# Sequencing in Isolation: Next-generation sequencing studies in founder populations

Arthur Leonard Gilly

Gonville & Caius College
University of Cambridge

September 2018

This dissertation is submitted for the degree of Doctor of Philosophy

Word count: 43,782
(excluding Preface and References)

# Abstract

**Introduction**. Although common variants are routinely assayed in populations, rare mutations and copy-number variants are understudied contributors to the aetiology of complex traits. Isolated populations hold the promise of increased power gains in detecting associations in rare and low-frequency variants that have drifted up in frequency due to founder events and geographical isolation. Population-specific imputation reference panels and very low-depth whole-genome sequencing have been proposed as ways to boost power in next-generation association studies while keeping sequencing costs low.

**Objective**. The aim of this work is to leverage the wealth of sequencing data generated as part of the HELIC project to study the allelic architecture of complex phenotypes and identify sequence variants associated with traits of medical relevance.

**Methods**. We develop METACARPA, a method that meta-analyses summary statistics from genome-wide association studies. We establish a robust pipeline for the imputation and refinement of 1x whole-genome sequencing data, as well as a quality control and association pipeline for cohort-wide high-depth sequencing. We examine variant selection and weighting methods for genome-wide burden testing of rare variants, and write several tools for the visualisation of single-point and aggregated association results. Finally, we develop UN-CNVc, a fast copy number variant caller optimised for population-wide sequencing data.

**Results**. Applying METACARPA to a 4-way multi-array and multi-cohort analysis of the HELIC array data allowed the discovery, among others, of two lipid-associated loci, including the cardioprotective low-frequency variant rs145556679. In our cohorts, 1x data provided access to more than 100,000 low-frequency variants not discovered using an imputed chip design, and allowed to replicate a burden of low-frequency and rare cardioprotective variants in the *APOC3* gene. We discover burdens of rare regulatory and coding variants independent of known common-variant associations at known loci, such as in the *ADIPOQ* gene for adiponectin or *GGT1* for gamma-glutamyltransferase, as well as novel associations entirely driven by rare variants, such as with triglycerides for the *FAM189B* gene. We describe

two complex gene deletions influencing serum levels of this genes' protein products, called using UN-CNVc.

**Conclusion**. Very low-depth whole-genome sequencing studies are a viable alternative to the imputed array design. Higher sequencing depths allow the extensive description of the contribution of rare variants to the allelic architecture of complex trait loci, and allows reliable calling of large variants influencing protein biomarkers.

# Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the Preface and specified in the text.

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

It does not exceed the prescribed word limit (60,000) for the relevant Degree Committee.

The HELIC project described in this thesis is a longstanding team effort involving scientists from various fields of study. As a consequence, the work presented here builds upon several previously published studies, and has also led to several article publications during the course of this PhD. The following articles provided the foundation for this work:

- Hatzikotoulas K, Gilly A, Zeggini E, Using population isolates in genetic association studies, Brief. Funct. Genomics, Sep. 2014
- Panoutsopoulou K. et al., Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants., Nat. Commun., Nov. 2014

Part of this work has been presented in the following publications:

- Gilly A. et al., Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation., Hum Mol Genet., Jun 2016
- Southam L., Gilly A. et al., Whole genome sequencing and imputation in two Greek isolated populations identifies associations with complex traits of medical importance., Nat. Commun., 2017

- Gilly A., Suveges D., Kuchenbaecker K. et al., Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits., *Nat. Commun. 2018.*

The following article was submitted to the BioRXiv preprint server ahead of submission:

- Gilly A., et al., Very low depth whole genome sequencing in complex trait association studies., BioRXiv, deposited July 28, 2017, doi 10.1101/169789. *Bioinformatics - accepted.*

In several projects, such as the ones described in Chapter 5, Chapter 6 and Chapter 7, I undertook project-management duties and day-to-day supervision of analysts and students. To better reflect the inherently collaborative nature of this work and of the current research environment in applied sciences, I indiscriminately use the first person singular and plural throughout this thesis. The specific work presented here is mine unless stated otherwise.

# Acknowledgments

I am indebted to Pr. Eleftheria Zeggini for allowing me to carry out this study, for providing access to invaluable datasets and technical training, as well as for her exceptional support for my personal and professional development. I thank my supervisor Dr Angela Wood for her support and encouragement. Special thanks also go to my colleagues at the Wellcome Sanger Institute, whose help was invaluable in carrying out this project, with a particular mention for Dr.-to-be Sophie Hackinger, Ms. Lorraine Southam, Dr. Konstantinos Hatzikotoulas and Dr. Chris Finan. I also thank Grace Png, whose help was truly invaluable with the CNV project. Lastly, I would like to extend my gratitude to Kerstin Brauneder for her patience and fortitude during my moments of doubt, as well as my apologies for the numerous occasions where I brought work back into our home.

# Nomenclature

Acronyms/Abbreviations

| | |
|---|---|
| GATK | Genome Analysis Toolkit [3] |
| GC | Genomic Control |
| GRCh37, hg19 | Genome Reference Consortium human reference, build 37 |
| GRCh38 | Genome Reference Consortium, human reference, build 38 |
| GWAS | Genome-wide association study |
| HRC | Haplotype Reference Consortium[4] |
| HWE | Hardy-Weinberg Equilibrium |
| LD | Linkage Disequilibrium |
| LMM | Linear Mixed Model |
| LoF | Loss of function (variant) |
| MHC | Major Histocompatibility Complex |
| RVAS | Rare variant association study |
| SNP | Single-nucleotide polymorphism |
| SNV | Single-nucleotide variant. It is sometimes used in lieu of SNP to include rare variants. Both abbreviations, when used in this thesis, refer to the full allelic spectrum. |
| VQSR | Variant Quality Score Recalibration [3] |
| WES | Whole-Exome sequencing |
| WGS | Whole-Genome sequencing |

Gene Names

| | |
|---|---|
| *ABCA12* | ATP Binding Cassette Subfamily A Member 12 |
| *ADAM22, ADAM30* | A Disintegrin And Metalloproteinase Domain-Containing Protein 22,30 |
| *ADAMTS19* | A Disintegrin And Metalloproteinase Domain-Containing Protein with thrombospondin type 1 motif 19 |
| *AGPAT3* | 1-acylglycerol-3-phosphate O-acyltransferase 3 |
| *AHRR* | Aryl-Hydrocarbon Receptor Repressor |
| *ANG* | angiogenin |
| *APOA1* | apolipoprotein A1 |
| *APOC3* | apolipoprotein C3 |
| *ARVCF* | Armadillo Repeat Gene Deleted In Velocardiofacial Syndrome, delta catenin family member |
| *BCRP3* | breakpoint cluster region pseudogene 3 |
| *C1orf56* | Chromosome 1 Open Reading Frame 56 |
| *C3orf62* | Chromosome 3 Open Reading Frame 62 |
| *CCDC170* | Coiled-Coil Domain Containing 170 |
| *CCL3, CCL4* | C-C Motif Chemokine Ligand 3, 4 |
| *CCL3L3* | C-C Motif Chemokine Ligand 3 Like 3 |
| *CCNK* | Cyclin K |
| *CETP* | Cholesteryl Ester Transfer Protein |
| *CYP2C18, CYP2C9* | Cytochrome P450 Family 2 Subfamily C Members 18 and 9 |
| *DGKD* | Diacyclglycerol Kinase Delta |
| *DLEU1* | deleted in lymphocytic leukemia 1 |
| *DTX1* | deltex E3 ubiquitin ligase 1 |
| *EIF3B* | Eukaryotic Translation Initiation Factor 3 Subunit B |
| *FAM189B* | Family With Sequence Similarity 189 Member B |
| *GABRD* | Gamma-Aminobutyric Acid Type A Receptor Delta Subunit |

| | |
|---|---|
| *GRAMD1B* | GRAM Domain Containing 1B |
| *GPT* | Glutamic-Pyruvic Transaminase/Alanine-aminotransferase |
| *GGT* | Gamma-Glutamyltransferase |
| *HBB* | haemoglobin subunit beta |
| *HLA* | Human Leucocyte Antigen |
| *HLADRB5* | major histocompatibility complex, class II, DR beta 5 |
| *HMGA2* | high mobility group AT-hook 2 |
| *HNF1A* | HNF1 homeobox A |
| *ISL1* | ISL LIM Homeobox 1 |
| *JAKMIP1* | janus kinase and microtubule interacting protein 1 |
| *KALRN* | Kalirin RhoGEF kinase |
| *KATNAL1* | katanin catalytic subunit A1 like 1 |
| *KCNC3* | potassium voltage-gated channel subfamily C member 3 |
| *LCAT* | lecithin-cholesterol acyltransferase |
| *LTBP1* | latent transforming growth factor beta binding protein 1 |
| *LY6G5C* | lymphocyte antigen 6 family member G5C |
| *MAEL* | maelstrom spermatogenic transposon silencer |
| *MBIP* | *MAP3K12* binding inhibitory protein 1 |
| *MMP26* | matrix metallopeptidase 26 |
| *MTRF1* | mitochondrial translation release factor 1 |
| *NOMO1, NOMO2, NOMO3* | NODAL Modulator 1-3 |
| *OR51M1* | olfactory receptor family 51 subfamily M member 1 |
| *OR52I2* | olfactory receptor family 52 subfamily I member 2 |
| *PAFAH1B2* | platelet activating factor acetylhydrolase 1b catalytic subunit 2 |
| *PCSK7* | proprotein convertase subtilisin/kexin type 7 |
| *PQLC2* | PQ loop repeat containing 2 |

| | |
|---|---|
| *PTK2B* | protein tyrosine kinase 2 beta |
| *R3HDM1* | R3H domain containing 1 |
| *RECQL4* | RecQ like helicase 4 |
| *RNASE4* | ribonuclease A family member 4 |
| *SHISA9* | Shisa Family Member 9 |
| *SIK3* | SIK family kinase 3 |
| *SIDT2* | *SID1* transmembrane family member 2 |
| *SLC24A3* | solute carrier family 24 member 3 |
| *SLC9A3* | solute carrier family 9 member 3 |
| *SNPH* | syntaphilin |
| *STKLD1* | serine/threonine kinase like domain containing 1 |
| *SUGP1* | SURP And G-Patch Domain Containing 1 |
| *TAGLN* | transgelin |
| *TBC1D3B* | TBC1 Domain Family Member 3B |
| *TMPRSS6* | transmembrane serine protease 6 |
| *TFDP2* | transcription factor Dp-2 |
| *TFF3* | Trefoil Factor 3 |
| *UBE3A, UBE4A* | ubiquitin protein ligase E3A, E4A |
| *UGT1A1-UGT1A10* | UDP glucuronosyltransferase family 1 member A1-10 |
| *ULK1* | unc-51 like autophagy activating kinase 1 |
| *USP40, USP47* | ubiquitin specific peptidase 40, 47 |
| *ZNF44* | zinc finger protein 44 |

# Table of Contents

# Figures

# Tables

# Chapter 1. Introduction

## 1.1.Genetic variation and complex traits

### 1.1.1.Background

Inherited traits have historically been categorised into monogenic phenotypes, which are influenced by a single gene and usually inherited in a Mendelian fashion, and polygenic or complex ones, which result from variants in multiple genes and their interaction with environmental factors. This thesis focuses on quantitative traits, which are important indicators of health and act as biomarkers or risk factors for many human disorders.

Single-nucleotide polymorphisms (SNP) or single-nucleotide variants (SNV) are single base pair changes that exert their effect on complex traits either directly, by disrupting gene function, or indirectly, by affecting regulatory mechanisms. These effects can be characterised by their direction, magnitude and the frequency of the effect allele in the study population. In general, variant effects are thought to be inversely proportional to their frequency: if a variant has severe effect, it is likely to have been under selective pressure thereby making it rarer. Non-deleterious variants on the other hand can increase in frequency purely due to random sampling, a phenomenon known as drift. Although drift of rare alleles occurs at random and equally at all sites, selection promotes beneficial alleles and represses fitness-decreasing ones.

### 1.1.2.Structural variants

SNVs are not the only class of variants to influence complex traits. Structural variants (SV) are an important class of large changes in the genetic sequence of an individual. These events can have a large influence on multiple diseases and traits[5,6], and range from duplication or deletion of large parts (or even entire) chromosomes, inversions, translocations, duplications and deletions[7]. The latter two are usually grouped together under the term "copy number variants" (CNV) as they are able to change the representation

of a sequence to integers different from those expected by the ploidy of the organism. When these copy number changes affect a gene, they can increase or suppress expression, with potentially large downstream phenotypic effects.

### 1.1.3.Methodological aspect of complex trait association studies

The study of complex traits has been tightly linked with the availability of affordable genotyping methods. Historically, linkage studies examined complex traits with familial aggregation by focusing on reduced sets of about 1,000 SNPs or microsatellite markers, which were enough to tag the reduced amount of recombination events in family studies[8]. Candidate gene studies on the other hand focussed on small numbers of variants in single genes, which were *a priori* believed to play a role in the aetiology of the trait of interest. However, both of these methods, which were widely used until the 2010s, have produced results that have been difficult to replicate[9] and are now believed to have been mostly underpowered[10,11]. The widespread adoption of cost-effective commercial genotyping chips that assay hundreds of thousands, or even millions of SNVs, has powered the advent of large-scale, genome-wide association studies (GWAS), which have in the last ten years been successful in identifying common-frequency single-nucleotide variants associated with complex traits and diseases[12,13].

Statistical imputation allows the inference of genotypes unobserved in a sparse array from denser panels that describe the haplotypic diversity of a population. Falling prices and improvements in sequencing technologies have made it possible to assemble collections of whole-genome sequences at low depth into large imputation reference panels, which have successfully opened up access to low-frequency variant genotypes at the population level[14]. Genotyping rare variants as well as detecting structural variants requires deep whole-exome or whole-genome sequencing, which remains challenging to implement at the population level due to elevated financial costs.

Although modern genotyping methods allow access to a sufficiently large variant pool to, in theory, explain most of the observed variance of complex traits[15], common and rare variants detected to date using association studies only explain a small fraction of that amount, a

phenomenon known as "missing heritability". This phenomenon is thought to be due to a combination of factors, including phenotypic heterogeneity, epistasis, the use of testing methods or study designs with insufficient power and other methodological factors, and last but not least, the presence of a large number of variants of all frequencies with undetectably small effect sizes. This has promoted efforts to assemble large sequencing and imputed GWAS cohorts to detect both common variants with even smaller effect sizes as well as rarer variants with modest to large effects.

However the identification of effects at such sites requires very large sample sizes. At constant sample size, power can be increased by leveraging the unique characteristics of isolated, or founder, populations in genetic association studies.

## 1.2.Genetic studies in population isolates

Population isolates can be defined as subpopulations derived from a small number of individuals who became isolated because of a founding event (e.g. settlement of a new territory, famine, war, environmental disruption, infectious disease epidemics, social and/or cultural barriers) and have stayed so for many generations. Genomes tend to show higher homogeneity in isolates compared with cosmopolitan populations, which is reflected by a reduced effective population size ($N_e$ or the effective number of individuals required to explain the observed genetic variability) [16].

Population isolates also exhibit environmental and cultural homogeneity. Individuals from an isolated population tend to share a common lifestyle, diet and physical activity levels. They are exposed to similar environmental and sanitary conditions and disease vectors.

### 1.2.1.Genetic consequences of isolation

#### *1.2.1.1.Reduced haplotype diversity*

Linkage disequilibrium (LD) is the tendency of two or more variants to be co-inherited more often than expected by chance. In isolated populations, LD tends to extend over longer

distances compared with non-isolated populations. As expected through ancestral recombination, the LD intervals of older isolates tend to be shorter than those of younger isolates [17]. Longer stretches of LD in isolates mean longer haplotype sharing, which makes imputation of untyped sites easier to perform [18]. On the other hand, such high levels of correlation among sequence variants make it harder to pinpoint the exact causal variants within a wide association peak signal, a process known as fine-mapping.

#### *1.2.1.2.Reduced allelic variability and genetic drift*

Due to the enrichment of some rare alleles resulting from the combined effect of endogamy, bottlenecks, genetic drift, recurrent mutation and selection, isolates have been shown to potentially exhibit an increased incidence of recessive disorders. Each isolate shows a unique profile of rare disease alleles [19], which can be expressed through a higher prevalence of some diseases and lower incidence of others [20,21].

In population isolates, certain alleles reach fixation or extinction at a particular locus, while certain others are lost [22]. Some variants that contribute to complex traits/diseases are rare in the parent population and drifted to higher frequency in the isolate. The enrichment of low-frequency alleles in the study population can empower the identification of these variants with smaller discovery sets. The phenomenon of reduced allelic variability, combined with extended LD, is expected to improve power for trait association at rare variants compared with populations with wider allelic diversity.

### 1.2.2.Previous evidence of complex trait locus discovery in isolates

Several studies have successfully leveraged these special population characteristics to enable the discovery of medically-relevant associated loci. In the Icelandic founder population, low-frequency and rare variants were found to be associated with sick sinus syndrome, gout, prostate cancer and Alzheimer's disease [23-26]. In a recent study in Finnish isolates, four novel loci were found to be associated with saccular intracranial aneurysms, a complex trait with a sporadic and a familial form[27]. One of these variants has drifted up 15

times in frequency compared with the Dutch general population and is virtually non-existent in other populations from the 1000 Genomes Project. A genome-wide significant risk locus for schizophrenia and bipolar disorder has been identified in an ethnically homogeneous cohort of Ashkenazi Jewish individuals[28]. The top signal (rs11098403) is an intergenic variant located in NDST3 and was replicated in 11 independent cohorts of varying ethnicities. Recently, a Greek isolated population replicated a genome-wide significant association between R19X, a cardioprotective variant in APOC3, and low blood triglyceride levels [29]. This study also demonstrated that associations discovered in population isolates can be generalizable, as the same variant (R19X) had previously been discovered in the Amish founder population [30]. These success stories have been enabled by careful study design and use of a methodology adapted to the genetic characteristics of the population, both being important considerations in order to maximise detection power when studying isolates.

### 1.2.3.Study design considerations

#### *1.2.3.1.Population matters*

Population choice is an important consideration in designing a genetic association study focusing on isolates. Factors such as the number of founding haplotypes, age of divergence from the parent population, effective sample size and degree of admixture with neighbouring populations, all play a role in the population's allelic architecture. For GWAS, the study of young founder populations with recent expansion is a powerful strategy [31] because of their higher degree of LD and reduced genetic diversity. It has been suggested that small populations that have remained stable throughout most of their history could lead to more economical locus discovery efforts due to a stronger effect of drift[32]. The power of a genetic association study in a population isolate will depend on the enrichment of alleles that are relevant to a phenotype of interest[33-35]. An association study in an isolate can often be motivated by a suspected higher prevalence of a trait or disease in that particular population.

#### *1.2.3.2.Sequencing: how many samples, at what depth?*

GWAS arrays assay variants selected to represent common, and increasingly, low-frequency variation across the genome. The exome chip, which focuses on likely functional coding low-frequency and rare variants, has been used successfully in founder populations, for example to associate rare variants with proinsulin[36] and HDL cholesterol levels. The decreasing cost of sequencing makes it increasingly easier to study the complete variation landscape irrespective of allele frequency [37].

Whole exome sequencing has the advantage of reduced cost compared with whole-genome sequencing, but is restricted to exonic regions. Previous experience from GWAS strongly indicates that the majority of complex trait signals reside outside of genes. High-depth sequencing is considered necessary to call high-quality variants across the allele frequency spectrum [38]. However, it has been shown that in the context of a population study, whole-genome sequencing many individuals at low depth can have variant detection power advantages over fewer individuals sequenced at higher depth [39,40].

When not all samples can be sequenced, whole-genome sequencing of a subset of cases and controls following an initial GWAS has proven to be a successful strategy for empowering rare variant association in complex trait studies [23,41]. Variants from the sequenced samples are phased using long-range haplotype phasing, then imputed back into the whole sample set, using the sequenced subset as a reference panel for imputation.

### 1.2.4.Analytical considerations

#### *1.2.4.1.Relatedness*

One intrinsic consequence of genetic isolation is relatedness among individuals, which can conflict with the assumptions of independence of many commonly used analysis tools and inflate test statistics affecting association signals. An efficient approach is to account for relatedness in the association analysis through the use of a linear mixed model (LMM), where the phenotype $y$ is expressed for each SNP in $n$ individuals as:

$$\boldsymbol{y} = \boldsymbol{W\alpha} + \boldsymbol{x}\beta + \boldsymbol{Zu} + \boldsymbol{\epsilon}$$

where $\boldsymbol{y}$ is a $n \times 1$ vector of phenotypes, $\boldsymbol{W}$ is a $n \times c$ matrix of (fixed effects) covariates including the intercept and $\boldsymbol{\alpha}$ is a $c \times 1$ vector of corresponding effects, $\boldsymbol{x}$ is a $n \times 1$ vector of genotypes, $\beta$ is a scalar, the effect of the SNP on the phenotype. $\boldsymbol{Z}$ is a $n \times m$ loading matrix and $\boldsymbol{u}$ is a $m \times 1$ vector of random effects. In the absence of known family structure, $m = n$, otherwise the analysis can be stratified according to the $m$ pedigrees. $\boldsymbol{\epsilon}$ the error of the model, is distributed $\mathcal{N}(0, \sigma \boldsymbol{I_n})$ where $\boldsymbol{I_n}$ is the $n \times n$ identity matrix. $\boldsymbol{u}$ follows $\mathcal{N}(0, \lambda\sigma \boldsymbol{K})$ where $\lambda$ is a scaling factor (the ratio between the two variances) and $\boldsymbol{K}$ is a $m \times m$ relatedness matrix between groups. In the absence of a known pedigree, $m = n$ and $\boldsymbol{K}$ is $n \times n$, the relatedness matrix between each individual.

Until recently, computation of an exact association test statistic for a significant non-null effect $\beta \neq 0$, such as the Wald statistic or likelihood ratio (implemented in EMMA [42]) was computationally impractical. Tools that compute approximate solutions have been developed [43-45], however optimized versions of the exact test, such as GEMMA [1] or FaST-LMM [46] are now widely used.

Several methods have been proposed to improve on the power of single-point tests for rare variants by combining information across multiple variants[47-50]. Relatedness information can be incorporated in the model, such as in famSKAT [51], MONSTER [52] or other tools [52-55].

#### *1.2.4.2.Imputation*

When performing association based on genotyping arrays, it is common practice to impute untyped variants based on a reference panel (e.g. the 1000 Genomes Project (www.1000genomes.org), the UK10K study data (www.uk10k.org, or more recently the Haplotype Reference Consortium[14]) to enhance the resolution of the study. This approach, where positions that were not genotyped in the sample are added using phase information in the reference set, is also relevant to refining genotype calls for low-depth sequencing data.

Imputation is closely related to phasing, a procedure that infers haplotypes based on identity by state (IBS), with other phased individuals. Relatedness is helpful for phasing because it increases the likelihood of finding a long IBS string of variants; the more related the samples, the more certain it is that these IBS sequences are actually inherited identical-by-descent (IBD) [56,57].

#### *1.2.4.3.Meta-analysis*

The synthesis of data through meta-analysis can increase the power of association studies. Two different classes of methods have been typically applied in traditional meta-analysis of GWAS: P-value-based tests and effect size-based methods, which can be further subdivided into fixed or random effects models [58]. Fixed effects models assume that the same underlying effect is present in all studies, whereas random effects models allow for effect sizes to be different. These approaches can be applied to meta-analysing data across isolates. However, in the era of rare variant association testing, allelic heterogeneity can decrease power either because of the presence of similarly associated multiple rare variants or different ethnic backgrounds in the populations being meta-analyzed [59]. In addition, meta-analysis generally assumes independence of the study samples, which does not hold in the case of within-isolate meta-analysis. Research in this field is still ongoing [60], and a continued effort in method development is needed.

## 1.3.Objectives

The objective of this PhD thesis will be to identify sequence variants associated with quantitative traits of medical relevance, using founder populations to boost power. This will be done by :

- putting in place robust sequence calling and quality control pipelines, with wide applications in low-depth (1x) and high-depth (>15x) sequence-based association studies;

- identifying challenges associated with analysing these data and developing statistical methods to address them;
- identifying robustly replicating association signals for traits of medial relevance.

# Chapter 2. The HELIC study: genotype and phenotype datasets

## 2.1.Background

The HELIC (HELlenic Isolated Cohorts) study is composed of two cohort-based datasets. The HELIC-MANOLIS (Minoan isolates), collection, which was named in honour of Manolis Giannakakis (1978–2010), comprises individuals from the mountainous Mylopotamos villages, including Anogia, Zoniana, Livadia and Gonies (estimated population size of 6,000 in total). Residents of the Mylopotamos villages have over the centuries preserved their customs and traditions, which date back to Minoan times (2700-1200 BC), and speak their own dialect. The HELIC-Pomak collection comprises individuals from the Pomak villages located in the regional unit of Xanthi (estimated to be 25,000 in total) (Figure 1). The Pomaks represent a small minority of Muslims in Greece with uncertain historical origins, inhabiting the regional units of Xanthi, Rhodope and Evros [61]. They are presumed to be either descendants of natives of the Rhodope mountains (on the border of Bulgaria and Greece [62]) or to have come from Asia [63]. The Pomak language is thought to be either a Bulgarian or a Slavic dialect; however, today most Pomaks are also fluent in Turkish and Greek. After World War II and through the end of 1995, most Pomaks lived in a military-restricted zone, access to which required special permission [64].

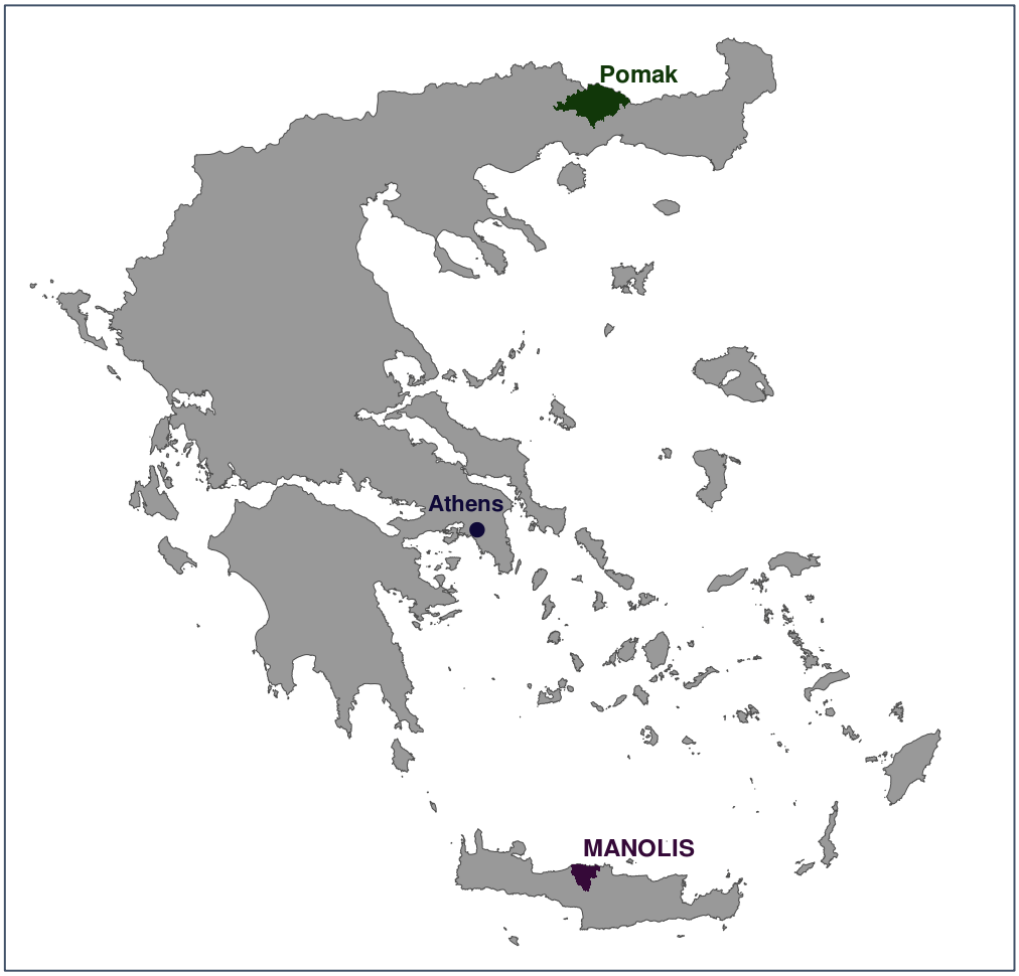



Figure 1: Location of both collection areas on a map of Greece.

## 2.2.Phenotype collection

A range of quantitative and discrete phenotypes were measured in both cohorts. For association studies, we concentrate on 50 biochemical, anthropometric and blood pressure traits (Table 1: Quantitative phenotypes selected for association analysis in HELIC). 12 of these phenotypes are also relevant when adjusted for BMI, and were hence analysed independently after regressing them on that variable. Biochemical traits are further subdivided in the “Glucose and Insulin”, “Haematological”, “Lipid” and “Other” profiles. Several phenotypes are not commonly measured in association studies, notably bone-growth related markers (BGP) as well as various endocrine and organ function traits (FT4, gammaGT, SGP, leptin, adiponectin, Bilirubin, Fe_iron and Ferritin).

| trait family | trait abreviation | trait name | unit | BMI adjustment |
|---|---|---|---|---|
| Anthropometric | BMI | Body Mass Index | $kg.m^{-2}$ | |
| | Height | height | cm | |
| | Hip | hip circumference | cm | ✓ |
| | Waist | waist circumference | cm | ✓ |
| | Weight | weight | $kg.m^{-2}$ | |
| | WHR | waist/hip ratio | none | ✓ |
| Blood pressure | DBP | diastolic blood pressure | mmHg | ✓ |
| | SBP | systolic blood pressure | mmHg | ✓ |
| Glucose and insulin | HOMA_b | homeostasis model assessment, beta cell function | | ✓ |
| | HOMA_ir | homeostasis model assessment, insulin resistance | | ✓ |
| | RG | random glucose | $mmol.L^{-1}$ | ✓ |
| | RI | random insulin | $\mu IU.ml^{-1}$ | ✓ |
| | FG | fasting glucose | $mmol.L^{-1}$ | ✓ |
| | FI | fasting insulin | $uIU.ml^{-1}$ | ✓ |
| Haematological | GRAN | granular cell count | $10^9.L^{-1}$ | |
| | GRANPC | granular cell percent | % | |
| | HCT | haematocrit | % | |
| | HGB | haemoglobin | $g.dL^{-1}$ | |
| | LPCR | large platelet concentration ratio | % | |
| | LYM | lymphocytes | $10^9.L^{-1}$ | |
| | LYMPC | lymphocyte percent | % | |
| | MCH | mean corpuscular haemoglobin | Pg | |
| | MCHC | mean corpuscular haemoglobin concentration | $g.dL^{-1}$ | |
| | MCV | mean corpuscular volume | fL | |
| | MID | monocytes, eosinophils, basophils | $10^9.L^{-1}$ | |
| | MIDPC | monocytes, eosinophils, basophils, percent | % | |
| | MPV | mean platelet volume | fL | |
| | NEU | neutrophils | $10^3.L^{-1}$ | |
| | PDW | platelet distribution width | dL | |
| | PCT | plateletcrit | $\mu g.L^{-1}$ | |
| | PLT | platelets | $10^9.L^{-1}$ | |
| | RBC | red blood cells | $10^{12}.L^{-1}$ | |
| | RDW | red cell distribution width | fL | |
| | RDWPC | red cell distribution width, percent | % | |
| | WBC | white blood cells | $10^9.L^{-1}$ | |
| Lipid | HDL | high-density lipoprotein | $mmol.L^{-1}$ | |
| | LDL | low density lipoprotein | $mmol.L^{-1}$ | |
| | TC | total cholesterol | $mmol.L^{-1}$ | |
| | TG | triglycerides | $mmol.L^{-1}$ | |
| | VLDL | very low density lipoprotein | $mmol.L^{-1}$ | |
| Other biochemical | adiponectin | adiponectin | $\mu g.mL^{-1}$ | ✓ |
| | BGP | bone growth protein (osteocalcin) | $ng.mL^{-1}$ | |
| | Bilirubin | bilirubin | $mg.dL^{-1}$ | |
| | CRP | C-reactive protein | $mg.L^{-1}$ | |
| | Fe_iron | Iron levels | $mmol.L^{-1}$ | |
| | Ferritin | ferritin levels | $pmol.L^{-1}$ | |
| | FT4 | free thyroxin | $ng.dL^{-1}$ | |
| | gammaGT | gamma glutamyl transpeptidase | $IU.L^{-1}$ | |
| | SGP | alanine aminotransferase | $IU.L^{-1}$ | |
| | TSH | thyroid stimulating hormone | $IU.mL^{-1}$ | |
| | leptin | leptin | $ng.mL^{-1}$ | |

Table 1: Quantitative phenotypes selected for association analysis in HELIC

Due to different protocols being used for different genotype datasets, not all phenotypes were included in all analyses presented in this thesis, which primarily focuses on non-haematological traits. Furthermore, 273 quantitative proteomics phenotypes from the Cardiovascular II (https://www.olink.com/products/cvd-ii-panel/), Cardiovascular III (https://www.olink.com/products/cvd-iii-panel/) and Metabolism (https://www.olink.com/products/metabolism-panel/) panels provided by OLINK were generated in a subset of 1,455 individuals in the MANOLIS cohort. 1,328 of these also had high-depth sequencing data available.

