

[88] Visscher PM, Wray NR, Zhang Q, Sklar P, McCarthy MI, et al. (2017) 10 Years of GWAS Discovery: Biology, Function, and Translation. *Am J Hum Genet* 101(1):5–22.

[89] Ward LD, Kellis M (2012) HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res* 40(D1):D930–D934.

[90] Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, et al. (2015) A global reference for human genetic variation. *Nature* 526(7571):68–74.

[91] Eu-Ahsunthornwattana J, Miller EN, Fakiola M, Wellcome Trust Case Control Consortium 2 WTCCC, Jeronimo SMB, et al. (2014) Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet* 10(7):e1004445.

[92] Chen W-M, Abecasis GR (2007) Family-based association tests for genomewide association scans. *Am J Hum Genet* 81(5):913–26.

[93] Haller T, Kals M, Esko T, Magi R, Fischer K (2015) RegScan: a GWAS tool for quick estimation of allele effects on continuous traits and their combinations. *Brief Bioinform* 16(1):39–44.

[94] Svishcheva GR, Axenovich TI, Belonogova NM, van Duijn CM, Aulchenko YS (2012) Rapid variance components?based method for whole-genome association analysis. *Nat Genet* 44(10):1166–1170.

[95] Willer CJ, Li Y, Abecasis GR (2010) METAL: fast and efficient meta-analysis of genomewide association scans. *Bioinformatics* 26(17):2190–1.

[96] Vitart V, Bencić G, Hayward C, Skunca Herman J, Huffman J, et al. (2010) New loci associated with central cornea thickness include COL5A1, AKAP13 and AVGR8. *Hum Mol Genet* 19(21):4304–11.

[97] Surakka I, Horikoshi M, Mägi R, Sarin A-P, Mahajan A, et al. (2015) The impact of low-frequency and rare variants on lipid levels. *Nat Genet* 47(6):589–597.

[98] van Leeuwen EM, Sabo A, Bis JC, Huffman JE, Manichaikul A, et al. (2016) Meta-analysis of 49 549 individuals imputed with the 1000 Genomes Project reveals an exonic damaging variant in ANGPTL4 determining fasting TG levels. *J Med Genet* 53(7):441–9.

[99] Sanna S, Jackson AU, Nagaraja R, Willer CJ, Chen W-M, et al. (2008) Common variants in the GDF5-UQCC region are associated with variation in human height. *Nat Genet* 40(2):198–203.

[100] Smemo S, Tena JJ, Kim K-H, Gamazon ER, Sakabe NJ, et al. (2014) Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature* 507(7492):371–375.

[101] El Ghouzzi V, Dagoneau N, Kinning E, Thauvin-Robinet C, Chemaitilly W, et al. (2003) Mutations in a novel gene Dymeclin (FLJ20071) are responsible for Dyggve-Melchior-Clausen syndrome. *Hum Mol Genet* 12(3):357–64.

[102] Cohn DH, Ehtesham N, Krakow D, Unger S, Shanske A, et al. (2003) Mental retardation and abnormal skeletal development (Dyggve-Melchior-Clausen dysplasia) due to mutations in a novel, evolutionarily conserved gene. *Am J Hum Genet* 72(2):419–28.

[103] Heinemann HO, Goldring RM (1974) Bicarbonate and the regulation of ventilation. *Am J Med* 57(3):361–370.

[104] Okbay A, Beauchamp JP, Fontana MA, Lee JJ, Pers TH, et al. (2016) Genome-wide association study identifies 74 loci associated with educational attainment. *Nature* 533(7604):539–42.

[105] Enomoto A, Kimura H, Chairoungdua A, Shigeta Y, Jutabha P, et al. (2002) Molecular identification of a renal urate–anion exchanger that regulates blood urate levels. *Nature* 417(6887):447–52.

[106] Hagos Y, Stein D, Ugele B, Burckhardt G, Bahn A (2007) Human Renal Organic Anion Transporter 4 Operates as an Asymmetric Urate Transporter. *J Am Soc Nephrol* 18(2):430–439.

[107] Köttgen A, Albrecht E, Teumer A, Vitart V, Krumsiek J, et al. (2012) Genome-wide association analyses identify 18 new loci associated with serum urate concentrations. *Nat Genet* 45(2):145–154.

[108] Tall AR, Yvan-Charvet L (2015) Cholesterol, inflammation and innate immunity. *Nat Rev Immunol* 15(2):104–16.

[109] Rouch AJ, Kudo LH (2000) Role of PGE2 in α2-induced inhibition of AVP- and cAMP-stimulated H2O, Na+, and urea transport in rat IMCD. *Am J Physiol - Ren Physiol* 279(2).

[110] Walter K, Min JL, Huang J, Crooks L, Memari Y, et al. (2015) The UK10K project identifies rare variants in health and disease. *Nature* 526(7571):82–90.

[111] Simola K, Karli P, De La Chapelle A (1977) Two pericentric inversions of human chromosome 11. *J Med Genet* 14(5):371–4.

[112] Feuk L (2010) Inversion variants in the human genome: role in disease and genome architecture. *Genome Med* 2(2):11.

[113] Puig M, Casillas S, Villatoro S, Cáceres M (2015) Human inversions and their functional consequences. *Brief Funct Genomics* 14(5):369–79.

[114] Kolz M, Johnson T, Sanna S, Teumer A, Vitart V, et al. (2009) Meta-analysis of 28,141 individuals identifies common variants within five new loci that influence uric acid concentrations. *PLoS Genet* 5(6):e1000504.

[115] Kotowski IK, Pertsemlidis A, Luke A, Cooper RS, Vega GL, et al. (2006) A spectrum of PCSK9 alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am J Hum Genet* 78(3):410–22.

[116] Ott J, Hoh J (2000) Statistical approaches to gene mapping. *Am J Hum Genet* 67(2):289–94.

[117] Dawn Teare M, Barrett JH, Gallo A, Al. E, Kendler K, Bouchard C (2003) Genetic linkage studies. *Lancet* 366(9490):1036–44.

[118] Pulst SM (1999) Genetic Linkage Analysis. *Arch Neurol* 56(6):667–72.

[119] O'Connell JR (2001) Rapid multipoint linkage analysis via inheritance vectors in the Elston-Stewart algorithm. *Hum Hered* 51(4):226–40.

[120] Han L, Abney M (2011) Identity by descent estimation with dense genome-wide genotype data. *Genet Epidemiol* 35(6):557–67.

[121] Tsui L-C, Dorfman R (2013) The cystic fibrosis gene: a molecular genetic perspective. *Cold Spring Harb Perspect Med* 3(2):a009472.

[122] Almasy L, Blangero J (1998) Multipoint quantitative-trait linkage analysis in general pedigrees. *Am J Hum Genet* 62(5):1198–211.

[123] Kosambi DD (1943) The estimation of map distances from recombination values. *Ann Eugen* 12(1):172–175.

[124] Jacquard A (1974) *The Genetic Structure of Populations* (Springer Berlin Heidelberg, Berlin, Heidelberg).

[125] Fisher RA (2006) *Statistical methods for research workers* (Published by Cosmo Publications for Genesis Pub).

[126] Thompson EA (2013) Identity by descent: variation in meiosis, across genomes, and in populations. *Genetics* 194(2):301–26.

[127] Gusev A, Palamara PF, Aponte G, Zhuang Z, Darvasi A, et al. (2012) The architecture of long-range haplotypes shared within and across populations. *Mol Biol Evol* 29(2):473–86.

[128] Glodzik D, Navarro P, Vitart V, Hayward C, McQuillan R, et al. (2013) Inference of identity by descent in population isolates and optimal sequencing studies. *Eur J Hum Genet* 21(10):1140–1145.

[129] Kaback DB, Barber D, Mahon J, Lamb J, You J (1999) Chromosome size-dependent control of meiotic reciprocal recombination in Saccharomyces cerevisiae: the role of

crossover interference. *Genetics* 152(4):1475–86.

[130] King JS, Mortimer RK (1990) A polymerization model of chiasma interference and corresponding computer simulation. *Genetics* 126(4):1127–38.

[131] Solberg OD, Mack SJ, Lancaster AK, Single RM, Tsai Y, et al. (2008) Balancing selection and heterogeneity across the classical human leukocyte antigen loci: a meta-analytic review of 497 population studies. *Hum Immunol* 69(7):443–64.

[132] Baird DM, Coleman J, Rosser ZH, Royle NJ (2000) High Levels of Sequence Polymorphism and Linkage Disequilibrium at the Telomere of 12q: Implications for Telomere Biology and Human Evolution. *Am J Hum Genet* 66(1):235–250.

[133] Rodriguez JM, Bercovici S, Huang L, Frostig R, Batzoglou S (2015) Parente2: a fast and accurate method for detecting identity by descent. *Genome Res* 25(2):280–9.

[134] Albrechtsen A, Moltke I, Nielsen R (2010) Natural Selection and the Distribution of Identity-by-Descent in the Human Genome. *Genetics* 186(1).

[135] Westerlind H, Imrell K, Ramanujam R, Myhr K-M, Celius EG, et al. (2015) Identity-by-descent mapping in a Scandinavian multiple sclerosis cohort. *Eur J Hum Genet* 23(5):688–692.

[136] Mahtani MM, Willard HF (1998) Physical and genetic mapping of the human X chromosome centromere: repression of recombination. *Genome Res* 8(2):100–10.

[137] Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. *PLoS Biol* 4(3):e72.

[138] Amador C, Huffman J, Trochet H, Campbell A, Porteous D, et al. (2015) Recent genomic heritage in Scotland. *BMC Genomics* 16(1):437.

[139] Conrad DF, Hurles ME (2007) The population genetics of structural variation. *Nat Genet* 39(7 Suppl):S30-6.

[140] Tishkoff SA, Reed FA, Ranciaro A, Voight BF, Babbitt CC, et al. (2007) Convergent adaptation of human lactase persistence in Africa and Europe. *Nat Genet* 39(1):31–40.

[141] Blangero J, Williams JT, Almasy L (2003) Novel family-based approaches to genetic risk in thrombosis. *J Thromb Haemost* 1(7):1391–7.

[142] Visscher PM, Medland SE, Ferreira MAR, Morley KI, Zhu G, et al. (2006) Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genet* 2(3):e41.

[143] Shin Y-B, Lim J-E, Jin H-S, Hong K-W, Oh B (2012) Characterization of the ATP2B gene family in blood pressure. *Genes Genomics* 34(5):539–547.

[144] Rankinen T, Sung YJ, Sarzynski MA, Rice TK, Rao DC, Bouchard C (2012) Heritability of submaximal exercise heart rate response to exercise training is accounted for by nine SNPs. *J Appl Physiol* 112(5):892–7.

[145] Liu H, Tang Y, Liu X, Zhou Q, Xiao X, et al. (2014) 14-3-3 tau (YWHAQ) gene promoter hypermethylation in human placenta of preeclampsia. *Placenta* 35:981–988.

[146] Paschos P, Paletas K (2009) Non alcoholic fatty liver disease and metabolic syndrome. *Hippokratia* 13(1):9–19.

[147] Kasapoglu B, Turkay C, Bayram Y, Koca C (2010) Role of GGT in diagnosis of metabolic syndrome: a clinic-based cross-sectional survey. *Indian J Med Res* 132:56–61.

[148] Rietveld CA, Esko T, Davies G, Pers TH, Turley P, et al. (2014) Common genetic variants associated with cognitive performance identified using the proxy-phenotype method. *Proc Natl Acad Sci U S A* 111(38):13790–4.

[149] Berto S, Usui N, Konopka G, Fogel BL (2016) ELAVL2-regulated transcriptional and splicing networks in human neurons link neurodevelopment and autism. *Hum Mol Genet* 25(12):2451–2464.

[150] Scheckel C, Drapeau E, Frias MA, Park CY, Fak J, et al. (2016) Regulatory consequences of neuronal ELAV-like protein binding to coding and non-coding RNAs in human brain. *Elife* 5.

[151] Li Y, Nowotny P, Holmans P, Smemo S, Kauwe JSK, et al. (2004) Association of late-onset Alzheimer's disease with genetic variation in multiple members of the GAPD gene family. *Proc Natl Acad Sci U S A* 101(44):15688–93.

[152] Rogaeva E, Premkumar S, Song Y, Sorbi S, Brindle N, et al. (1998) Evidence for an Alzheimer disease susceptibility locus on chromosome 12 and for further locus heterogeneity. *JAMA* 280(7):614–8.

[153] Kehoe P, Wavrant-De Vrieze F, Crook R, Wu WS, Holmans P, et al. (1999) A full genome scan for late onset Alzheimer's disease. *Hum Mol Genet* 8(2):237–45.

[154] Harashima S, Horiuchi T, Wang Y, Notkins AL, Seino Y, Inagaki N (2012) Sorting nexin 19 regulates the number of dense core vesicles in pancreatic β-cells. *J Diabetes Investig* 3(1):52–61.

[155] DeFronzo RA, Cooke CR, Andres R, Faloona GR, Davis PJ (1975) The effect of insulin on renal handling of sodium, potassium, calcium, and phosphate in man. *J Clin Invest* 55(4):845–55.

[156] Tiwari S, Riazi S, Ecelbarger CA (2007) Insulin's impact on renal sodium transport and blood pressure in health, obesity, and diabetes. *Am J Physiol - Ren Physiol* 293(4):F974-84.

[157] Yuan Q, Bie J, Wang J, Ghosh SS, Ghosh S (2013) Cooperation between hepatic cholesteryl ester hydrolase and scavenger receptor BI for hydrolysis of HDL-CE. *J Lipid Res* 54(11):3078–84.

[158] Sato M, Kawakami T, Kondoh M, Takiguchi M, Kadota Y, et al. (2010) Development of high-fat-diet-induced obesity in female metallothionein-null mice. *FASEB J* 24(7):2375–2384.

[159] Newton-Cheh C, Johnson T, Gateva V, Tobin MD, Bochud M, et al. (2009) Genome-wide association study identifies eight loci associated with blood pressure. *Nat Genet* 41(6):666–676.

[160] Lim Y-H, Han C, Bae S, Hong Y-C (2017) Modulation of blood pressure in response to low ambient temperature: The role of DNA methylation of zinc finger genes. *Environ*

Res 153:106–111.

[161] Wang G, Zuo X, Liu J, Jiang L, Liu Y, et al. (2009) Expression of Mipu1 in response to myocardial infarction in rats. *Int J Mol Sci* 10(2):492–506.

[162] Voruganti VS, Göring HHH, Mottl A, Franceschini N, Haack K, et al. (2009) Genetic influence on variation in serum uric acid in American Indians: the strong heart family study. *Hum Genet* 126(5):667–76.

[163] Tin A, Woodward OM, Kao WHL, Liu C-T, Lu X, et al. (2011) Genome-wide association study for serum urate concentrations and gout among African Americans identifies genomic risk loci and a novel URAT1 loss-of-function allele. *Hum Mol Genet* 20(20):4056–68.

[164] Kamatani Y, Matsuda K, Okada Y, Kubo M, Hosono N, et al. (2010) Genome-wide association study of hematological and biochemical traits in a Japanese population. *Nat Genet* 42(3):210–5.

[165] Gill J, Endres-Brooks J, Bauer P, Marks WJ, Montgomery R (1987) The effect of ABO blood group on the diagnosis of von Willebrand disease. *Blood* 69(6):1691–1695.

[166] Smith NL, Chen M-H, Dehghan A, Strachan DP, Basu S, et al. (2010) Novel associations of multiple genetic loci with plasma levels of factor VII, factor VIII, and von Willebrand factor: The CHARGE (Cohorts for Heart and Aging Research in Genome Epidemiology) Consortium. *Circulation* 121(12):1382–92.

[167] Desch KC, Ozel AB, Siemieniak D, Kalish Y, Shavit JA, et al. (2013) Linkage analysis identifies a locus for plasma von Willebrand factor undetected by genome-wide association. *Proc Natl Acad Sci U S A* 110(2):588–93.

[168] Li Q, Wojciechowski R, Simpson CL, Hysi PG, Verhoeven VJM, et al. (2015) Genome-wide association study for refractive astigmatism reveals genetic co-determination with spherical equivalent refractive error: the CREAM consortium. *Hum Genet* 134(2):131–46.

[169] Pickrell JK, Berisa T, Liu JZ, Ségurel L, Tung JY, Hinds DA (2016) Detection and interpretation of shared genetic influences on 42 human traits. *Nat Genet* 48(7):709–17.

[170] Simpson CL, Wojciechowski R, Oexle K, Murgia F, Portas L, et al. (2014) Genome-wide meta-analysis of myopia and hyperopia provides evidence for replication of 11 loci. *PLoS One* 9(9):e107110.

[171] Verhoeven VJM, Hysi PG, Wojciechowski R, Fan Q, Guggenheim JA, et al. (2013) Genome-wide meta-analyses of multiancestry cohorts identify multiple new susceptibility loci for refractive error and myopia. *Nat Genet* 45(3):314–8.

[172] Ye XC, Pegado V, Patel MS, Wasserman WW (2014) Strabismus genetics across a spectrum of eye misalignment disorders. *Clin Genet* 86(2):103–11.

[173] Wagner AH, Anand VN, Wang W-H, Chatterton JE, Sun D, et al. (2013) Exon-level expression profiling of ocular tissues. *Exp Eye Res* 111:105–11.

[174] Wilkinson B, Chen JY-F, Han P, Rufner KM, Goularte OD, Kaye J (2002) TOX: an HMG box protein implicated in the regulation of thymocyte selection. *Nat Immunol*

3(3):272–280.

[175] Marquardt T, Shirasaki R, Ghosh S, Andrews SE, Carter N, et al. (2005) Coexpressed EphA receptors and ephrin-A ligands mediate opposing actions on growth cone navigation from distinct membrane domains. *Cell* 121(1):127–39.

[176] Summers Rada JA, Shelton S, Norton TT (2006) The sclera and myopia. *Exp Eye Res* 82(2):185–200.

[177] Zabaneh D, Krapohl E, Simpson MA, Miller MB, Iacono WG, et al. (2017) Fine mapping genetic associations between the HLA region and extremely high intelligence. *Sci Rep* 7:41182.

[178] Park M-H, Kwak SH, Kim KJ, Go MJ, Lee H-J, et al. (2013) Identification of a genetic locus on chromosome 4q34-35 for type 2 diabetes with overweight. *Exp Mol Med* 45(2):e7.

[179] Ahn MY, Jee SD, Lee BM, Yeon J-H, Park K-K, et al. (2010) Antidiabetic Effects and Gene Expression Profiling in Obese Mice Treated With *Isaria sinclairii* Over a 6-Month Period. *J Toxicol Environ Heal Part A* 73(21–22):1511–1520.

[180] Pettus BJ, Kitatani K, Chalfant CE, Taha TA, Kawamori T, et al. (2005) The Coordination of Prostaglandin E2 production by sphingosine-1-phosphate and ceramide-1-phosphate. *Mol Pharmacol* 68(2):330–5.

[181] Paajanen TA, Oksala NKJ, Kuukasjärvi P, Karhunen PJ (2010) Short stature is associated with coronary heart disease: a systematic review of the literature and a meta-analysis. *Eur Heart J* 31(14):1802–9.

[182] Nelson CP, Hamby SE, Saleheen D, Hopewell JC, Zeng L, et al. (2015) Genetically determined height and coronary artery disease. *N Engl J Med* 372(17):1608–18.

[183] Riggio V, Pong-Wong R (2014) Regional Heritability Mapping to identify loci underlying genetic variation of complex traits. *BMC Proc* 8(Suppl 5):S3.

[184] Zeng Y, Navarro P, Shirali M, Howard DM, Adams MJ, et al. (2016) Genome-wide Regional Heritability Mapping Identifies a Locus Within the TOX2 Gene Associated With Major Depressive Disorder. *Biol Psychiatry* 82(5):312–321.

[185] Teslovich TM, Musunuru K, Smith A V., Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466(7307):707–713.

[186] Cebamanos L, Gray a, Stewart I, Tenesa a (2014) Regional heritability advanced complex trait analysis for GPU and traditional parallel architectures. *Bioinformatics*:2–4.

[187] Canela-Xandri O, Law A, Gray A, Woolliams JA, Tenesa A (2015) A new tool called DISSECT for analysing large genomic data sets using a Big Data approach. *Nat Commun* 6:10162.

[188] Frayling TM, Timpson NJ, Weedon MN, Zeggini E, Freathy RM, et al. (2007) A Common Variant in the FTO Gene Is Associated with Body Mass Index and Predisposes to Childhood and Adult Obesity. *Science (80-)* 316(5826):889–894.

[189] Scuteri A, Sanna S, Chen W-M, Uda M, Albai G, et al. (2007) Genome-Wide Association

Scan Shows Genetic Variants in the FTO Gene Are Associated with Obesity-Related Traits. *PLoS Genet* 3(7):e115.

[190] Dina C, Meyre D, Gallina S, Durand E, K?rner A, et al. (2007) Variation in FTO contributes to childhood obesity and severe adult obesity. *Nat Genet* 39(6):724–726.

[191] Okada Y, Sim X, Go MJ, Wu J-Y, Gu D, et al. (2012) Meta-analysis identifies multiple loci associated with kidney function–related traits in east Asian populations. *Nat Genet* 44(8):904–909.

[192] Eppinga RN, Hagemeijer Y, Burgess S, Hinds DA, Stefansson K, et al. (2016) Identification of genomic loci associated with resting heart rate and shared genetic predictors with all-cause mortality. *Nat Genet* 48(12):1557–1563.

[193] Pattaro C, De Grandi A, Vitart V, Hayward C, Franke A, et al. (2010) A meta-analysis of genome-wide data from five European isolates reveals an association of COL22A1, SYT1, and GABRR2with serum creatinine level. *BMC Med Genet* 11(1):41.

[194] Chambers JC, Zhang W, Lord GM, van der Harst P, Lawlor DA, et al. (2010) Genetic loci influencing kidney function and chronic kidney disease. *Nat Genet* 42(5):373–375.

[195] Köttgen A, Pattaro C, B?ger CA, Fuchsberger C, Olden M, et al. (2010) New loci associated with kidney function and chronic kidney disease. *Nat Genet* 42(5):376–384.

[196] Pattaro C, Teumer A, Gorski M, Chu AY, Li M, et al. (2016) Genetic associations at 53 loci highlight cell types and biological pathways relevant for kidney function. *Nat Commun* 7:10023.

[197] Lander E, Kruglyak L (1995) Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 11(3):241–247.

[198] Karasik D, Dupuis J, Cho K, Cupples LA, Zhou Y, et al. (2010) Refined QTLs of osteoporosis-related traits by linkage analysis with genome-wide SNPs: Framingham SHARe. *Bone* 46(4):1114–1121.

[199] Yang J, Lee SH, Goddard ME, Visscher PM (2011) GCTA: a tool for genome-wide complex trait analysis. *Am J Hum Genet* 88(1):76–82.

[200] Speed D, Balding DJ (2014) Relatedness in the post-genomic era: is it still useful? *Nat Rev Genet* 16(1):33–44.

[201] Speed D, Hemani G, Johnson MR, Balding DJ, Koellinger PD, et al. (2012) Improved heritability estimation from genome-wide SNPs. *Am J Hum Genet* 91(6):1011–21.

[202] Astle W, Balding DJ (2009) Population Structure and Cryptic Relatedness in Genetic Association Studies. *Stat Sci* 24(4):451–471.

[203] Blangero J, Almasy L (1997) Multipoint oligogenic linkage analysis of quantitative traits. *Genet Epidemiol* 14(6):959–964.

[204] Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era — concepts and misconceptions. *Nat Rev Genet* 9(4):255–266.

[205] Ohira T, Cushman M, Tsai MY, Zhang Y, Heckbert SR, et al. (2007) ABO blood group, other risk factors and incidence of venous thromboembolism: the Longitudinal Investigation of Thromboembolism Etiology (LITE). *J Thromb Haemost* 5(7):1455–

1461.

[206] van Loon J, Dehghan A, Weihong T, Trompet S, McArdle WL, et al. (2016) Genome-wide association studies identify genetic loci for low von Willebrand factor levels. *Eur J Hum Genet* 24(7):1035–40.

[207] Sudlow C, Gallacher J, Allen N, Beral V, Burton P, et al. (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS Med* 12(3):e1001779.

[208] Bycroft C, Freeman C, Petkova D, Band G, Elliott LT, et al. (2017) Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv*.

[209] Canela-Xandri O, Rawlik K, Tenesa A (2017) An atlas of genetic associations in UK Biobank. *bioRxiv*.

[210] Xia C, Amador C, Huffman J, Trochet H, Campbell A, et al. (2016) Pedigree- and SNP-Associated Genetics and Recent Environment are the Major Contributors to Anthropometric and Cardiometabolic Trait Variation. *PLOS Genet* 12(2):e1005804.

[211] Marioni RE, Campbell A, Hagenaars SP, Nagy R, Amador C, et al. (2017) Genetic Stratification to Identify Risk Groups for Alzheimer's Disease. *J Alzheimer's Dis* 57(1):275–283.

[212] Marioni RE, Ritchie SJ, Joshi PK, Hagenaars SP, Okbay A, et al. (2016) Genetic variants linked to education predict longevity. *Proc Natl Acad Sci* 113(47):13366–13371.

[213] Zeng Y, Navarro P, Xia C, Amador C, Fernandez-Pujals AM, et al. (2016) Shared Genetics and Couple-Associated Environment Are Major Contributors to the Risk of Both Clinical and Self-Declared Depression. *EBioMedicine* 14:161–167.

[214] Amador C, Xia C, Nagy R, Campbell A, Porteous DJ, et al. (2017) Regional variation in health is predominantly driven by lifestyle rather than genetics. *Nat Commun* (In Press).

[215] Kirin M, Nagy R, MacGillivray TJ, Polašek O, Hayward C, et al. (2017) Determinants of retinal microvascular features and their relationships in two European populations. *J Hypertens* 35(8):1646–1659.

[216] Spiliopoulou A, Nagy R, Bermingham ML, Huffman JE, Hayward C, et al. (2015) Genomic prediction of complex human traits: relatedness, trait architecture and predictive meta-models. *Hum Mol Genet* 24(14):4167–4182.

Supplementary Tables

Supplementary Table 1 - Hits that exceed the suggestive but not the genome-wide significance threshold in cohort-specific GWAS following HRC imputation

This table summarizes the hits that passed the suggestive significance threshold (logP>7.3) in the GWAS using imputed genotypes, providing their -$\log_{10}$(*p-value*), effect size and its standard error (Beta and Beta_SE columns), the allele for which the effect size is reported (EA column) as well as the cohort and trait-specific frequency of this allele (EAF column). The nhits column indicates the number of SNPs within 500kb of the reported SNP that exceeded the cohort-specific GWS threshold in the GWAS using imputed genotypes. Within this 1Mb interval, the position and -log10 (p-value) of the SNP with the most significant test statistic in the GWAS using genotyped SNPs is shown (Pos_G and logP_G columns). The name of, and distance to, the gene closest to the reported SNP is indicated – the distance is 0 if the SNP lies within the gene itself. The final column indicates whether other GWAS have identified this hit before. The first value indicates whether any SNPs in the 1000 Genomes data that are in strong LD with the reported SNP (R^2 and D' > 0.8) have been identified with other GWAS, while the second value looks at all SNPs within 100kb of the reported SNP, regardless of LD.

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Orkney														
BMI	1	232053321	rs17821689	7.31	0.593	0.1087	G	0.0269	1	232477033	1.22	*DISC1*	0	0|0
CRP	1	159665921	rs2808624	7.78	-0.216	0.0383	G	0.3703	5	159652939	7.45	*CRP*	16156	1|1
CRP	6	33035969	rs3179779	7.37	-0.483	0.0883	G	0.0633	1	33026246	3.28	*HLA-DPA1*	0	0|0
CRP	19	45411941	rs429358	7.85	-0.278	0.0489	C	0.1806	2	45395619	6.69	*APOE*	0	1|1
Diastolic BP	7	5914756	rs146713555	8.44	6.451	1.0936	T	0.0217	3	6353371	2.05	*OCM*	5671	0|0
Forced Vital Capacity	9	76961908	rs62550438	7.99	0.336	0.0587	T	0.0922	2	76786297	3.14	*RORB*	150342	0|0
GGT	22	24996582	rs2330795	7.58	0.274	0.0492	A	0.3561	6	25368543	2.84	*GGT1*	0	1|1
Total Cholesterol	19	45412079	rs7412	8.65	-0.373	0.0624	T	0.0764	2	45395619	4.31	*APOE*	0	1|1
Triglycerides	11	116648917	rs964184	7.86	-0.143	0.0252	C	0.866	1	116621963	4.51	*ZNF259*	357	1|1
Uric acid1	11	49346641	rs187788041	8.36	-0.971	0.1654	A	0.0103	9	49627833	0.92	*FOLH1*	116418	0|0
Uric acid1	11	56679210	rs189737773	8.77	-0.991	0.1644	A	0.0117	3	56466099	1.64	*OR5AK2*	77177	0|0