## 2.3.Genotype array data

The MANOLIS and Pomak cohorts were each genotyped in two tranches: one on the Illumina HumanOmniExpress BeadChip and Illumina HumanExome BeadChip, and one on the Illumina HumanCoreExome beadchip. The genotypes from the OmniExpress and HumanExome chips were merged into a single dataset. This merged genotype dataset is referred to as the "OmniExome" dataset, which contained 1265 samples and 621,908 variants for the MANOLIS cohort and 1003 samples and 612,403 variants for HELIC Pomak. For the additional samples genotyped on the Illumina HumanCoreExome-12-v1.1, 211 samples and 529,604 variants passed QC in MANOLIS, 734 samples and 529,086 variants in Pomak.

## 2.4.Sequencing data

### 2.4.1.Low-depth sequencing

250 samples from the HELIC MANOLIS cohort were put through whole-genome sequencing (WGS) at 4x depth using Illumina HiSeq 2000 sequencers. The samples were selected from the total pool of 1,118 samples so as to be maximally unrelated using an identity-by-descent statistic ($\hat{\pi} < 0.15$). This ensured that the haplotypes present in these samples would best represent the whole cohort. Following sequencing, 1 ethnic outlier was excluded, bringing

the sample count to 249. The quality control work was performed by Jeremy Schwartzentruber and the sample selection by Ioanna Tachmazidou.

### 2.4.2.Imputation reference panel

An imputation reference panel was built using IMPUTE2[65] by Lorraine Southam. It contained the phased haplotypes of 1092 samples from the 1000 Genomes Project Phase 1 study, 3781 7x WGS samples from the UK10K TwinsUK [66] and ALSPAC [67] studies, as well as the 249 MANOLIS samples whole genome sequenced at 4x depth.

### 2.4.3.Very low depth WGS

990 and 1166 samples were sequenced at an average 1x depth from the MANOLIS and Pomak cohorts, respectively. Due to the sparseness of the 1x datasets, sample-level QC has to be performed after imputation (Chapter 4). 58 individuals were removed from the Pomak cohort due to contamination and sample swap issues (N=1108). 5 samples were excluded from the MANOLIS 1x cohort due to ethnicity issues (N=1239).

### 2.4.4.WES

Genotype quality is an important consideration for genome-wide association studies. In order to evaluate sensitivity and specificity in genic regions, we sequenced the exomes of 5 individuals from each cohort at an average 75x depth.

### 2.4.5.High-depth WGS

1485 samples from the MANOLIS cohort and 1647 from the Pomak cohort were selected for high-depth whole-genome sequencing on the basis of material availability and prior sample swap/duplication information. In MANOLIS, 3 samples failed sequencing QC and 25 further samples were removed for downstream genotype and sample QC reasons (N=1457). Details of sample quality control for high-depth WGS are given in Chapter 5.

## 2.5. Satellite datasets

### 2.5.1. TEENAGE

We use TEENAGE data for comparison and evaluation purposes in Chapter 4, and as a calling cohort in Chapter 7. The TEENs of Attica: Genes & Environment (TEENAGE) study focuses on 857 adolescents (mean age 13.4) recruited from secondary public schools in Athens, Greece. A range of phenotypes, notably dietary, were collected, however we do not make use of them in this work.

100 samples from the TEENAGE study were sequenced using Illumina HiSeq X Ten technology to a target depth of 30x. The sequencing protocol was identical to the one used for MANOLIS and Pomak, except for a final multiplexing step which allowed to halve the target depth in the latter two cohorts . Initial calling and QC was performed by Allan Daly and Martin Pollard. Martin wrote a downsampling script allowing to reduce the depth by an arbitrary factor and used it to downsample the dataset to 15x. I used it to downsample it to 22.5x, the depth achieved in the MANOLIS cohort, and performed variant quality score recalibration on both the 30x and downsampled datasets.

### 2.5.2.INTERVAL

I use data from the large cosmopolitan INTERVAL WGS project for replication in Chapter 4 and Chapter 6.

#### *2.5.2.1. Dataset preparation*

The INTERVAL randomized controlled trial is a large-scale study focusing on healthy blood donors [68] with sequencing data available. We use the first release of the INTERVAL whole-genome sequencing data to replicate burden signals observed in MANOLIS, Pomak and the meta-analysis. Sequencing, variant calling and quality control was performed for 3,762 INTERVAL participants using the same protocol and pipeline as for the MANOLIS sequences by Kousik Kundu under Klaudia Walter's supervision, and using the pipeline developed in Chapter 5. 38 samples were excluded on the basis of ethnicity, excessive relatedness (pi-

hat>0.125), excess heterozygosity and contamination. VQSR thresholds of 99% and 90% for SNVs and INDELs, respectively, were applied to variant calls.

#### *2.5.2.2.Phenotype preparation*

46 phenotypes were requested to the INTERVAL analysis team for replication through submission of a research proposal and 45 were delivered (osteocalcin could not be released due to licensing issues). 25 requested phenotypes were haematological, and presented almost no sample missingness. Leptin, adiponectin, thyroid-stimulating hormone, alanine aminotransferase and insulin were assayed on the SomaLogic platform and were only available in an almost non-overlapping sample set with the sequenced samples and were discarded. Further excluding VLDL (which was a linear transform of TG), age and the E-stereoisomer of bilirubin, we obtain 12 non-haematological phenotypes for replication. We note the absence of insulin, which prevent replication RI, FI, HOMA b and HOMA ir traits. The total time between request submission and reception of the data was 114 working days (160 calendar days).

# Chapter 3. Meta-analysis of summary statistics from genome-wide association studies in the presence of sample overlap

## 3.1. Background

As open data and data sharing policies increase within the scientific community, more and more studies are expected to make their results publicly available in the coming years. This opens exciting perspectives for meta-analysis studies, which can aggregate results of genome-wide association studies in order to boost power and discover potential new genetic associations. When performing meta-analysis, particular care needs to be given to sample selection, because meta-analysed studies are supposed to contain independent samples only. However, due to privacy concerns, researchers often do not publish raw genotype data, or if they do, they scramble sample IDs so that no genotype can be traced back to the actual person.

Here, I present METACARPA, which is designed for meta-analysing genetic association studies with overlapping or related samples, when details of the overlap or relatedness are unknown. It implements and expands a method first described by Province and Borecki[69], which described a p-value based meta-analysis. However what researchers are most interested in are effect sizes, i.e. the combined effect of a particular SNP across all studies. Lin and Sullivan[70] describe how this can be done when the degree of overlap is known; METACARPA combines both these methods, estimating the overlap using Province and Borecki's tetrachoric correlation approach, and using it to perform Lin and Sullivan's effect-size based meta-analysis. This enables meta-analysis of effect sizes when sample overlap is unknown. We test this method using simulation of meta-analysis studies under increasing degrees of overlap, and apply it to the cross-platform, cross-cohort meta-analysis of imputed GWAS datasets in both HELIC cohorts.

## 3.2.Methods

When meta-analysing GWAS, both the P values and effect sizes can be meta-analysed on a per-variant basis. Researchers usually favour effect-size based meta-analysis as it allows to quantify the magnitude and direction of the combined effect, however it requires effect sizes to be available in all participant cohorts. The fixed-effects meta-statistics for effect sizes is of the form:

$$\hat{\eta} = \sum_{k=1}^{K} w_k \widehat{\eta_k}$$

where $\hat{\eta}$ is the estimator of a common effect $\eta$ across all studies, $k \in \{1..K\}$ identifies the study among the K that should be meta-analysed, $\widehat{\eta_k}$ is the effect in study $k$ and $w_k$ is a study-specific weight.

If effect sizes are not available, we combine association p-values by converting them to z-scores using $z_k = \Phi^{-1}\left({}^{p_k}/_{2}\right) \times sgn(\eta_k)$ where $\Phi$ is the cumulative distribution of the standard normal, and then perform a weighted sum to obtain the meta-analysis test statistic:

$$\hat{z} = \sum_{k=1}^{K} w_k z_k$$

Then, Z-scores are transformed back to P with the complement of the previous transformation: $p_{meta} = 2\Phi_{0,\sigma}(-|\hat{z}|)$. In both cases, the variance $\sigma \neq 1$ needs to be derived. We have:

$$Var(\hat{\eta}) = \sum_{k=1}^{K} w_k^2 \; Var(\hat{\eta}_k) + 2 \sum_{k=1}^{K} \sum_{l=k+1}^{K} w_k w_l Cov(\widehat{\eta_k}, \widehat{\eta_l})$$

$$Var(\hat{z}) = \sum_{k=1}^{K} w_k^2 \; Var(z_k) + 2 \sum_{k=1}^{K} \sum_{l=k+1}^{K} w_k w_l Cov(z_k, z_l)$$

$Var(z_k) = 1$ and $Var(\hat{\eta}_k)$ is given by the input dataset along with $\widehat{\eta_k}$. In order to calculate the covariances, we build a $K \times K$ matrix describing inter-study relatedness. Lin and Sullivan[70,71] propose:

$$Corr(\widehat{\eta_k}, \widehat{\eta_l}) \approx \frac{n_{kl}}{\sqrt{n_k n_l}}$$

the number of overlapping individuals $n_{kl}$ in relation to the studies sample sizes $n_k$ and $n_l$. However, in many cases $n_{kl}$ is unknown, or the relatedness is subtler than a simple overlap. Province and Borecki[69] propose:

$$Corr(z_k, z_l) \approx r_{tetrachoric}(z_{k_{0|1}}, z_{l_{0|1}}) = r_{k,l}$$

where $z_{k_{0|1}} = \begin{cases} 1 \text{ if } z_k \geq 0 \\ 0 \text{ if } z_k < 0 \end{cases}$ and $r_{tetrachoric}$ is the tetrachoric correlation coefficient. We assume that $r_{k,l} = Corr(z_k, z_l) = Corr(\widehat{\eta_k}, \widehat{\eta_l})$.

In the presence of overlap, weights are[70] of the form:

$$w = \frac{1}{\mathbb{1}^T \Omega_\eta^{-1} \mathbb{1}} \times \mathbb{1}^T {\Omega_\eta}^{-1}$$

where $\mathbb{1}$ is the unity vector of size $K$ and $\Omega_\eta$ is the estimated covariance matrix above. For the P meta-analysis, we add the following weight vector:

$$w = \frac{1}{\mathbb{1}^T s} \times s$$

where $s$ is a vector containing the sample sizes of all studies.

## 3.3.Results

### 3.3.1.Software package

This method is implemented in the METACARPA software (META-analysis in C++ Accounting for Relatedness using arbitrary Precision Arithmetic). Binary and sources are freely available (https://github.com/wtsi-team144/metacarpa).

### 3.3.2.Application to meta-analysis of HELIC datasets

We applied METACARPA to the analysis of 13 cardiometabolic, 9 anthropometric and 9 haematological traits among those collected in the HELIC cohorts. We conducted association with array genotypes imputed up to a reference panel containing 10,422 haplotypes, including those called from the 249 4x WGS in MANOLIS[65]. METACARPA was

used to account for the potential relatedness between the MANOLIS OnmiExpress and ExomeChip (N=1,265), MANOLIS CoreExome (N=211), Pomak OmniExpress and ExomeChip (N=1,003) and Pomak CoreExome(N=734) datasets. This allowed the identification of a new association with high density lipoprotein cholesterol (HDL) at chr16:70790626 (beta=-1.71 (SE=0.25), p=1.57x$10^{-11}$, EAF=0.007), a variant present in the MANOLIS sequences only. We also identify a cardioprotective signal (rs145556679, EAF 0.013) associated with decreased triglycerides (TG) (beta -1.13 (SE 0.17), $P$ = 2.53 x $10^{-11}$) and with very low density lipoprotein cholesterol (VLDL) levels (beta -1.13 (SE 0.17), $P$ = 2.90 x $10^{-11}$) in MANOLIS. This work was performed by Lorraine Southam and has been published along with the method[72].

### 3.3.3.Simulation and benchmark

Prior to deployment, implementation was tested by simulation. METACARPA needs genome-wide statistics to estimate sample overlap, however, simulating haplotypes genome-wide is computationally costly. Instead, we repeatedly drew two sets of 2,000 samples each, from the UKHLS GWAS dataset (EGA accession EGAD00010000890), a large set of 9,967 individuals from the cosmopolitan UK population. Sample sets were drawn with several levels of overlap (e.g. 2x1,000 samples with 500 samples in common, giving 50% sample overlap). Sample phenotypes were simulated from a standard normal. The two studies were associated separately using GEMMA[73], then meta-analyzed using METACARPA, and the whole process was repeated 1,000 times for each level of overlap. An uncorrected fixed-effects, sample-size weighted P-value based meta-analysis[74] was implemented in the software for comparison, as well as an uncorrected inverse-variance weighted, effect size based meta-analysis. We used degrees of overlap ranging from 0.5% to 75% of the total sample size (Figure 2, Figure 3). We assessed the false positive rate calculated at a genome-wide significance threshold of 5.00 x $10^{-8}$, and the power to detect a single associated SNP. Effect SNPs were chosen randomly for each simulation, MAF and effect sizes were constrained so that the effect SNP explained 1% of phenotype variance.

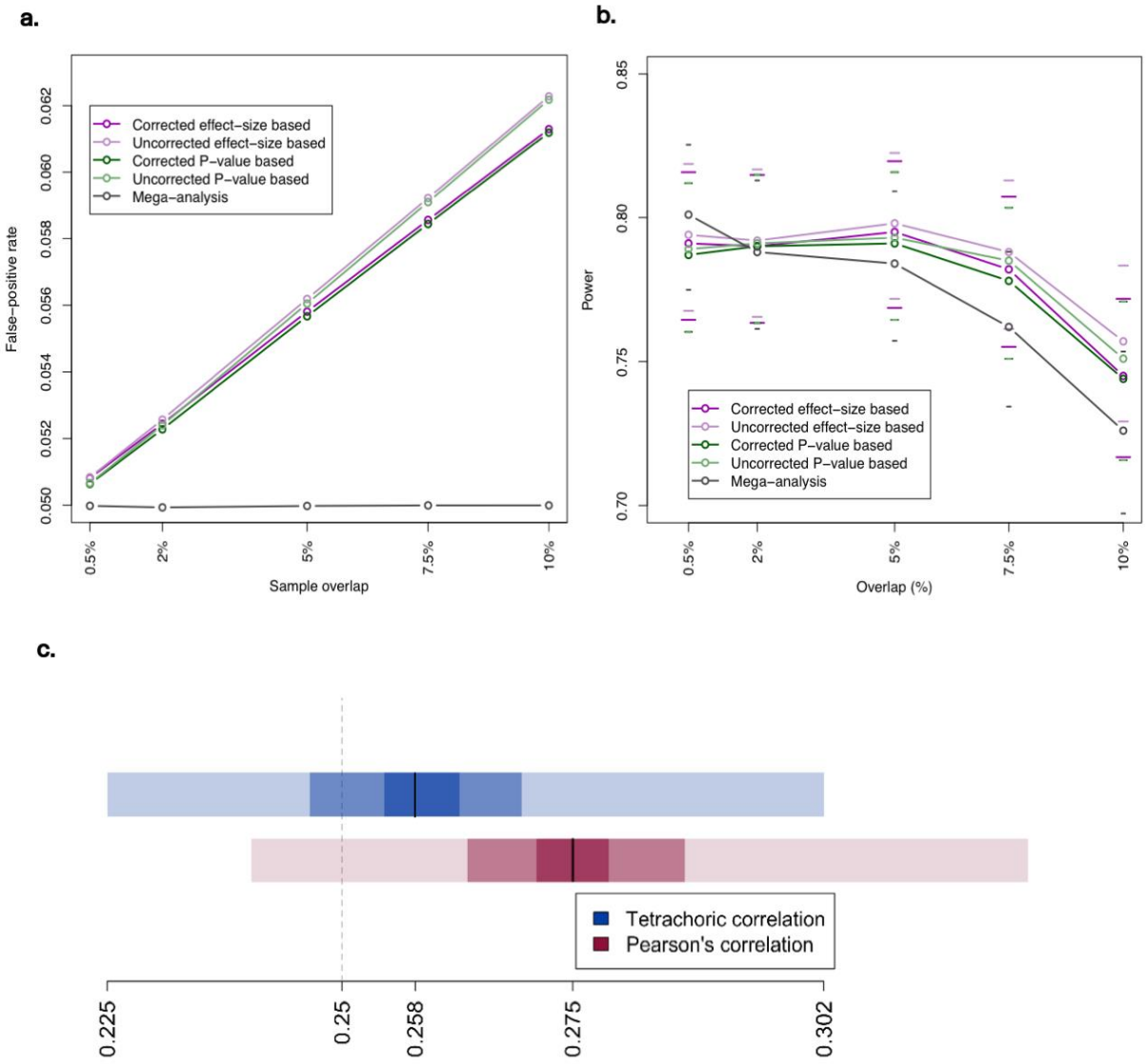



Figure 2: False positive rate and meta-analysis power in the presence of sample overlap using METACARPA. a. Empirical false-positive rate as a function of sample overlap in 1,000 repeats of a meta-analysis of two studies including 2,000 samples each, at a significance threshold of 5.00 x 10-8. b. Empirical power of the four tests implemented in METACARPA as a function of sample overlap in the same simulation setting. Power is calculated as the discovery rate of a SNP explaining 1% of a standard normal phenotype under the same simulation scenario (e.g. a MAF of 1% and an effect size of 0.705, or a MAF of 20% and an effect size of 0.176). c. Compared accuracy of Digby's estimate of tetrachoric correlation and Pearson's correlation for a true (dashed line) 25% overlap under a polygenic burden, with 10,000 SNPs affecting a quantitative trait with 20% heritability. Estimates of correlation for both methods are calculated over 300 genome-wide simulations. The black line indicates the median, shaded rectangles represent the interquintile ranges.

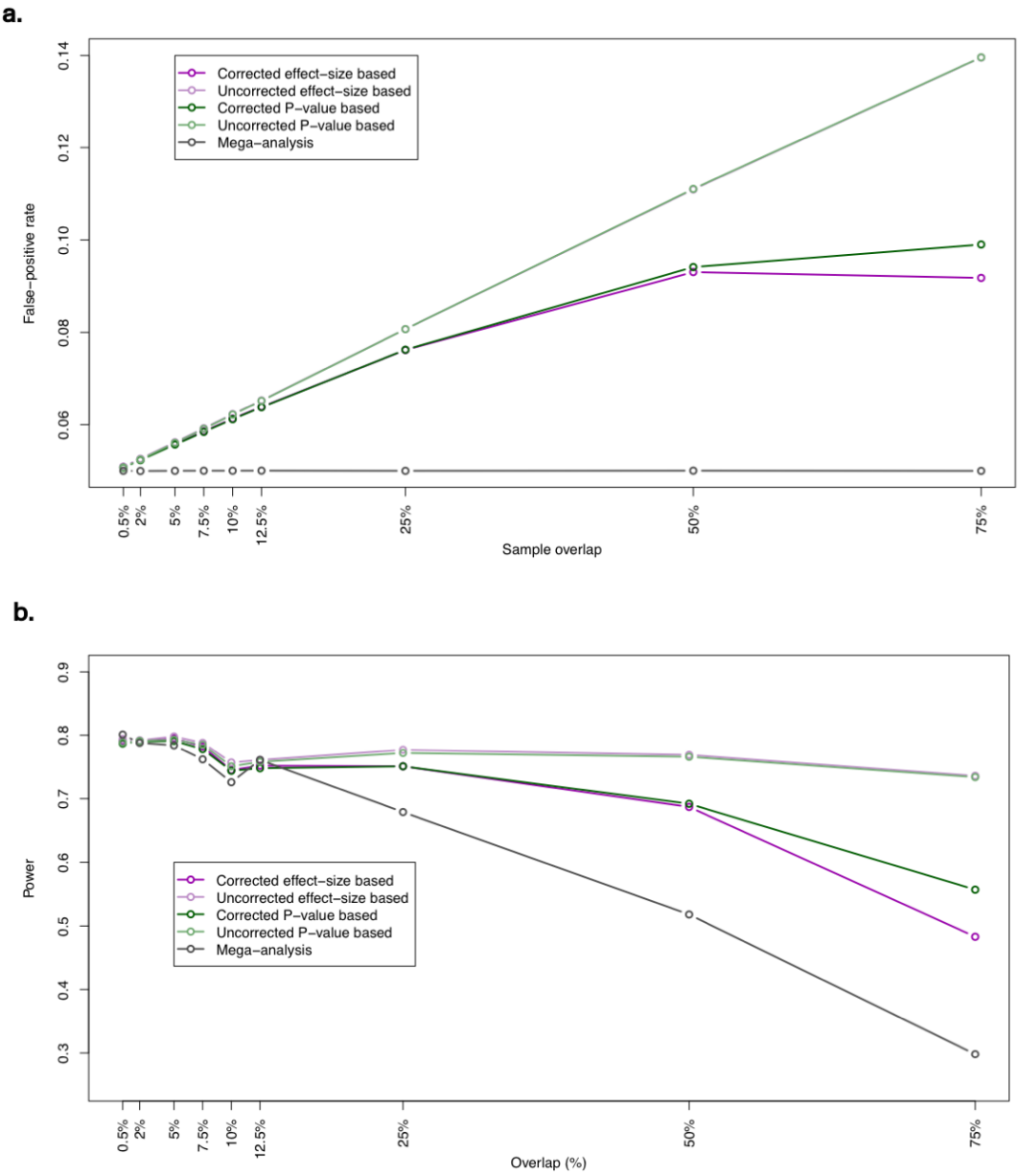



Figure 3 : False positive rate and power under large and extensive overlap.

We also added a genotype-level mega-analysis for comparison, which is the gold standard when individual level data are available and maintains the type-I error rate at nominal significance. For typical to substantial levels of overlap (0.5 - 10%), false positive rate grows linearly for both the two uncorrected and the two corrected methods (Figure 2.a). However, for the latter, the growth rate is reduced from $6 \times 10^{-5}$%/sample to $5.5 \times 10^{-5}$%/sample (8.3%). While for typical (0.5 - 5%) levels of overlap, power to detect a single SNP is conserved, for

substantial levels of overlap (5 - 10%) it drops at an approximate rate of 0.05%/sample (Figure 2.b.). For extensive levels of overlap (10 - 75%), the increase in false positive rate slows further and stabilizes around 9% for overlaps greater than 50% for both corrected methods (Figure 3.a), whereas uncorrected methods keep growing at an unchanged rate. Due to the reduction in effective sample size, power decreases to below 60% for very high levels of overlaps. At the levels of overlap inferred in the HELIC datasets (1.96% and 1.84%), power is decreased by 0.1% and false-positive rate is decreased by 0.2% between the corrected and uncorrected effect-size based meta-analyses. These small differences were confirmed by comparing METACARPA to a mega-analysis of the genotype-level data, as well as a summary-level meta-analysis not accounting for relatedness, as implemented in the GWAMA software[75]. All three methods yielded similar median statistics ($\lambda$=0.985±0.015) for association with HDL, confirming that they are all robust to the moderate levels of relatedness observed between the datasets of the HELIC study (Figure 4).

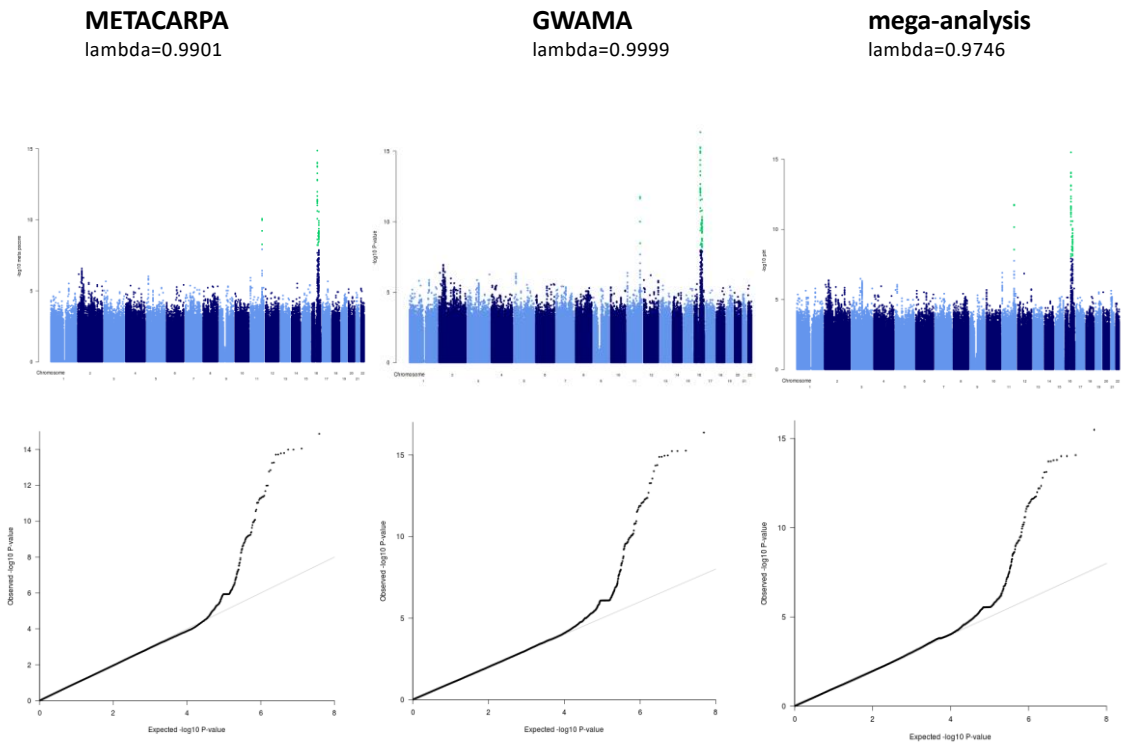



Figure 4: Comparison between three meta-analysis strategies on the HELIC MANOLIS dataset. The trait being meta-analysed is HDL. The top 3 panels are the Manhattan plots for the 3 analysis approaches taken, with the corresponding qq-plots in the lower panels. Plot by Lorraine Southam.

### 3.3.4.Accuracy of tetrachoric correlation

Tetrachoric correlation should provide a better estimation of true sample overlap than Pearson's correlation of z-scores in the presence of signal. As expected, it performs poorly under the null (Figure 5), but performs better than Pearson's under a simulated polygenic burden across 10,000 SNPs for a trait that is 20% heritable under 25% sample overlap (Figure 2.c.). This suggests that tetrachoric correlation is able to correct for a relatively high number of truly associated, correlated SNPs, a likely scenario when analysing highly polygenic traits.

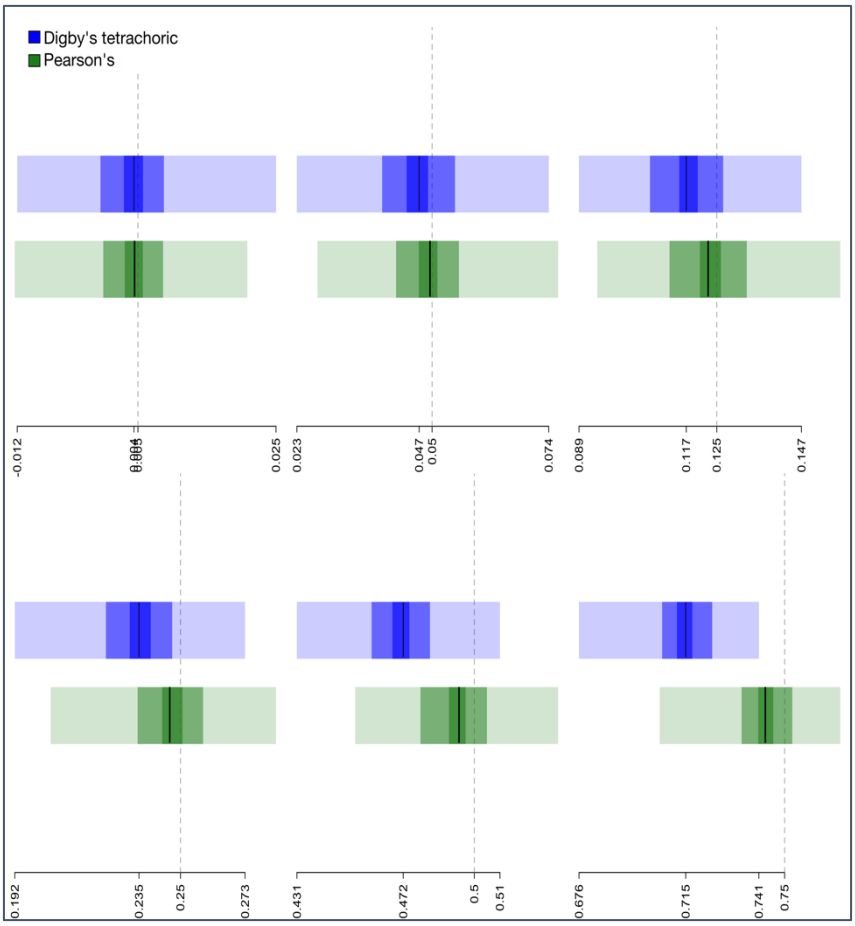



Figure 5: Estimator Accuracy when all SNPs are under the null. Green boxes represent the quintiles of the correlation estimates for Pearson's correlation of z-scores, blue boxes the quintiles of Digby's estimator of tetrachoric correlation. The simulated correlation is represented by the dashed grey line.

## 3.4.Conclusion

Through simulation, we have shown that METACARPA controls type-I error rate, albeit imperfectly, and that it conserves power for large overlaps. Tetrachoric correlation is an improvement compared to Pearson's correlation, especially under a polygenic burden of variants. We have applied the method to an analysis of several dozen quantitative traits, and have shown that it performs similarly to a mega-analysis of individual participant data. Additional theoretical research is required to discover more specific estimators, that provide an even better type-I error correction under all signal architectures.

# Chapter 4. Very low depth whole-genome sequencing in complex trait association studies

## 4.1.Background

Improvements in calling algorithms have enabled robust genotyping of whole genome sequences at low depths (4x-8x), leading to the creation of large reference panels [4,14], as well as completion of cohort-wide sequencing-based association studies [76,77].

Since sample size and haplotype diversity, not depth, are key to increasing power in association studies [39], it should be possible to leverage very low depth (<2x) sequencing data for association with medically-relevant traits. Simulations in whole-exome studies have shown that extremely low sequencing depths (0.1-0.5x) capture almost as many SNPs as imputed GWAS arrays in the common (MAF>5%) and low-frequency (1%<MAF<5%) categories[78]. The CONVERGE consortium performed the first successful case-control study of major depressive disorder in 4,509 cases and 5,337 controls using 1x very low-depth whole-genome sequencing[79]. However, no systematic evaluation of very low depth WGS based genotyping is available to date. This chapter describes a pipeline for calling, refining and imputing 1x WGS data, demonstrates the feasibility of both single-point and burden associations with quantitative traits using low sequencing depths, and examines the effects of this genotyping method on discovery power in quantitative trait association studies.

## 4.2.Methods

### 4.2.1.Datasets and QC

For MANOLIS, we merge the 4x (n=249) and 1x (n=990) datasets for a total of 1,239 samples with low and very low depth WGS, and for Pomak, we use the 1,166 samples with 1x WGS. Due to the sparseness of very low depth data and the unreliability of raw genotype calls, sample QC is performed after variant-level QC, genotype refinement and imputation to simplify filtering procedures.

#### 4.2.2.Variant Quality Score Recalibration

We perform variant-level QC before genotype refinement and imputation. Most common variant callers (GATK and samtools, notably) output a variant quality score, which is saved in the QUAL record of the VCF. The variant confidence score in QUAL is the phred-scaled ($-10.log_{10}$) probability that the variant in question is a false-positive, and is assigned by the caller. Traditionally, analysts have used QUAL, as well as other variant-level metrics (called variant annotations) to filter out variants that are likely to be false positives.

However, using QUAL as a metric to assess and filter variant quality has several limitations. First, callers can cap this quality above a certain threshold, making it impossible to determine which among two relatively good-quality variants the caller is most confident in. The score is also biased, as it assumes the underlying reads are produced with the same average quality, and that the read mapping quality metric can be trusted. However, this is not strictly true, as it has been known for some time that read quality information is differently assessed depending on sequencing cycle, independently of the actual sequencing accuracy of the read bases. It is therefore advised to use annotations in the VCF, such as read depth, strand bias, Hardy-Weinberg equilibrium and inbreeding coefficient to determine whether a variant is trustable or not. This process suffers from multiple arbitrary thresholds, and therefore leads to different "best practices" among analysts dealing with sequencing data. Alternative models aim to integrate many different annotations in a single holistic score that accurately predict variant quality, using standard machine-learning tools such as Gaussian mixture models[3] or support vector machines [80].

With VQSR (Variant Quality Score Recalibration), the GATK software suite [3] aims to correct this by providing an analysis software that automates this process, integrating all useful annotations at all variant sites, and aggregating them into a single score, the VQSLOD (variant quality score logg-odds). This score represents the odds ratio of being a true variant versus being false. It is estimated by normalising the annotations, and then by fitting a Gaussian mixture model to a set of "gold standard" variants. The full documentation of VQSR

can be found online at https://gatkforums.broadinstitute.org/gatk/discussion/39/variant-quality-score-recalibration-vqsr.

Variants are then grouped into VQSLOD "quality tranches", groups of variants with similar quality. The tranche boundaries are determined by the user, and the idea behind them is that in order to gain access to a high percentage of true positive sites, for example 99%, one has to include many more false positives than if one is interested only in 80% of variants. Variant tranches are then assigned so that, for every tranche, the percentage of true positives obtained when all tranches above the one in question are removed is equal to the value of the tranche. For example, all tranches below 99% will contain 99% of true positive variants, whereas the 99%-100% tranche will contain the remaining 1%, but also a majority of false positives.

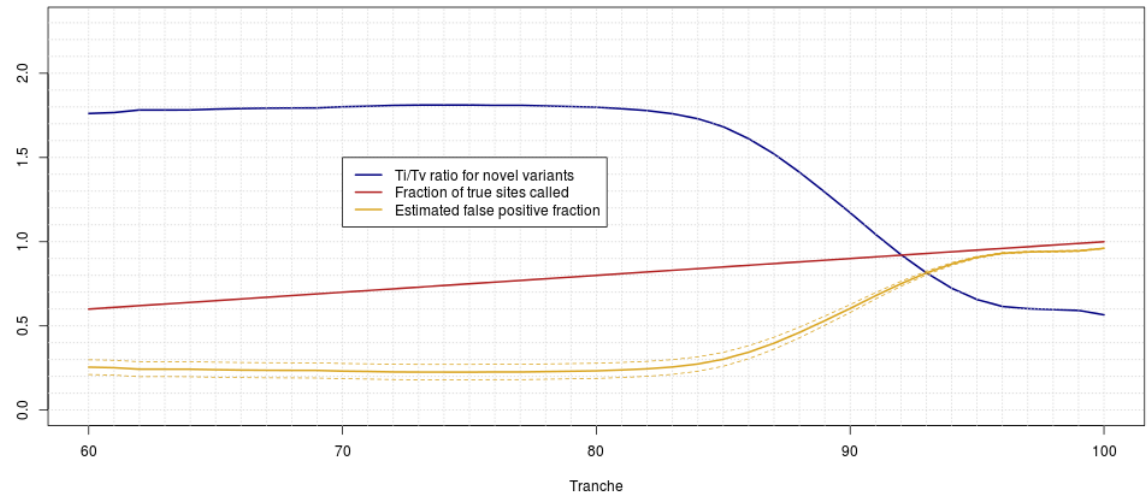



Figure 6 :VQSR recalibration curves for the MANOLIS cohort, using VQSR version 3.1.1. and default parameters. Ti/Tv ratio for novel variants is given by VQSR. The fraction of true sites called is the proportion of variants in the positive training set called at that tranche or below. The estimated false positive rate is calculated as $FP = \frac{TiTv_{expected} - TiTv_{novel}}{TiTv_{expected} - TiTv_{FP}}$ where $TiTv_{FP} = 0.5$ and $TiTv_{expected}$ is 2.6 for the lower dashed curve, 2.1 for the upper dashed curve, and 2.3 for the solid curve.

We evaluated VQSR v.2.7 and v.3.1.1 to perform variant quality score recalibration. In both cases, using the default parameters for the VQSR mixture model yields poor filtering, with a Ti/Tv ratio dropoff at 83% percent sensitivity and a Ti/Tv ratio of 1.8 for high-quality tranches (Figure 6). We therefore ran exploratory runs of VQSR across a range of values for the model

parameters, using the dropoff point of the transition/transversion (Ti/Tv) ratio below 2.0 as an indicator of good fit (Figure 7). This was implemented in the VARECH script (available at https://github.com/wtsi-team144/VeryLowDepthSequencing/blob/master/varech.pl), which varies the following parameters automatically:

| parameter | Description | Values tested |
|---|---|---|
| --badLodCutoff | Variants with a LOD score below this will be used to build the negative model. | {−2,−5} |
| --maxGaussians | Maximum number of components in the Gaussian model. | {3, 5, 8, 10, 12} |
| --maxNegativeGaussians | Similar, for the negative model. | {2, 4} |
| --minNumBadVariants | Minimum number of variants for building the negative model. | {1000, 5000, 10000, 12000, 15000, 20000} |
| | The prior for dbSNP variants (v.130) | {2, 5} |
| --qualThreshold | Variants with a quality score below this will not be included to build the model. | {0, 50, 100, 150, 200, 250, 300} |

A small number of configurations outperformed all others, which allowed us to select an optimal set of parameters. For the chosen set of parameters, false positive rate is estimated at 10%±5% (Figure 8). Indels were excluded from the dataset out of concerns for genotype quality. Variant quality score recalibration was performed using the QualByDepth, HaplotypeScore, MQRankSum, ReadPosRankSum, FS, InbreedingCoeff and DP

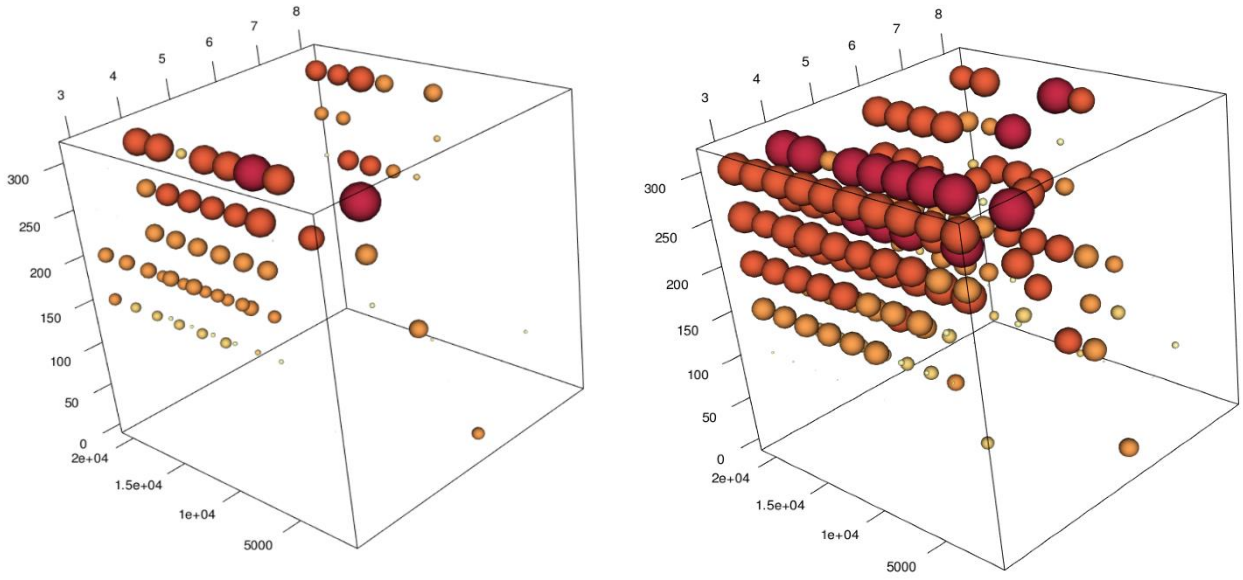



Figure 7 : VQSR optimisation grid for two versions of VQSR. Left: version 2.7.2, right: version 3.1.1. The bottom axis represents the minimum number of variants in the negative training set (minNumBadVariants), the vertical axis is the quality field threshold (qualThreshold) and the top axis is the maximum number of Gaussian distributions in the mixture model (maxGaussians). All other parameters are constant.

annotations. After the main analysis for this work was already completed, we tested a newer version, GATK v.3.6, where a different set of annotations was recommended (HaplotypeScore was removed from the model, and MQ was added).

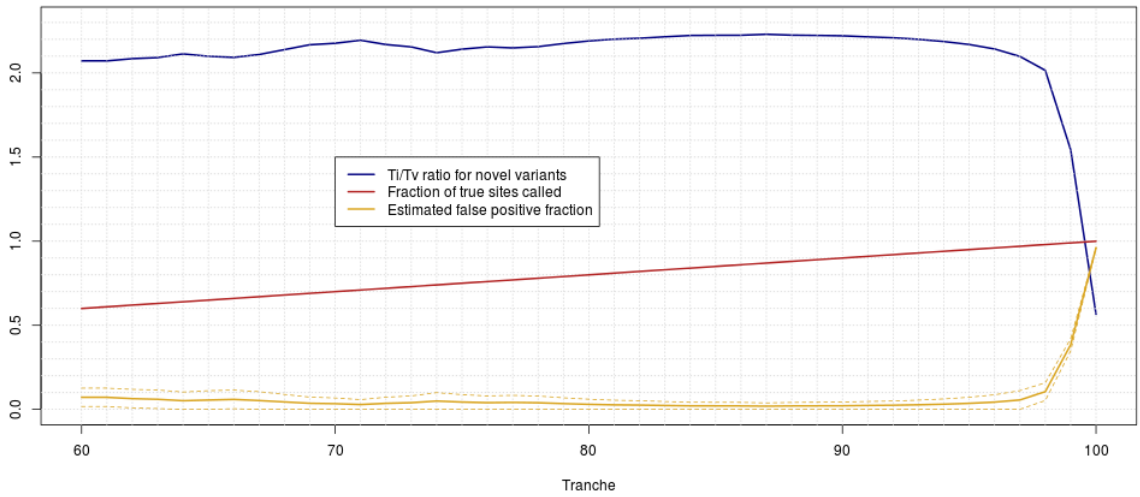



Figure 8 : VQSR recalibration plots for one of the optimal runs selected after the optimisation procedure.

With identical input, a change in VQSR version had large consequences in the recalibration output. This yielded very poor performance, with identical parameters, to the optimal run selected in v.3.1., with an increasing Ti/Tv by tranche curve instead of a strictly decreasing one. We then restored the set of annotations to HaplotypeScore, BaseQRankSum, MQRankSum, ReadPosRankSum, FS and InbreedingCoeff and obtained a very good performance, with a 99.4% sensitivity at the 2.0 Ti/Tv cutoff point, while conserving the 10% type-I error rate. This is a much better sensitivity/specificity compromise than the cut-off chosen at the analysis stage of this work, and highlights the extreme variability of recalibration results for very-low-depth sequencing data. The non-negligible effect of the version of VQSR is worrying, and highlights the dependence of researchers handling WGS data on a single algorithm that due to its complexity and rapid development pace, proves relatively hard to troubleshoot. This effect is likely to be reinforced with the introduction of even more complex, “black-box”-type algorithms such as the upcoming convoluted neural network approach (2D CNN) used for variant quality score recalibration in GATK4.

### 4.2.3.Pre-refinement genotyping accuracy

For quality control purposes, reads from 17 of the well-characterised Platinum Genomes sequenced by Illumina at 50x depth [14], and downsampled to 1x depth using samtools [64] were included in the merged BAM files. VQSR-filtered calls were then compared to the high-confidence call sets made available by Illumina for those samples. 524,331 of the 4,348,092 non-monomorphic variant sites called from 1x data were not present in the high-confidence calls, whereas 1,246,403 of the 5,070,164 non-monomorphic high-confidence were not recapitulated in the 1x data. This corresponds to an estimated false positive rate of 12% and false negative rate of 24.6% pre-imputation. Both unique sets had a much higher proportion of singletons (corresponding to MAF < 2.9%) than the entire sets (57.9% vs 19.9% of singletons among 1x calls and 51% vs 18.1% among high-confidence calls), which suggests that a large fraction of the erroneous sites lies in the low-frequency and rare part of the allelic spectrum. However, genotype accuracy is poor across the MAC range, to the point where it distorts the distribution of allele counts (Figure 9).

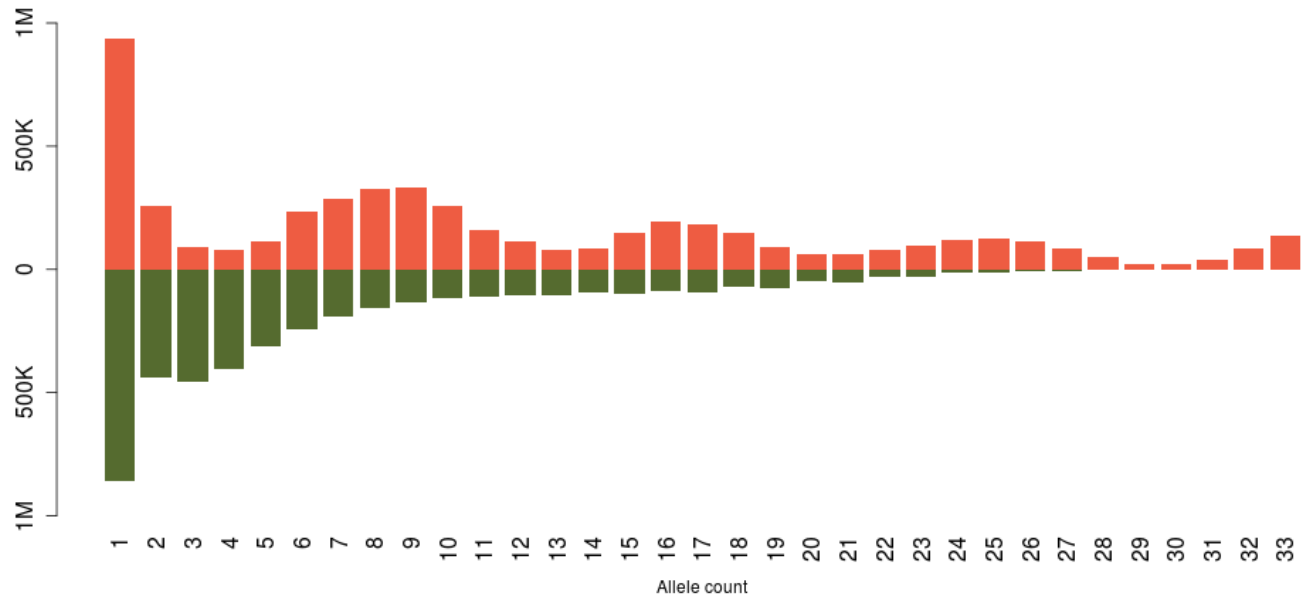



Figure 9 : Allelic spectra for non-monomorphic sites in the Platinum Genomes samples. Orange bars are the high-confidence calls from high-depth WGS, green bars are calls from downsampled BAMs at 1x. Differences in the shape of the distribution are likely due to the scarcity of information provided by the 1x data, which does not allow to recapitulate the family structure in the CEPH/Utah pedigree 1463, which composes the Platinum Genomes.

In order to improve these numbers, we performed genotype refinement and imputation using the reference panel described in Chapter 2. Due to the Platinum samples having been included in the 1000 Genomes Project and hence our reference panel, we remove the 17 Platinum Genomes prior to imputation.

#### 4.2.4.Genotype refinement pipeline

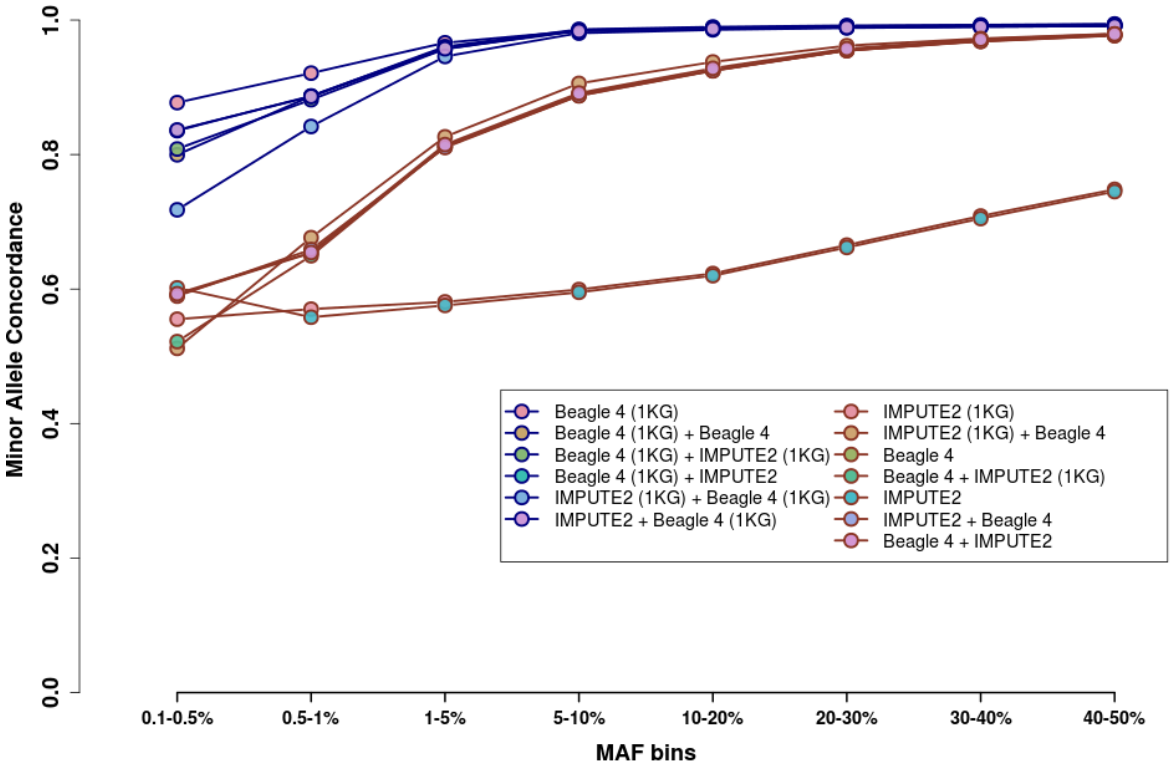



Figure 10 : Minor allele concordance for genotype refinement pipelines. Chromosome 11 data, genotype refinement without imputation. Pipelines using Beagle v4 with a reference panel are drawn in blue, pipelines not involving Beagle or not using a reference panel are coloured red. 1KG=1000 Genomes phase I reference panel.

Standard pipelines for GWAS data imputation often contain a pre-phasing step. The authors of SHAPEIT2[81], a widely used phasing software, advise to phase whole chromosomes when performing pre-phasing in order to preserve downstream imputation quality. This approach is computationally intractable for the 1x datasets, where the smallest chromosomes contain

almost 7 times more variants than the largest chromosomes in a GWAS dataset, we therefore do not perform phasing prior to imputation.

For benchmarking purposes, we designed 13 genotype refinement pipelines involving Beagle v4.0 [82] and SHAPEIT2 [81] using a 1000 Genomes phase 1 reference panel, which we evaluated against minor allele concordance against OmniExpress and ExomeChip data. All

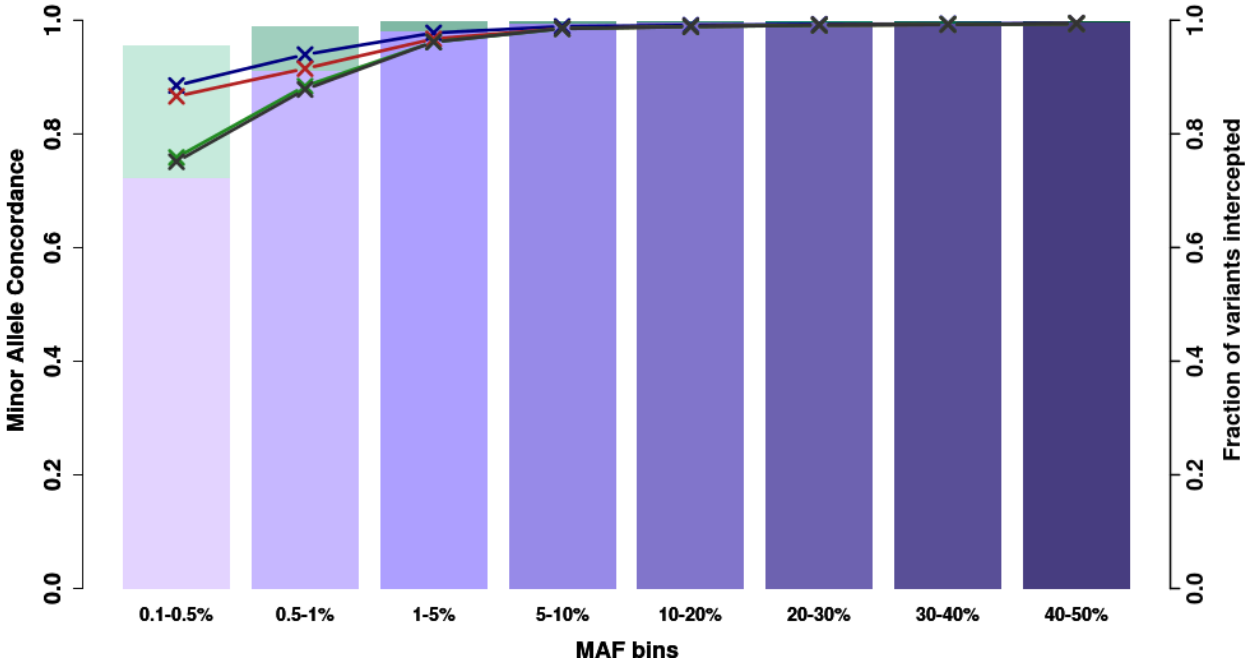



Figure 11 : Effect of decreasing chunk size and adding imputed variants on minor allele concordance and variant overlap. Concordance and overlap are evaluated against GWAS data. Purple bars: variant overlap in the refined-only dataset. Green bars: Imputed variants. Blue curve: Minor allele concordance, refined variants only using the 3-way reference panel, large chunk size (3,000 variants, 1000 flanking variants). Red curve: Refined variants only using the 3-way reference panel, small chunk size (1,500 variants, 500 flanking variants). Green curve: Imputed variants added, small chunk size. Blue curve: Beagle phasing run added (no missingness).

pipelines were run using the vr-runner scripts authored by Petr Danecek (https://github.com/VertebrateResequencing/vr-runner).

Pipelines involving Beagle with the use of a reference panel ranked consistently better (Figure 10), with a single run of reference-based refinement using Beagle outperforming all other runs. IMPUTE2 performed worst on its own, whether with or without reference panel;

in fact the addition of a reference panel did not improve genotype quality massively. Phasing with Beagle without an imputation panel improved genotype quality, before or after IMPUTE2.

After the benchmark, we used a reference panel composed of 10,244 haplotypes from the 1000 Genomes Project Phase 1 (n=1,092), UK10K [83] TwinsUK [66] and ALSPAC [67] (n=3,781, 7x WGS), and 249 MANOLIS samples sequenced at 4x depth, which has been described before [65] (Chapter 3). Alleles in the reference panel were matched to the reference allele in the called dataset. Positions where the alleles differed between the called and reference datasets were removed from both sources. Indels were filtered out due to poor calling quality.

Benchmarking imputation pipelines showed that compute time for the first round of imputation is a main bottleneck. Halving the number of SNVs per refinement chunk (including 500 flanking positions) from the 4,000 recommended by the vr pipelines resulted in only a modest loss of genotype quality in the rare part of the allelic spectrum (Figure 11), while allowing for a twofold increase in refinement speed. Genotype quality dropped noticeably for rare variants when imputation was turned on (Figure 11), but remained high for low-frequency and common ones. A reference-free run of Beagle allowed to phase all positions and remove genotype missingness with no major impact on quality and a low computational cost. We also tested thunderVCF [30] for phasing sites, however, the program took more than 2 days to run on a 5,000 SNV chunk and was abandoned.

Concordance rates were likely overestimated in the best performing run (Figure 10). Since a great number of low-frequency and rare variants are not refined, their missingness remains high, biasing estimates of MAF and concordance. In the end, the pipeline with best minor allele concordance across the board used Beagle v.4 [82] to perform a first round of imputation-based genotype refinement on 1,239 HELIC MANOLIS variant callsets, using the aforementioned reference panel. This was followed by a second round of reference-free

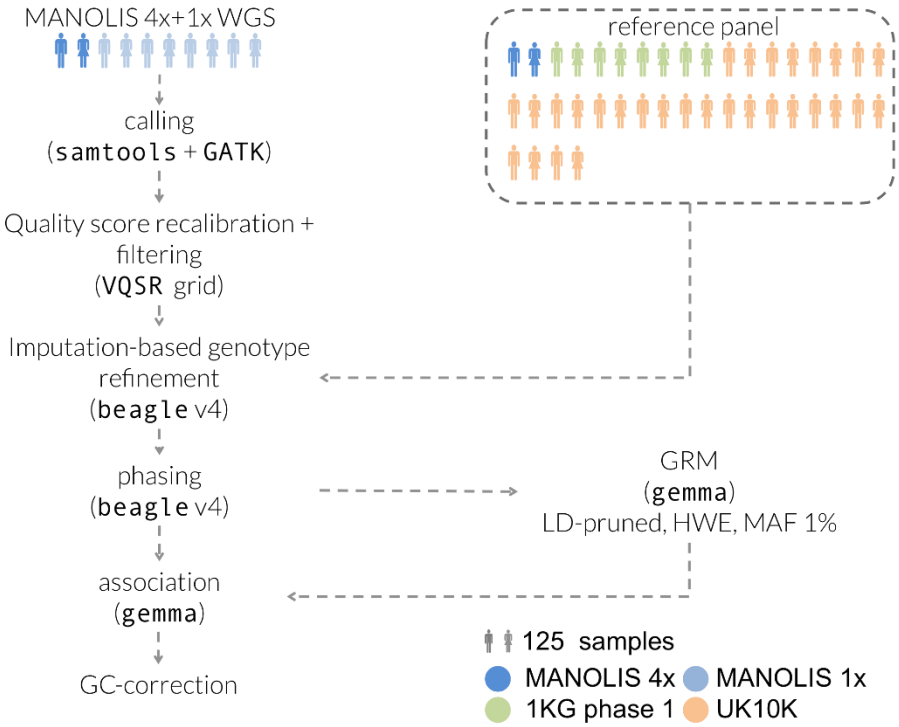



Figure 12 : Processing pipeline for the MANOLIS 1x data.

imputation, using the same software. The pipeline is represented in Figure 12 along with the reference panel composition and the association steps.