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Uric acid1	11	59901327	rs181316004	8.11	-0.887	0.1536	C	0.0134	1	59824526	1.58	*MS4A2*	35386	0|0
Uric acid1	11	71909897	rs117521416	7.78	-0.9	0.1595	A	0.0112	2	72107188	2.45	*FOLR1*	2529	0|0
Uric acid2	11	46397253	rs147604011	7.48	-0.974	0.1762	G	0.0119	1	45975130	1.91	*DGKZ*	0	0|0
Uric acid2	11	50720455	rs186997623	7.69	-0.902	0.1608	C	0.0136	1	50522307	0.34	*LOC646813*	340652	0|0
Uric acid2	11	57989542	rs202065992	8.09	-0.885	0.1535	A	0.0193	1	58004159	1.58	*OR10Q1*	5810	0|0
Uric acid2	11	59901327	rs181316004	7.44	-0.889	0.1613	C	0.0141	1	59738667	1.21	*MS4A2*	35386	0|0
Uric acid2	11	67512382	rs144771122	8.31	-0.822	0.1404	T	0.0179	1	67423892	2.35	*LOC645332*	46854	0|0
Uric acid2	11	71120029	rs117991852	8.02	-0.892	0.1554	T	0.0145	1	70819038	2.15	*LOC339902*	0	0|0
Uric acid2	11	72900306	rs79750124	7.97	-0.835	0.1460	T	0.017	1	72414189	3.33	*P2RY2*	29036	0|0
Waist	11	65356754	rs948578	7.55	0.186	0.0335	T	0.4952	1	65494987	3.68	*EHBP1L1*	0	0|0
Vis														
Calcium	10	77964415	rs74584091	7.78	0.487	0.0863	T	0.1369	1	77675236	3.19	*C10orf11*	0	0|0
GGT	14	103569748	rs72706640	7.56	0.329	0.0592	C	0.2337	1	103756099	2.23	*EXOC3L4*	0	1|1
GGT	22	24996582	rs2330795	7.52	0.293	0.0529	A	0.4144	8	24991144	1.77	*GGT1*	0	1|1
Glucose	20	22847511	rs978411	7.62	0.726	0.1302	C	0.0553	1	22372227	2.4	*SSTR4*	168544	0|0
Glucose_nodiab	12	51504172	rs10876137	7.82	0.455	0.0804	T	0.1086	3	51479682	2.19	*TFCP2*	0	0|0
GPT	11	76660793	rs56789915	8.3	0.276	0.0472	C	0.5814	35	76414893	8.1	*ACER3*	0	0|0
Height	5	101606071	rs2600831	7.81	0.018	0.0031	A	0.3915	7	101688812	4.31	*SLCO4C1*	0	0|0
LDL	19	45412079	rs7412	8.17	-0.505	0.0871	T	0.095	1	45879279	2.19	*APOE*	0	1|1
Shetland														
Alcohol consumption	12	107068030	rs149828296	7.68	-0.583	0.1041	C	0.0259	1	106931703	2.66	*RFX4*	0	0|0
BMI	21	22896160	rs233799	8.08	0.201	0.0349	A	0.6637	2	22896160	7.96	*NCAM2*	0	0|0
Educational Attainment	1	105502253	rs74954432	7.57	3.541	0.6367	G	0.0099	1	105067105	2.18	*LOC100129138*	882559	0|0
Forced Vital Capacity	10	106409752	rs2864026	7.48	-0.204	0.0370	G	0.6238	1	106409545	6.09	*SORCS3*	0	0|0

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
GGT	22	24994708	rs2017188	8.47	0.204	0.0345	C	0.3369	15	24992266	7.09	*GGT1*	0	1|1
HbA1c	6	162977834	rs116884606	7.59	0.487	0.0875	T	0.0401	2	163270143	3.55	*PARK2*	0	0|0
Heart Rate	3	31330858	rs9859166	7.47	4.363	0.7904	T	0.0409	6	31329185	7.04	*STT3B*	243631	0|0
Height	3	141082990	rs13100711	8.54	0.012	0.0020	T	0.3259	8	141074962	7.67	*ZBTB38*	0	0|1
Height	13	29559059	rs147382386	7.89	0.054	0.0096	G	0.0107	1	29063969	2.64	*MTUS2*	39687	0|0
Waist	19	17483973	rs79607374	8.7	-0.587	0.0978	A	0.0314	1	17000632	4.69	*PLVAP*	0	0|0
Korčula														
Albumin	4	73515987	rs115058973	8.17	0.85	0.1467	A	0.0103	1	73618440	2.12	*ADAMTS3*	81470	0|0
Albumin	6	8344289	rs149984974	7.51	0.655	0.1183	C	0.0184	1	7924658	2.8	*SLC35B3*	69010	0|0
Albumin	15	49771796	rs12901446	7.48	-0.492	0.0890	T	0.0266	1	49562266	1.95	*FGF7*	0	0|0
Central Corneal Thickness	16	88331515	rs9934580	7.31	0.285	0.0523	G	0.5972	1	88298124	6.63	*ZNF469*	162362	1|1
fev1perfvc	9	78732604	rs114200252	8.18	0.357	0.0615	A	0.1044	1	78688079	1.98	*PCSK5*	0	0|0
fev1perfvc	10	16711372	rs7092141	7.64	-0.272	0.0486	C	0.1104	1	16556710	1.82	*RSU1*	0	0|0
fev1perfvc	12	94417315	rs12146740	8.35	-0.359	0.0612	A	0.0621	1	94374805	1.73	*PLXNC1*	125182	0|0
fev1perfvc	14	68861465	rs7145422	8.47	-0.224	0.0379	T	0.7028	3	69203145	1.67	*RAD51L1*	0	0|0
fev1perfvc	16	31023414	rs72800849	7.62	0.186	0.0333	A	0.4262	1	31289396	1.27	*STX1B*	1584	0|0
Triglycerides	9	88921734	rs117621872	7.39	-0.236	0.0430	A	0.0319	1	89035754	1.43	*ZCCHC6*	0	0|0
GS														
BMI	2	228006255	rs10498218	7.4	-0.942	0.1715	G	0.0012	1	228336315	2.43	*COL4A4*	0	0|0
BMI	15	39654281	rs149913955	7.66	-0.463	0.0827	A	0.0059	2	39458415	2.36	*C15orf54*	107232	0|0
BMI	16	17177110	rs571835655	8.18	-1.06	0.1827	A	0.0011	1	17065473	2.09	*XYLT1*	18516	0|0
Body fat	1	192893075	rs10921288	7.98	1.466	0.2561	C	0.0237	7	192891048	7.8	*UCHL5*	88419	0|0
Body fat	2	227660641	rs142101835	7.49	5.537	1.0020	G	0.0022	1	227521410	2.12	*IRS1*	0	0|0
Body fat	16	17177110	rs571835655	7.69	-7.252	1.2930	A	0.0011	1	17549067	2.35	*XYLT1*	18516	0|0

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Body fat	18	46837578	rs141793746	7.48	4.318	0.7815	T	0.0029	1	46554787	2.66	*DYM*	0	0|0
Creatinine	1	234420503	rs573421908	7.87	0.726	0.1278	A	0.0027	1	234452980	3.12	*SLC35F3*	0	0|0
Creatinine	4	9542350	rs62412107	7.73	0.152	0.0270	A	0.0653	2	9443894	1.93	*MIR548I2*	15437	0|0
Educational Attainment	2	113518931	rs187728767	7.43	1.979	0.3595	A	0.0027	1	113807144	3.13	*CKAP2L*	0	0|0
Educational Attainment	9	99112276	rs753676304	7.49	2.302	0.4163	G	0.002	1	99085363	2.68	*SLC35D2*	0	0|0
Educational Attainment	12	54690428	rs150528113	7.44	2.279	0.4138	C	0.0022	1	54568590	3.99	*NFE2*	0	0|0
FEV1	19	54243719	rs117764849	7.63	0.339	0.0607	C	0.0155	1	53939690	4.09	*MIR517C*	846	0|0
Forced Vital Capacity	2	186162246	rs182694754	8.5	-0.598	0.1010	T	0.0041	15	185926550	1.94	*ZNF804A*	358031	0|0
Forced Vital Capacity	9	19468530	rs7021552	7.32	-0.071	0.0130	T	0.308	1	19468530	7.26	*ACER2*	16029	0|0
Glucose	10	3328318	rs533883198	7.41	0.683	0.1243	A	0.0026	1	3442533	2.77	*PITRM1*	113315	0|0
Glucose	13	28499962	rs7981781	7.86	0.079	0.0139	A	0.2334	22	28491198	7.5	*PDX1*	0	1|1
Glucose	21	44276432	rs370189685	8.14	-1.176	0.2033	C	0.0013	1	44711272	4.93	*WDR4*	0	0|0
Glucose_nodiab	1	214150821	rs79687284	7.73	0.204	0.0363	C	0.0364	1	214173840	2.61	*PROX1*	11037	0|1
Glucose_nodiab	2	27741105	rs780095	8.09	0.071	0.0123	G	0.5758	11	27741237	7.53	*GCKR*	0	1|1
Glucose_nodiab	9	96182703	rs143399767	7.85	0.362	0.0638	C	0.0103	1	96682611	1.96	*FAM120AOS*	26077	0|0
Glucose_nodiab	21	44276432	rs370189685	7.56	-1.15	0.2070	C	0.0013	1	44711272	3.29	*WDR4*	0	0|0
HDL	1	8233270	rs149963466	7.5	0.868	0.1569	A	0.0016	1	8280517	2.19	*ERRFI1*	146876	0|0
HDL	2	43255645	rs76183280	7.38	-0.498	0.0908	C	0.0048	1	43314843	2.33	*ZFP36L2*	193894	0|0
HDL	8	9183596	rs4841132	7.97	0.109	0.0191	G	0.9079	4	9177732	6.14	*LOC157273*	0	1|1
HDL	9	107661742	rs2740488	7.6	-0.069	0.0124	C	0.2743	3	107666513	7.53	*ABCA1*	0	1|1
HDL	14	106824367	rs2338129	8.16	-0.653	0.1127	A	0.0038	1	106846077	2.18	*NCRNA00226*	79399	0|0
HDL	18	47106028	rs149615216	8.5	0.302	0.0509	T	0.0121	5	47179516	5.41	*LIPG*	0	0|1
HDL	19	54809343	rs453755	7.49	0.074	0.0133	G	0.2489	2	54797848	6.08	*LILRA3*	5077	0|0

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
HDL	20	44538484	rs435306	7.54	-0.07	0.0127	T	0.7452	1	44545048	7.11	*PLTP*	0	0|1
Heart Rate	1	6296238	rs9970334	7.36	0.697	0.1272	G	0.4474	1	6278414	3.79	*ICMT*	193	0|1
Heart Rate	3	87751558	rs755291044	7.74	8.6	1.5280	A	0.0017	1	87910931	2.52	*HTR1F*	288288	0|0
Heart Rate	8	4102424	rs145669495	7.7	7.665	1.3660	G	0.0022	1	3797879	2.39	*CSMD1*	0	0|0
Heart Rate	8	62481520	rs142916219	7.66	5.971	1.0670	G	0.0038	1	62517963	2.34	*ASPH*	0	0|0
Height	1	177404510	rs146949893	7.35	0.039	0.0071	C	0.0031	1	177129230	2.43	*FAM5B*	152951	0|0
Height	1	182706092	rs558671668	7.6	0.027	0.0048	C	0.0062	1	182971816	2.78	*NPL*	52336	0|0
Height	4	17994298	rs35362908	8.15	-0.008	0.0013	C	0.1005	1	17957354	6.44	*LCORL*	0	1|1
Height	4	86496830	rs552283803	7.4	-0.053	0.0096	C	0.0016	1	86751384	2.81	*AHRGAP24*	0	0|0
Height	5	32773275	rs72742734	7.44	0.008	0.0015	G	0.0538	1	32765457	4.69	*NPR3*	0	0|1
Height	7	113267859	rs184469050	7.8	-0.022	0.0038	G	0.0088	1	113333114	2.96	*PPP1R3A*	249021	0|0
Height	11	75282052	rs634552	7.64	-0.006	0.0010	G	0.8635	5	75276178	7.35	*SERPINH1*	0	1|1
Height	12	4384844	rs76895963	7.46	0.012	0.0023	G	0.0285	1	4071901	2.02	*CCND2*	0	0|0
Height	12	93956672	rs11614062	7.68	0.005	0.0009	T	0.1944	4	93962959	7.12	*LOC144481*	2730	0|1
Sodium	5	175615619	rs144466366	7.5	0.84	0.1517	T	0.0017	1	175248342	1.85	*LOC643201*	0	0|0
Sodium	9	93886686	rs140344432	8.55	0.98	0.1649	C	0.0015	1	94179978	3.01	*LOC100129316*	49271	0|0
Sodium	14	103031952	rs118075232	8.44	-0.367	0.0622	A	0.0102	1	103244840	3.53	*MIR4309*	25888	0|0
Sodium	16	69548788	rs7200764	8.12	-0.085	0.0147	C	0.8403	25	69559696	8.02	*CYB5B*	48620	0|0
Total Cholesterol	6	160997118	rs74617384	8.52	0.121	0.0204	T	0.084	4	160882029	3.89	*LPA*	0	1|1
Total Cholesterol	16	72108093	rs2000999	8.15	0.083	0.0143	A	0.1781	5	72059149	5.61	*HPR*	0	1|1
Urea	1	155178782	rs760077	8.2	-0.064	0.0110	T	0.5758	3	155053719	5.74	*MTX1*	0	0|1
Urea	5	40635920	rs112647987	7.51	-0.119	0.0215	T	0.0686	1	40681254	5.82	*PTGER4*	44110	0|0
Urea	7	1270699	rs6950388	7.8	0.08	0.0141	A	0.8122	1	1274582	7.3	*UNCX*	1953	1|1
Urea	7	151413194	rs10224210	8.24	0.073	0.0125	C	0.2798	8	151405818	6.74	*PRKAG2*	0	1|1
Waist	2	211962722	rs556444167	7.39	-0.846	0.1541	T	0.0017	1	211731145	2.59	*ERBB4*	277724	0|0

Trait	Chr	Pos	rsID	logP	Beta	Beta_SE	EA	EAF	nhits	Pos_G	logP_G	Gene	Dist	GWAS
Waist Hip Ratio	9	105853716	rs149924309	7.43	0.693	0.1258	T	0.0023	1	105788456	2.47	*CYLC2*	72945	0|0
Waist Hip Ratio	14	94762510	rs187209742	7.31	-0.721	0.1322	G	0.0024	1	94971978	2	*SERPINA10*	2913	0|0

Supplementary Table 2 - Hits that exceed the suggestive but not the genome-wide significance threshold in the GWAS meta-analysis

This table summarizes the hits that passed the suggestive significance threshold (logP=7.3) in the GWAS meta-analysis, providing their -$\log_{10}$(*p-value*), effect size (Beta) of the effect allele (EA) and its standard error (Beta_SE) as well as the direction of the effect in each cohort (O=Orkney, V=Vis, K=Korčula, S=Shetland, G=GS; - or + values indicate effect size direction, ? indicates that this SNP was not available in the cohort, x indicates that the trait was not analysed in the cohort). The nhits column indicates the number of SNPs within 500kb of the reported SNP that exceeded the GWS threshold in the meta-analysis. The name of, and distance to, the gene closest to the reported SNP is indicated – the distance is 0 if the SNP lies within the gene itself. The final column indicates whether other GWAS have identified this hit before. The first column indicates whether any SNPs in the 1000 Genomes data that are in strong LD with the reported SNP (R^2 and D' > 0.8) have been identified with other GWAS, while the second value looks at all SNPs within 100kb of the reported SNP, regardless of LD.

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
Albumin	19	15756663	rs575873404	0.1425	7.31	C	-0.2323	0.0426	1	2676	?	?	-	?	x	*CYP4F3*	0	0|0
Axial length1	4	4718847	rs62290301	0.0309	7.79	T	0.417	0.0738	2	3566	+	+	?	+	x	*LOC100507266*	6182	0|0
Axial length2	4	4718847	rs62290301	0.0303	7.41	T	0.4126	0.0751	1	3517	+	+	?	+	x	*LOC100507266*	6182	0|0
BMI	2	651349	rs2867112	0.1639	8.6	T	0.0783	0.0131	149	27488	-	+	+	+	+	*TMEM18*	16622	1|1
BMI	4	45182527	rs10938397	0.4402	8.19	A	-0.056	0.0096	5	27488	-	-	-	-	-	*GNPDA2*	453914	1|1
BMI	16	17177110	rs571835655	0.0011	7.61	A	-1.06	0.1901	1	19900	?	?	?	?	-	*XYLT1*	18516	0|0
BMI	18	57896742	rs17175643	0.2511	7.54	T	0.0605	0.0109	6	27488	+	+	+	+	+	*MC4R*	141820	1|1
Central Corneal Thickness	13	41655900	rs78169557	0.1307	7.39	T	0.1983	0.0361	1	4416	+	+	+	+	x	*WBP4*	0	0|0
Cortisol	6	150495762	rs62440055	0.2371	7.36	T	-0.1816	0.0332	1	2921	-	-	x	x	x	*PPP1R14C*	0	0|0
Creatinine	1	15914135	rs6679323	0.1459	7.58	A	-0.0994	0.0179	1	23950	-	-	-	-	-	*AGMAT*	2529	0|0
Creatinine	16	20392332	rs77924615	0.1915	7.55	A	-0.072	0.013	1	23950	-	-	-	-	-	*PDILT*	0	0|1
Creatinine	1	234420503	.	0.0027	7.44	A	0.7259	0.1318	1	16347	?	?	?	?	+	*SLC35F3*	0	0|0

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
Creatinine	1	168754413	rs548873184	0.0011	7.4	A	1.06	0.193	1	16347	?	?	?	?	+	*DPT*	89916	0|0
CRP	1	66155515	rs4655584	0.3574	8.23	T	0.1465	0.0252	67	4970	+	+	x	+	x	*LEPR*	52338	1|1
Diastolic BP	15	76295311	rs187680191	0.0006	8.28	A	-18.58	3.1825	1	19429	?	?	?	?	-	*NRG4*	0	0|0
Diastolic BP	3	136063325	.	0.0054	8.04	C	12.0257	2.0934	1	1933	+	?	?	?	?	*STAG1*	0	0|0
Diastolic BP	1	60787284	rs80303520	0.005	7.95	A	12.8615	2.2515	1	2090	?	?	?	+	?	*C1orf87*	247841	0|0
Diastolic BP	4	20875931	rs79488205	0.0841	7.59	A	-0.917	0.1647	2	26948	-	-	-	-	-	*KCNIP4*	0	0|0
Diastolic BP	3	193362377	rs528908640	0.0005	7.49	T	-14.67	2.6539	1	19429	?	?	?	?	-	*OPA1*	0	0|0
Diastolic BP	3	147599147	rs187240817	0.0053	7.38	T	-13.04	2.3785	1	1933	-	?	?	?	?	*ZIC1*	464640	0|0
Diastolic BP	9	109802437	rs568998724	0.0006	7.32	A	13.49	2.472	1	19429	?	?	?	?	+	*ZNF462*	177059	0|0
FEV1	4	106828795	rs6823809	0.3102	7.38	T	0.0639	0.0116	1	22682	+	-	-	+	+	*NPNT*	0	0|1
fev1perfvc	4	145454964	rs1489762	0.4041	8.71	T	-0.0607	0.0101	65	22682	-	-	-	-	-	*LOC646576*	109102	1|1
fev1perfvc	9	78732604	rs114200252	0.1044	8.04	A	0.3566	0.062	1	2397	?	?	+	?	?	*PCSK5*	0	0|0
Fibrinogen	4	155416635	rs78318334	0.0222	8.08	A	-0.5253	0.0912	2	4007	-	-	x	-	x	*DCHS2*	3704	0|1
Forced Vital Capacity	6	7807702	rs1225986	0.1689	8.41	T	0.0819	0.0139	16	22717	+	-	+	+	+	*BMP6*	0	0|1
Forced Vital Capacity	2	185767897	rs146201345	0.004	7.99	A	0.5929	0.1035	1	15867	?	?	?	?	+	*ZNF804A*	0	0|0
GGT	14	103565080	rs36027406	0.2233	8.44	T	-0.1926	0.0326	14	3075	-	x	x	-	x	*EXOC3L4*	1399	1|1
GGT	12	121424574	rs2393775	0.365	7.71	A	0.1549	0.0276	7	3075	+	x	x	+	x	*HNF1A*	0	1|1
Glucose	21	44276432	rs370189685	0.0013	7.94	C	-1.176	0.206	1	16076	?	?	?	?	-	*WDR4*	0	0|0
Glucose	2	27730940	rs1260326	0.3856	7.42	T	-0.0551	0.01	2	23631	-	-	+	-	-	*GCKR*	0	1|1
Glucose_nodiab	2	27741237	rs780094	0.3757	8.01	T	-0.0599	0.0104	5	22080	-	-	-	-	-	*GCKR*	0	1|1
Glucose_nodiab	1	214150821	rs79687284	0.037	7.49	C	0.1705	0.0308	1	22080	+	-	+	+	+	*PROX1*	11037	1|1
Glucose_nodiab	9	7698812	rs149779116	0.0055	7.44	A	-0.4663	0.0847	1	22080	?	?	-	-	-	*C9orf123*	97677	0|0
Glucose_nodiab	21	44276432	rs370189685	0.0013	7.39	C	-1.15	0.2096	1	15141	?	?	?	?	-	*WDR4*	0	0|0
HDL	1	230307222	rs4846923	0.2909	8.57	T	-0.0628	0.0106	10	26920	-	-	-	-	-	*GALNT2*	0	0|1

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
HDL	6	127825034	rs150971345	0.0169	8.55	A	0.2502	0.0421	1	26920	+	+	+	+	+	*C6orf174*	0	0|0
HDL	2	21208211	rs7557067	0.2268	7.92	A	-0.0643	0.0113	11	26920	-	-	-	-	-	*APOB*	16088	1|1
HDL	8	116610180	rs2737205	0.4354	7.75	T	-0.0541	0.0096	6	26920	-	-	-	-	-	*TRPS1*		0|1
HDL	12	125353810	rs67053123	0.1346	7.52	A	0.0789	0.0142	2	26920	-	+	+	+	+	*SCARB1*	5290	0|1
Heart Rate	20	5376623	#N/A	0.0008	8.14	A	18.54	3.2045	1	19798	x	x	?	?	+	*PROKR2*	94306	0|0
Heart Rate	3	87751558	.	0.0017	7.42	A	8.6	1.5639	1	19798	x	x	?	?	+	*HTR1F*	288288	0|0
Heart Rate	8	62481520	rs142916219	0.0038	7.34	A	-5.971	1.0921	1	19798	x	x	?	?	-	*ASPH*	0	0|0
Height	7	92237426	rs42032	0.2575	8.75	A	0.0041	0.0007	8	27555	+	+	+	+	+	*CDK6*	0	0|1
Height	6	19838447	rs147448400	0.0612	8.71	A	0.0078	0.0013	1	27555	+	+	+	+	+	*ID4*	0	0|1
Height	1	17309718	rs3754512	0.4738	8.63	T	0.0035	0.0006	11	27555	+	+	+	+	+	*MFAP2*	1636	1|1
Height	4	82296362	rs28814370	0.2986	8.53	T	0.0038	0.0006	38	27555	+	+	+	+	+	*RASGEF1B*	51855	0|1
Height	13	46341501	rs75061684	0.0006	8.15	A	-0.1252	0.0216	1	19965	?	?	?	?	-	*SIAH3*	12913	0|0
Height	12	93972987	rs11107115	0.1984	8.12	A	-0.0043	0.0007	19	27555	-	-	-	-	-	*SOCS2*	3008	0|1
Height	4	17994298	rs35362908	0.0981	8.06	T	0.0065	0.0011	5	27555	+	+	+	+	+	*LCORL*	0	1|1
Height	1	149892872	rs11205277	0.4281	7.96	A	-0.0035	0.0006	2	27555	-	-	-	-	-	*SF3B4*	2337	0|1
Height	19	8670147	rs62621197	0.0391	7.72	T	-0.0099	0.0018	1	27555	-	-	-	-	-	*ADAMTS10*	0	0|1
Height	10	130682872	.	0.001	7.62	T	-0.0919	0.0165	1	19965	?	?	?	?	-	*MGMT*	582576	0|0
Height	2	232790491	rs11900720	0.1244	7.54	T	-0.005	0.0009	2	27555	-	-	-	-	-	*NPPC*	0	0|1
Height	15	100692953	rs72755233	0.1271	7.49	A	-0.0054	0.001	1	27555	-	-	+	-	-	*ADAMTS17*	0	0|1
Height	6	126851160	rs1490384	0.4985	7.4	T	0.0032	0.0006	1	27555	+	-	+	-	+	*CENPW*	181405	1|1
Insulin	19	8917703	rs188218737	0.0053	7.62	A	1.3849	0.2482	1	2089	?	?	x	+	x	*ZNF558*	2677	0|0
LDL	1	109821307	rs583104	0.2278	8.12	T	0.1161	0.0201	10	7663	+	+	+	+	x	*PSRC1*	867	1|1
LDL	5	74656539	rs12916	0.3932	7.31	T	-0.0932	0.0171	1	7663	-	-	-	-	x	*HMGCR*	0	1|1
Spherical Equivalent Refraction	14	76779969	rs188808208	0.0092	7.83	C	-1.1541	0.2038	1	1924	?	?	?	-	x	*ESRRB*	57719	0|0

Trait	Chr	Pos	rsID	MAF	logP	EA	Beta	BetaSE	nhits	n	O	V	K	S	G	Gene	Dist	GWAS
Total Cholesterol	16	72108093	rs2000999	0.1816	8.45	A	0.0716	0.0121	6	26960	+	+	+	-	+	*HPR*	0	1|1
Total Cholesterol	2	179387808	rs186622371	0.0068	7.56	C	1.2565	0.2261	3	2007	+	?	?	?	?	*LOC100506866*	0	0|0
Triglycerides	5	83471386	rs143440237	0.0045	8.38	T	0.7517	0.1279	1	2674	?	?	+	?	x	*EDIL3*	0	0|0
Triglycerides	2	27730940	rs1260326	0.4054	8.19	T	0.0501	0.0086	4	7698	+	+	+	+	x	*GCKR*	0	1|1
Urea	1	155722506	rs12134456	0.4122	7.58	C	-0.0607	0.0109	1	23316	-	-	x	-	-	*GON4L*	0	0|0
Uric acid1	11	69924331	rs531520064	0.0074	8.71	A	1.2364	0.2061	1	2003	+	?	?	?	x	*ANO1*	77	0|0
Uric acid1	11	73682274	rs539867601	0.0067	8.5	A	-1.2073	0.2039	1	2003	-	?	?	?	x	*DNAJB13*	0	0|0
Uric acid1	11	56102498	rs182906282	0.0091	8.25	T	-1.017	0.1746	1	2003	-	?	?	?	x	*OR8K3*	16715	0|0
Uric acid1	4	89039082	rs1481012	0.0923	7.93	A	-0.1669	0.0293	9	7715	-	-	-	-	x	*ABCG2*	0	1|1
Uric acid1	11	46559649	rs148042064	0.0085	7.85	T	-1.04	0.1833	1	2003	-	?	?	?	x	*AMBRA1*	0	0|0
Uric acid1	5	72425458	rs13182742	0.2564	7.59	T	0.1142	0.0205	1	7715	+	+	+	+	x	*TMEM171*	0	0|1
Uric acid2	11	47081384	rs757626652	0.0089	7.76	A	-1.1139	0.1976	1	1715	-	?	?	?	x	*C11orf49*	0	0|0
Uric acid2	11	50740928	rs117993891	0.0083	7.71	A	-1.1122	0.198	3	1715	-	?	?	?	x	*LOC646813*	361125	0|0
Uric acid2	11	50751235	.	0.0083	7.71	A	-1.1122	0.198	3	1715	-	?	?	?	x	*OR4A5*	660143	0|0
Uric acid2	11	56102498	rs182906282	0.0087	7.66	T	-1.0641	0.1901	1	1715	-	?	?	?	x	*OR8K3*	16715	0|0
Uric acid2	11	57113738	.	0.0088	7.65	A	-1.0909	0.1951	1	1715	-	?	?	?	x	*P2RX3*	0	0|0
Uric acid2	11	71674670	rs183077143	0.0087	7.62	A	1.1201	0.2008	4	1715	+	?	?	?	x	*RNF121*	0	0|0
Uric acid2	11	69924331	rs531520064	0.0074	7.57	A	1.2224	0.2198	1	1715	+	?	?	?	x	*ANO1*	77	0|0
Uric acid2	11	44713083	rs536771513	0.0104	7.4	C	-1.0594	0.1929	1	1715	-	?	?	?	x	*TSPAN18*	34932	0|0
Uric acid2	11	73682274	rs539867601	0.0067	7.4	A	-1.2069	0.2198	1	1715	-	?	?	?	x	*DNAJB13*	0	0|0
Waist	15	100516472	rs11634977	0.3406	8.38	A	0.0585	0.01	29	27212	+	+	+	+	+	*ADAMTS17*	0	0|0
Waist Hip Ratio	15	70439786	rs751156121	0.0006	7.58	A	-1.42	0.2551	1	19645	?	?	?	?	-	*TLE3*	49271	0|0

Supplementary Table 3 - Loci that exceed the suggestive but not the genome-wide significance threshold with pedigree-based linkage analysis, using IBD coefficients calculated by IBDLD

The rsID of the SNP at which the highest LOD score was obtained is shown, as is the chromosome (Chr column) and position (Pos column) where this SNP is located. The start and end positions of the interval where LOD scores that are within a 2-LOD drop of the top hit are shown, in Megabases (2-LOD drop column). The total trait heritability (h^2) and heritability explained by the hit (h^2QTL), as output by SOLAR, are shown. The 'Gene' and 'Gene_Dist' columns indicate the gene nearest the top hit, as well as the distance to this gene from the top hit (this distance is 0 when the top hit is within the gene itself).