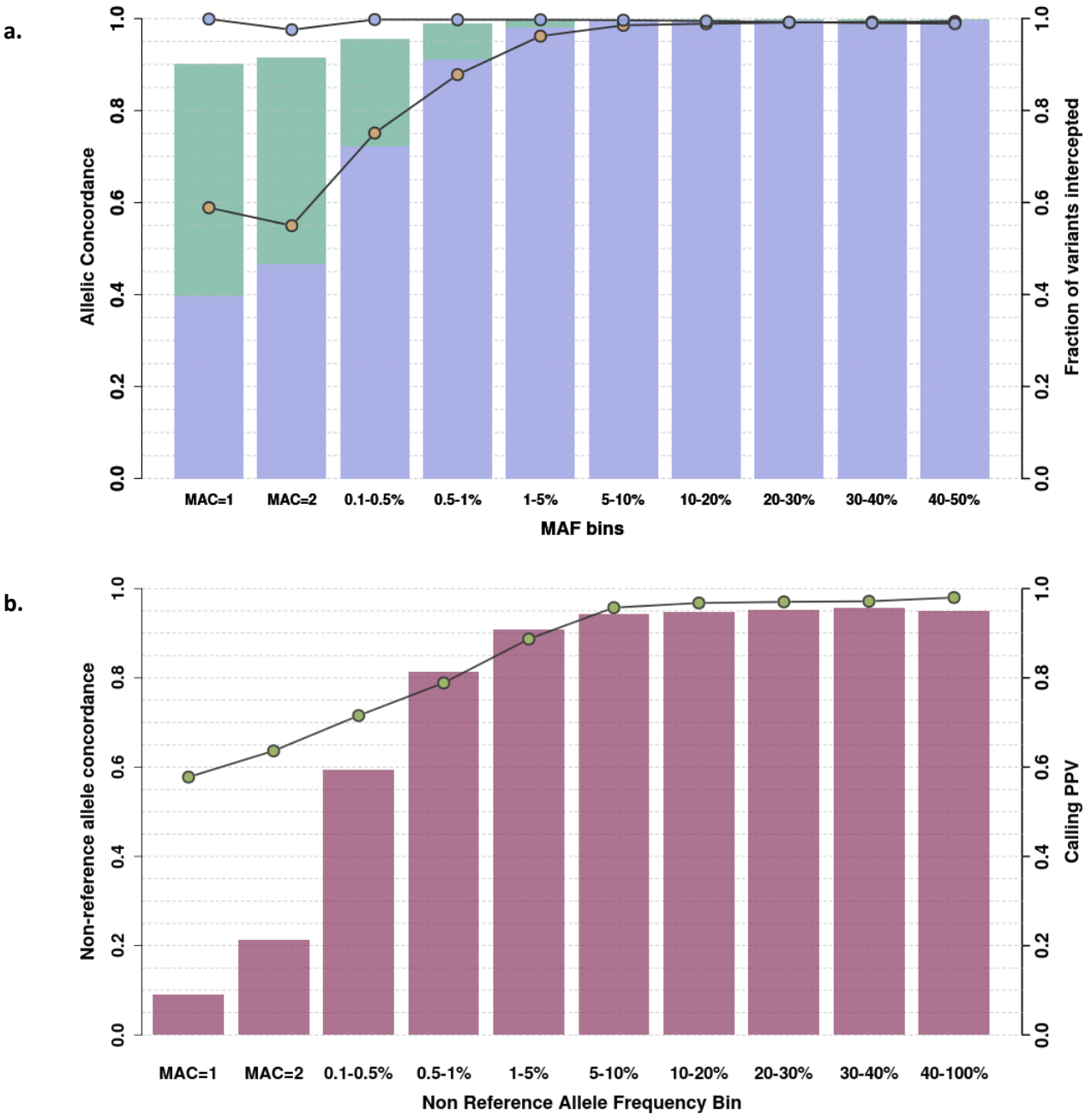



Figure 13 : Concordance and call rate for very low depth WGS genotypes. a. Genotype (blue circles) and minor allele (yellow circles) concordance is computed for 1,239 samples in MANOLIS against merged OmniExpress and ExomeChip data. Call rate is assessed for the refined (purple) and refined plus imputed (green) datasets. b. Non-reference allele concordance (green circles) and positive predictive value (PPV) (fuchsia bars) is computed for 1,225 MANOLIS samples with both 22x WGS and low-depth calls.

After imputation and QC, we captured 95% of rare, 99.7% of low-frequency and 99.9% of common variants present in the GWAS data, with an average minor allele concordance of 97% across the allele frequency spectrum (Figure 13 a.). 79.7% of 1x WGS variants were found using high-depth WGS at 22x in a subset of the MANOLIS samples (n=1,225, see Chapter 5 for dataset preparation and variant calling), although this positive predictive value varied across the MAF spectrum, from 8.9% for singletons to 95.1% for common variants (Figure 13 b.), and genotype concordance was similar, although slightly lower, than with the chip variants.

#### 4.2.5. Comparison with WES and false positive/negative rates in exonic regions

Due to the 22x data being aligned to a different build, we were unable to compute genome-wide false positive rates, as a number of regions do not map between builds, which risks biasing our estimates. However, WES data is available for a reduced set of individuals in both cohorts on the same build, making it possible to estimate sensitivity and specificity in exonic regions.

A set of high confidence genotypes was generated for the 5 exomes in MANOLIS using filters for variant quality (QUAL>200), call rate (AN=10, 100%) and depth (250x). These filters were derived from the respective distributions of quality metrics. When compared to 5 whole-exome sequences from each cohort, imputed 1x calls recapitulated 77.2% of non-monomorphic, high-quality exome sequencing calls. Concordance was high, with only 3.5%

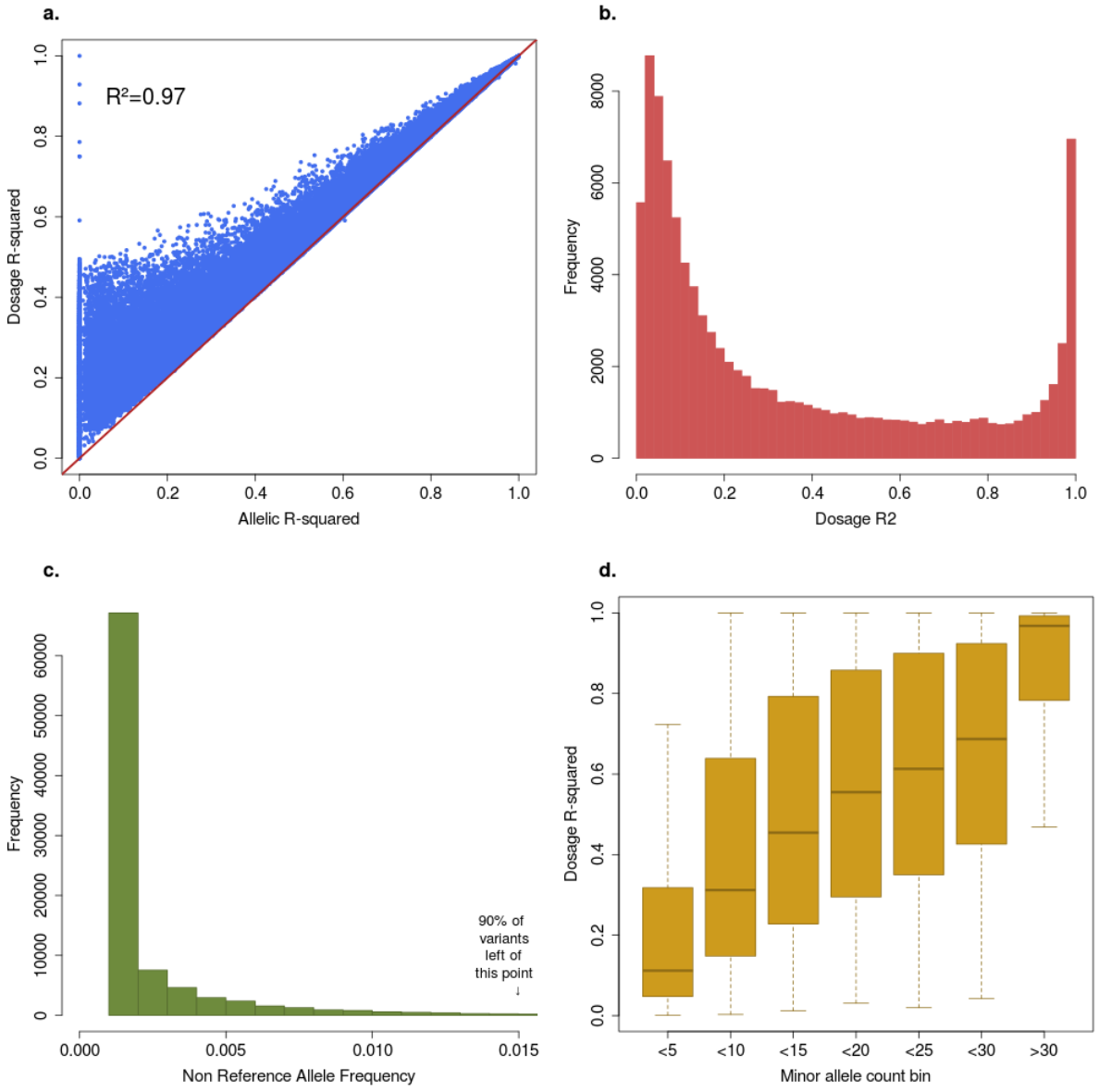



Figure 14 : Distributions and relationships of the two imputation accuracy measures provided by Beagle. Chromosome 11 data, MANOLIS, imputed positions only. a. Correlation of Beagle imputation quality metrics. b. Distribution of Dosage R-squared. c. Allele frequencies of imputed variants. d. Dosage R-squared and allele frequency.

of the overlapping positions exhibiting some form of allelic mismatch. When restricting the analysis to singletons, 9105 (58%) of the 15,626 high-quality singletons in the 10 exomes were captured, with 21% of the captured positions exhibiting false positive genotypes (AC>1). To assess false positive call rate, we extracted 1x variants falling within the 71,627 regions targeted by the Agilent design file for WES in overlapping samples, and compared them to those present in the unfiltered WES dataset. 103,717 variants were called in these regions

from WES sequences, compared to 58,666 non-monomorphic positions in the 1x calls. 1,419 (2.4%) of these positions were unique to the 1x dataset, indicating a low false-positive rate in exonic regions post-imputation.

#### 4.2.6.Post-imputation variant quality control

Beagle provides two variant-level imputation metrics, allelic R-squared(AR2) and dosage R-squared (DR2). Both measures are highly correlated (Figure 14.a). Values between 0.3 and 0.8 are typically used for filtering [84]. In both 1x datasets 59% and 91% of imputed variants lie below those two thresholds, respectively. The shape of the distribution of scores does not suggest an obvious filtering threshold (Figure 14.b). Since most imputed variants are rare and R-squared measures are highly correlated with MAF, filtering by AR2 and DR2 would be similar to imposing a MAF threshold (Figure 14.c and d.).

Due to an issue in the version of the vr-pipeline used for imputing and phasing the very low depth data, AR2 and DR2 were not reported for positions that were present both in the panel and the sequencing datasets (i.e. positions that underwent genotype refinement), only for imputed ones. This makes post-imputation QC particularly difficult in our study (more recent versions of the vr-pipelines do report imputation quality metrics for refined positions). Given these considerations, we chose not to apply any filters based on imputation quality prior to association, relying on post-association filtering and experimental genotype validation to confirm any potential signal.

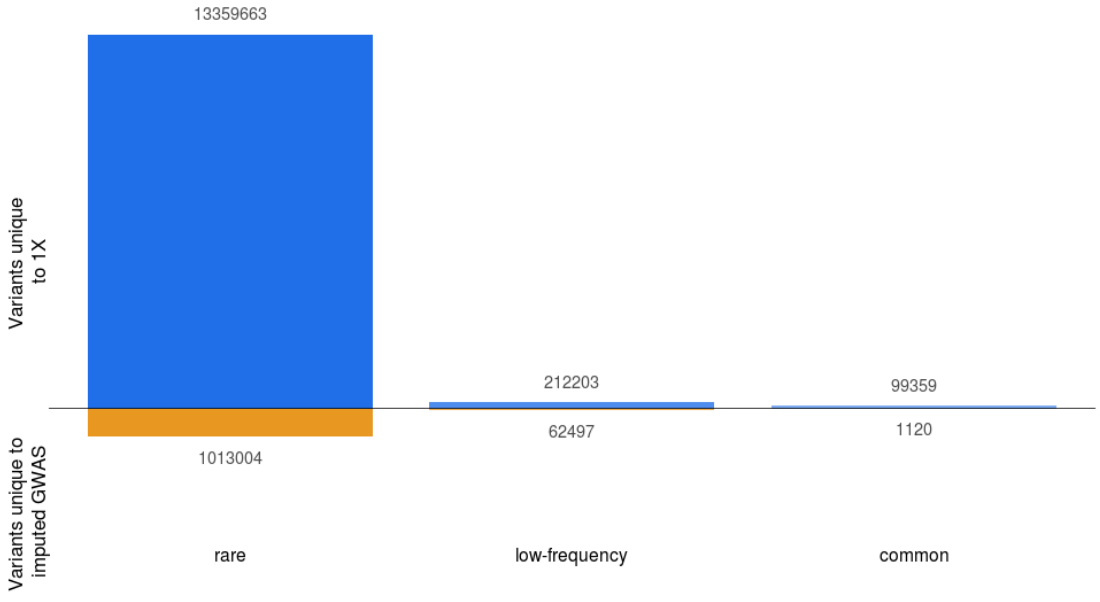
#### 4.2.7.Sample-level QC

Most individuals with low-depth sequencing have been previously genotyped on the OmniExpress and ExomeChip arrays, on the CoreExome array, or both. For each cohort, we merge array genotypes with low-depth sequencing data and perform pi-hat estimation using PLINK (see Chapter 5) to identify sample swaps and duplicates. The 4x dataset had already been QCed, and 1 sample had been removed as part of the reference panel preparation performed by Lorraine Southam. No further samples were excluded in the 1x data in the MANOLIS cohort. In the Pomak 1x dataset, 58 samples were excluded due to

widespread contamination, which was likely due to a handling incident in the automated sequencing pipeline. This quality control was performed by Rachel Moore as part of her PhD rotation in our group.

#### 4.2.8.Comparison of variants to the imputed GWAS

The genotype refinement and imputation step yielded 30,483,136 and 29,740,259 non-monomorphic SNPs in 1,239 MANOLIS and 1,108 Pomak individuals, respectively. In a subset of 982 MANOLIS individuals sequenced at 1x depth, for which a previously-described GWAS dataset imputed up to the same panel[65] was also available, we called 25,673,116 non-monomorphic SNPs using 1x WGS data. This is compared to 13,078,518 non-monomorphic SNPs in the imputed GWAS dataset, equivalent to a 96% increase in the number of variants called. Each dataset presents a small non-overlapping number of low-frequency and common variants, but the main source of differences stems from the rare variant category




(Figure 15). 13,359,663 of the variants called in the refined 1x WGS but not the imputed GWAS are rare. Excluding 82% of singletons and doubletons, 2,143,187 rare variants were not found in the imputed GWAS data and passed QC.

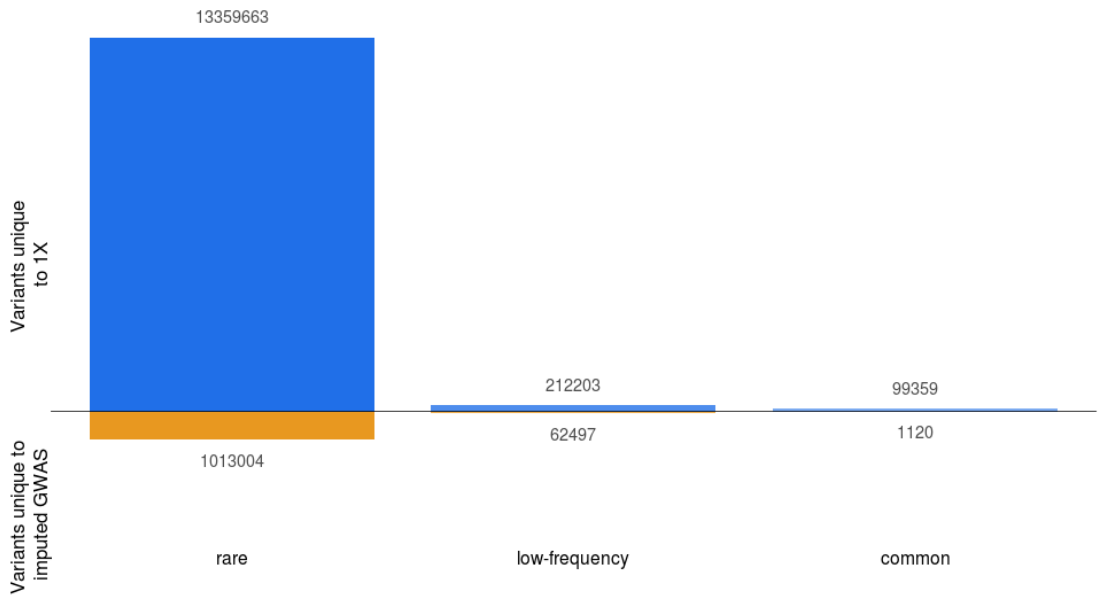



Figure 15: Unique variants called by sequencing and imputed GWAS per MAF bin. Both datasets are unfiltered apart from monomorphics, which are excluded. MAF categories: rare (MAF<1%), low-frequency (1%<MAF<5%), common (MAF>5%).

The presence of common variants unique to either dataset is surprising, especially given the use of an identical reference panel, as imputation should be particularly accurate for this frequency class. We were not able to pinpoint a single cause for the presence of unique variants in either set.

Approximately 50% of the unique common variation in the 1x dataset was explained by the accurate calling of variants not present in the reference panel cohorts (which includes 1000 Genomes Phase 1), but present in the more recent (and more dense) 1000 Genomes Phase 3 panel. The remaining half is composed either of fully novel variants (without an rsID) or variants previously reported in other populations (with an rsID) but not present in the reference panel.

Most (915) of the common variants unique to the imputed GWAS data were directly typed variants not present in the reference panel. The remaining 205 were indeed present in the panel but were filtered out from both sets due to them having unreconcilable reference and alternate alleles.

A crucial question is the proportion of true positives among these additional SNVs not found by GWAS and imputation. By comparing their positions and alleles with high-depth WGS in the same samples, we find that the PPV profile for these variants is much lower compared to when all variants are examined (Figure 16 and Figure 13). As expected, PPV is almost zero for additional singletons and doubletons, and just above 40% for the few additional common variants. 62% of low-frequency variants unique to the 1x are true positives, which corresponds to 140,844 low-frequency variants with high genotyping quality that are missed

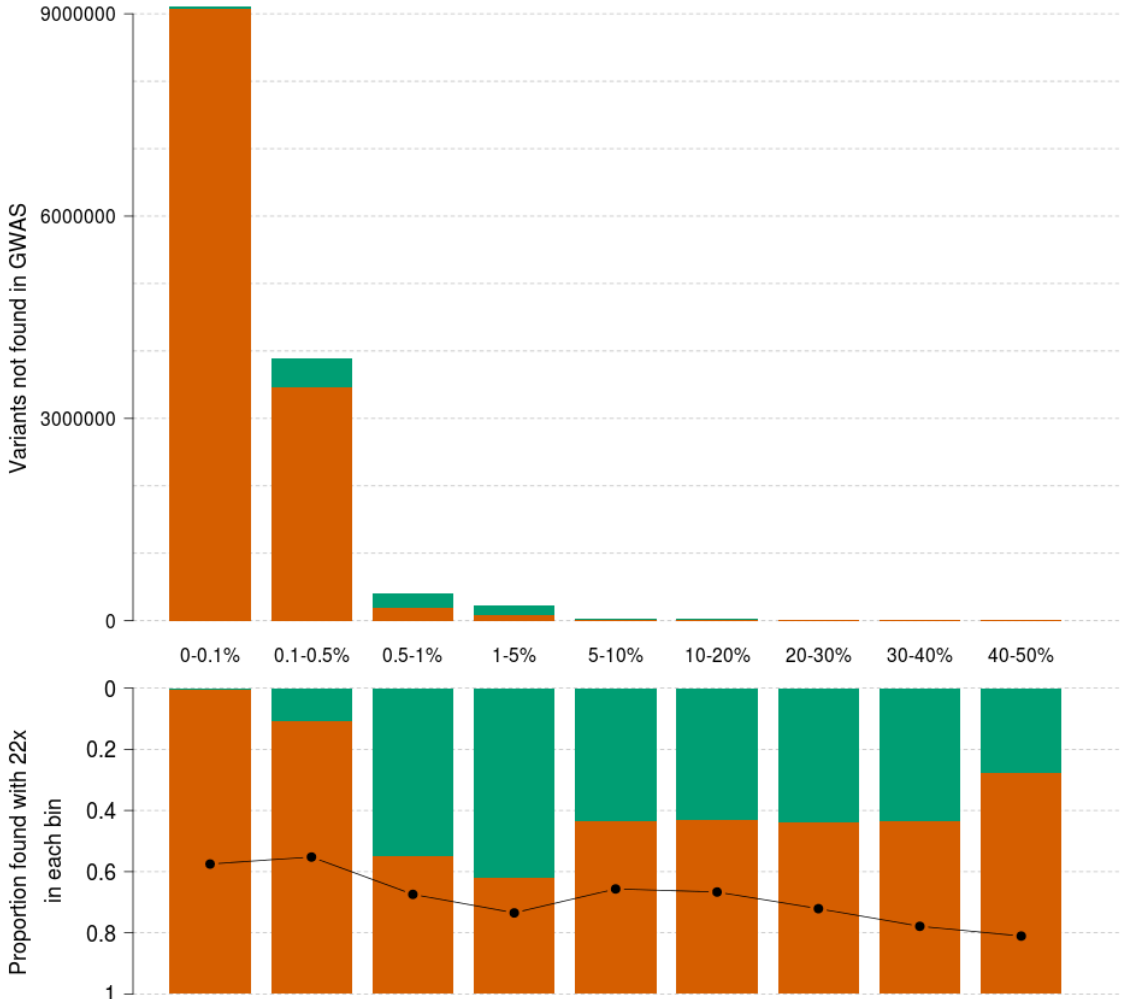



Figure 16 : Positive predictive value of additional variants called in 1x sequencing. 1x variants not found in the GWAS data, arranged by MAF bin, in raw numbers (top). Green bars count variants recapitulated in the 22x (true positives). The proportion of these over the total (positive predictive value) is displayed in each bin in the bottom panel. The black line indicates minor allele concordance for true positive variants. The first category (0-0.1%) contains singletons and doubletons only.

by the imputed GWAS. Minor allele concordance is lower than for all variants, with a lower bound at 55% for rare variants and reaching 73% for novel low-frequency variants.

#### 4.2.9.Genetic relatedness matrix

In order to correct for genetic relatedness when performing association within the two isolated cohorts, we calculated a genetic relatedness matrix using GEMMA [1], testing different variant sets to calculate kinship coefficients. Using the unfiltered 1x variant dataset produced the lowest coefficients (Figure 17), whereas a well-behaved set of common SNVs [85] produced the highest. However, these differences only had a marginal impact on association statistics, with a lambda median statistic difference of 0.02 between the two most extreme estimates of relatedness (Figure 17). This may suggest that in cohorts with less or comparable relatedness than in HELIC, the choice of variants included for the calculation of the relatedness matrix is more of a computational and/or resource problem (since using filtered sequencing data for the matrix requires markedly more time to compute) rather than a means to ensure optimal correction of test statistics.

For our association study, we used LD-pruned 1x variants filtered for MAF<1% and Hardy Weinberg equilibrium $P<1x10^{-5}$, which is the variant set that yields the smallest relatedness estimates apart from the unfiltered set of all called variants. Inflation of the test statistic was below 0.1 in both cohorts, except for a few anthropometric traits in Pomak, where outliers were present. Residual levels of inflation, defined here as $\lambda_{GC} - 1 > 0.05$, were present in most traits in both cohorts. Neither using a different set of variants to compute the relatedness matrix, accounting for additional covariates in the phenotype transformations, nor filtering out rare variants from the analysis completely attenuated the inflation. We therefore applied genomic control (GC) correction[86] to all traits as a final step of the association pipeline.

### 4.2.10.Estimating the significance threshold

For single-point association, strict Bonferroni correction for the number of tested variants and traits would result in significant loss of power, as it does not account for LD between variants and correlation between traits. We determine the significance threshold by

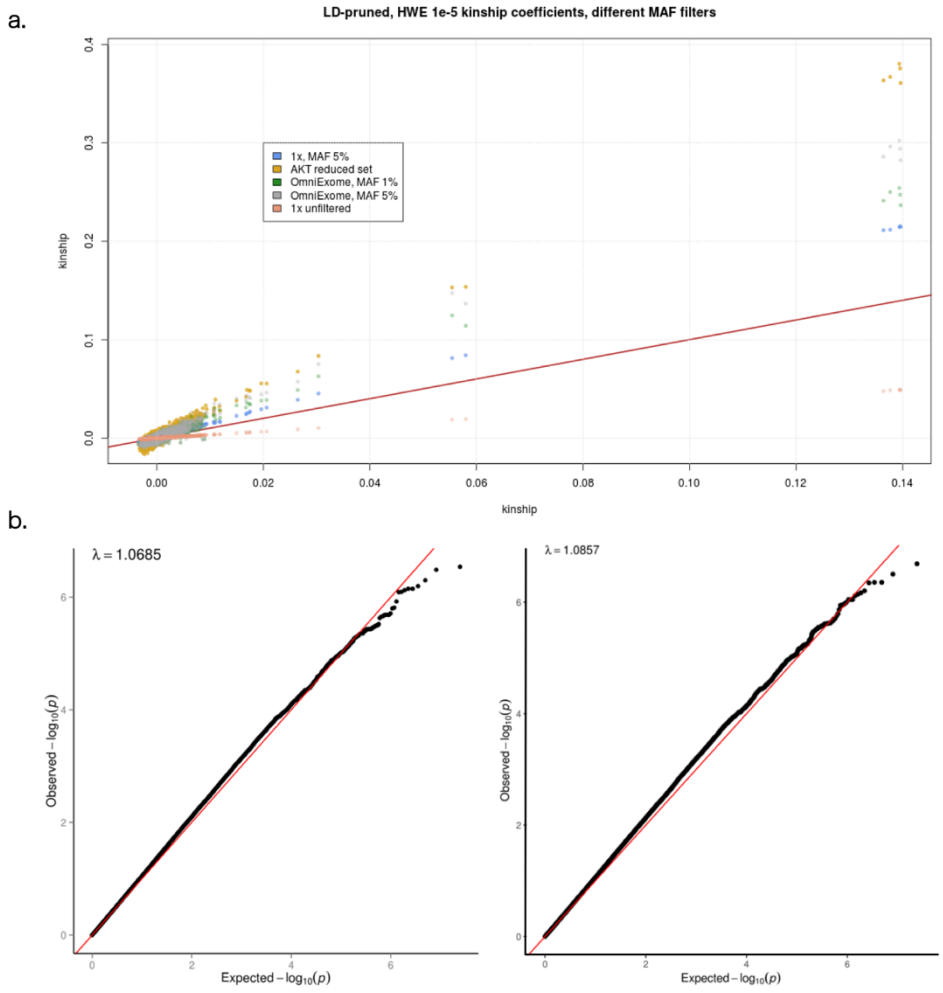



Figure 17: a. Kinship coefficients in the genetic relatedness matrix across 5 sets of sites compared to the coefficients used in the analysis (red line, 20,667,092 sites, 1x data, MAF>1%) for the POMAK dataset. All datasets are LD-pruned and filtered for Hardy-Weinberg $p<1x10^{-5}$. b. Effect of these coefficients on QQ-plots of a triglycerides (TG) association. Left: GRM calculated using the 1x unfiltered dataset (lowest values in a.). Right: GRM calculated on the AKT variant set (highest values in a.).

calculating $\alpha_{adj} = \frac{0.05}{N_{eff} \times M_{eff}}$, where $N_{eff}$ is the effective number of SNVs after correcting for LD and $M_{eff}$ is the effective number of traits tested after correcting for correlation. We estimated $M_{eff}$ using two different methods. The first method selects the number of principal components (PCs) in a principal component analysis (PCA) of standardised, normalised traits that explain 95% of total trait variance. This yielded $M_{eff} = 24$ ($M = 64$). The second method uses the Kaiser method on the eigenvalues of the trait correlation matrix to calculate $M_{eff}$ [87], and gives $M_{eff} = 31$. This approximately corresponds to the point where the variance explained curve from the PCA saturates. We use $M_{eff} = 24$, which includes both haematological and non-haematological traits.

For $N_{eff}$, we extrapolate the number of SNVs based on calibration curves[88] that provide the number of independent SNVs given the total number of tested SNVs (assuming MAF>0.005). This gives $N_{eff} = 5{,}145{,}236$ for MANOLIS ($\alpha_{adj} = 4.05 \times 10^{-10}$) and $N_{eff} = 5{,}361{,}759$ for Pomak ($\alpha_{adj} = 3.89 \times 10^{-10}$). Performing LD-pruning using PLINK [89] yields 8,123,367 variants with MAC>2 for Pomak ($\alpha_{adj} = 2.56 \times 10^{-10}$) and 6,833,823 variants for MANOLIS ($\alpha_{adj} = 3.05 \times 10^{-10}$). For MANOLIS this yields significance thresholds between We use a more relaxed threshold to declare genome-wide significance, at $\alpha_{adj} = 1.0 \times 10^{-9}$ for both cohorts.

## 4.3.Results

### 4.3.1. Comparison of association summary statistics with imputed GWAS

1x WGS calls a larger number of variants and is noisier than imputed GWAS in the same samples. To evaluate how this difference affects association study power, we performed genome-wide association of 57 quantitative traits in 1,225 overlapping samples with both imputed OmniExome and 1x WGS using both sources of genotype data in the MANOLIS cohort. We then compared independent suggestively associated signals at $p<5x10^{-7}$. These signals were then cross-referenced with a larger (n=1,457) study based on 22x WGS on the same traits in the same cohort[90] as described in Chapter 5.

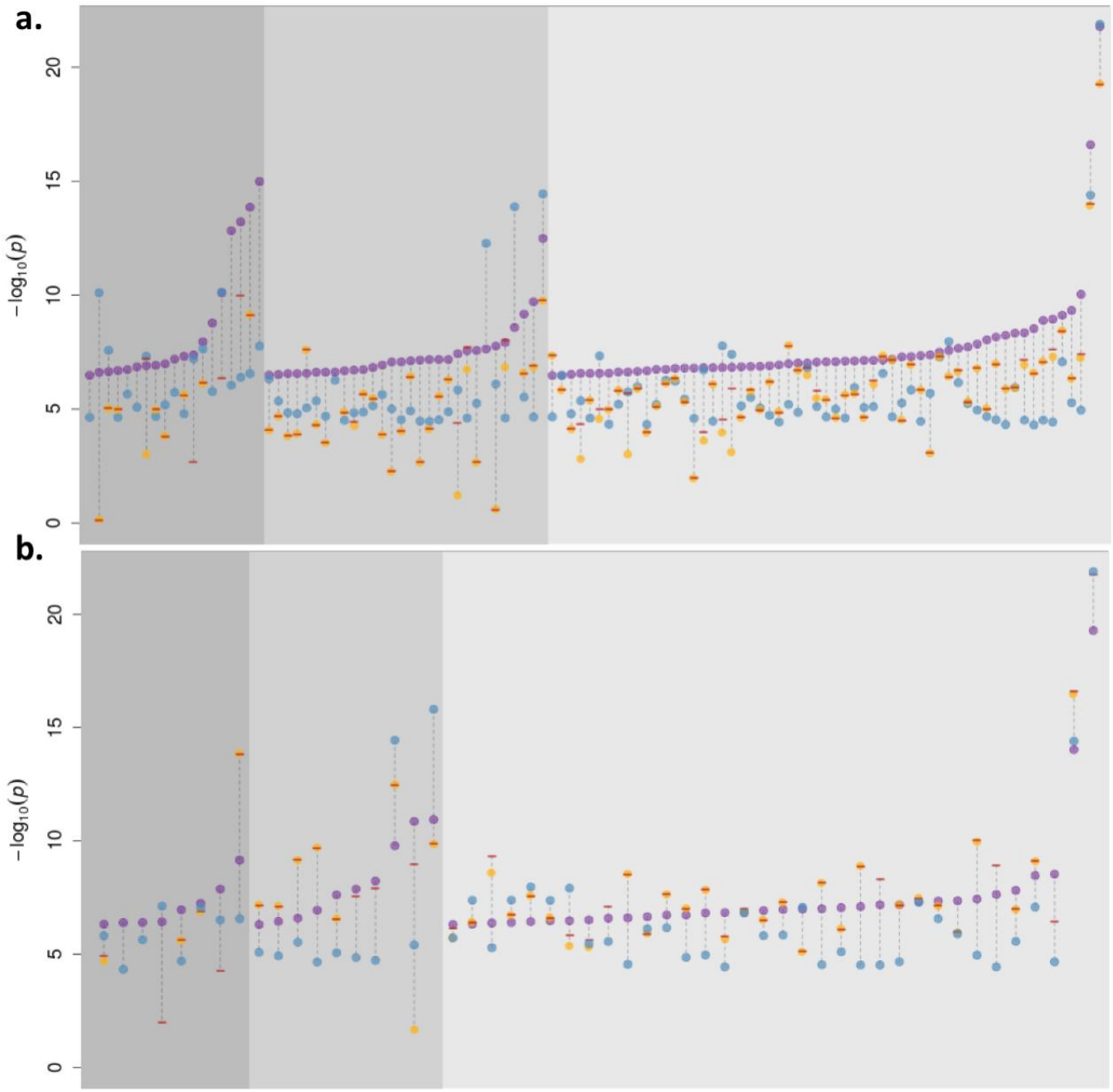



Figure 18 : Association signals in the 1x WGS and imputed GWAS at $p<5x10^{-7}$ for 57 quantitative traits in 1,225 samples. Purple dots represent significant results in the 1x WGS (a.) and imputed GWAS (b.) analysis. Orange dots, if present, denote the p-value of the same SNP in the other study. Blue dots represent the association p-value in a larger (n=1,457) association study based on 22x WGS. Signals with a 22x WGS p-value above $5x10^{-5}$ were considered as false positives in both studies and excluded from the plot. Red dashes indicate the minimum p-value among all tagging SNVs in the other dataset ($r^2>0.8$). Absence of an orange dot and/or a red dash means that the variant was not present and/or no tagging variant could be found for that signal in the other study.