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
Orkney										
BMI	16	rs8045775	58867855	3.4169	0.43	0.24	56.57-64.44	q12.2-q21	*GOT2*	99608
Cortisol	3	rs1564760	47582724	4.1181	0.29	0.23	42.33-54.81	p22.1-p14.3	*CSPG5*	21002
Cortisol	3	rs1910236	59434420	3.9671	0.31	0.26	56.79-60.09	p14.3-p14.2	*FHIT*	300614
Educational Attainment	12	rs10847446	123160304	4.1442	0.48	0.27	120.27-126.01	q24.23-q24.32	*HCAR2*	25534
Educational Attainment	15	rs7171078	29900518	3.6896	0.47	0.27	28.15-32.19	q13.1-q13.3	*FAM189A1*	37590
FEV1	12	rs11609579	101843240	3.7871	0.4	0.29	96.75-103.54	q23.1-q23.2	*SPIC*	28093
Forced Vital Capacity	12	rs4964460	106704974	3.9456	0.34	0.28	105.43-108.51	q23.3	*TCP11L2*	0
Glucose	4	rs2191684	14392330	3.4498	0.33	0.24	6.9-18.94	p16.1-p15.31	*LOC152742*	250653
Insulin	22	rs8135371	40757228	4.0937	0.31	0.26	37.93-42.87	q13.1-q13.2	*ADSL*	0
tPA	7	rs10248098	111636575	3.7611	0.32	0.32	101.63-129.66	q22.1-q32.2	*DOCK4*	0
Uric acid1	15	rs1608962	23750690	3.4771	0.4	0.23	20.16-25.73	q11.1-q12	*MIR4508*	56517
Uric acid2	17	rs9897025	63860895	3.4816	0.4	0.27	60.98-67.22	q23.2-q24.3	*CEP112*	0
vWF	1	rs238099	229508905	3.4887	0.59	0.35	228.61-232.09	q42.13-q42.2	*C1orf96*	30216

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
Vis										
Central Corneal Thickness	1	rs2494467	182283352	3.8436	0.79	0.71	179.55-185.43	q25.2-q25.3	*GLUL*	67485
Central Corneal Thickness	2	rs1507705	25714916	4.2914	0.7	0.7	23.75-26.81	p24.1-p23.3	*DTNB*	0
Central Corneal Thickness	4	rs980038	31669915	3.7106	0.83	0.54	28.87-35.65	p15.1	*PCDH7*	521491
Lens Thickness	5	rs2052852	31993222	3.7721	0.66	0.63	31.64-32.01	p13.3	*PDZD2*	0
Lens Thickness	13	rs9544368	36016609	3.569	0.53	0.53	32.89-39.48	q13.1-q13.3	*NBEA*	0
Systolic BP	9	rs7038500	38553816	3.7598	0.53	0.28	38.51-38.56	p13.1	*ANKRD18A*	17543
Shetland										
Axial length1	9	rs10970529	3189170	4.0379	0.81	0.26	2.57-3.88	p24.2	*RFX3*	35475
Axial length2	9	rs10970529	3189170	4.019	0.82	0.25	2.5-3.88	p24.2	*RFX3*	35475
BMI	8	rs4732980	29473431	3.4223	0.57	0.25	28.06-30.81	p21.1-p12	*C8orf75*	105343
Pulse Pressure	1	rs6429360	242654482	3.6847	0.21	0.21	242.2-245.02	q43-q44	*PLD5*	0
Creatinine	9	rs6475020	16275107	3.649	0.41	0.25	14.72-18.17	p22.3-p22.2	*BNC2*	134392
Diastolic BP	1	rs318405	83014374	4.0586	0.28	0.25	81.49-84.03	p31.1	*LPHN2*	556266
Diastolic BP	19	rs10417057	58177308	4.2744	0.28	0.26	56.9-58.84	q13.43	*ZSCAN4*	2993
Diastolic BP	19	rs12709954	56716579	3.4142	0.26	0.26	54.85-58.91	q13.42-q13.43	*ZSCAN5B*	12157
Educational Attainment	2	rs7572507	111490731	3.5059	0.33	0.31	109.74-111.98	q12.3-q13	*ACOXL*	0
Glucose	2	rs2177250	176346170	3.9739	0.28	0.27	175.21-177.57	q31.1	*ATP5G3*	299679
Glucose	4	rs17064578	178221198	4.5595	0.3	0.3	177.34-178.99	q34.2-q34.3	*NEIL3*	9791
Glucose	4	rs13102637	173491064	4.5233	0.3	0.3	172.9-176.57	q34.1-q34.2	*GALNTL6*	0
Glucose	4	rs4692709	170228549	3.6024	0.3	0.26	168.99-180.55	q32.3-q34.3	*SH3RF1*	36299
Glucose_nodiab	4	rs2555674	175466101	4.4746	0.29	0.29	173.01-176.67	q34.1-q34.2	*HPGD*	22056

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
Glucose_nodiab	4	rs17064578	178221198	4.4017	0.28	0.28	177.39-178.79	q34.2-q34.3	*NEIL3*	9791
Glucose_nodiab	4	rs4692709	170228549	3.5242	0.27	0.25	169.6-180.32	q32.3-q34.3	*SH3RF1*	36299
HbA1c	1	rs3134613	40364803	3.8234	0.42	0.27	39.39-41.24	p34.3-p34.2	*MYCL1*	0
HbA1c	2	rs10203413	3629178	3.4611	0.45	0.26	3.15-3.87	p25.3	*RPS7*	668
HbA1c	8	rs1825074	106510518	3.9771	0.43	0.27	105.39-107.34	q22.3-q23.1	*ZFPM2*	0
HbA1c	8	rs2443786	109028325	3.7722	0.43	0.27	108.7-110.1	q23.1	*RSPO2*	0
HbA1c	8	rs10093934	77331219	3.4919	0.44	0.25	75.85-87.25	q21.11-q21.3	*LOC100192378*	191893
HDL	18	rs8099543	61130942	3.7127	0.45	0.26	60.69-61.33	q21.33	*SERPINB5*	13200
IntraOcular Pressure	19	rs7251022	49534958	3.9219	0.27	0.27	48.53-51.1	q13.33	*SNAR-G2*	0
IntraOcular Pressure	19	rs3859451	51548531	3.7747	0.26	0.26	51.28-51.87	q13.33-q13.41	*KLK12*	10382
IntraOcular Pressure	19	rs873286	47163774	3.509	0.26	0.26	46.54-52.05	q13.32-q13.41	*DACT3*	0
Triglycerides	1	rs11121374	9396660	4.1537	0.33	0.24	8.19-11.02	p36.23-p36.22	*SPSB1*	0
Waist	5	rs11749791	166806876	4.1131	0.52	0.26	165.94-167.38	q34	*ODZ2*	0
Korčula										
Axial length1	6	rs4706308	88825342	3.5187	0.63	0.63	88.52-95.34	q15-q16.1	*CNR1*	24241
Axial length2	6	rs6454673	88871049	3.4546	0.63	0.63	88.52-90.63	q15	*CNR1*	0
Calcium	15	rs11071215	56072934	3.9449	0.46	0.25	49.03-57.74	q21.1-q21.3	*PRTG*	37756
Calcium	15	rs6495130	33810168	3.7299	0.47	0.25	30.16-36.7	q13.1-q14	*RYR3*	0
Educational Attainment	3	rs3937932	2459320	4.1637	0.54	0.32	1.3-3.1	p26.3-p26.2	*CNTN4*	0
Fibrinogen	17	rs2535609	15856740	3.4982	0.61	0.29	14.51-20.84	p12-p11.2	*ADORA2B*	0
Fibrinogen	20	rs2273534	56285540	3.6765	0.63	0.35	55.42-56.74	q13.31-q13.32	*PMEPA1*	0
Glucose_nodiab	10	rs1610357	30194393	4.6484	0.31	0.31	29.31-31.2	p12.1-p11.23	*KIAA1462*	107334

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
HbA1c	6	rs7768422	110796405	4.0879	0.44	0.32	109.29-113.71	q21	*SLC22A16*	0
Uric acid2	4	rs10025159	37575251	3.6789	0.37	0.26	35.81-38.37	p14	*C4orf19*	0
Waist Hip Ratio	4	rs13114765	189390949	3.6316	0.44	0.27	187.54-190.72	q35.2	*LOC401164*	0
GS										
Alcohol consumption	2	rs2037221	18559177	3.7288	0.29	0.1	18.26-19.18	p24.2	*RDH14*	176810
Alcohol consumption	6	rs416622	32973281	3.4785	0.28	0.09	32.74-33.31	p21.32	*HLA-DOA*	0
Alcohol consumption	7	rs11981366	37316506	4.472	0.29	0.11	36.53-37.87	p14.2-p14.1	*ELMO1*	0
Alcohol consumption	7	rs727162	34874038	3.4667	0.29	0.1	32.86-35.49	p14.3-p14.2	*NPSR1*	0
BMI	11	rs11603598	12064576	3.6082	0.55	0.09	11.36-12.15	p15.3	*DKK3*	33658
Body fat	7	rs2191261	12950387	3.5695	0.53	0.09	11.77-14.18	p21.3-p21.2	*ARL4A*	219828
Pulse Pressure	1	rs11240777	798959	4.0299	0.11	0.11	0.8-1.06	p36.33	*FAM41C*	4490
Pulse Pressure	1	rs513287	170664237	3.7903	0.11	0.11	169.58-172.3	q24.2-q24.3	*PRRX1*	0
Pulse Pressure	11	rs704664	44787201	3.8152	0.11	0.1	44.64-44.92	p11.2	*TSPAN18*	0
Pulse Pressure	11	rs2186580	101524873	3.6579	0.11	0.09	101.4-102.33	q22.1-q22.2	*TRPC6*	70213
Creatinine	1	rs10913949	180178695	3.6621	0.53	0.12	177.93-181.76	q25.2-q25.3	*FLJ23867*	8835
Educational Attainment	6	rs2523949	29917591	5.6092	0.49	0.09	29.34-30.77	p22.1-p21.33	*HLA-A*	3929
Educational Attainment	9	rs17262780	20304873	4.0182	0.51	0.12	18.75-27.84	p22.1-p21.2	*MLLT3*	40093
Educational Attainment	9	rs3808745	17627988	3.4247	0.51	0.11	16.36-28	p22.3-p21.1	*SH3GL2*	0
Educational Attainment	11	rs4435039	118297391	4.5771	0.5	0.1	118.03-118.74	q23.3	*MLL*	9812

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
Educational Attainment	11	rs4936417	117903836	4.4071	0.5	0.11	117.82-118.02	q23.3	*LOC100526771*	0
Educational Attainment	11	rs11022262	12260355	3.5824	0.5	0.09	12.17-12.28	p15.3	*MICAL2*	0
Educational Attainment	11	rs1379170	113866551	3.4758	0.5	0.1	113.32-114.12	q23.2	*HTR3A*	5516
Educational Attainment	11	rs666290	117476142	3.4307	0.5	0.09	116.55-118.9	q23.3	*DSCAML1*	0
Educational Attainment	12	rs10877461	61258329	3.9672	0.51	0.11	58.52-66.66	q14.1-q14.3	*FAM19A2*	843710
Educational Attainment	12	rs1458614	75487456	3.7045	0.51	0.12	73.35-77.29	q21.1-q21.2	*KCNC2*	0
Educational Attainment	12	rs1921051	81700111	3.631	0.5	0.11	77.81-89.68	q21.2-q21.33	*PPFIA2*	0
Educational Attainment	20	rs1033859	16076697	4.1988	0.5	0.12	15.69-16.72	p12.1	*MACROD2*	42855
Forced Vital Capacity	9	rs1867283	87450766	4.5389	0.42	0.15	86.87-88.11	q21.32-q21.33	*NTRK2*	0
Forced Vital Capacity	9	rs357642	80793672	4.4117	0.41	0.14	79.79-82.45	q21.2-q21.31	*CEP78*	57317
Forced Vital Capacity	9	rs11139921	85748119	4.3688	0.42	0.14	84.01-86.76	q21.31-q21.32	*RASEF*	70075
Forced Vital Capacity	9	rs6420275	79219334	4.0177	0.42	0.13	76.95-82.52	q21.13-q21.31	*PRUNE2*	6956
Forced Vital Capacity	21	rs2823687	17624130	4.3307	0.41	0.14	15.37-21.44	q11.2-q21.1	*C21orf34*	0
Glucose	22	rs3827409	46875743	3.976	0.25	0.13	45.29-47.79	q13.31	*CELSR1*	0
Glucose_nodiab	12	rs1919450	119303805	3.5119	0.29	0.13	117.08-119.94	q24.22-q24.23	*SRRM4*	115493
HDL	5	rs6876176	79495456	3.7446	0.59	0.1	79-80.26	q14.1	*SERINC5*	0

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
HDL	5	rs10050614	82154979	3.5508	0.59	0.1	73.84-84.09	q13.3-q14.3	*MIR3977*	18935
HDL	11	rs3924413	134404193	3.5489	0.58	0.09	133.76-134.75	q25	*LOC283177*	28637
Height	1	rs182358	97463150	3.8635	0.92	0.07	96.4-97.73	p21.3	*DPYD*	80148
Height	1	rs492220	94542569	3.456	0.92	0.07	90.31-98.72	p22.2-p21.3	*ABCA4*	0
Height	2	rs4667393	150714840	4.0427	0.92	0.08	143.26-151.73	q22.2-q23.3	*MMADHC*	270509
Height	2	rs7572208	35584461	3.4896	0.92	0.08	34.97-36.56	p22.3	*LOC100288911*	997429
Height	5	rs7705193	41133587	3.9552	0.92	0.08	40.18-41.44	p13.1	*C6*	8747
Height	7	rs2107945	68285442	3.8846	0.92	0.08	67.28-70.55	q11.22	*AUTS2*	778461
Height	7	rs2909668	131300258	3.8741	0.92	0.08	128.55-137.97	q32.1-q33	*PODXL*	58881
Height	7	rs1534131	45523448	3.6936	0.92	0.08	43.48-52.76	p13-p12.1	*ADCY1*	90675
Height	8	rs7459445	119683126	3.4781	0.92	0.07	118.47-120.08	q24.11-q24.12	*NCRNA00252*	0
Height	10	rs11001471	77453942	4.242	0.92	0.08	76.81-80.03	q22.2-q22.3	*C10orf11*	88575
Height	10	rs10822936	68649352	3.7031	0.92	0.08	68.08-73.34	q21.3-q22.1	*CTNNA3*	0
Height	10	rs7090884	63962026	3.6903	0.92	0.07	63.46-67.76	q21.2-q21.3	*RTKN2*	0
Height	11	rs10897109	60518182	4.5188	0.92	0.09	60.11-61.26	q12.2	*MS4A15*	6156
Height	11	rs7130134	45719881	3.9585	0.92	0.07	45.5-46.33	p11.2	*CHST1*	32708
Height	11	rs10835931	32515997	3.6427	0.92	0.06	32.13-32.6	p13	*WT1-AS*	54376
Height	11	rs2902421	45016421	3.597	0.92	0.08	44.29-45.22	p11.2	*LOC221122*	16842
Height	11	rs680999	79799533	3.4828	0.92	0.06	79.05-80.37	q14.1	*ODZ4*	647837
Height	12	rs2258342	48690363	3.5106	0.92	0.07	43.06-49.92	q12-q13.12	*H1FNT*	32398
Height	13	rs2329177	81725976	3.8325	0.92	0.08	76.6-83.83	q22.2-q31.1	*SPRY2*	810889
Height	13	rs4622388	94428051	3.5256	0.92	0.08	88.19-94.9	q31.2-q31.3	*GPC6*	0
Height	15	rs8024133	61130639	3.6337	0.92	0.08	60.02-61.58	q22.2	*RORA*	0
Height	17	rs9891296	58275994	4.5367	0.92	0.09	54.98-60.54	q22-q23.2	*USP32*	0
Systolic BP	3	rs1549110	175178296	4.0233	0.25	0.12	174.37-179.79	q26.31-q26.33	*NAALADL2*	0