We only considered signals to be true if they displayed evidence for association with at most a two order of magnitude attenuation compared to our suggestive significance threshold ($P<5x10^{-5}$). According to this metric, 52 of 182 independent signals (28.5%) were true in the imputed GWAS, in contrast to 108 of 462 (23.4%) in the 1x study (Figure 18). With an equal sample size and identically transformed traits, 1x therefore allowed to discover twice as many independent GWAS signals with almost identical truth sensitivity. Seven rare and three suggestive low-frequency variant associations in the 1x WGS data (9.2% of all signals) were driven by a variant not present and without a tagging SNP at $r^2>0.8$ in the imputed GWAS, whereas the converse is true for only two rare variants in the imputed GWAS. Four rare, 11 low-frequency and 5 common SNV associations detected in the 1x (19% of total) are not seen associated below that threshold, either directly or through a tagging variant, in the imputed GWAS. As expected, there are significantly fewer (3.8%, P=0.01, one-sided chi-square proportion test) true associations in the imputed GWAS not recapitulated by the 1x study.

#### 4.3.2.Single-point association analysis

We performed single-point association of 14,948,665 and 15,564,905 variants with MAC>2 in MANOLIS and Pomak, respectively, with 63 and 55 quantitative traits available in each cohort. We also performed meta-analysis for 52 overlapping traits using METACARPA(Chapter 3).

We developed Peakplotter(https://github.com/wtsi-team144/peakplotter), an automated peak detection and annotation pipeline, and used it to aggregate signals across all three analyses.

In the meta-analysis, only one signal was present at study-wide significance. The common (MAF=0.24 in MANOLIS and MAF=0.25 in Pomak) intronic *ABCA12* variant rs60395874 was associated with iron levels ($p=8.00x10^{-5}$ in MANOLIS $p=2.56x10^{-4}$ in Pomak and $p=4.79x10^{-10}$ in the meta-analysis with a concordant negative direction of effect). Direct genotyping of this variant using Agena Biosciences MassARRAY technology revealed poor genotyping in the 1x dataset (minor allele concordance 0.97, but positive predictive value 0.36) which resulted in

a strongly attenuated signal in both cohorts using experimentally validated genotypes (p=0.19, p=0.065 and p=0.02 in the three analyses, respectively).

#### *4.3.2.1. Extended regions of association*

In the single-cohort analyses, two signals were driven by multiple variants located several megabase pairs away from each other, but linked by extended LD, which caused Peakplotter to break them up into multiple signals. The first is an 8Mbp large region on chromosome 16 (16:66-74M) in which 14 low-frequency variants (MAF=0.011 to 0.024) in linkage disequilibrium (LD) are associated with high-density lipoprotein (but not any other lipid trait) at study-wide significance in MANOLIS. This region had already been identified in the imputed GWAS study performed by Lorraine Southam[72]. The variant with strongest evidence for association is chr16:73480635 T/A (MAF=0.011, $p=6.45x10^{-14}$), a novel variant not present in 1000 Genomes phase 3 or gnomAD located in an intergenic region. All of the 13 remaining variants have similar p-values, making it difficult to pinpoint the main driver of this signal.

The second is a 5Mbp region on chromosome 11 (11:5.5M-8.5M) containing 26 rare variants associated at study-wide significance with several red blood cell traits in MANOLIS. The most strongly associated variant is rs35004220, a known thalassemia-causing non-coding transcript exon variant close to the *HBB* gene, which reaches $2.71x10^{-21}$ for red cell distribution width in MANOLIS. An attenuated association is also present in Pomak at the same locus with a similarly extended LD structure, which is driven by the intronic low-frequency *MMP26* variant rs76678043 (MAF=0.03, $p=5.3x10^{-10}$).

Such extensive LD structures have been described before at sickle cell loci[91], and may indicate the presence of a structural event, or constitute signatures of recent selection. Disentangling the haplotype structure across several megabases of DNA would require a more in-depth exploration that is outside the scope of this work. Performing haplotyping, assembly, structural variation calling and *other in*-silico methods may shed further light on the nature of these loci in the future. If evidence for large structural variants is confirmed,

long-read sequencing methods, such as the ones provided by PacBio, Oxford Nanopore or 10x Genomics could help validate these events and help pinpoint their exact boundaries.

#### *4.3.2.2.Associations at known loci*

In MANOLIS, we further recapitulate three previously-reported associations. The common intronic *UGT1A10* variant rs4148325 is associated with increased bilirubin levels (MAF=0.28, $p=3.71x10^{-19}$). The common intronic *CETP* variant rs1532624 is associated with increased HDL levels (MAF=0.41, $p=2.04x10^{-13}$). Finally, the cardioprotective *APOC3* R19X variant rs76353203 is associated with triglycerides and HDL, an association previously discovered in this cohort using imputed GWAS data[29]. In Pomak, the association between rs4148325 and bilirubin levels ($p=4.20x10^{-30}$) is the only other association below study-wide significance.

#### *4.3.2.3.Experimental validation and in-silico replication for potentially novel loci*

We further examined suggestively significant ($p<5x10^{-7}$) signals in all three analyses. Out of 213 independent potentially novel signals identified by Peakplotter, we prioritised 64 variants (21 common, 23 low-frequency and 19 rare) for direct genotyping, based on variant level metrics such as genotyping, refinement or imputation scores, minor allele frequency and sequencing quality. Experimental genotyping was carried out using the Agena Biosciences MassARRAY technology in a subset of 1087 and 859 samples in the MANOLIS and Pomak cohorts, respectively. In 42 cases (64%), the signal was not attenuated below suggestive significance when using the direct genotypes to validate the association. On average, minor allele concordance was 76% and positive predictive value was 82%. As expected, these values differ between MAF categories (Table 2), however they were in line with those computed genome-wide between 1x calls and GWAS data.

Table 2 : Average minor allele concordance and positive predictive value for experimentally validated variant genotypes. Rare: MAF<1%, low-frequency: MAF 1-5%, common: MAF>5%.

| | rare | low-frequency | common |
|---|---|---|---|
| minor allele concordance | 69.30% | 82.16% | 95.20% |
| positive predictive value | 84.20% | 89.04% | 97.50% |
| genotype concordance | 96.80% | 98.80% | 99.40% |

We carried forward for replication all validated SNPs associated in either analysis with non-haematological traits for which phenotype measurements were available in external cohorts. First, we sought replication in several population isolate studies with access to genotyping array or sequencing data as part of the Sequencing Isolates (SILC) consortium (ORCADES[92](n~1,600, HRC imputed), METSIM[93] (n~10,000 GWAS data, 1000 Genomes and GO-T2D imputed), SardiNIA[94] (n~6,000, GWAS data, 1000 Genomes imputed), Friul-Venezia Giulia[95] (n~600, 1000 Genomes and UK10K imputed) and Kibbutzim Family Study [96] (n~350, WGS)). The phenome coverage varied across cohorts, hence whenever a given SNV-trait pair was available in more than one cohort, we performed inverse-variance, effect-size based meta-analysis, and considered the meta-analysis p-value as replicating at p=0.05. None of the 64 variants submitted passed our replication criteria, whereas for 34 trait-variant pairs, either the variant was not present altogether or the relevant trait was not measured.

The high SNV density found in 1x WGS makes replication particularly difficult in GWAS studies that use a small reference panel. As a consequence, we queried single-point association results for 12 phenotypes in the INTERVAL 15x single-point association results (TG, LDL, HDL, TC, Height, Weight, BMI, bilirubin, random glucose, iron, ferritin and C-reactive protein), performed using an association pipeline identical to the one used for the 1x data. A single common intronic *CCDC170* variant, rs4870030, which was associated with decreased height in the HELIC 1x meta-analysis (MAF=0.32, $p=1.82 \times 10^{-8}$), weakly replicated in INTERVAL (p=0.03). The discovery and replication meta-analysis did not reach study-wide significance.

Genome-wide association results were also available for lipid traits (TG, LDL, HDL, TC), height, BMI and Waist-Hip ratio adjusted for BMI in 10,484 individuals from the UK-based

Understanding Society (UKHLS) study (https://www.understandingsociety.ac.uk/) and in 2,871 individuals from the Genetic Overlap between Metabolic and Psychiatric traits (GOMAP) study. This genotyping array data, imputed using the HRC reference panel, is being contributed to the GIANT meta-analysis of anthropometric traits by Lorraine Southam. Given the lower density of imputed data, we extended replication to a 1Mbp window either side of every association signal arising from the 1x data, and looked for tagging variants at r2>0.5 in the replication cohorts. Four variants associated with lipid measurements in MANOLIS weakly replicated in these studies (minimum replication p-value 0.017), however replication always occurred either in GOMAP or in UKHLS data, but not in the combined dataset. Furthermore, three of these signals were attenuated above suggestive significance in the larger MANOLIS high-depth WGS association study, further casting doubt on the robustness of these associations.

#### 4.3.3.Rare variant burden in *APOC3*

We identified 57 SNVs in the *APOC3* gene. Two variants were significant at the $1x10^{-8}$ level, the null mutation R19X ($\beta$=-1.09,$\sigma$=0.164, $p=1.01x10^{-10,}$), which is a C/T substitution in exon 2 that changes codon 19 into a premature stop codon, and the splice donor variant rs138326449, located 1 base pair downstream ($\beta$=-1.17,$\sigma$=0.189, $p=1.37x10^{-9}$), which disrupts the donor splice site in intron 2. These two variants are in very low LD ($r^2<0.0001$). We use SKAT[97] to perform rare variant burden testing on the 4 rare or low-frequency (MAF<5%) variants that lie in exons or the essential splice sites in the canonical transcript of *APOC3* (*APOC3-001*) (This burden decreased triglyceride levels ($p=3.0x10^{-18}$) and increased high-density lipoprotein levels (HDL, $p=4.8x10^{-16}$). Association remained strong after removing R19X from the model ($p_{trig}=6.15x10^{-10}$). R19X is not in LD with rs138326449, 11:116701489 or rs187628630, therefore R19X does not drive the association alone. When rs187628630 and 11:116701489 are removed from the model, the significance of the association with triglycerides stays unchanged ($p=4.3x10^{-18}$), but when both R19X and rs138326449 are removed, the association is fully attenuated (p=0.49). We identified a single heterozygous individual for both minor alleles, with relatively low triglyceride levels (0.509 $mmol.L^{-1}$), but not as low as other individuals heterozygous for the minor allele of one or the other loss-of-function variant.

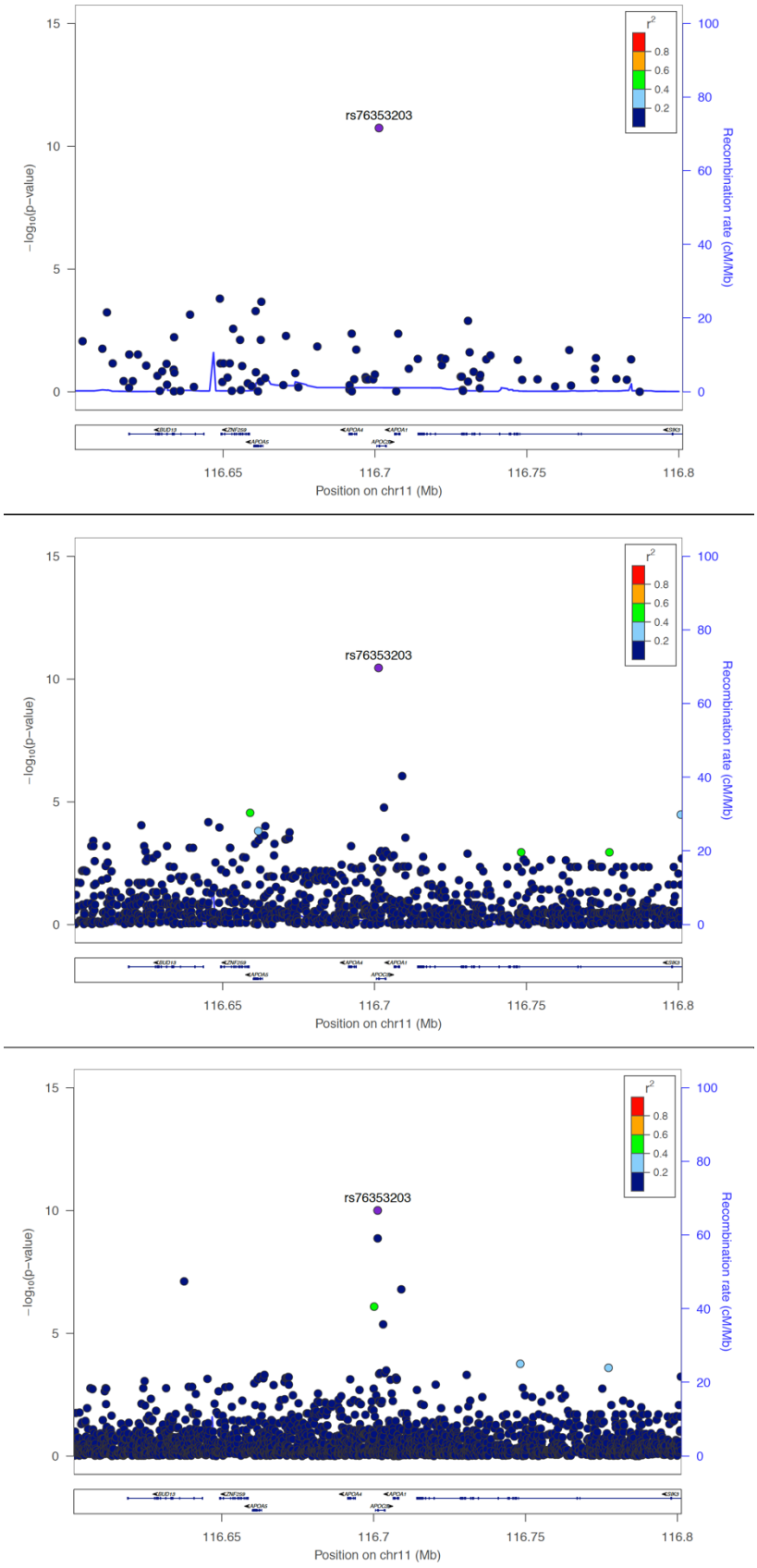
Table 3). These included the two LoF variants R19X and rs138326449. We additionally identified a single carrier of a novel missense variant (11:116701489) also in codon 19, but in exon 3 as the intron falls between the first and second bases of the codon. The resulting amino acid substitution (R19L) is predicted to be deleterious by SIFT [98] and appears to be a novel variant - it is not observed in 1000 Genomes Project phase 3 [99], Exome Sequencing Project [100], or Exome Aggregation Consortium data [101]. Lastly, rs187628630, a rare variant that displays little annotation, resides in the 3' UTR of *APOC3-001* and was thus also included in the burden.

This burden decreased triglyceride levels ($p=3.0x10^{-18}$) and increased high-density lipoprotein levels (HDL, $p=4.8x10^{-16}$). Association remained strong after removing R19X from the model ($p_{trig}=6.15x10^{-10}$). R19X is not in LD with rs138326449, 11:116701489 or rs187628630, therefore R19X does not drive the association alone. When rs187628630 and 11:116701489 are removed from the model, the significance of the association with triglycerides stays unchanged ($p=4.3x10^{-18}$), but when both R19X and rs138326449 are removed, the association is fully attenuated (p=0.49). We identified a single heterozygous individual for both minor alleles, with relatively low triglyceride levels (0.509 $mmol.L^{-1}$), but not as low as other individuals heterozygous for the minor allele of one or the other loss-of-function variant.

Table 3 : Rare variants in APOC3 and blood lipid levels. type: Ensembl VEP[102] annotation for the non-reference allele in transcript APOC3-001/ ENST00000227667, carriers: number of carriers of effect allele in MANOLIS. Mean TG level: mean triglyceride levels in carriers, expressed in $mmol.L^{-1}$. Mean HDL level: mean high-density lipoprotein levels in carriers, expressed in $mmol.L^{-1}$. Numbers in parentheses denote standard deviations. P-values are calculated using SKAT on sex-stratified log-transformed values for TG, and on sex-stratified, inverse-normal transformed, age and age-squared adjusted values for HDL. MANOLIS MAF: minor allele frequency (MAF) in MANOLIS, 1KG P3 EUR MAF: MAF in the EUR (European) population from phase 3 of the 1000 Genomes Project, ESP-EA MAF: MAF in the EA (European-American) population from the Exome Sequencing Project, ExAC MAF: MAF in all samples from the Exome Aggregation Consortium (all external resources were accessed in April 2015). Single-point p-value is the score test p-value calculated using GEMMA on sex stratified and log-transformed triglyceride levels.

| rsID | Position (GRCh37) | type | carriers | Mean TG level | Mean HDL level | MANOLIS MAF | 1KG P3 EUR MAF | ESP-EA MAF | ExAC MAF | Single point p-value |
|---|---|---|---|---|---|---|---|---|---|---|

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| rs76353203 | 116701353 | stop gained | 34 | 0.8471471 | 1.6835 | 1.42% | - | 0.03% | 0.07% | $1.01x10^{-10}$ |
| rs138326449 | 116701354 | splice donor | 28 | 0.8767931 | 1.564759 | 1.17% | 0.30% | 0.18% | 0.14% | $1.37x10^{-9}$ |
| - | 116701489 | missense | 1 | 1.424 | 0.932 | 0.04% | - | - | - | 0.944 |
| rs187628630 | 116703739 | 3' UTR | 4 | 1.2226 | 1.4972 | 0.16% | 0.40% | - | - | 0.631 |
| Total *APOC3* carriers | | | 67 | 0.9008529 (±0.3957894) | 1.598206 (±0.3627785) | | | | | |
| Total *APOC3* non-carriers | | | 1125 | 1.656739 (±1.205513) | 1.261232 (±0.3448805) | | | | | |
| Carriers v non-carriers (%) | | | | -45.6 | +26.7 | | | | | |
| P value | | | | $3.0x10^{-18}$ | $4.8x10^{-16}$ | | | | | |




Both loss-of-function variants, but none of the other variants identified in HELIC, are included in the rare variant burden associations with triglyceride levels discovered in *APOC3* by two recent large-scale exome sequencing studies. The first, by Crosby et al. from the TG and HDL Working Group of the Exome Sequencing Project[103], includes two other variants, a missense variant, A43T, in exon 3 (position 116701560), and another splice variant at the donor splice site of intron 3 (position 116701613). The second, by Li et al.[104], includes rs140621530, a rare splice donor variant, and the novel singleton frameshift indel 11:116703578. These 4 variants are all absent from the HELIC-MANOLIS cohort. These differences

Figure 19: Regional association plots for the region flanking R19X (rs76353203) in APOC3. Top panel: Merged OmniExpress and ExomeChip data. Middle panel: OmniExpress and ExomeChip data imputed up to a merged reference panel. Bottom panel: WGS data (n=1178 overlapping samples).

demonstrate the expected allelic heterogeneity underlying rare variant burden signals that traverse populations and highlight the importance of seeking replication at the locus rather than at the constituent variant level.

A combined meta-analysis of the *APOC3* burden signals identified across the exome sequencing study by Crosby et al [103], the MANOLIS WGS finding described here and the exome sequencing study by Li et al [104] using Stouffer's method yields strong evidence for association with triglyceride levels ($p=3.23x10^{-31}$, n=13,480).

Regional association plots of the 100kb region flanking *APOC3* show a clear gain in variant resolution when comparing WGS data with genotype array data (), whereas imputed GWAS data show similar variant density to WGS. Despite being imputed up to a large reference panel including WGS from the same founder population, the only signal above genome-wide significance in the imputed GWAS dataset is R19X, which is associated with triglycerides to a similar level of magnitude ($p=3.48x10^{-11}$) as in the low-depth WGS data ($p=1.01x10^{-10}$). Three of the four rare variants included in the sequence-based burden test are present in the imputed data, with the exception of 11:116701489. R19X is directly typed on the genotyping array, and imputation quality scores for rs138326449 and rs187628630 are 0.49 and 0.70, respectively. The lipid-associated burden of these three variants ($p=6x10^{-13}$) is fully attenuated when R19X is removed (p=0.11, nine orders of magnitude higher compared to the low-depth sequence data).

## 4.4.Discussion

The advantages of 1x sequencing come at a certain compute and financial cost. As of summer 2018, 1x WGS on the HiSeq 4000 platform was approximately half of the cost of a dense GWAS array (e.g. Illumina Infinium Omni 2.5Exome-8 array), the same cost as a sparser chip such as the Illumina HumanCoreExome array, and half of the cost of WES at 50x depth. By comparison, 30x WGS was 23 or 15 times more costly depending on the

sequencing platform (Illumina HiSeq 4000 or HiSeqX, respectively). The number of variants called by 1x WGS is lower than high-depth WGS, but is in the same order of magnitude, suggesting comparable disk storage requirements for variant calls. However, storage of the reads required an average 650Mb per sample for CRAMs, and 1.3Gb per sample for BAMs.

A computationally costly step is added with genome-wide refinement and imputation, since the number of input markers is close to 50 times higher than for a GWAS chip. The complexity of the imputation and phasing algorithms used in this study is linear in the number of markers and in the number of target samples, and quadratic in the number of reference samples [105], which results in a 50-fold increase in total processing time compared to an imputed GWAS study for equal sample sizes. In MANOLIS the genome was divided in 13,276 chunks containing equal number of SNVs, which took an average of 31 hours each to refine and impute. The total processing time was 47 core-years (Figure 20). Parallelisation allowed to process the 1,239 MANOLIS samples in just under a month, although the effective running time will depend on cluster size and computational load. As imputation software continue to grow more efficient [106], future pipelines should greatly simplify postprocessing of very low depth sequencing data.

In our study of 57 quantitative traits, we show that an 1x-based design allows the discovery of twice as many of the signals suggestively associated in the more accurate 22x WGS study, compared to the imputed GWAS design. Almost 10% of the suggestive signals arising in the 1x data are not discoverable in the imputed GWAS, but the great majority (96%) of imputed GWAS signals is found using the 1x.

The 1x-based study seems to discover more signals than the imputed GWAS across the MAF spectrum, and this remains true whether or not the signals are filtered for suggestive association p-value in the more accurate 22x based study. At first glance this suggests 1x WGS has better detection power than the imputed GWAS across the MAF spectrum, however it is unlikely that this is true for common variants, which are reliably imputed using chip data. Instead, this phenomenon is likely due to a slightly less accurate imputation than

in the GWAS dataset caused by a noisier raw genotype input.This effect is marginal, as evidenced by genome-wide concordance measures which are very high at the common end of the MAF spectrum. However, it is important to note that this slightly less accurate imputation can attenuate some signals as well as boost others. For this reason, we would recommend relaxing the discovery significance threshold in 1x studies in order to capture those less well imputed, signal-harbouring variants, followed by rigorous replication in larger cohorts and direct validation of genotypes.

We focused on the performance of commonly used general-purpose tools for low-depth sequencing data in isolates, both for genotype calling (GATK) and imputation (BEAGLE, IMPUTE). There are ongoing efforts to leverage the specificities of both low-depth sequencing [107,108](https://www.gencove.com) and of isolated populations [109]. The popularity and long-term support of established generic methods is an advantage when running complex study designs, as has been shown in other isolate studies [110].

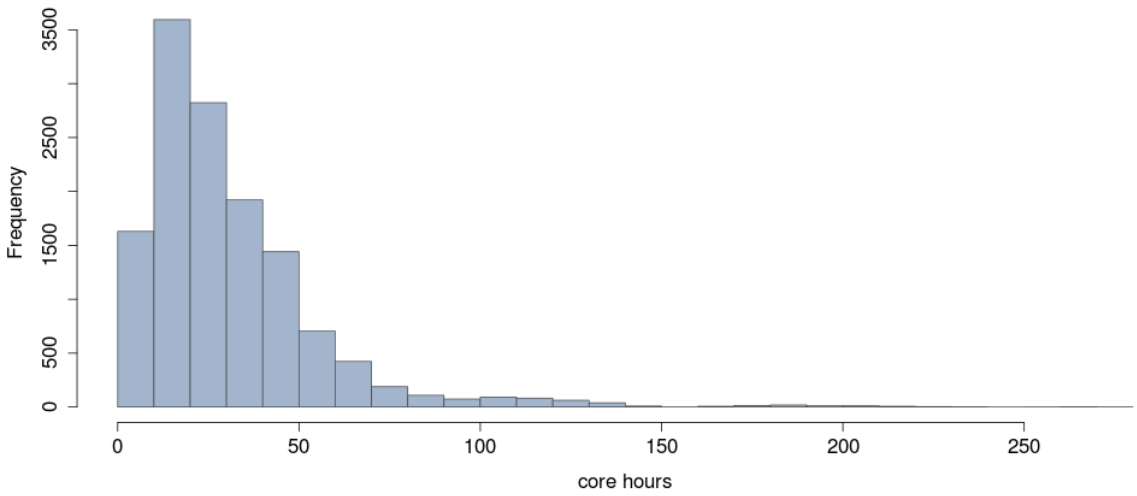



Figure 20 : Genotype refinement and imputation compute time. Measured in core hours, on chunks of 2,000 SNPs for the MANOLIS cohort. Genome-wide data. Some chunks were located in hard-to-impute regions (e.g. HLA) took up to 271 hours (11 days) to complete.

## 4.5.Conclusion

Very low-depth whole-genome sequencing allows the accurate assessment of most common and low-frequency variants captured by imputed GWAS designs. It also provides a denser coverage of the low-frequency and rare end of the allelic spectrum compared to an imputed GWAS dataset, albeit at an increased computational cost and with a much higher error rate. This allows very-low depth sequencing studies to recapitulate most findings from imputed chip-based efforts, and to discover significantly associated variants missed by GWAS imputation [111]. 1x WGS allowed the discovery of a burden of low-frequency and rare variants in the *APOC3* gene, associated with blood triglycerides. The burden is missed by genome-wide imputation of GWAS data, and adds to a body of evidence linking apolipoprotein C-III with the regulation of triglycerides.

# Chapter 5. Quality control and single-point association of cohort-wide 15x whole-genome sequencing data

## 5.1. Introduction

Like all genotyping methods, whole-genome sequencing is prone to measurement and human error. Quality control (QC) is an important first step to ensure that data of a high enough quality is carried over to the analysis stage. As with genotyping array data, quality control comprises two main steps: variant-level QC and sample-level QC. The former is used to exclude bad quality variants which might lead to both false positives and superfluous tests in downstream association analyses (which, in turn, has an effect on the p-value threshold). The latter is to ensure that the samples tested are indeed the samples that were sent for sequencing, without duplicates or sample swaps, and that the samples have good overall quality, with minimal traces of contamination or noise. Consequences of imperfect sample QC are varied, from inflation of the test statistics to loss of power and the presence of spurious associations.

As opposed to array-based genotyping, where sample QC can be performed before variant QC, it is advised to first perform variant QC when dealing with WGS data. This is because the typical WGS-based variant calling pipeline creates a large number of false positive variants due to various biases both at the sequencing, alignment and calling steps, which can greatly perturb sample-level metrics.

Here, we describe the checks performed as part of the quality control pipeline for the MANOLIS and Pomak high-depth WGS datasets. We then briefly present an overview of variant quality and composition, and then perform single- and cross-cohort single point analysis in the two cohorts.

## 5.2.Methods

### 5.2.1.Variant-level QC

#### *5.2.1.1.Variant Quality Score Recalibration*

We first run Variant Quality Score Recalibrator, a short description of which is given in chapter 4. Running VQSR on thousands of samples of high-depth WGS corresponds to an ideal use case, therefore QC was much easier to run for high-depth than for the previous low-depth dataset. In order to run the software, we first clean our VCF files by removing star alleles and separating multi-allelic variants into separate records. Star alleles are a new type of variant called from GATK 3.5 onwards, which represent the absence of a SNP when it overlaps a deletion. Thus, we replace this type of record:

```
chr1  1234  G     A,T,*
```

into two separate records:

```
chr1  1234  G     A
chr1  1234  G     T
```

This allows each alternate allele to be given a separate score if allele-specific annotations are present. We also split indels from SNPs, as VQSR uses different metrics for each type of variant.

We then run VQSR in the usual two-step process:

- A model-fitting step using VariantRecalibrator, which produces the Gaussian Mixture Model (this step is run genome-wide);
- An annotation and filtering step which adds the VQSLOD metric to the VCF and allows to filter variants above a certain tranche (this step is run chromosome-wide).

This first step allows to produce a tranche plot (Figure 21), where we concurrently plot the proportion of "true" dbSNP variants intercepted in our callset (red), as well as the estimated proportion of true positives among calls (blue). There is a clear point of inflexion in the latter

curve around 99%. We choose a tranche threshold of 99.4%, which corresponds to an estimated false positive rate of 5% and a true positive rate of 95%.