Trait	Chr	rsID	Pos	LOD	h^2	h^2QTL	2-LOD drop	Band	Gene	Gene_Dist
Sodium	2	rs6545786	60403018	3.6665	0.26	0.1	59.79-62.02	p16.1-p15	*MIR4432*	211477
Sodium	7	rs7811904	143425361	3.6296	0.26	0.11	141.57-144.64	q34-q35	*FAM115C*	0
Sodium	11	rs10791302	133521772	4.3264	0.25	0.11	133.22-133.76	q25	*OPCML*	119368
Sodium	11	rs11221452	128607061	3.4208	0.26	0.1	127.9-128.8	q24.3	*FLI1*	0
Sodium	21	rs81481	42771362	3.5041	0.26	0.1	41.31-43.88	q22.2-q22.3	*MX2*	0
Total Cholesterol	1	rs6604877	78289045	3.624	0.3	0.1	72.99-79.25	p31.1	*FAM73A*	0
Total Cholesterol	1	rs11206690	56407227	3.4669	0.3	0.1	48.98-57.27	p33-p32.2	*PPAP2B*	553190
Total Cholesterol	6	rs10946765	25361055	4.5571	0.3	0.11	24.51-26.33	p22.3-p22.2	*LRRC16A*	0
Total Cholesterol	6	rs9381214	42693959	3.9993	0.3	0.12	42.28-44.47	p21.1	*ATP6V0CP3*	1353
Total Cholesterol	6	rs927657	45164355	3.6764	0.3	0.11	44.5-46.02	p21.1	*SUPT3H*	0
Total Cholesterol	6	rs12215142	65509089	3.5148	0.3	0.1	53.35-67.05	p12.1-q12	*EYS*	0
Total Cholesterol	11	rs512932	121973541	3.777	0.3	0.1	121.77-122.59	q24.1	*MIR100HG*	0
Total Cholesterol	13	rs10459371	41307789	3.6759	0.3	0.11	40.71-43.28	q14.11	*MRPS31*	0
Total Cholesterol	13	rs7325834	33852798	3.5909	0.3	0.11	32.89-34.72	q13.1-q13.2	*STARD13*	0
Waist	8	rs6999082	65807063	4.0009	0.42	0.1	62-66.32	q12.2-q13.1	*CYP7B1*	95714
Waist	8	rs11996294	67114738	3.6286	0.42	0.09	62-69.69	q12.2-q13.2	*LOC100505659*	5183
Waist	8	rs6985810	73429848	3.4568	0.41	0.09	73.07-74.27	q13.3-q21.11	*KCNB2*	19776
Waist Hip Ratio	1	rs4920295	18438239	4.0873	0.28	0.11	17.53-19.04	p36.13	*IGSF21*	0
Waist Hip Ratio	11	rs2062208	83696133	3.5213	0.27	0.08	83.58-84.22	q14.1	*DLG2*	0

Supplementary Table 4 - Meta-analysis results that exceed the suggestive but not genome-wide significance threshold in pedigree-based linkage analysis

The meta-analysis $-\log_{10}(p\text{-}value)$ is indicated (logP column) for each peak. These peaks represent a 0.1 cM interval that starts at the cM position indicated (cM column). The start and end positions of the region surrounding these peaks where the meta-analysis test statistic continuously exceeded the suggestive significance threshold (logP>4.43) is indicated in cM (Reg_cM) and Mbp (Reg_Mbp), as is the chromosome band where these regions can be found. The per-cohort LOD scores in the 0.1 cM peak region are indicated in the last columns (O = Orkney, S=Shetland, G=Generation Scotland, V=Vis, K=Korčula).

Trait	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G	V	K
Axial length1	2	138.8	4.67	q14.1	138.7-139	118-118.41	0.37	1.95	NA	1.78	0.47
Axial length1	2	138.3	4.60	q14.1	138.2-138.5	117.45-117.82	0.48	2.35	NA	1.22	0.45
Axial length2	2	138.8	4.52	q14.1	138.7-138.9	118-118.37	0.39	1.86	NA	1.83	0.36
BMI	11	21.8	5.50	p15.3	21.7-21.9	11.15-11.25	0	0.27	5.74	0.16	0.22
Central Corneal Thickness	4	52	5.87	p15.1	51.5-53.3	31.16-32.59	0	0.95	NA	3.68	1.49
Central Corneal Thickness	3	184.7	5.77	q26.31	183.4-185.2	172.32-173.16	1.21	0.09	NA	2.14	2.35
Central Corneal Thickness	4	50.3	5.40	p15.1	49.7-50.7	29.13-30.2	0	0.66	NA	2.63	2.34
Central Corneal Thickness	12	156.8	4.91	q24.32	156.7-156.9	127.88-127.92	0	0.52	NA	5.09	0
Central Corneal Thickness	12	157.8	4.83	q24.32	157.4-157.9	128.05-128.16	0	0.4	NA	5.15	0
Central Corneal Thickness	4	55.4	4.82	p15.1	55.3-55.6	35.15-35.33	0	1.07	NA	2.69	1.22
Central Corneal Thickness	4	62.1	4.80	p14	61.8-62.6	40.02-40.4	0	0.45	NA	3.12	1.45
Central Corneal Thickness	7	92.1	4.66	q11.23	91.7-92.5	75.24-75.59	0.57	0.68	NA	1.38	1.86
Central Corneal Thickness	4	63.1	4.64	p14	62.9-63.5	40.47-40.61	0	0.9	NA	2.95	0.96
Central Corneal Thickness	4	68	4.62	p12	67.9-68.3	45.02-46.12	0	0.49	NA	1.47	2.87
Central Corneal Thickness	1	197.5	4.58	q25.2	197.4-197.6	176.44-176.71	0	0	NA	5.08	0.24
Central Corneal Thickness	4	62.7	4.43	p14	61.7-63.6	39.75-40.66	0	0.93	NA	2.59	1.05
Cortisol	17	78.2	4.75	q22	78.1-78.4	53.3-53.35	2.46	NA	NA	1.62	NA

Trait	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G	V	K
Creatinine	6	172.2	5.49	q25.3	171.2-172.5	156.13-156.87	0.07	1.36	2.56	1.57	0.24
Creatinine	9	41.8	5.14	p21.3	41.4-42.2	20.18-20.49	1.33	3.23	0.2	0	0.87
Creatinine	4	203.7	4.74	q35.1	203.6-204.5	184.54-184.76	0.28	1.93	1.83	0.69	0.19
Creatinine	6	170.3	4.50	q25.2-q25.3	170.2-170.4	155.49-155.53	0.25	1.25	1.91	1.21	0.07
Creatinine	6	124.5	4.48	q22.1	124.3-124.6	117.57-117.63	1.27	0.68	2.02	0.27	0.35
Diastolic BP	2	22.7	4.74	p25.1	22.3-23.3	10.09-10.5	6.23	0.01	0	0	0
Diastolic BP	2	20.9	4.71	p25.1	20.5-21.4	9.35-9.91	6.26	0	0	0	0
Educational Attainment	9	46.7	5.22	p21.3	46.5-47.2	23.68-24.72	0	0	6.14	-0.06	0.4
Educational Attainment	11	54.1	5.18	p13	53.9-54.2	35.21-35.4	0	2.18	2.71	0.06	0.8
Educational Attainment	11	128.8	4.97	q23.3	128.7-128.9	118.13-118.39	0.18	0.62	4.58	0.1	0.1
Educational Attainment	2	79.1	4.71	p16.3	78.9-79.3	51.24-51.84	0.45	0.21	1.61	-0.11	2.93
Educational Attainment	6	49.9	4.68	p22.1-p21.33	49.8-51.4	28.26-31.43	0.93	0.03	3.5	0.03	0.7
Educational Attainment	12	149.7	4.66	q24.31	149.3-150	125.09-125.34	3.2	1.13	0.55	-0.01	0.24
Educational Attainment	11	128.3	4.64	q23.3	128.2-128.4	117.9-117.98	0.02	0.97	4.41	0.03	0.02
Educational Attainment	9	47.8	4.59	p21.3	47.5-47.9	24.83-25	0	0	5.73	-0.06	0.16
Educational Attainment	12	147	4.57	q24.31	146.8-147.5	123.15-123.93	4.1	0.29	0.4	0.02	0.29
Educational Attainment	9	47.3	4.56	p21.3	46.4-48	23.65-25.03	0	0	5.41	0	0.36
Educational Attainment	12	146.7	4.52	q24.31	146.6-147.6	122.67-124.01	4.01	0.33	0.46	0.01	0.25
FEV1	1	267.5	4.65	q43	267.3-267.8	236.68-236.81	0.13	1.18	2.72	0.76	0.13
Fibrinogen	3	46.6	4.59	p24.2	46.5-46.9	25.46-25.51	1.1	0.16	NA	1.27	1.97
Forced Vital Capacity	9	101.8	5.36	q21.33	101.7-102.9	87.23-87.81	0.54	0.97	4.36	0.07	0.02
Forced Vital Capacity	9	92	5.10	q21.2	91.6-92.2	79.46-79.64	0.52	0.21	3.77	0.32	0.6
Forced Vital Capacity	9	90.5	5.07	q21.13	90.1-90.9	78.81-79.15	0.38	0	3.81	0.87	0.53
Forced Vital Capacity	9	89.1	4.90	q21.13	88.9-89.2	78.47-78.55	0.08	0	3.33	1.46	0.59
Forced Vital Capacity	9	91.1	4.79	q21.13-q21.2	90-91.2	78.79-79.23	0.25	0.01	4.02	0.47	0.55

Trait	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G	V	K
Forced Vital Capacity	9	89.6	4.66	q21.13	89.4-89.7	78.64-78.67	0.16	0	2.95	1.28	0.72
GGT	1	208.8	5.84	q31.1	208.7-208.9	188.77-189.38	0.01	0	NA	6.31	NA
GGT	19	96.6	5.44	q13.42	96.5-97.4	54.93-55.2	0.6	0	NA	4.97	NA
GGT	19	95.5	4.54	q13.42	95.4-95.7	54.74-54.81	0.51	0.06	NA	3.9	NA
Glucose	12	145.2	4.47	q24.31	145.1-145.3	120.91-121.1	0.33	1.16	2.26	0	1.06
Glucose	2	195.5	4.43	q31.1	195.4-195.6	176.64-176.77	0	3.78	0.18	0	1.26
Glucose_nodiab	10	57.4	4.97	p11.23	56.9-58	30.01-30.42	0.46	0.48	0.02	0.04	4.65
Glucose_nodiab	10	55.6	4.82	p12.1	55.3-56	29.32-29.6	1.2	1.21	0.02	0	2.95
Glucose_nodiab	4	185.3	4.74	q34.1	184.8-185.6	175.28-175.72	1.2	4.41	0	0	0.05
Glucose_nodiab	10	55	4.66	p12.1	54.8-55.2	29.13-29.32	1.31	1.39	0	0	2.58
Glucose_nodiab	20	42.1	4.63	p12.1	41.8-42.3	17.61-17.83	0	1.66	0	2.35	1.23
Glucose_nodiab	20	42.8	4.54	p12.1-p11.23	42.6-42.9	17.87-18.02	0	1.58	0	2.56	1.01
Glucose_nodiab	2	172.8	4.50	q23.3	172.7-172.9	153.47-153.92	0.01	3.21	0.24	1.03	0.4
Glucose_nodiab	4	182.3	4.50	q34.1	182.2-182.5	173.22-173.8	0.5	3.93	0	0.82	0
Glucose_nodiab	20	41.6	4.48	p12.1	41.5-41.7	17.49-17.54	0	1.63	0	2.36	1.09
HbA1c	8	106.2	4.91	q21.13-q21.2	106-106.4	84.54-85.1	0.89	4.36	NA	0.03	0.03
HbA1c	1	67.8	4.74	p34.2	67.5-67.9	40.29-40.76	1.01	3.75	NA	0	0.29
HbA1c	8	101.6	4.72	q21.11-q21.12	101.5-101.7	78.02-78.63	0.01	4.92	NA	0.39	0
HbA1c	8	105.6	4.63	q21.13	105.3-105.7	83.43-83.99	0.46	4.4	NA	0.16	0.01
HbA1c	2	6.2	4.61	p25.3	6.1-6.3	3.5-3.65	0	3.46	NA	0	1.67
HbA1c	20	45.3	4.54	p11.23	45.1-45.4	19.61-19.71	2.99	2.05	NA	0	0
HbA1c	1	68	4.48	p34.2	67.4-68.1	40.25-40.81	1.13	3.21	NA	0	0.37
HbA1c	20	45.7	4.46	p11.23	45.6-45.8	19.81-19.88	2.81	2.02	NA	0.02	0
HDL	16	71.5	5.46	q12.2	70.2-72.3	55.72-56.6	1.2	1.74	0.76	0.03	1.99
HDL	11	160.1	4.66	q25	159.9-160.2	134.26-134.35	0.05	0.51	4.98	0.03	0

Trait	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G	V	K
Height	10	96.2	5.83	q22.1	95.6-97.3	72.23-73.05	0	0.84	5	0	0.88
Height	17	86	5.76	q22-q23.2	84.9-87.8	56.18-60.2	1.9	0.01	4.54	0	0.19
Height	5	61.2	5.54	p13.1	61-61.5	40.36-41.3	0.28	2.26	3.64	0.07	0
Height	2	167.8	5.46	q22.3	167.5-168.2	147.25-148.37	1.23	0	5.14	0.11	0
Height	13	79.4	5.39	q31.1	79.1-80.8	79.81-82.05	0.51	0	3.71	0.98	0.68
Height	15	75.8	5.20	q21.3	75.7-76.3	56.79-57.64	0	2.8	3.41	0	0
Height	17	83.8	5.11	q22	82.3-84	55.28-55.8	1.59	0	4.38	0	0.09
Height	15	86.5	5.08	q22.2-q22.31	86.1-87.7	62.86-64.14	0.16	0.32	4.63	0	0.67
Height	2	78.1	4.84	p16.3	77.9-78.2	50.14-50.56	0.09	1.39	2.31	0	1.51
Height	7	73	4.82	p12.3	72.7-73.3	47.13-47.43	0	0.94	5.01	0	0
Height	15	77.1	4.82	q21.3	76.9-77.4	57.77-58.04	0.03	2.55	3.16	0	0
Height	10	94.5	4.70	q22.1	94.4-94.7	71.64-71.68	0	0.01	5.08	0	0.72
Height	15	88.1	4.68	q22.31	87.8-88.2	64.19-65.1	0.19	0.05	4.03	0	1.1
Height	7	77.1	4.67	p12.1	76.7-77.5	50.63-51.45	0	0.67	4.97	0.04	0
Height	13	81.3	4.67	q31.1	81.2-81.5	82.58-82.95	0.55	0.3	3.61	0.31	0.23
Height	1	241.5	4.65	q41	241.4-241.8	217.58-218.1	0.6	2.32	1.48	0.17	0.27
Height	2	170	4.64	q23.2-q23.3	169.7-170.2	150.48-151.03	1.24	0	4.04	0.19	0
Height	6	174.6	4.62	q25.3	174.5-174.7	158.35-158.49	1.26	2.62	1.37	0	0
Height	2	168.4	4.62	q22.3-q23.1	168.3-168.6	148.54-149.21	1.15	0	4.23	0.11	0
Height	17	99.9	4.56	q24.3	99.8-100	69.39-69.5	3.15	0	1.84	0	0.3
Height	2	165.5	4.55	q22.3	165.3-165.6	144.14-144.51	0.87	0	3.97	0.47	0
Height	2	166.8	4.54	q22.3	166.7-167	145.82-146.11	0.76	0	4.66	0.07	0
Height	7	80.6	4.53	p11.2	80.5-80.7	54.75-54.93	0.09	2.22	2.7	0	0.14
Height	2	168.8	4.53	q22.3-q23.1	167.4-168.9	146.94-149.64	1.53	0	3.66	0.15	0
Height	15	90.8	4.49	q22.32	90.6-90.9	67.22-67.28	0.19	0	4.25	0.28	0.41

Trait	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G	V	K
Height	11	4.3	4.49	p15.5	4.2-4.4	1.8-1.86	0.84	0	3.28	-0.01	1.03
Height	17	99.5	4.48	q24.3	99.4-99.6	69.19-69.24	3.27	0	1.59	0	0.34
Height	2	169	4.48	q22.3-q23.2	167.3-169.1	146.66-149.93	1.63	0	3.45	0.17	0
Height	7	77.9	4.47	p12.1	77.8-78	51.68-51.82	0	0.64	4.78	0.04	0
Height	15	91.2	4.47	q22.33	91.1-91.3	67.41-67.47	0.16	0	3.9	0.47	0.5
Height	17	82.2	4.46	q22	82.1-84.1	55.18-55.84	1.91	0	3.17	0	0.13
Insulin	22	50	4.83	q13.1-q13.2	49.6-50.2	40.26-42.53	4.03	0.04	NA	0.71	NA
IntraOcular Pressure	19	83.9	4.58	q13.41	83.8-84.1	51.53-51.57	0.3	3.77	NA	NA	NA
Pulse Pressure	3	83.1	4.98	p14.2	82.9-83.3	62.44-62.59	0.19	2.7	2.26	0	0.36
Systolic BP	3	188.9	5.15	q26.31	188.5-189.4	175.05-175.51	0.27	0	4.02	1.49	0.07
Systolic BP	3	195.4	4.92	q26.33	195.1-195.5	180.04-181.24	0.27	0	2.55	2.63	0.08
Systolic BP	3	187.5	4.80	q26.31	187.3-188.3	174.68-175.02	0.08	0	3.56	1.75	0.11
Systolic BP	3	190.1	4.56	q26.32	190-190.2	176.03-176.27	0.14	0	2.95	2.06	0.06
Systolic BP	3	190.4	4.49	q26.32	190.3-190.5	176.27-176.41	0.17	0	2.42	2.39	0.09
Triglycerides	1	22.1	4.62	p36.22	21.8-22.2	10.88-10.99	1.42	3.13	NA	0	0.31
Triglycerides	1	24	4.58	p36.22	23.9-24.1	11.82-11.99	0.97	2.56	NA	0	1.2
Triglycerides	1	23.4	4.52	p36.22	23.1-23.5	11.63-11.75	1	2.38	NA	0	1.27
Urea	9	63.7	5.60	p13.2-p13.1	63.6-64.3	38.25-38.44	0.07	0	0.12	6.18	NA
Urea	21	27.3	5.01	q21.3	26.5-27.7	28.63-29.12	3.25	1.87	0.21	0	NA
Urea	21	29	4.71	q21.3	28.8-29.2	29.95-30.84	2.89	2.2	0.03	0	NA
Urea	21	28.7	4.49	q21.3	28.6-29.3	29.77-30.85	2.96	1.7	0.13	0	NA
Urea	21	28.5	4.43	q21.3	28.4-29.4	29.52-30.96	2.78	1.92	0.07	0	NA

Supplementary Table 5 - Meta-analysis results that exceed the suggestive but not the genome-wide significance threshold in pedigree-based linkage analysis, with cohorts grouped by geographical location

The meta-analysis -$\log_{10}$(*p-value*) is indicated (logP column) for each peak. These peaks represent a 0.1 cM interval that starts at the cM position indicated (cM column). The start and end positions of the region surrounding these peaks where the meta-analysis test statistic continuously exceeded the suggestive significance threshold (logP>4.43) is indicated in cM (Reg_cM) and Mbp (Reg_Mbp), as is the chromosome band where these regions can be found. The per-cohort LOD scores in the 0.1 cM peak region are indicated in the last columns (O = Orkney, S=Shetland, G=Generation Scotland, V=Vis, K=Korčula).

Trait - Croatia	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	V	K	
Central Corneal Thickness	1	197.5	5.78	q25.2	197.4-197.6	176.44-176.71	5.08	0.24	
Central Corneal Thickness	1	203	4.59	q25.3	202.8-203.3	182.04-182.39	3.84	0.25	
Central Corneal Thickness	2	46.1	4.51	p23.3	46-46.2	25.6-25.9	4.29	0	
Central Corneal Thickness	3	184.7	5.18	q26.31	183.9-185.2	172.59-173.16	2.14	2.35	
Central Corneal Thickness	4	52.3	5.84	p15.1	51.5-54.9	31.16-34.36	3.63	1.55	
Central Corneal Thickness	4	50.3	5.66	p15.1	49.4-51.2	28.94-30.7	2.63	2.34	
Central Corneal Thickness	4	62.1	5.24	p14	61.2-62.7	39.02-40.43	3.12	1.45	
Central Corneal Thickness	4	68.6	5.09	p13-p12	66.8-69	43.07-47.57	1.67	2.74	
Central Corneal Thickness	4	70	4.83	q12	69.9-70.2	53.88-54.55	1.94	2.22	
Central Corneal Thickness	4	55.4	4.56	p15.1	55.3-55.6	35.15-35.33	2.69	1.22	

Trait - Croatia	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	V	K	
Central Corneal Thickness	4	63.1	4.55	p14	62.9-63.2	40.47-40.57	2.95	0.96	
Central Corneal Thickness	4	69.7	4.55	q12	69.5-69.8	53.27-53.88	1.9	1.97	
Central Corneal Thickness	12	157.8	5.34	q24.32	157.4-157.9	128.05-128.16	5.15	0	
Central Corneal Thickness	12	156.8	5.29	q24.32	156.7-156.9	127.88-127.92	5.09	0	
Educational Attainment	3	4.9	4.58	p26.3	4.7-5	2.43-2.53	0.05	4.16	
Glucose	20	11.2	4.67	p13	10.9-11.6	3.08-3.68	1.25	2.76	
Glucose_nodiab	10	57.4	5.02	p11.23	56.9-58.2	30.01-30.65	0.04	4.65	
Glucose_nodiab	10	58.4	4.44	p11.23	58.3-58.5	30.66-30.69	0.18	3.8	
Systolic BP	9	64.6	4.68	p13.1	64.4-64.7	38.51-38.57	3.76	0.38	
Uric acid1	20	32.5	4.96	p12.2-p12.1	32.1-32.9	12.01-12.63	3.22	1.1	
Uric acid1	20	31.9	4.50	p12.2	31.8-32	11.78-12	3.15	0.74	

Trait - Scotland	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G
BMI	11	21.8	5.77	p15.3	21.6-21.9	11.12-11.25	0	0.27	5.74
BMI	16	76.2	4.92	q21	75.8-76.9	58.01-58.95	3.21	0	1.69
Pulse Pressure	1	281.9	5.07	q43	281-282	242.28-242.77	0.01	3.59	1.4
Pulse Pressure	3	83.1	5.40	p14.2	82.8-83.4	62.41-62.6	0.19	2.7	2.26
Pulse Pressure	11	110.1	4.72	q22.1	110-110.3	101.4-101.9	0.8	0.11	3.66
Creatinine	9	41.8	5.01	p21.3	41.4-42.2	20.18-20.49	1.33	3.23	0.2
Diastolic BP	2	22.7	5.77	p25.1	19.2-23.6	8.64-10.57	6.23	0.01	0
Diastolic BP	2	239.4	4.47	q36.1	239.3-239.5	222.86-222.98	0.38	2.11	1.63
Diastolic BP	2	24.4	4.45	p25.1	24.3-24.5	10.87-10.96	4.88	0	0

Trait - Scotland	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G
Diastolic BP	2	24.7	4.45	p25.1	24.6-24.9	10.98-11.01	4.87	0	0
Educational Attainment	6	50.8	5.17	p21.33	50.7-50.9	30.77-30.99	1.32	0	3.85
Educational Attainment	9	46.9	5.61	p21.3	46.3-48.1	23.64-25.04	0	0	6.14
Educational Attainment	9	49.9	4.73	p21.2	49.7-50.3	25.89-25.96	0	0	5.18
Educational Attainment	11	128.8	5.48	q23.3	128.7-128.9	118.13-118.39	0.18	0.62	4.58
Educational Attainment	11	128.3	5.41	q23.3	128.1-128.5	117.82-118.02	0.02	0.97	4.41
Educational Attainment	11	54.1	4.92	p13	53.9-54.2	35.21-35.4	0	2.18	2.71
Educational Attainment	12	149.7	5.20	q24.31	148.3-150.9	124.73-125.77	3.2	1.13	0.55
Educational Attainment	12	76.3	5.06	q14.2	76-76.6	63.24-63.64	0	0	5.54
Educational Attainment	12	146.7	4.97	q24.31	146.3-148	122.2-124.63	4.01	0.33	0.46
Educational Attainment	12	77.6	4.59	q14.2	77.4-77.7	64.43-64.91	0	0	5.03
Educational Attainment	12	148.1	4.49	q24.31	146.2-148.2	121.82-124.72	3.32	0.38	0.54
Educational Attainment	15	99.9	4.53	q24.1	99.8-100.1	74.45-74.55	0.19	2.29	1.77
Educational Attainment	20	39.2	5.24	p12.1	39.1-39.5	16.06-16.22	0.74	0.16	4.2
FEV1	5	89.4	4.54	q13.3	89.3-89.5	76.54-76.72	0.55	2.22	1.4
FEV1	10	169.5	4.51	q26.2	169.4-169.6	130.07-130.21	0	2.35	2.1
fev1perfvc	10	0	4.60	p15.3	0-0.1	0.07-0.37	0.45	0.86	2.98
Forced Vital Capacity	9	93.2	4.97	q21.2	92.9-93.5	80.22-81.07	0.64	0	4.41
Forced Vital Capacity	9	92	4.67	q21.2	91.8-92.2	79.5-79.64	0.52	0.21	3.77
Forced Vital Capacity	9	103.9	4.66	q21.33	103.7-104	88.36-88.87	0.47	0.68	3.24
Forced Vital Capacity	9	99.4	4.58	q21.32	99.3-99.5	85.59-85.8	0.26	0.02	4.37
Forced Vital Capacity	9	105.4	4.52	q21.33	105.3-105.6	89.6-89.95	1.13	0.15	3.01
Forced Vital Capacity	12	126.1	4.87	q23.3	125.9-126.2	106.66-107.27	3.8	0	1.08
Glucose	4	185	5.37	q34.1	183.4-185.7	174.68-175.79	1.06	4.36	0
Glucose	4	182.7	4.92	q34.1	182.2-182.8	173.22-174.29	0.59	4.42	0

Trait - Scotland	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G
Glucose	4	183	4.66	q34.1	182.1-185.8	173.01-175.87	0.69	4.03	0
Glucose	20	77.3	4.51	q13.2	77.2-77.4	50.01-50.08	0.8	2.69	0.67
Glucose_nodiab	4	185.3	5.55	q34.1	183.5-185.9	174.74-175.96	1.2	4.41	0
Glucose_nodiab	4	183	4.57	q34.1	182.8-183.1	174.29-174.48	0.86	3.73	0
Glucose_nodiab	4	183.2	4.50	q34.1	182.7-183.4	174.06-174.68	0.87	3.64	0
Glucose_nodiab	4	186	4.45	q34.1-q34.2	182.6-186.3	173.95-176.38	0.98	3.47	0
HDL	11	160.1	5.53	q25	159.6-160.3	134.2-134.35	0.05	0.51	4.98
HDL	11	159.3	4.72	q25	159.2-159.4	133.9-133.93	0.03	0.16	4.64
HDL	18	92.2	4.81	q22.1	92.1-92.5	64.95-65.29	0.08	2.14	2.39
Height	1	241.5	4.77	q41	241.4-241.8	217.58-218.1	0.6	2.32	1.48
Height	1	246.3	4.46	q41	246.2-246.4	222.59-222.74	0.61	1.86	1.61
Height	1	246.3	4.46	q41	246.2-246.4	222.61-222.74	0.61	1.86	1.61
Height	1	242.2	4.45	q41	242.1-242.3	218.3-218.39	0.48	1.44	2.16
Height	2	169.9	5.25	q23.2-q23.3	169.4-170.5	150.19-151.27	1.24	0	4.04
Height	2	165.5	4.80	q22.2-q22.3	165.1-165.7	144.06-144.63	0.87	0	3.97
Height	2	193.5	4.51	q31.1	193.4-193.6	174.48-174.55	1.72	0	2.74
Height	2	166.1	4.44	q22.3	166-166.2	144.94-145.22	0.85	0	3.6
Height	5	63.6	4.74	q11.2	63.1-63.8	51.61-52.23	0.38	1.56	2.46
Height	6	174.6	5.63	q25.3	172.8-175	157.05-158.96	1.26	2.62	1.37
Height	6	172.6	4.59	q25.3	172.5-175.1	156.87-159.05	0.6	2.34	1.28
Height	7	73	5.86	p12.3	72.4-73.6	46.8-47.57	0	0.94	5.01
Height	7	77.1	5.52	p12.2-p12.1	76.6-78.5	50.42-52.21	0	0.67	4.97
Height	7	80.6	5.19	p11.2	80.3-80.8	54.66-55.1	0.09	2.22	2.7
Height	7	76	4.80	p12.2	75.4-76.3	49.29-50.25	0	0.63	4.24
Height	7	73.7	4.64	p12.3	72.3-73.8	46.7-47.7	0	0.15	4.7