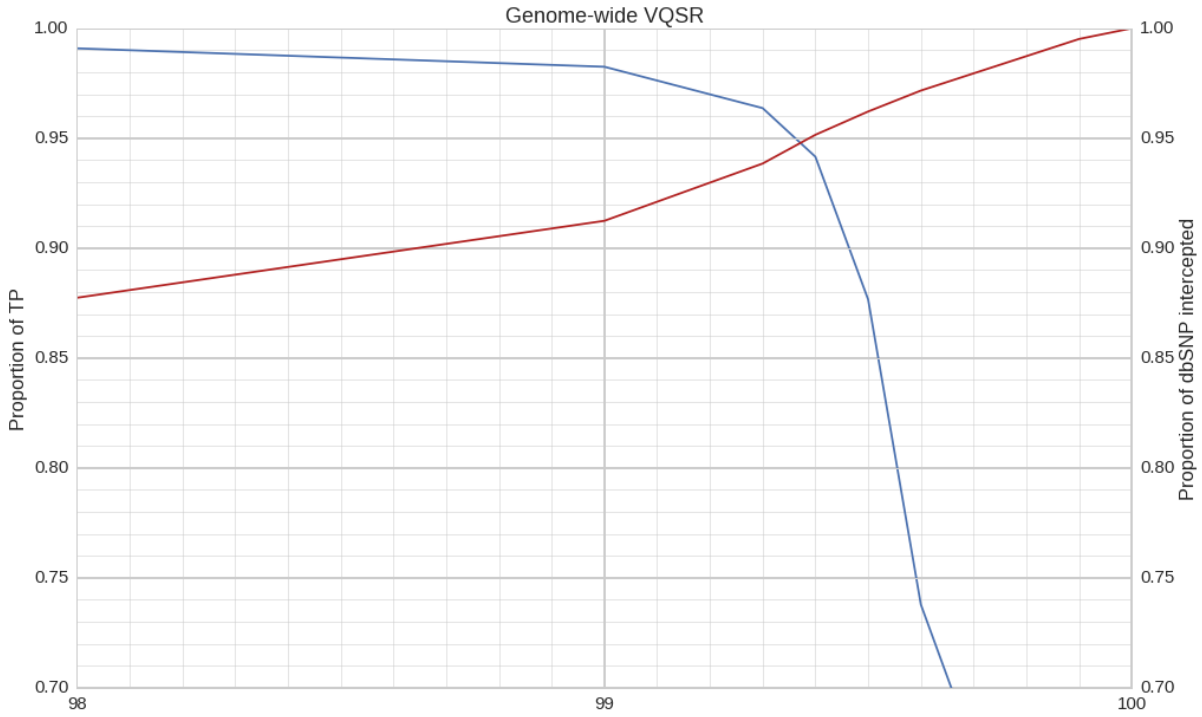



Figure 21: Tranche plot from genome-wide VQSR analysis of 1,482 MANOLIS samples sequenced at a target depth of 15x.

The procedure is repeated for INDELs, with the difference that since there is a lack of a robust training set of multi-nucleotide polymorphisms, there are no diagnostic plots or variant tranches, and a fixed threshold of 99% is used as per the tool's recommendations. After filtering, multiallelics are collapsed using bcftools and SNPs and INDELs are merged.

#### *5.2.1.2.Missingness*

Variant callers will call a genotype as missing if the read evidence is not sufficient to support any confident genotype call, which is evidenced when comparing the distribution of sequencing depths when missingness is present and when it is not (

Figure 22). Further missingness is added in the previous steps when the star alleles from GATK are removed from the VCF files.

For most purposes, including single-point and burden tests, a small amount of missingness in genotype calls is acceptable. For example, MONSTER, the variant aggregation software used for running multi-variant tests in Chapter 6, is able to impute missing genotypes in an ad-hoc manner, albeit by using a relatively crude algorithm. It is common practice to remove variants with missingness above an empirically determined threshold. The alternative (which is computationally impractical with WGS data, as demonstrated by chapter 4) is to impute genotypes using a haplotype reference panel. For the high-depth WGS data in MANOLIS, we chose the former option.

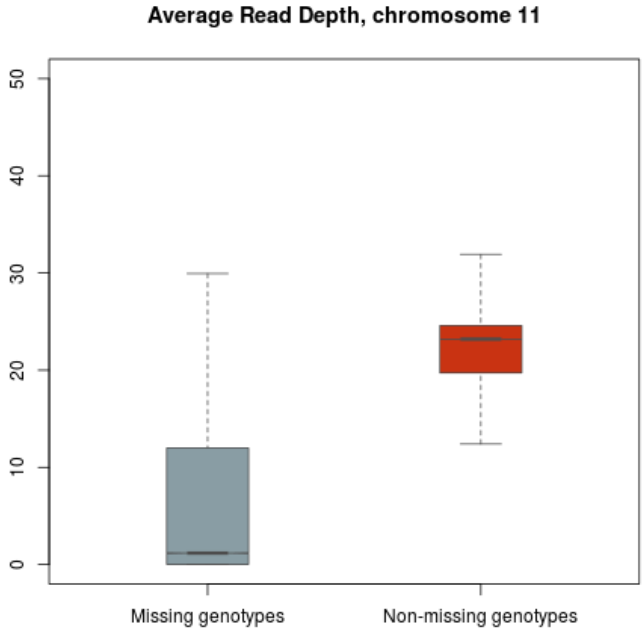



Figure 22 : Depth distribution for missing versus non-missing genotypes on chromosome 11 in the MANOLIS data.

In the MANOLIS data, 20% of variants exhibit some kind of missingness, however the distribution of missingness has a very long tail and does not saturate quickly (Figure 23), which makes filtering difficult, as we have no point of saturation at which to set the

threshold. We choose a threshold of 1% (15 missing genotypes), which will remove around 69.5% of all variants with missingness and 14% of all variants in total.

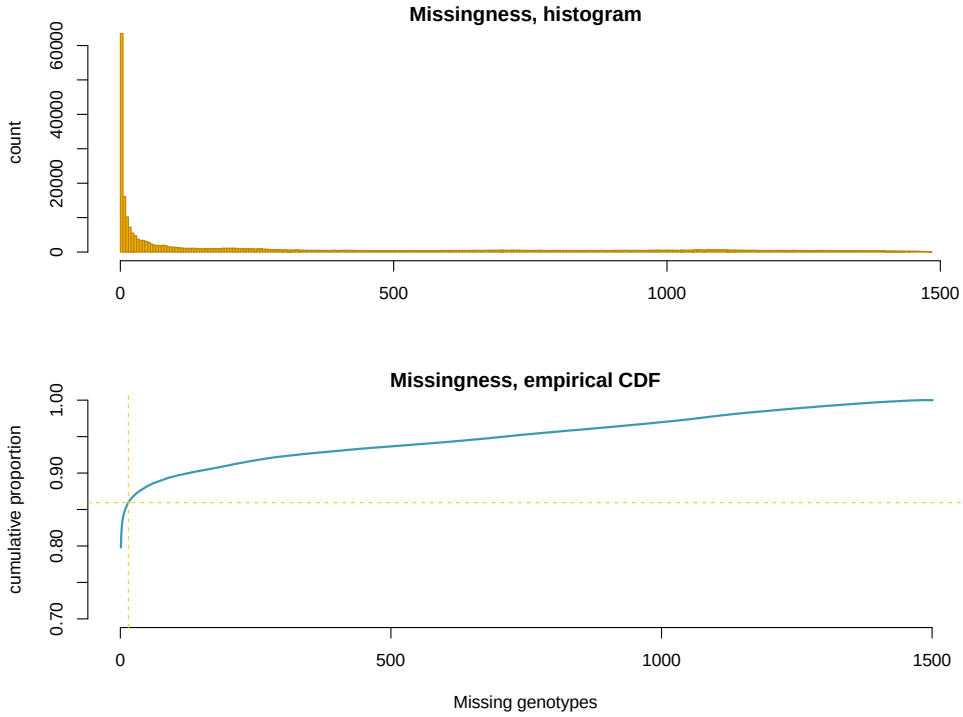



Figure 23 : Histogram of missing genotypes (top) and empirical cumulative distribution function (bottom) for missing genotype count on chromosome 11, MANOLIS data. 1% missingness threshold is showed by the dashed yellow cross.

We apply this filter using bcftools view.

#### *5.2.1.3.Low-complexity regions*

Regions of the reference genome are said to be hard to access when the region could not be assembled, when too many or too few reads tend to map to them, or if read mappings located in them have a consistently low quality. The 1000 Genomes Project has provided genome accessibility masks for the reference genome, by aligning whole-genome sequencing reads to it and assessing mapping coverage and quality.

We wanted to know whether missingness and read depth were dependent on region accessibility. At the time of initial QC, the genome accessibility mask was not available for

the reference build used in the HELIC high-depth data (GRCh38). We therefore used a community-generated GRCh38 version of the Low-Complexity Region map (https://github.com/lh3/varcmp/tree/master/scripts), which only intercepts a subset of the inaccessible genome (low-complexity regions are repetitive, and therefore hard to assemble and map). We find no significant difference in missingness between the two types of regions on chromosome 11 (Figure 24).

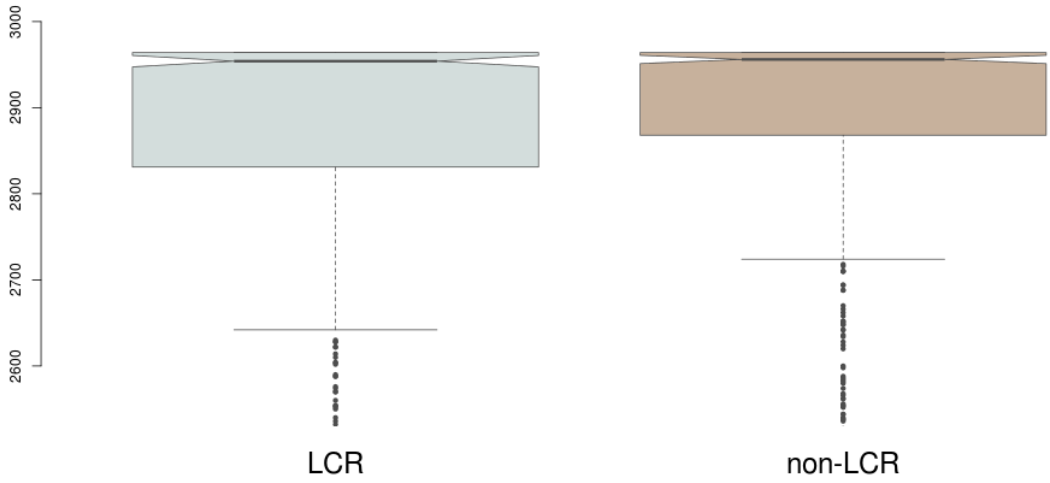



Figure 24: Non-missing genotypes in LCR and non-LCR regions, chromosome 11, MANOLIS.

In the meantime, in July 2017 a new alignment of 1000 genomes sequences to the most recent GRCh38 build became available[112], in which the authors reproduced the methodology previously applied to the GRCh37 mask. We used the conservative definition of accessible regions, which labels more regions as hard to access. Inaccessible regions spanned 30.2Mbp on chromosome 11, as opposed to a mappable 107.2Mbp. Although there were no marked differences between the two types of regions in terms of SNP density (0.008 variants per base in inaccessible as opposed to 0.007 in accessible regions), the INDEL and multiallelic concentration was 3.5 times higher in the inaccessible genome (0.0028 vs 0.0007). This likely reflects the statistical pitfalls that remain when calling more complex

variants such as indels, but may also indicate that SNP calls are relatively reliable at this depth and sample size, even in difficult genomic regions.

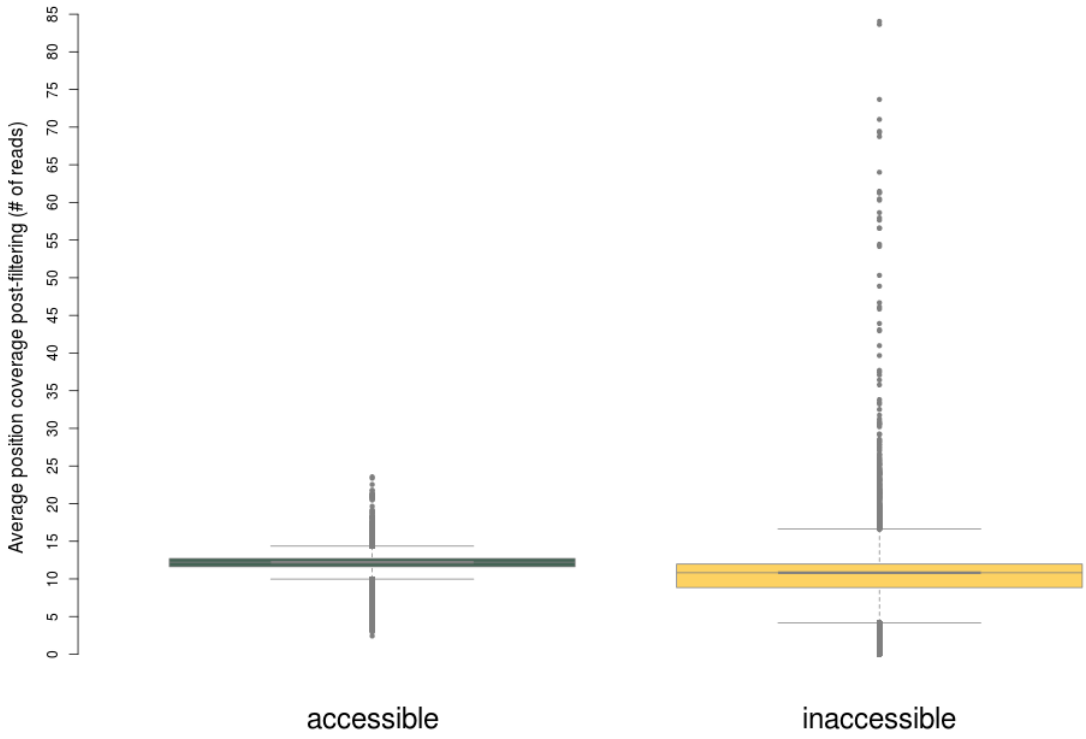



Figure 25: Average read coverage per position, chromosome 11, accessible vs. inaccessible genome, in the MANOLIS data.

Other metrics strongly differentiate accessible and inaccessible regions. The read coverage distribution is both narrower and has a higher mean in the accessible compared to the inaccessible genome (one-sided t-test p-value<$1x10^{-20000}$) (Figure 25). This translates into a much higher number of positions having missing genotypes (Figure 26) in the inaccessible genome, and a higher proportion of variants with extreme missingness (100 missing genotypes or above, which translates approximately to 5% of genotypes).

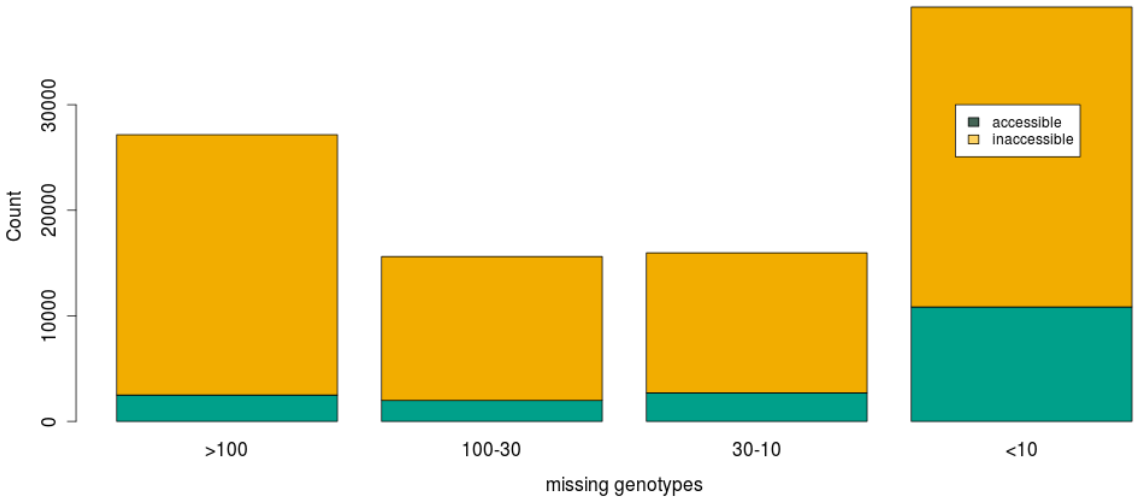



Figure 26: Missing genotypes on chromosome 11 in the MANOLIS data, inaccessible vs. accessible genome.

#### *5.2.1.4.Analysis-specific variant filtering*

Single-point association testing only has power to detect signals above a certain variant frequency, that will be dependent on study sample size. For the study of single-point association signals in the high-depth WGS data, we filter out variants with a MAC below 8, missingness greater than 1% or a Hardy-Weinberg equilibrium (with mid-p adjustment) p-value below $1x10^{-5}$.

For rare variant burden testing (Chapter 6), we apply the same filters except for MAC threshold, which is replaced by an upper MAF threshold of 5% to only include rare and low-frequency variants.

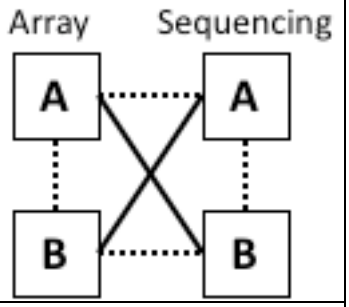
For computation of the genetic relatedness matrix, we use the same filters as for single-point association testing after applying LD-based pruning using PLINK's default parameters and applying a more stringent MAF threshold of 5%. A detailed discussion on the effects of different filtering thresholds on kinship coefficients and inflation is available in Chapter 4 and Chapter 6.

### 5.2.2.Sample-level QC

The objective of sample-level QC is twofold: eliminate samples with low-quality genotypes across the genome, and remove or correct sample mix-ups. The former can occur, for example, due to sample contamination at the sequencing stage (material from another sample is mixed to a particular sample's DNA) or improper storage which degrades the sample, whereas the latter occurs due to manipulation errors, such as mislabelling of the phenotype files, DNA samples or genotype file metadata. Sample mix-ups can involve material from samples from the same study or from a different one.

Various metrics are available to the researcher in order to disentangle the different scenarios that can occur. In HELIC, sample QC is facilitated by the many genotypic datasets that are available and previously QC-ed, which allow to compare concordance (genotype correlation) between successive genotyping assays of the same person.

For a sample of interest A and a second sample B, the signatures of the main type of sample QC failures are detailed below.

| **Error type** | **Signature** |
|---|---|
| Sample swap | + Clear sex check fail (if swapped samples have different genders)<br>+ High inter-sample correlation measures: $A_{seq}$ correlates with $B_{array}$ and conversely.<br>  |
| Foreign sample swap | + same as above, but sample B is not genotyped as part of the study<br>+ if B is from a different source population (often the case with isolates), A will be an outlier on a PCA plot |

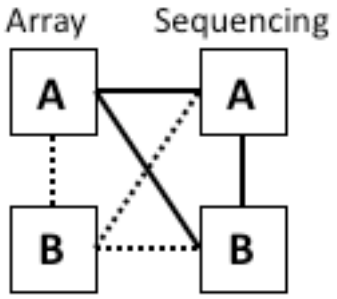
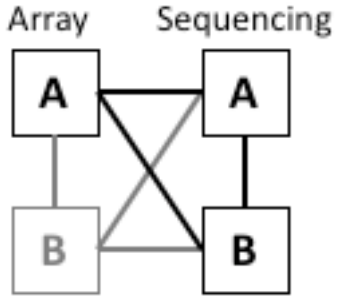
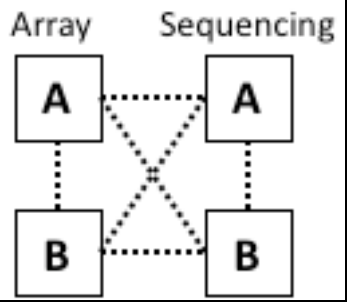
| | | |
|---|---|---|
| Genetic sample duplicate | + Duplicate correlation signature: $A_{seq}$ correlates with both $A_{array}$ and $B_{seq}$, $B_{seq}$ and $B_{array}$ have low concordance. |  |
| Phenotypic sample duplicate | + same as above, but samples A and B are actually the same sample sequenced twice. This is unlikely when both A and B have been genotyped previously, as QC would have picked it up. |  |
| Sample contamination<br>or<br>Low sample quality | + the main sign of contamination or sample degradation is lower self-correlation between $A_{seq}$ and $A_{array}$<br>+ in case of cross-contamination, this will be supplemented by higher correlation with both $B_{seq}$ and $B_{array}$ beyond what can be expected through the kinship between A and B |  |

Figure 27 : Example of sample QC failures with their consequences on cross-sample genotype concordance.

#### *5.2.2.1.Concordance and contamination checks (Chipmix, Freemix, het-rate, NRD, $\hat{\pi}$)*

The HELIC cohorts are well positioned in that they have been assayed using multiple genotyping methods. In particular, high-quality QCed genotyping chip data is available for a large subset of sequenced samples. A variety of methods exist for comparing two sets of genotypes, some specifically designed to analyse data from the same sample (CHIPMIX), others more generic and able to simply compute an aggregate measure of similarity between two genome-wide genotyped samples.

Perhaps the most useful measure of quality for sequenced samples with chip array data available is $\hat{\pi}$, a per-sample measure of genome-wide genotype concordance. It is calculated by PLINK, and provides an easy to interpret measure of genotype correlation

between two samples. Therefore, it can be used both to quantify sequencing quality by assessing $\hat{\boldsymbol{\pi}}$ between the sequencing and chip-based genotypes for each samples, and to identify potential sample mixups or contamination (if a high $\hat{\boldsymbol{\pi}}$ is observed between the sequencing genotypes and chip data of one or more other sample). In the HELIC cohorts where isolatedness is higher, this latter examination is made somewhat harder by the existence of close family relationships between members of the cohort. It is defined as:

$$\hat{\boldsymbol{\pi}} = P(IBD = 2) + 2.P(IBD = 1)$$

where both probabilities are estimated using empirical identity-by-state probabilities genome-wide (compounded by a factor that accounts for allele frequencies at all shared loci)[89].

A related measure is non-reference allele discordance (NRD) which is calculated by bcftools as follows:

$$NRD = \mathbb{E}\left[\frac{FP + FN}{FP + TP + FN}\right]$$

where FP, FN and TP are the number of false positive, false negative and true positive alternate allele calls at shared loci and $\mathbb{E}$ denotes the expected value over all shared loci. This method has the advantage of not being too optimistic, as would be the case if we took into account concordance at homozygous reference calls[113].

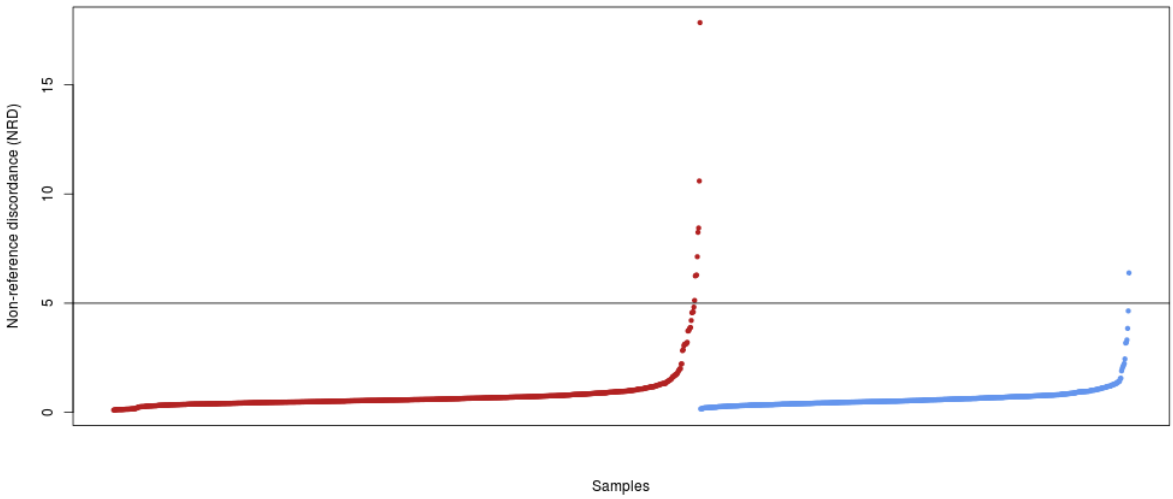



Figure 28 : non-reference allele discordance in the HELIC MANOLIS sequencing data, for samples typed using the OmniExpress (red) and CoreExome(blue) chips. The horizontal line represents 5% NRD, which if used as a threshold, would lead us to exclude 9 samples.

CHIPMIX is another tool, part of the verifyBAMid suite[114], that uses read information to assess the likelihood that the observed reads at a given locus give rise to a known genotype from another source, such as array data. Since it works on read alignments, it can be used earlier on in the pipeline, before variant calling happens, however that also makes it extremely slow compared to genotype-based methods.

Freemix is a metric implemented in the same package, that performs estimations of contamination based on sequencing reads in the absence of genotype calls, as well as an externally supplied population allele frequency, often taken from a VCF's AF INFO filter. It uses excess heterozygosity to detect potential sample contamination, and is therefore very similar to heterozygosity rate, another commonly used sample-level check. We indeed observe a strong correlation between these two measures, especially in the range (high Freemix, high heterozygosity rate) that would be indicative of contamination (Figure 29).

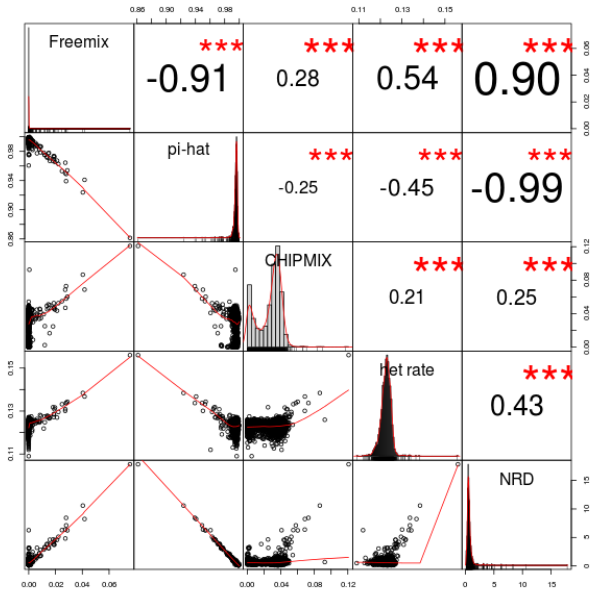



Figure 30 : Correlogram of contamination metrics in the Pomak WGS dataset.

In general, we observed a strong correlation among outliers in all measures of contamination, which would tend to suggest that for future studies, if time or compute resources are a significant bottleneck, not all the measures we calculated on the two

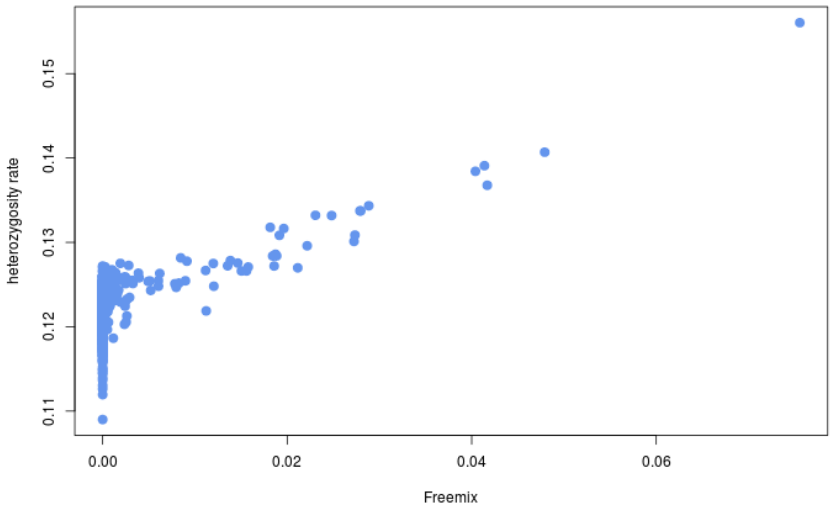



Figure 29 : Freemix and heterozygosity rate in the Pomak cohort, based on WGS data.

isolated cohorts are required. The strongest correlation was observed between Freemix, pi-hat and NRD (Figure 30). Of these three measures, pi-hat also provides cross-sample similarity, and should therefore be preferentially used as it also allows to identify sample swaps.

#### *5.2.2.2.Sex check*

Checking sample sex is one of the main quality control measures that help flag sample abnormalities. The idea is to compare phenotypic gender information with the sex inferred from genotypic information on the sex chromosomes. Within the HELIC project there are several sources of phenotypical gender information: the master phenotype file, the sequencing manifest (which is filled manually when blood samples are processed) and the raw field team data, which we request in case the former two disagree or if we suspect potential sex issues. The simplest way to impute sex assignments from genotype data is to use the fact that for males, no heterozygosity should be observed in the haploid regions of chromosome X. Since the haploid flag was not set for non-autosomal regions when calling variants with GATK, some heterozygosity will be present, however it should be notably inferior to that of females.

We use PLINK's sex check option, which is based on the F-statistic computed on non-pseudoautosomal regions of chromosome X:

$$\hat{F} = \frac{\#hom_{observed} - \#hom_{expected}}{M - \#hom_{expected}}$$

where $\#hom_{expected}$ is estimated across all available biallelic SNPs by:

$$\widehat{\#hom_{expected}} = \sum_{SNP\ j} 1 - 2f_j.(1 - f_j)\frac{2n}{2n-1}$$

where $M$ is the number of biallelic SNPs, $n$ is the number of samples and $f_j$ is the minor allele frequency of SNP $j$. For males in non-pseudoautosomal regions, we expect $\#hom_{observed}$ to be close to $M$, which results in a value very close to 1, whereas in females the F statistic is a regular inbreeding coefficient, and varies depending on population characteristics. F-values

lower than 0.2 are usually assigned to females, whereas values greater than 0.8 are assigned to males.

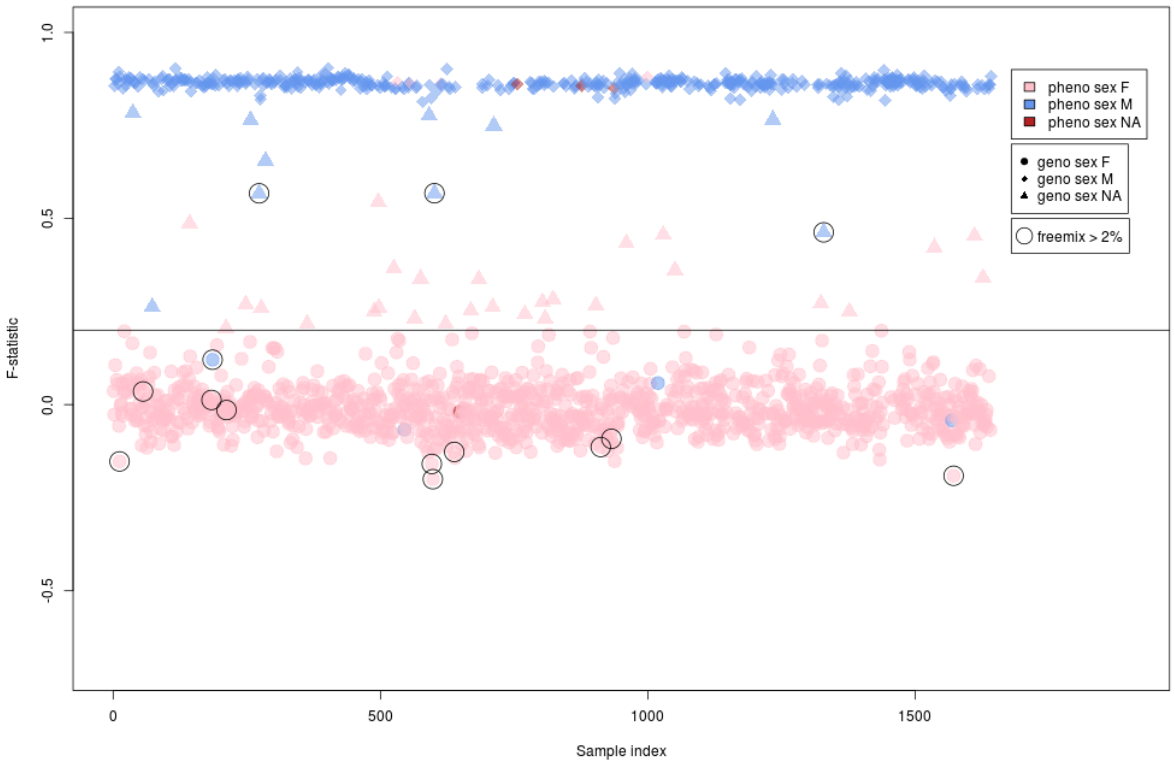



Figure 31 : F-statistic and phenotypic sex for the Pomak WGS sequencing data. The horizontal line denotes the 0.2 line usually used to separate female calls. Samples with potentially high contamination are circled for comparison.