Trait - Scotland	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G
Height	10	96.2	5.73	q22.1	95.9-96.9	72.34-72.82	0	0.84	5
Height	10	94.5	4.69	q22.1	94.4-94.7	71.64-71.68	0	0.01	5.08
Height	10	112.9	4.58	q23.1	112.8-113	86.01-86.46	0	2.12	2.4
Height	11	85.8	4.93	q14.1	85.7-86	78.35-78.48	0.32	1.14	3.19
Height	11	67.5	4.69	q12.1	67.4-67.6	57.35-58.31	0.23	0	4.64
Height	11	6.4	4.44	p15.5	6.3-6.5	2.55-2.6	1.14	0	3.28
Height	13	80.6	4.83	q31.1	80.5-80.8	81.68-82.05	0.31	0.5	3.83
Height	13	81.3	4.67	q31.1	81.2-81.5	82.58-82.95	0.55	0.3	3.61
Height	15	87	5.37	q22.2	86.1-87.5	62.86-63.69	0.05	0.18	5.24
Height	15	89.8	5.36	q22.31	89.1-90.2	66.36-67.02	0.17	0.04	5.29
Height	15	88.9	4.51	q22.31	88.8-89	65.69-66.33	0.14	0.2	4.13
Height	17	83.8	5.92	q22-q23.2	81.7-88.2	55.09-60.78	1.59	0	4.38
Height	17	99.9	5.01	q24.3	99.4-100.2	69.19-69.54	3.15	0	1.84
Height	17	101.4	4.91	q24.3	101-101.8	70.18-70.41	2.64	0	2.24
Height	17	100.4	4.63	q24.3	99.3-100.9	69.08-70.16	3.11	0	1.44
Height	17	90.1	4.56	q24.1	89.9-90.3	62.78-63.08	1.65	0.07	2.63
Height	17	90.1	4.56	q24.1	89.9-90.3	62.79-63.08	1.65	0.07	2.63
Height	17	90.1	4.56	q24.1	89.9-90.3	62.81-63.08	1.65	0.07	2.63
Height	17	90.1	4.56	q24.1	89.9-90.3	62.82-63.08	1.65	0.07	2.63
Height	17	90.1	4.56	q24.1	89.9-90.3	62.84-63.08	1.65	0.07	2.63
Height	17	90.1	4.56	q24.1	89.9-90.3	62.86-63.08	1.65	0.07	2.63
Height	17	90.1	4.56	q24.1	89.9-90.3	62.86-63.08	1.65	0.07	2.63
Height	17	89.1	4.53	q23.3	89-89.3	62.11-62.41	1.62	0	2.84
Total Cholesterol	1	107.3	5.34	p31.1	107-107.5	77.69-78.7	0	1.73	3.62
Total Cholesterol	1	104.4	4.61	p31.1	104.3-104.6	74.98-75.15	0	1	3.61

Trait - Scotland	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	G
Total Cholesterol	6	81.1	5.88	p11.2-q12	80.7-81.3	58.15-64.61	0.39	0	5.7
Total Cholesterol	6	48.8	5.10	p22.3-p22.2	48.5-49.1	25.14-25.6	0.47	0.05	4.56
Total Cholesterol	6	66.6	4.52	p21.1	66.3-66.7	42.45-42.83	0.23	0.24	3.94
Total Cholesterol	11	135.1	4.81	q24.1	134.8-135.2	121.77-122.03	0.99	0.01	3.78
Total Cholesterol	11	137.5	4.62	q24.1	137.3-137.6	123.07-123.16	1.45	0.41	2.42
Urea	21	27.3	5.56	q21.3	25.8-27.9	28.39-29.29	3.25	1.87	0.21
Urea	21	29	5.24	q21.3	28.2-29.4	29.45-30.96	2.89	2.2	0.03
Urea	21	28	4.51	q21.3	25.7-28.1	28.39-29.42	2.63	1.71	0.02
Waist	5	178.4	4.58	q34	178.3-178.6	166.67-166.86	0.42	4.11	0.04

Trait - Scottish Isles	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	
Axial length1	9	9.3	4.65	p24.2	9.1-9.6	3.09-3.35	0.17	4.04	
Axial length2	9	9.3	4.77	p24.2	9-9.6	3.03-3.35	0.26	4.02	
Creatinine	9	41.8	5.22	p21.3	41.4-42.2	20.18-20.49	1.33	3.23	
Diastolic BP	2	26	4.59	p25.1	19.1-26.1	8.63-11.52	4.37	0	
Diastolic BP	19	105.7	4.50	q13.43	105.6-105.8	58.06-58.38	0	4.27	
Educational Attainment	12	149.7	4.97	q24.31	148.9-150.5	124.97-125.54	3.2	1.13	
Educational Attainment	12	146.9	4.90	q24.31	146.3-147.5	122.2-123.93	4.14	0.27	
Educational Attainment	12	148.7	4.46	q24.31	148.4-148.8	124.76-124.95	3.1	0.74	
Glucose	4	189.9	5.01	q34.3	189.2-190.2	177.98-178.3	0.07	4.56	
Glucose	4	190.4	4.62	q34.3	189.1-190.5	177.97-178.35	0.08	4.15	
Glucose	4	179.7	4.57	q32.3-q33	179.5-179.8	169.99-170.53	0.42	3.6	
Glucose_nodiab	4	189.9	4.83	q34.3	189.7-190.2	178.15-178.3	0.06	4.4	
Glucose_nodiab	4	179.7	4.51	q32.3-q33	179.5-179.8	169.99-170.53	0.44	3.52	

Trait - Scottish Isles	Chr	cM	logP	Band	Reg_cM	Reg_Mbp	O	S	
HbA1c	1	67.8	5.38	p34.2	67.5-68.3	40.29-40.93	1.01	3.75	
HbA1c	8	101.6	5.19	q21.11-q21.12	101.2-101.7	77.7-78.63	0.01	4.92	
HbA1c	20	45.3	5.73	p11.23	44.2-46.3	19-19.95	2.99	2.05	
HbA1c	20	51.5	4.83	p11.21	51.1-51.6	24.49-25.18	3.24	0.96	
HDL	18	87.7	4.50	q21.33	87.5-87.8	61.05-61.18	0.28	3.71	
Height	6	174.6	4.53	q25.3	174.5-174.7	158.35-158.49	1.26	2.62	
Insulin	22	49.9	4.45	q13.1-q13.2	49.8-50	40.53-41.41	4.09	0.03	
IntraOcular Pressure	19	83.9	4.58	q13.41	83.8-84.1	51.53-51.57	0.3	3.77	
Triglycerides	1	22.1	5.22	p36.22	21.6-22.4	10.84-11.03	1.42	3.13	
Triglycerides	1	19.4	5.18	p36.23-p36.22	18.5-19.9	9.07-9.54	0.51	4.11	
Triglycerides	1	18.2	4.98	p36.23-p36.22	18-20	8.96-9.58	0.47	3.97	
Triglycerides	1	20.6	4.73	p36.22	20.1-20.7	9.69-10.23	0.95	3.15	
Urea	21	27.3	5.80	q21.3	25.7-29.4	28.39-30.96	3.25	1.87	
Uric acid1	13	116.3	4.48	q33.3	116.2-116.5	108.74-108.83	2.56	1.26	
Uric acid1	13	116.7	4.45	q33.3	116.6-116.8	108.92-108.93	2.61	1.19	
Uric acid1	17	92.2	4.66	q24.1-q24.2	91.9-92.3	64.2-64.57	2.61	1.38	
Waist	5	178.4	5.06	q34	178.1-179	166.56-166.94	0.42	4.11	

Supplementary Table 6 - Regional heritability results that exceeded the suggestive but not the genome-wide significance threshold in individual cohorts

The RH -$\log_{10}$(*p-value*) of each region is displayed in the logP column. The chromosome and start and end positions of each region (in Mbp) are provided. Additionally, the start of each 0.3 cM region is also indicated in cM. The total trait heritability (h^2) and heritability explained by the region (h^2reg), are shown, as well as the chromosome band (Band) where the region is located.

Trait	Chr	cM	Mbp	logP	h^2	h^2reg	Band
Orkney							
Albumin	5	15	5.83-5.97	5.47	0.28	0.019	p15.32
CRP	19	70.5	45.34-45.43	6.25	0.3	0.022	q13.32
Vis							
CRP	1	175.2	159.53-159.76	5.68	0.54	0.075	q23.2
GPT	11	84.3	76.51-76.75	5.59	0.19	0.031	q13.5
Uric acid1	4	23.4	9.91-10.54	5.99	0.31	0.033	p16.1
Uric acid2	4	23.4	9.91-10.54	5.55	0.31	0.032	p16.1
GS							
BMI	2	0.6	0.47-0.68	6.14	0.47	0.002	p25.3
BMI	13	27	31.16-31.38	5.93	0.47	0.002	q12.3
Body fat	12	67.8	53.64-54.24	5.70	0.44	0.001	q13.13
Body fat	14	100.8	97.59-97.64	5.96	0.44	0.002	q32.2
Creatinine	4	90.3	77.14-77.44	5.56	0.44	0.002	q21.1
Forced Vital Capacity	9	39.9	19.45-19.47	5.56	0.34	0.001	p22.1
Glucose	13	20.7	28.45-28.59	6.19	0.23	0.002	q12.2
HDL	12	149.7	125.26-125.34	5.58	0.5	0.003	q24.31
HDL	16	84.9	66.94-68.43	5.57	0.5	0.002	q22.1
HDL	18	68.7	46.56-47.18	6.07	0.5	0.003	q21.1
HDL	20	66.9	44.14-44.68	6.23	0.5	0.004	q13.12
Heart Rate	20	55.8	36.74-36.9	5.70	0.25	0.002	q11.23
Height	1	204.9	183.98-184.35	5.50	0.82	0.002	q25.3
Height	2	253.2	232.79-233.32	5.89	0.82	0.002	q37.1
Height	5	181.5	168.21-168.33	5.39	0.82	0.001	q34
Height	6	92.1	80.92-81.77	6.01	0.82	0.003	q14.1
Height	6	164.4	152.2-152.42	5.66	0.82	0.002	q25.1
Height	10	106.8	81.06-81.18	6.61	0.82	0.002	q22.3
Height	15	108.6	84.36-85.45	6.36	0.82	0.002	q25.2-q25.3
Sodium	16	85.8	69.54-70.76	5.44	0.19	0.003	q22.1
Urea	1	169.8	154.97-156.09	6.36	0.22	0.003	q21.3-q22
Urea	18	63	43.13-43.3	6.96	0.22	0.004	q12.3

Supplementary Table 7 - RH meta-analysis results that pass the suggestive but not genome-wide significance threshold

The meta-analysis -$\log_{10}$(*p-value*) is indicated (logP column) for each peak. These peaks represent 0.3 cM regions that start at the cM position indicated (cM column). The start and end positions of these regions are also shown in Mbp, as is the chromosome band (Band) where these regions can be found. The per-cohort -$\log_{10}$(*p-values*) at the peak region are shown (O = Orkney, S=Shetland, G=GS, V=Vis, K=Korčula). The final column shows the genes in each region (or on the same chromosome band) that have been implicated in GWAS of the corresponding trait reported in the literature.

Trait	Chr	cM	Mbp	logP	Band	O	S	G	V	K	GWAS
Calcium	3	134.1	121.82-122.29	6.54	q13.33-q21.1	3.37	4.91	NA	0.30	1.31	*CASR*
Creatinine	5	200.1	176.78-176.96	5.97	q35.2-q35.3	0.30	0.58	7.45	1.45	0.36	*SLC34A1*
CRP	19	70.5	45.34-45.43	5.55	q13.32	6.25	1.25	NA	0.3	NA	*TOMM40*, *APOC2*, *APOE*, *APOC4*, *APOC1*
Glucose	13	20.7	28.45-28.59	5.58	q12.2	0.3	0.7	6.19	1.94	0.53	*PDX1*
Glucose_nodiab	3	179.7	170.38-170.77	5.63	q26.2	0.36	0.33	8.17	0.3	0.58	*SLC2A2*
Glucose_nodiab	13	20.7	28.45-28.59	5.43	q12.2	0.3	0.58	7.75	0.56	0.3	*PDX1*
HDL	11	126	116.55-117.1	6.65	q23.3	0.30	0.64	8.86	0.84	0.31	*APOA1,APOA5,APOA4,APOC3*
HDL	12	149.7	125.26-125.34	5.94	q24.31	0.76	1.96	5.58	1.34	0.46	*SCARB1*
HDL	1	219.6	201.72-201.98	5.83	q32.1	1.32	2.19	3.1	0.77	2.59	*NA*
Heart Rate	2	246.3	228.29-228.52	6.43	q36.3	NA	0.50	0.68	NA	7.60	*COL4A3*
Heart Rate	20	55.8	36.74-36.9	5.56	q11.23	NA	1.79	5.7	NA	0.32	*KIAA1755*
Height	6	49.2	25.69-26.67	6.84	p22.2	0.30	0.30	9.16	0.65	0.76	*HIST cluster*
Height	12	79.2	66.3-66.42	6.73	q14.3	0.72	1.77	7.31	0.41	0.83	*HMGA2*
Height	6	54.3	34.04-34.24	6.66	p21.31	0.36	1.03	7.99	NA	0.65	*HMGA1*
Height	10	106.8	81.06-81.18	6.58	q22.3	1.55	0.30	6.61	1.76	0.65	*ZMIZ1, PPIF*
Height	1	33.9	17.19-17.5	6.29	p36.13	0.30	1.45	6.89	0.39	1.49	*MFAP2*
Height	15	108.6	84.36-85.45	5.95	q25.2-q25.3	1.83	0.49	6.36	0.53	0.90	*ADAMTSL3*

Trait	**Chr**	**cM**	**Mbp**	**logP**	**Band**	**O**	**S**	**G**	**V**	**K**	**GWAS**
Height	2	253.2	232.79-233.32	5.83	q37.1	2.13	0.3	5.89	0.74	0.91	*DIS3L2, NPPC, ALPP*
Height	1	150.6	118.84-119.31	5.68	p12	0.42	1.19	6.99	NA	0.3	*SPAG17, PHGDH*
Height	6	132.6	126.44-127.53	5.55	q22.31-q22.33	0.44	0.3	7.53	NA	0.48	*CENPW*
Height	4	33.6	17.74-18.27	5.5	p15.32-p15.31	0.54	0.43	7.42	NA	0.3	*FAM184B, NCAPG, LCORL*
Height	15	146.1	100.78-100.87	5.49	q26.3	0.3	1.17	5.01	2.78	0.3	*IGF1R, ADAMTS17*
Total Cholesterol	1	92.1	62.83-63.38	5.77	p31.3	0.57	0.3	8.02	0.3	0.7	*ANGPTL3, DOCK7*
Urea	1	169.8	154.97-156.09	5.9	q21.3-q22	0.3	0.79	6.36	1.7	NA	*MTX1, GBA*