In isolated populations, we typically observe a higher upper limit for the F-statistic in females (Figure 31), which is likely due to an average higher genome-wide F-statistic than in cosmopolitan populations. This leads us to ignore female sex assignments that are flagged as suspicious (F>0.2) if they do not stand out for any other QC metric. Similarly, we keep the 0.8 threshold for males, but only excluded samples if another metric also suggested a sample anomaly. For example, in Figure 31, we identify 13 suspicious individuals with male phenotypic sex (blue circles or triangles). Three are marked as having contamination issues (Freemix >2%) and are removed. 6 have a low F, but are not otherwise remarkable, and are kept as males. The remaining have conflicting phenotypic and genotypic sex information,

with 4 of them having F<0.2 and one failing the 0.2 threshold but still belonging to the expected female range. These males are flagged for sample swaps or sex reassignment.

In some cases, the samples had missing phenotypic sex information (red glyphs on the graph). If the sample was not flagged by any other metric, we assigned the sex imputed by the F statistic. We did not encounter the case where a missing phenotypic sex could not be imputed (i.e. had an F-statistic between 0.2 and 0.8).

### Chromosome X relative depth

Another way of imputing sex, which is specific to whole-genome sequencing data, is to look at relative read depth on chromosome X. For this, we compute autosomal average read depth using bcftools stats, then separately compute average read depth on either the non-pseudoautosomal or the entire X chromosome (due to the reduced sizes of these regions, the effect of inclusion/exclusion is minimal on chromosome-wide depth measurements).

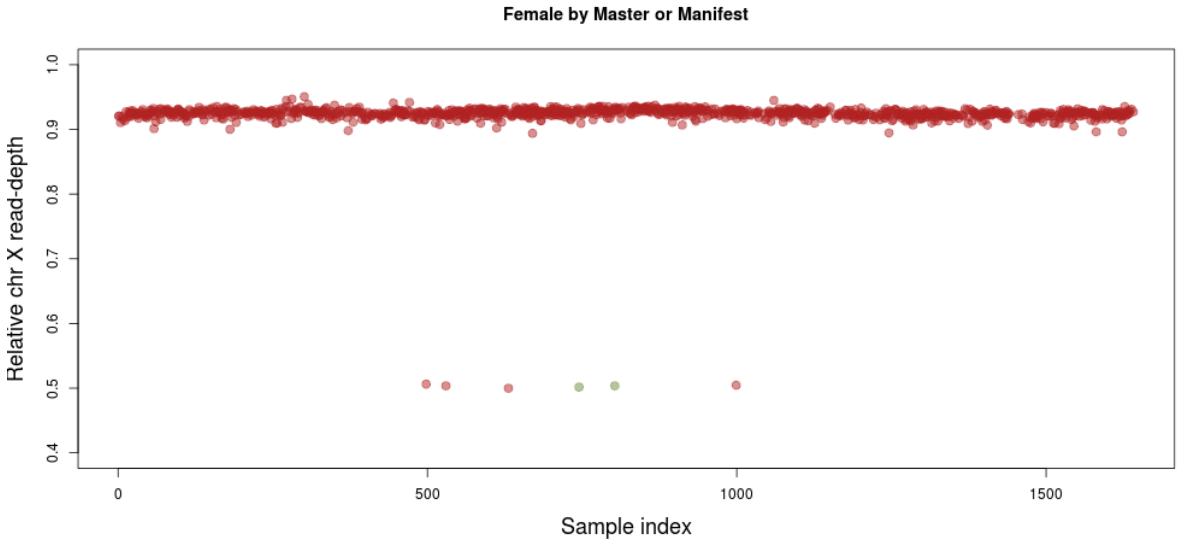



Figure 32 : chromosome X relative depth for all samples annotated as female by the sample manifest or phenotype master files (points are grey where the two sources disagree). Six samples exhibit a depth of 0.5, indicative of a sex mislabelling or sample swap.

Being based on read depth only, this method has the advantage of being independent of genotype data itself. It should not therefore be compounded by other effects, such as heterozygosity, contamination, and generally on the quality of the sample. This results in

extremely well-differentiated clusters for sex separation (Figure 32), which makes this method well suited for sex checking whenever sequencing data is available, as it is independent of population factors. In contrast, the F statistic exhibits a negative correlation with contamination measures such as Freemix for both males and females (Figure 33).

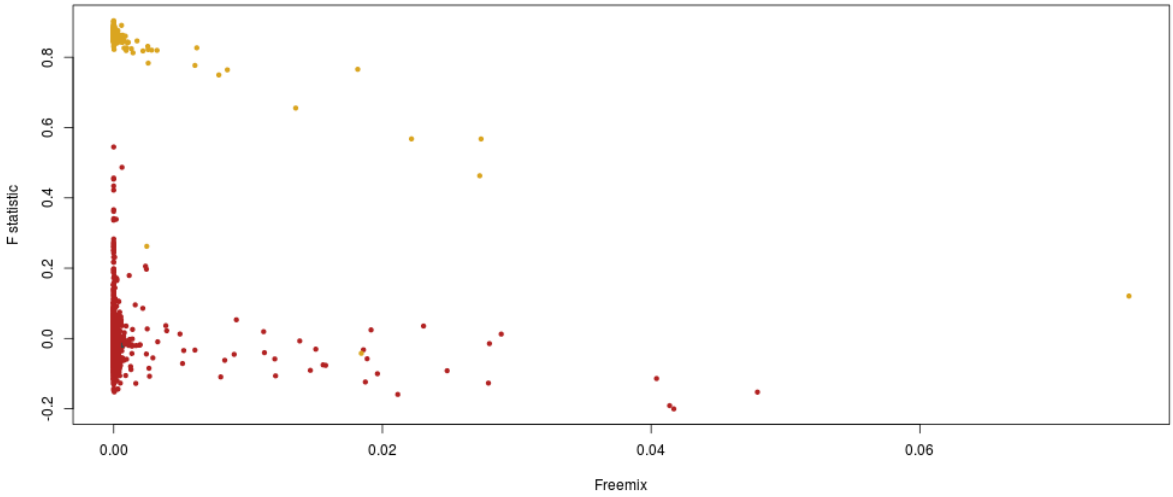



Figure 33 : Freemix is negatively correlated with the F statistic on chromosome X independently of sex, as evidenced by the two negatively-sloped groups of points.

#### *5.2.2.3.Missingness and depth by sample*

Read depth is an important determinant of both variant quality and number, specifically at the rare end of the frequency spectrum: a lower depth outlier will result in less sensitivity and specificity, whereas the reverse is true for a sample with abnormally high depth. At very low depths, average per-sample missingness is also expected to rise as the caller will struggle to deal with the uncertainty caused by lack of reads.

Computing the average depth and missingness genome-wide is a computationally intense task when undertaken by a single process. Furthermore, it cannot readily be applied to single chromosomes, which would be a natural way to reduce the number of markers, because some anomalies are large enough to perturb averages chromosome-wide (Figure 34).

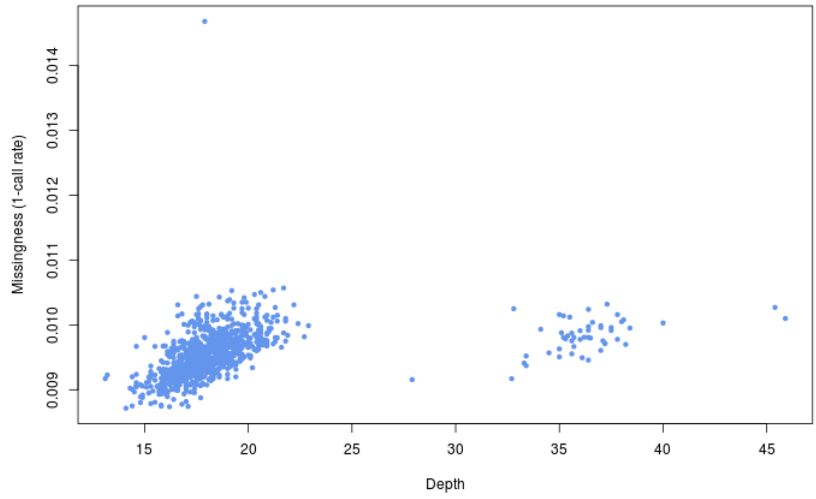



Figure 34 : Depth versus missingness in the Pomak WGS sample on chromosome 11. The upper missingness outlier is the sole carrier of a homozygous deletion spanning several hundreds of kilobases (See Chapter 7).

We therefore compute missingness (using PLINK) and depth per-chromosome (using bcftools) and then compute genome-wide averages weighted by the number of variants.

In both cohorts, several samples were flagged during the sequencing pipeline as having an insufficient number of reads, and additional lanes were requested for these samples. This resulted in a fat-tailed distribution of per-sample depth, with the bulk of samples exhibiting depths around 20x, and dozens (for Pomak) or hundreds (for MANOLIS) of samples with depths between 26x and 40x. This batch effect was not found to impact variant quality or sensitivity for SNVs, in keeping with down-sampling simulations which showed a negligible effect of depth increases between 15x and 30x [115]. In particular we found no correlation

between depth and singleton rate, as opposed to ethnicity (village of origin), which was the strongest per-sample predictor of singleton rate.

#### *5.2.2.4.Ethnicity checks*

Ethnicity checks take advantage of frequency differences in variations across human populations, to easily flag a sample that seems to originate from a different population than that which has been sequenced. The most common reason for such a discrepancy is sample mishandling at the sequencing facility, where DNA from one study gets mislabelled with an ID belonging to another.

Representing the sequenced samples on a PCA exposes the structure present in the isolates (Figure 35) but the absence of a reference population makes this unsuitable for ethnicity QC. In addition, we therefore perform two PCAs per cohort, one including all samples from the build 38 version of the 1000 Genomes VCF, the second including only the samples belonging to the EUR subpopulation.

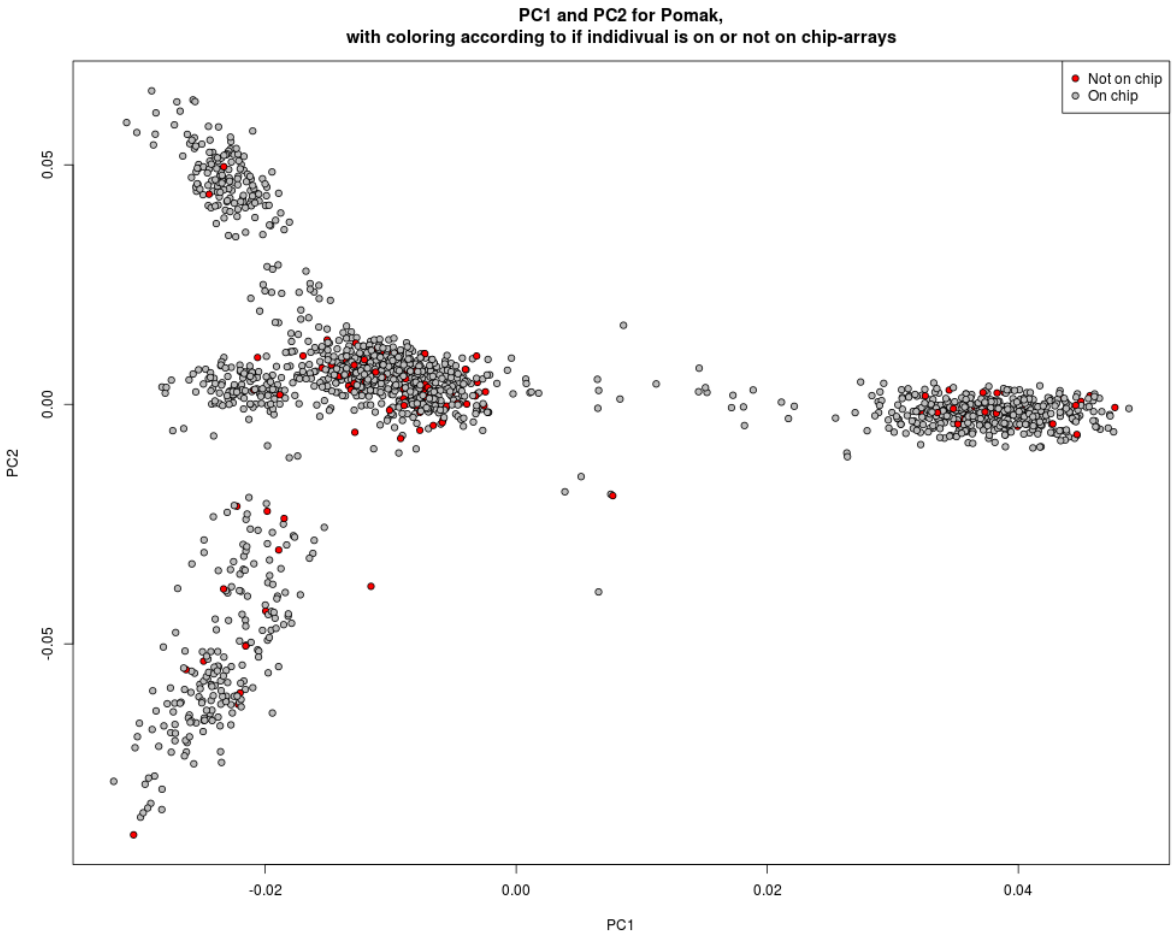



Figure 35 : First two principal components in the PCA of all sequenced individuals in the Pomak cohort.

In all cohorts with high-depth sequencing data analysed, no samples failed ethnicity checks using this method.

## 5.3.Results

### 5.3.1.Quality control results

For both cohorts, we used a threshold of 2% for Freemix, 5% for NRD and CHIPMIX, and a 4 standard deviation threshold for heterozygosity rate. We also excluded samples with less than 10x genome-wide average depth (only one sample in the Pomak cohort).

In total, we exclude 25 samples in MANOLIS and 27 samples in Pomak, respectively. In MANOLIS, we identify 10 samples with strong pi-hat with other samples; in 9 of those cases

we are able to resolve the duplication due to the presence of chip data. In Pomak, we remove three such duplicates.

In 100 samples from the TEENAGE cosmopolitan cohort used for comparison purposes, we apply the same quality control pipeline and do not exclude any samples based on sample-level QC.

### 5.3.2.Variant description

In the 1x dataset, we observed a genotype quality that decreased at the lower end of the MAF spectrum when compared to genotypes from array data. As expected, rare and low-frequency variants do not exhibit such a drop in the post-QC high-depth datasets (Figure 36).

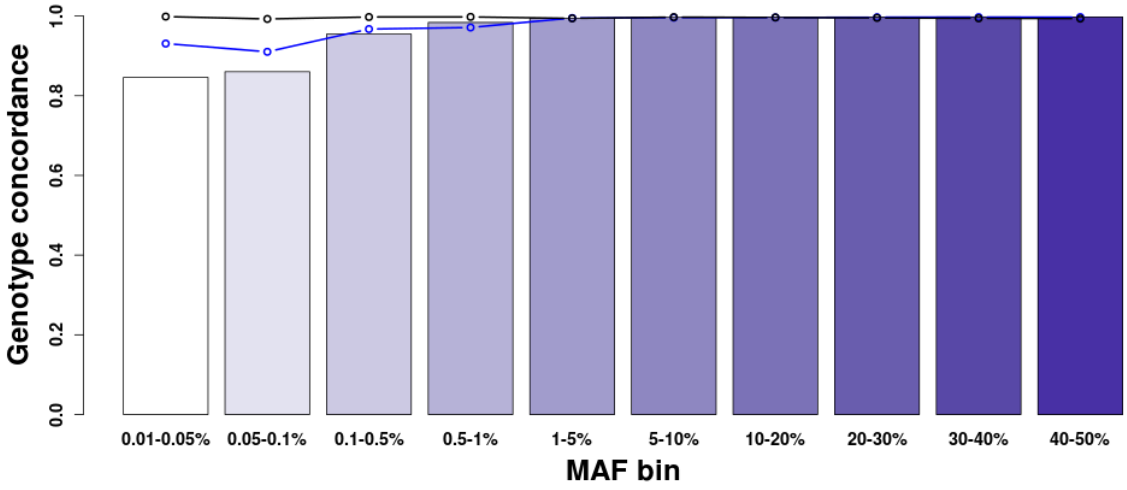



Figure 36 : Genotype (black) and minor allele (blue) concordance in the post-QC MANOLIS 22x dataset. Bars represent the proportion of variants intercepted when compared to the OmniExpress and ExomeChip data in the same samples. The first two MAF bins correspond to singletons and doubletons, respectively.

Further examining variants in the MANOLIS cohort, we count 24,163,896 non-monomorphic SNVs and INDELs, 97.9% of which are biallelic. 14,281,180 (60.31%) of the biallelic SNVs are rare (MAF<0.01); 3,103,273 (13.1%) are low-frequency (MAF 0.01-0.05); and 6,292,726 (26.57%) are common (MAF>0.05). We call 8,294 non-monomorphic variants annotated as loss-of-function (LoF) with low-confidence (LC)by Loftee[116], and 438 variants annotated as LoF with high-confidence (HC). On average, each individual carries 405 (standard deviation σ=19) LC LoF variants and 31 (σ=6) HC LoF variants, compared to 149 LoF variants per sample in a whole genome sequencing study of 2,636 Icelanders[117]. 0.6% and 1% of HC and LC LoF carrier genotypes are homozygous, respectively. Consistently with previous reports, INDELs are significantly more frequent among LoF variants, with 53.2% and 76% in the low- and high-confidence sets, respectively, compared to 13.5% genome-wide. We also observe an enrichment of rare variants in the loss-of function and "other coding" categories (Figure 37) which is still significant when merging all coding and splice variant categories ($p$=9.5x10$^{-16}$).

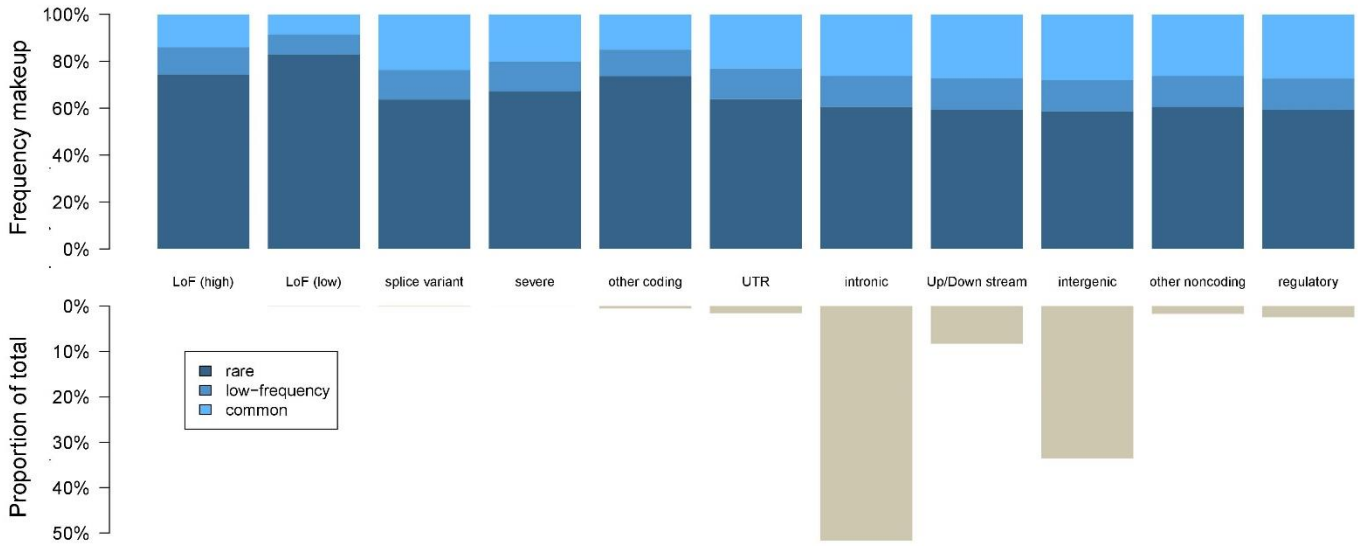



Figure 37 : Variant count proportions and minor allele frequency bin by functional class. Functional consequences are taken from Ensembl Variant Effecct Predictor (VEP)

Comparing singleton and doubleton rate with those observed in the general Greek population is not straightforward, as the sample size is an order of magnitude lower in the TEENAGE dataset. We randomly draw 1,000 sets of 100 samples from the MANOLIS study and count singletons and doubletons. We then build a distribution and compute empirical quantiles for the singleton and doubleton counts calculated in TEENAGE. We count 270,916 singletons and 61,690 doubletons in TEENAGE, compared with a median of 179,100 ($p=1.4\times10^{-94}$) and 75,280 ($p=3.0\times10^{-19}$), respectively, in MANOLIS (Figure 38). This reflects a lower rate of

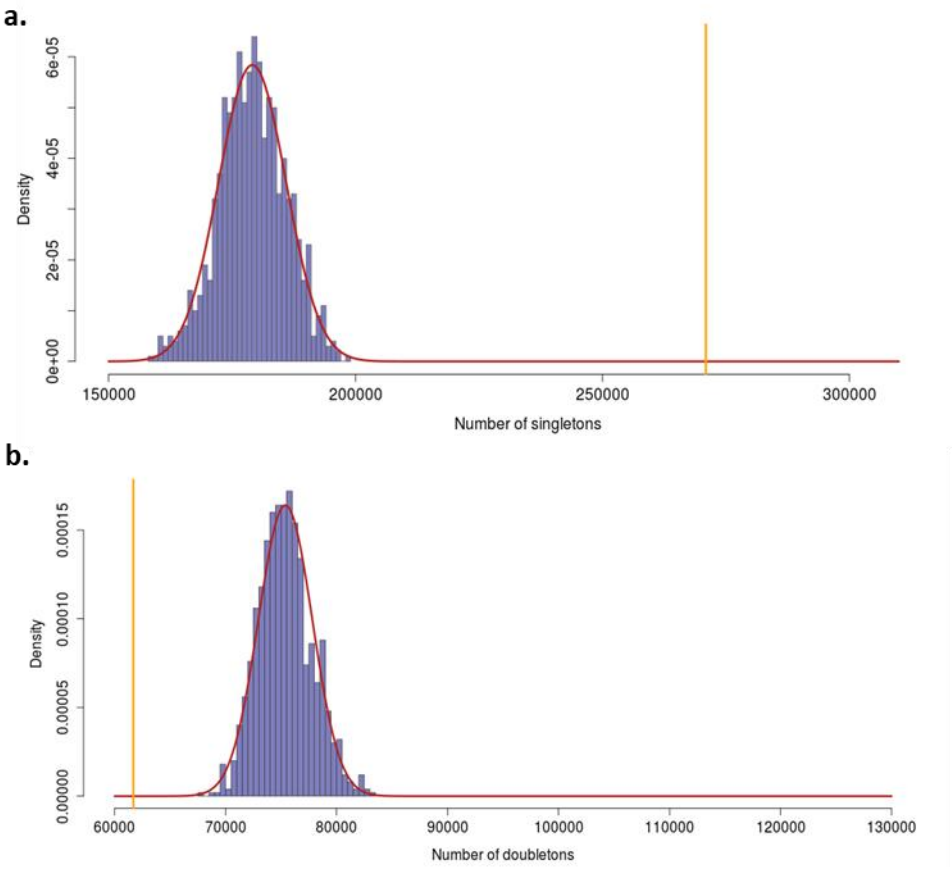



Figure 38 : Distributions of (a) singleton and (b) doubleton counts in 1,000 draws of 100 MANOLIS samples. The vertical orange line indicates the observed count in 100 TEENAGE samples downsampled to 22.5x.

singletons compared to the general Greek population ($p=10^{-167}$, one-sided empirical *P*-value), in keeping with the isolated nature of this Cretan population. Among the 5,102,175 novel

biallelic variants (not present in gnomAD[118] or Ensembl release 84[119]), 4,394,678 are SNVs, and the majority are rare (Figure 39).

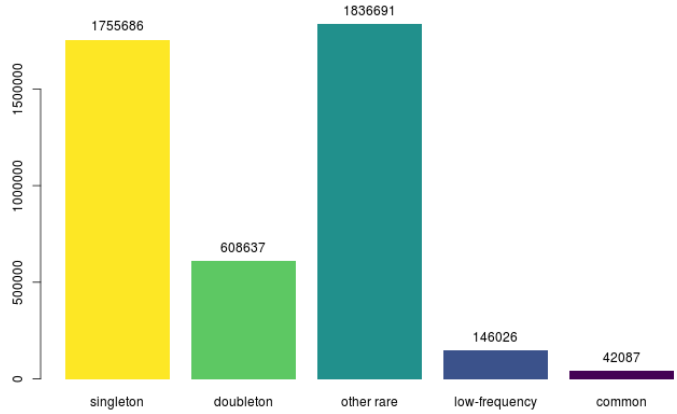



Figure 39: (a) Frequencies of all novel variants in the MANOLIS whole genome sequence data. Novelty is established by comparing variant location and alleles to Ensembl VEP annotation as well as gnomAD genomic variants lifted-over to build 38

### 5.3.3.Single-point association analysis

We performed single-point association using the same wrapper script for GEMMA that we had used for analysis of the 1x data. The genome was split into 50,000 marker chunks, and we used an empirical relatedness matrix calculated using GEMMA after applying the filters mentioned in 5.2.1.4.

Since most of the common and low-frequency single-point signals would have been expected to be intercepted using the 1x and imputed GWAS association studies, the high-depth WGS-based single-point was not used as an association discovery dataset. Rather, its main use was to support the display and quality control of results from the rare variant burden study (Chapter 6).

We examine the MANOLIS single-point results to confirm this hypothesis. We identify 26 independent signals at study-wide significance. This includes 5 previously reported signals, all of which were found in the 1x study, except one which was attenuated to $9x10^{-4}$(Table 4).

Table 4 : P-values in the 1x data for all suggestively significant ($5x10^{-7}$) known signals in the 22x MANOLIS data.

| Trait | Variants (build 38) | Variants (build 37) | Frequency class | $5x10^{-7}$ in 1x |
|---|---|---|---|---|
| Bilirubin | chr2:233755940 | chr2:234664586 | Common | yes |
| MPV, LPCR, PDW | chr12:121779004 | chr12:122216910 | Common | no ($1x10^{-4}$) |
| HGB, MCH, MCHC, HCT, RDW, RDWPC, MCV, RBC | chr11:6223409, chr11:11102466, chr11:5213949, chr11:6813230, chr11:249151, chr11:10524276, chr11:9899824, chr11:1120073, chr11:9267314, chr11:8257597, chr11:7368326, chr11:1704128, chr11:8246586, chr11:2763925, chr11:12919966, chr11:3037439, chr11:3601909, chr11:10003705, chr11:2115539, chr11:5226820, chr11:3905668, chr11:5193583 | chr11:6244639, chr11:11124013, chr11:5235179, chr11:6834461, chr11:249151, chr11:10545823, chr11:9921371, chr11:1113981, chr11:9288861, chr11:8279144, chr11:7389557, chr11:1725358-1725358, chr11:8268133, chr11:2785155, chr11:12941513, chr11:3058669, chr11:3623139, chr11:10025252, chr11:2136769, chr11:5214813 | low-frequency | yes |
| TG, HDL | chr11:116830637, chr11:117332401 | chr11:116701353, chr11:117203117, chr11:117279094 | low-frequency | yes |
| HDL | chr16:67661066, chr16:56970977, chr16:67057964, chr16:66457064, chr16:68206426 | chr16:67694969, chr16:57004889, chr16:67091867, chr16:66490967, chr16:68240329 | Common | yes |

Similarly, for the 21 novel signals for which we could not find a relevant previous association for the variant itself, variants in LD or in a 1Mbp window either side of the driver, all except two are present in the 1x study, 5 are genome-wide significant, and a further 5 have single-point p-values smaller than $5x10^{-5}$ in the 1x data. In all cases except for a single association of a rare variant with FT4 on chromosome 9, we are able to find a suggestively associated tagging variant with association p-value smaller than $10^{-4}$.

Table 5 : P-values in the 1x data for all suggestively significant ($5x10^{-7}$) novel signals in the 22x MANOLIS data.

| Trait | Trait Group | Consequences | Closest protein coding genes | frequency category | MAF | Position (build38) | Position (build37) | P-value in 22x data | P-value in 1x data | Proxy p-value (1Mb window) |
|---|---|---|---|---|---|---|---|---|---|---|
| Hip | anthropometric | intergenic variant | *C2orf91*(8.3kb) | Rare | 0.006 | chr2:41962658 | chr2:42189798 | 2.25E-08 | 7.00E-03 | 1.00E-05 |
| gammaGT | Biochemical | intergenic variant | *ADRA2C*(100.6kb) | Rare | 0.005 | chr4:3869191 | chr4:3870918 | 1.16E-08 | 8.70E-05 | 6.80E-08 |
| TSH | Biochemical | intron variant | *BRAF*(0kb) | Rare | 0.009 | chr7:140878126 | chr7:140577926 | 2.77E-08 | 5.70E-04 | 5.80E-05 |
| FT4 | Biochemical | upstream gene variant | *ERCC6L2*(80.4kb) | Rare | 0.003 | chr9:96094983 | chr9:98857265 | 6.03E-09 | 2.83E-02 | 4.00E-05 |
| PLT | Blood | intron variant | *LINC00959*(0kb) | low-freq | 0.014 | chr10:130079049 | chr10:131877313 | 2.01E-08 | 8.20E-04 | 5.70E-05 |
| HGB, HCT | Blood | intron variant | *CHRM3*(0kb) | Rare | 0.008 | chr1:239836658 | chr1:239999958 | 4.91E-08 | 1.20E-06 | 6.40E-08 |
| WBC | Blood | intergenic variant | *VRK1*(438.6kb) | low-freq | 0.03 | chr14:97370421 | chr14:97836758 | 2.18E-08 | 7.20E-05 | 9.70E-06 |
| MCH, MCV, RBC | Blood | 3 prime UTR variant | *RAB11FIP3*(0kb) | low-freq | 0.017 | chr16:521731 | chr16:571731 | 5.33E-14 | 2.50E-06 | 6.40E-08 |
| LPCR, PDW | Blood | intron variant | *VAT1L*(0kb) | common | 0.285 | chr16:77899776 | chr16:77933673 | 1.97E-08 | 2.70E-06 | 1.20E-06 |
| GRAN | Blood | intron variant | *NPHS1*(0kb) | common | 0.086 | chr19:35837072 | chr19:36327974 | 2.07E-08 | 3.10E-03 | 8.00E-06 |
| LPCR | Blood | intergenic variant | *SLC19A3*(36.9kb) | common | 0.057 | chr2:227754935 | chr2:228619651 | 2.80E-08 | 2.00E-07 | 2.00E-07 |
| HGB | Blood | intergenic variant | *GOLIM4*(64.9kb) | low-freq | 0.013 | chr3:167943703 | chr3:167661491 | 2.75E-08 | 1.70E-05 | 1.70E-05 |
| WBC | Blood | intron variant | *MICU3*(12.9kb) | low-freq | 0.024 | chr8:17135610 | chr8:16993119 | 2.50E-08 | 3.70E-05 | 2.50E-05 |
| PLT | Blood | 5 prime UTR variant | *CTSV*(0kb) | Rare | 0.005 | chr9:97039116 | chr9:99801398 | 4.52E-08 | not present | 7.60E-04 |
| HDL | Lipid | intron variant | *DYNLRB2*(401kb) | Rare | 0.004 | chr16:80139703 | chr16:80173600 | 4.57E-08 | 3.20E-03 | 1.00E-05 |
| adiponectin | Lipid | intron variant | *ATP8B1*(11.5kb) | Rare | 0.007 | chr18:57634877 | chr18:55302109 | 3.99E-08 | 2.43E-06 | 1.18E-07 |
| leptin | Lipid | upstream gene variant | *ATP11B*(50.9kb) | common | 0.488 | chr3:182742527 | chr3:182460315 | 3.50E-08 | 2.10E-05 | 1.90E-05 |
| HDL | Lipid | intron variant | *CXCL13*(0kb) | low-freq | 0.017 | chr4:77554679 | chr4:78475833 | 2.22E-08 | 1.80E-02 | 1.14E-05 |
| HDL | Lipid | intron variant | *TBXAS1*(0kb) | Rare | 0.008 | chr7:139991715 | chr7:139691514 | 8.23E-09 | 1.00E-03 | 3.90E-05 |
| HDL | Lipid | intergenic variant | *CSMD3*(353.1kb) | Rare | 0.006 | chr8:111869821 | chr8:112882050 | 7.49E-08 | 4.00E-06 | 4.00E-06 |
| RDW, RBC | Blood | downstream gene variant, intergenic variant | *OR4F4*(8.3kb) | Rare | 0.009 | chr15:101913651, chr15:101904736 | chr15:102453854, chr15:102444939 | 1.13E-07 | not present | 2.20E-05 |

### 5.3.4.Extended regions of association

Among the known signals, we recapitulate two extended LD regions found associated using the 1x data, one on chromosome 11 for blood traits, and another on chromosome 16 for lipid traits. Interestingly, the structure of the chromosome 16 association with high-density lipoprotein is slightly different when using the more accurate high-depth WGS genotypes. With approximate coordinates of 16:66143117-68618173 and a span of 2Mbp, the region is much shorter than the 8Mbp long segment identified using 1x sequencing. The LD structure is also different, with a string of 15 rare (MAF=0.007) novel variants in perfect LD driving the signal ($p=4.08x10^{-9}$). We list 21 carriers of this rare haplotype, of whom all except two originate from Anogia, the village with the largest amount of samples in MANOLIS.

### 5.3.5.Single-point cross-cohort meta-analysis using METACARPA

We performed single-point meta-analysis of 50 shared traits using METACARPA. We then produced quantile-quantile and Manhattan plots for all analysed traits, and performed peak calling using `peakplotter` using an exploratory threshold of $5x10^{-8}$. We report 120 peaks, which are distributed very unevenly between traits: 83% of traits have less than 5 peaks, whereas waist-hip ratio adjusted for BMI has 19. At a more stringent genome-wide significance threshold of $2x10^{-9}$, 46 peaks remain among which 30 belong to the haematology trait cluster on chromosome 11.

Table 6 : Signals in the single-point meta-analysis at various thresholds. The greyed out traits harbour peaks that survive the more stringent $2x10^{-9}$ and are discussed in the text.

| trait | Significant at $5x10^{-8}$ | Significant at $2x10^{-9}$ | Among which known |
|---|---|---|---|
| CRP | 1 | 1 | 1 |
| DBPBMIadj | 1 | 0 | 0 |
| Fe_iron | 1 | 1 | 1 |
| FG | 1 | 0 | 0 |
| Hip | 1 | 0 | 0 |
| HipBMIadj | 1 | 0 | 0 |
| HOMA_irBMIadj | 1 | 0 | 0 |
| LYM | 1 | 0 | 0 |
| SBP | 1 | 0 | 0 |
| SBPBMIadj | 1 | 0 | 0 |
| SGP | 1 | 0 | 0 |
| TSH | 1 | 1 | 1 |
| Weight | 1 | 0 | 0 |
| WHR | 1 | 0 | 0 |
| Bilirubin | 2 | 1 | 1 |
| Ferritin | 2 | 0 | 0 |
| gammaGT | 2 | 2 | 2 |
| PLT | 2 | 0 | 0 |
| RG | 2 | 1 | 1 |
| TC | 2 | 0 | 0 |
| Waist | 2 | 0 | 0 |
| FT4 | 3 | 0 | 0 |
| LDL | 4 | 1 | 1 |
| TG | 4 | 1 | 1 |
| HDL | 5 | 2 | 2 |
| WHRBMIadj | 19 | 5 | 0 |
| PLT | 2 | 0 | 0 |
| MPV | 8 | 4 | 4 |
| HGB | 7 | 2 | 1 |
| LPCR | 8 | 4 | 4 |
| PDW | 7 | 4 | 4 |
| LYM | 1 | 0 | 0 |
| HCT | 3 | 1 | 1 |
| WBC | 11 | 2 | 1 |
| MCHC | 4 | 3 | 1 |
| MCV | 3 | 3 | 1 |
| RBC | 6 | 4 | 1 |

The novelty of the signals reflects the attention that has been given to some of our traits in the GWAS literature so far, as well as the rarity of the driving SNVs: whereas all signals observed for triglycerides, Bilirubin, gamma-glutamyltransferase, C-reactive protein, iron and thyroid-stimulating hormone are known, all signals of association with FT4 (Free thyroxine) are novel and driven by rare intergenic SNVs. Both signals for ferritin are also novel and both driven by low-frequency variants, however none of the signals for FT4 or ferritin pass the more stringent $2x10^{-9}$ threshold. Briefly, we recapitulate:

- A known association between iron levels and rs4820268 in the *TMPRSS6* gene[120],
- two associations for gammaGT in *GGT1* and *HNF1A* genes. The first involves rs148246158, an intron variant in the nearby *BCRP3* gene, and an independent, documented association with rs2073398[121]. The second is driven by rs11065385, a common (MAF=0.35) intronic *HNF1A* variant previously associated with low-density lipoprotein[74]. *HNF1A* is a well-known locus, having been robustly associated both with gamma glutamyltransferase levels, C-reactive protein, LDL cholesterol, insulin resistance and type-II diabetes.
- rs4600067, a common intergenic variant (MAF=0.22), associated with TSH and in strong LD (r2=0.76) with rs10799824, a variant previously associated with thyroid-stimulating hormone[122].
- A similar association between an intergenic variant rs4546916, close to the *CRP* gene, in very high LD (r2=0.994) with rs7553007, a known[123,124] CRP variant.
- The APOC3 locus for both triglycerides and high-density lipoprotein [111].
- The common (MAF=0.38) *CETP* splice region variant rs1532625, associated with HDL, confirming numerous previous associations at this SNP and locus[74].
- The common (MAF=0.24) *UGT1A9* intron variant rs929596, associated with bilirubin levels, confirming numerous previous associations at this SNP and locus[125].
- The common (MAF=0.9) intronic *SUGP1* variant rs57962361, associated with LDL levels. It is in high LD (r2=0.79) with rs11668104, which has been associated with LDL before[74]. *SUGP1* plays a key role in the regulation of cholesterol metabolism.

The remaining associations which pass the $p=2x10^{-9}$ inclusion threshold are novel and driven by rare variants. One affects random glucose and is driven by a novel, rare intronic *DLEU1* variant. However given the extremely low MAF (singleton in MANOLIS, MAC=4 in Pomak) of the driver variant, this association must be regaded with caution. Waist-hip ratio adjusted for BMI, which harbours the remainder of non-haematological signals in the HELIC 15x meta-analysis, is associated exclusively with this class of rare variant in novel or undocumented loci: all five associations (2 intergenic, 3 intronic in *JAKMIP1, ADAM22* and *HMGA2*) are driven by single rare variants. This trait displays no measurable inflation ($\lambda_{GC} = 1.003$), although it does display a peculiar deviation at the upper end of the QQ-plot, likely carried over from the Pomak single-cohort analysis. This pattern persists when filtering out minor allele counts smaller than 5. In light of this observation, we exclude signals arising from this trait from downstream analyses in the Pomak-only analysis and the meta-analysis. A full list of the unfiltered Manhattan plots, QQ plots and signals files is available at https://github.com/wtsi-team144/HELICexplorer.

## 5.4. Discussion

We have established a variant-based and sample-based quality control pipeline for high-depth cohort wide sequencing, and have demonstrated that some metrics, like chromosome X relative depth, were more efficient than other variables traditionally used in the QC of genotyping array data. As opposed to very low depth WGS, variant-level QC is much simpler, as large number of high-depth samples is the ideal use case for VQSR, the main quality control tool used to filter variants. The direct availability of a high number of reads per sample and per position further allow to perform a much more complete quality control process than for the sparse 1x dataset, where sample QC was performed after imputation and mainly consisted in examining concordance with array data.

The isolated nature of the population also confers it special characteristics which need to be considered when performing sample quality control. In particular, variables assessing homozygosity or relatedness may exhibit higher average values in isolates. Since the extent of this increase will essentially depend on the population history of the studied cohort, it is

advised to base filtering criteria on the specific features of the distributions observed in an isolated study, rather than pre-determined arbitrary thresholds. In the HELIC datasets, cross-referencing the wealth of phenotypic and genotypic data allowed to disentangle sample quality from relatedness and perform rigorous sample QC.

For association, we re-used most of the procedures established with the 1x sequencing project, demonstrating the flexibility of the tools employed (chunking script, association script, peakplotter).

We demonstrate that for single-point association, all common and almost all low-frequency and rare variant driven signals found from high-depth WGS had already been discovered using 1x sequencing. In a meta-analysis of single-point association results between MANOLIS and Pomak, we report 16 independent signals passing the $p=1.9x10^{-9}$ mark, most of which are at previously reported loci. This is a staunch departure from what had been observed with 1x sequencing where most signals were novel and only two were significant at $p=1.9x10^{-9}$. The less noisy signal provided by high sequencing depths therefore seems to confer a definite power advantage in meta-analysis of association studies.

# Chapter 6. High-depth WGS-based rare variant aggregation tests

## 6.1. Background

The role of rare and/or noncoding variants in complex traits remains poorly understood. When several variants affect the same trait at the same locus, individual variant effects can be considered separately in a single-point framework, or in a combined fashion through variant aggregation methods. Although the constitution of very large reference panels for imputation such as HRC[14] and increasingly large sequencing-based projects[126] have enabled increased access to the rare end of the allelic spectrum, prohibitive costs still remain a significant obstacle to large-scale whole-genome-sequencing based studies of very rare variants. As a consequence, there is currently a lack of best practices regarding rare variant burden testing based on whole-genome WGS data and a number of unanswered questions remain.

First, there are study design challenges. One key question is to decide at which depth to sequence, in other words which sequencing depth confers the best sensitivity and specificity for rare variants while keeping sequencing costs to a minimum. Second, once sequencing data are obtained and passed through stringent quality control (Chapter 5), several analytical challenges present themselves, which are listed below.

Defining the unit of testing: Window-based approaches, in which variants across the genome are pooled based on proximity alone, suffer from loss of power due to the high number of tests and dramatic influence of overlap and size parameters[127]. Most rare variant studies so far, including our study of the *APOC3* gene, have focussed on exonic regions, as WES data remains the go-to choice for designing sequencing studies in many cases. Furthermore, coding variants have traditionally been thought to be more likely functional or more easily interpreted. Perhaps as a consequence, the function of rare noncoding variation remains unclear. WGS further allows to accurately genotype rare intronic, regulatory and intergenic variants. These regions can be considered independently of each other or can be linked together as part of a gene-centric approach.

Variant filtering: In common-variant GWAS, rare variants are often filtered out due to poor calling or imputation quality. Conversely for rare variant studies, variants with high MAF can be excluded to simplify the model and reduce execution time. To avoid loss of power, geneticists have sought to reduce the analysed set of variants to include only those most likely to contribute to an association signal. The study of exonic variation, as well as the constraints imposed by burden testing models, naturally put emphasis on selecting the variants with the highest likelihood of severely impacting gene function. Many methods exist to assess this functionality, some of them categorical[116], some of them providing a continuous score[128]. In the latter case, variants with a score below a certain threshold can be filtered out. Assessing the functionality of regulatory variation is not as straightforward, making the use of a binary inclusion/exclusion approach much harder in the context of WGS-based rare variant association studies.

Weighting scheme: Instead of, or in addition to the filtering approach above, it is possible to assign a different weight to each variant, which reflects its predicted consequence. Functional weighting scores now number in the hundreds, and although early efforts focussed on single indicators of variant function (such as the degree of evolutionary conservation, the impact on protein function or the regulatory impact), researchers now often use the "meta-score" approach, combining several existing metrics in order to summarize all available evidence at this locus[129]. Like the filtering method above, the performance of these methods in assessing variant functionality remains understudied, mainly due to lack of a robust list of positive controls; and this is especially true for regulatory variants. Choosing which weighting scheme to apply is therefore a challenging task with consequences for study power.

Linking regulatory regions to genes: A given gene might have several genomic regions regulating its function. It is therefore important to be able to aggregate rare variant effects across these regions and consider them as a single testing unit. One option is to include all regulatory regions within a certain distance of a gene of interest, without any guarantee that these elements actually effect gene expression. Another approach is to try and link regulatory regions to genes, for example when the regions overlap an eQTL for the gene.

eQTL, in turn, are not active in all tissues, which should prompt us to only consider expression in tissues relevant to the trait of interest. In many quantitative traits however, the relevant tissue is non-trivial or entirely unknown, and attempting to define tissues of interest based on current biological knowledge, especially on phenome-wide studies where the number of variables to consider is high, might bias results as well as increase false negative rates.

Meta-analysis and replication: Meta-analysing single-point association results is straightforward. In the case of rare variant burden tests, the same variants are unlikely to be present in all analysed cohorts, due to allelic heterogeneity. Moreover, even if some variants are shared between the cohorts, their frequencies and hence their effect might differ markedly. More generally, given the myriad of selection and weighting methods for variant inclusion, there is an open question about what constitutes a significant replication or meta-analysis in the context of burden studies. For example, if regulatory variants contribute to a burden of exonic variants in one cohort but not another, an exon-focused replication or meta-analysis strategy might fail to discover an actual gene-phenotype association. Here again, the lack of robust trans-cohort rare variant burden studies that would shed light on allelic heterogeneities at associated loci precludes us from making important choices a priori for our analysis pipeline.

## 6.2.Methods

### 6.2.1.Positive control analysis

An initial approach for solving analytical and study design issues was to focus on positive controls (regions where burden signals were expected to be found), and calibrate the parameters of the model so as to confer maximum discovery power in these documented cases. Inclusion of documented common variant loci would have been a way to further increase the number of positive controls. Such a positive control analysis is based on the hypotheses that:

- loci with common variant associations also harbour rare variant signals, *and/or*

- loci with documented rare variant associations will replicate as rare variant association in another cohort.

Both these hypotheses were unproven at the time this analysis was carried out, and subsequent results from the HELIC analyses actually contributed to disprove them. Furthermore, the list of documented rare variant associations that were both established and reliable was quite thin, effectively reducing the training set and putting more emphasis on the first hypothesis above. Furthermore, even if this method proved successful, the resulting study design might have become akin to an overfitted statistical model, that is, trained to recognise only the type of signal that it has already seen.

For these reasons, we instead focussed on trying to design a reasonable search space for all parameters of the analysis (so as not to excessively inflate the number of tests thereby reducing power), followed by genome-wide testing and stringent signal QC. This allowed us to discover patterns in the data at hand rather than overly relying on prior hypotheses, keeping in mind that a full comparison of all available variant weighting schemes for example, while much needed, is outside the scope of our applied analysis.

#### 6.2.2.Choice of association model

Two broad classes of methods are available for rare variant aggregation testing. The first family, commonly known as “burden tests” and exemplified by the Zeggini-Morris test[130], assume a concordant effect direction for all rare variants analysed. For a set of rare variants, and for each individual, they construct an aggregate genotype by either counting the number of minor alleles carried, or by recording whether a sample carries any number of the rare alleles. Because these methods essentially reduce to a single-point model, it is possible to evaluate the effect of the rare variants tested, since each additional rare allele will have an additive effect in the first case, and a dominant effect in the second. A drawback of this method is its tendency to quickly lose power if most tested variants have no true effect on the phenotype or have discordant directions of effect. The former caveat can be addressed by giving a lot of attention to the selection of variants included in the test, in order to reduce

the number of potentially non-functional variants as much as possible. Extensions of this test, such as FamBT[51] exist that account for family structure or relatedness.

The second family comprises so-called kernel or variance-component methods, which were pioneered by the Sequence Kernel Association Test (SKAT)[97], and which we have used in Chapter 4. The rare variants included in this type of test can have different effect directions and still contribute synergistically to the test statistic. This class of method has benefitted from a lot of methodological attention in the past decade, and a number of extensions of the method have been developed, notably for taking into account known family relationships[51].

Although these two models both have theoretical advantages in specific configurations of allelic architectures, it is impossible to know *a priori* which method will fit the data best and is likely to confer the most power in association discovery. Recently, a number of methods have attempted to bridge this gap, by using heuristics to decide which model is more likely at any given locus. The textbook case for these unified approaches is the SKAT-O method [131], which maintains near-optimal power in both types of allelic architectures.

Throughout this work, we have shown that accounting for increased relatedness is the main specificity of isolate studies compared to cosmopolitan cohorts, and this remains true when studying the effect of aggregations of rare variants on complex traits. Until not so long ago, accounting for relatedness had to be approximated by the inclusion of population covariates such as principal components, however in recent years methods were developed which could factor in relatedness matrices in the same way as linear mixed models could for single-point analyses, such as A-SKAT[55].

Ideally within the scope of the HELIC study, we would require a SKAT-O-type unified method, with the possibility to account for an empirical genetic relatedness matrix (as opposed to theoretical kinship derived from pedigree data), and which would allow to assign functional weights to variants to reflect their predicted consequence. At the time when the study design was carried out, only two such methods were available, FFBSKAT[132] and MONSTER[52]. We choose MONSTER for its speed and ease of use, despite the fact that FFBSKAT offers the advanced flexibility of choosing the type of kernel used in the model (MONSTER uses

only a linear kernel). In addition to the test p-value, MONSTER outputs a $\rho$ statistic, which is easily interpretable through the expression of its test statistic as the probability that the locus follows a burden, rather than a kernel-based model:

$$T_\rho = (1 - \rho).famSKAT + \rho.famBT$$

Where $famSKAT$ and $famBT$ are the respective test statistics of the kinship-adjusted SKAT and burden tests.

In this thesis, we use “burden testing” for brevity, even though the underlying parameter might give more weight to the variance component method.

### 6.2.3.Genetic relatedness matrix

Several methods are available to estimate the genetic relatedness present in isolated cohorts such as HELIC-MANOLIS[133]. We compared methods proposed in GEMMA[1], EMMAX[45], KING[134] and PLINK[89], and found that the kinship coefficients reported by each method were highly correlated, but on a different scale from each other (Figure 40). For consistency with single-point studies performed on the same samples, we calculated a genetic relatedness matrix using GEMMA [1] after filtering for MAF<0.05, missingness <1% and LD-based pruning.

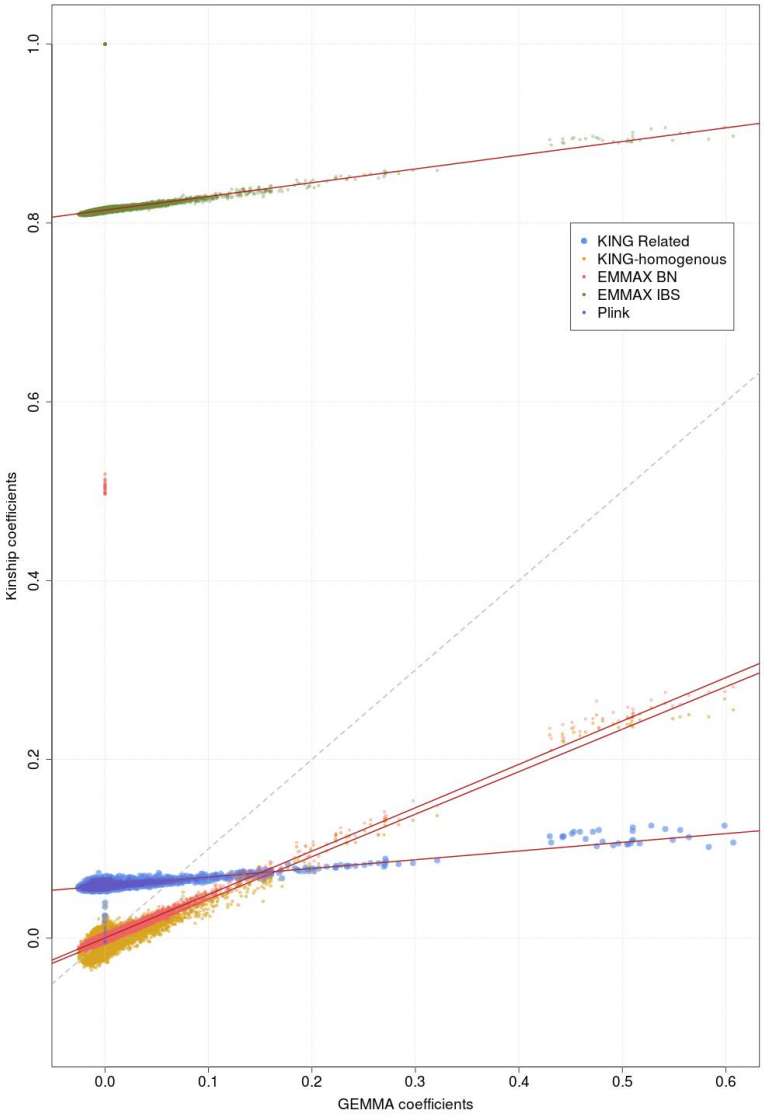



Figure 40: Comparison of kinship coefficients produced by KING and EMMAX to those produced by GEMMA[1]. All kinship coefficients are calculated using the same dataset (MAF>5%, missingness<1%, LD-pruned). IBS coefficients (KING Related, EMMAX IBS and Plink) are both higher on average and less sensitive to increased relatedness than their Balding-Nichols[2] counterparts (GEMMA, EMMAX BN and KING homogenous). Red lines represent per-dataset OLS regression slopes.

In addition, MONSTER requires self-kinship coefficients on the diagonal of the relatedness

matrix, which we calculated using the $\widehat{F_1}$ metric from PLINK 1.9. The matrix was then converted to the long format using the reshape2 R package.

#### 6.2.4.Conditional p-value

One defining criterion for a burden test is whether it is driven by a single variant. There are two methods to control for single-driver status: leave one out and conditional analysis. The first one repeats the burden test as many times as there are variants, by excluding one single variant each time. The second also repeats the burden analysis, but including all variants and conditioning the burden on the genotypes of every variant successively. Both methods have drawbacks: the first one is insensitive to LD, whereas MONSTER itself takes LD into account. This is problematic in cases such as when two variants are in high LD and are the sole drivers of a burden: removing one will leave the other in the test and leave the p-value mostly unchanged. Although rare variant genotypes are usually thought to be less correlated than those of common variants, this is mostly true for moderate levels of LD, whereas more extreme correlations can indeed be observed for rare variants, especially when several very rare variants sit on the same haplotype. The second conditional method is robust to this type of scenario, but is not as powerful when testing whether a low-frequency and a very rare variant are independent drivers. Indeed, if the covariate represents the genotypes of a singleton or doubleton, it contains almost no information and is likely to leave the p-value unchanged.

The choice of a conditional testing method is compounded by the additional difficulty that in our design, we run more than 10 different flavours of burden tests for every trait pair. This can create situations in which the most significant condition might fail chosen criteria for conditional analysis but less significant ones might pass. For example, if the most significant condition is in an exonic test and is driven by a single variant, an exon and regulatory run with a slightly higher p-value might survive conditional analysis. In theory, this scenario corresponds to an attenuation of an exonic single-point signal by a regulatory variant, however the slight difference in p-value might also be due to the use of a different score for the two analyses.

Given the necessary limitations imposed by compute time, we use a method that is likely to be broad-sweeping: we perform conditional analysis on the SNV with the most significant p-value only. We compute this conditional both in the run with the most significant burden p-value and in the run where the conditional p-value is maximal. We use a threshold of 0.05 for either p-value to qualify for the suggestive significance stage.

In practice, we found that the limit cases described above rarely occurred, and that the maximum conditional p-value was always close, rarely more than one order of magnitude away, from the conditioned most significant burden p-value. We then updated our conditional p-value threshold to $5x10^{-3}$ for either the top or maximum conditional test.

#### 6.2.5.Benchmarking of variant selection and weighting methods

We used a gene-centric approach, whereby burden testing was performed using MONSTER[52], across all 18,997 protein-coding genes defined in GENCODE v25 using 14 different conditions, i.e. combinations of regions of interest (coding regions only, coding and regulatory regions and regulatory regions only), variant filters (inclusion criteria based on severity of predicted consequence) and weighting schemes (Table 7).

Table 7 : Region definition, variant selection and weighting systems used to define testing conditions for burden analysis.

| Burden analysis condition | Weighting system | Criterion | Exons | Regulatory Regions | Number of genes (average all phenotypes, MANOLIS) |
|---|---|---|---|---|---|
| LOFTEE HC | none | Predicted LoF by LOFTEE with high confidence | yes | no | 85 |
| LOFTEE LC | none | Predicted LoF by LOFTEE, both high and low confidence | yes | no | 1,727 |
| Exon severe | none | Ensembl most severe consequence more severe than missense | yes | no | 7,660 |
| Exon CADD | CADD | none | yes | no | 18,428 |
| Exon CADD median | CADD | CADD>5.851 | yes | no | 18,138 |
| Exon+50 CADD | CADD | none | yes extended by 50bp | no | 18,551 |
| Exon Linsight | Linsight | none | yes extended by 50bp | no | 18,581 |
| Exon+Regulatory Eigen | Eigen (raw score + 1) | Eigen>0 | yes extended by 50bp | yes | 18,961 |
| Exon+Regulatory EigenPhred | Phred-transformed Eigen | none | yes extended by 50bp | yes | 18,660 |
| Exon+Regulatory EigenPCPhred | Phred-transformed EigenPC | none | yes extended by 50bp | yes | 18,722 |
| Exon+Regulatory mixed | CADD for coding variants, Eigen for all others, phred-scaled | none | yes extended by 50bp | yes | 18,650 |
| Exon+Regulatory Linsight | Linsight | none | yes extended by 50bp | yes | 18,688 |
| Regulatory only EigenPhred | Phred-transformed Eigen | none | no | yes | 17,607 |
| Regulatory only Linsight | Linsight | none | no | yes | 17,610 |

#### *6.2.5.1.Exonic-only runs*

First, we extracted exonic coordinates for all protein-coding genes, which defines the region of interest for strictly exonic variants. These regions of interest were used in combination with 5 different variant filtering and weighting schemes. First, we included only variants predicted as high-confidence (HC) loss-of-function (LoF) by LOFTEE[116] that reside in the exons of protein-coding genes (LOFTEE HC). As only 460 variants in 85 genes passed this inclusion criterion (all figures presented are for MANOLIS), we performed an additional analysis including 8,570 low-confidence (LC) loss-of-function variants spread across 1,727 genes (LOFTEE LC). Stop-gained and frameshift mutations were the largest contributors to both the LC and HC sets. However, the LC set also includes a large number of splice donor and splice acceptor variants. We further performed an analysis with more relaxed inclusion criteria, including all exonic variants for which the Ensembl most severe consequence was more damaging than missense as predicted by the Variant Effect Predictor[135] (Exon severe). We also employed Combined Annotation Dependent Depletion (CADD)[128] scores, either to weigh all exonic variants (Exon CADD) or to filter out variants with CADD scores below the genome-wide median (Exon CADD median). Finally, we extended exon boundaries as defined above with 50 base pairs either side, to account for cases where potentially damaging variants occur on the edges of exons, as has been shown to happen for previously identified rare variant burdens[111]. These regions of interest were used in combination with one variant weighting scheme only (Exon+50 CADD).

#### *6.2.5.2.Regulatory runs*

We extracted regulatory regions (promoters, enhancers and transcription-factor binding sites) from Ensembl build 84[119]. We assigned regulatory regions to genes if they directly overlapped or if the regulatory region overlapped with an eQTL for the gene based on the GTEx database[136]. If an eQTL was reported for several genes, overlapping variants were assigned to all of them. We did not take tissue specificity into account. For selecting variants, we either used the coordinates of the regulatory features alone, or regulatory features plus the extended exons. We used Eigen, an aggregate score that combines

information from multiple regulatory annotation tracks[129], to weigh variants in all tests that include regulatory variants. In addition to raw Eigen scores, the authors also proposed EigenPC, a score derived from the first eigenvector of the correlation matrix of annotations. Both scores were available as is, or transformed using phred-scaling, which maps a distribution's support to $[1, +\infty[$, thereby guaranteeing inclusion and relative up-weighting of all variants. In the regulatory regions plus exon analyses we used both the raw Eigen scores, shifted by 1 unit to the right, with negative scores set to the smallest possible floating point number $0 + \epsilon$ (Exon and regulatory Eigen), and the Phred-transformed Eigen and EigenPC scores (Exon and regulatory EigenPhred and EigenPCPhred). This transformation was a technical requirement as MONSTER could only read weights belonging to $]0, +\infty[$. For phred-transformed scores, we repeated the transformation as several chunks were found to be missing in the files provided on the authors' website. In the analyses containing the regulatory regions only, variants were weighted using the phred-scaled Eigen scores (regulatory only EigenPhred). Finally, to eliminate any potential bias of either method towards coding or regulatory variants, we also combined the phred-transformed Eigen scores and phred-transformed CADD scores by using the latter for coding variants, the former for every other variant.

As a comparison, we also included the Linsight[137] weighting system, a recently published competitor to Eigen, in our benchmark. We used it instead of CADD to weigh exonic analyses (Exon Linsight), and instead of Eigen for weighting both exon and regulatory (Exon+Regulatory Linsight) and regulatory only (Regulatory only Linsight) analyses.

Independently of all other parameters, we only performed a test if at least two SNVs passed the inclusion criteria for a given condition. After preparing the weights as described above, burden testing was then run genome-wide on 57 phenotypes for MANOLIS and 51 phenotypes for Pomak using these 14 configurations.

## 6.3.Results

### 6.3.1.Benchmark of testing methods

First, we use the MANOLIS test results to examine how different sets of parameters influence the results of rare variant association studies. Overall, p values correlate highly within three distinct clusters (). Among exonic-only runs, rare variant tests that only include unweighted high-consequence variants (exon LoF and exon severe) cluster separately from ones where variants are weighted according to their functionality scores (exon CADD, exon CADD median, exon+50bp CADD, exon Linsight), whereas the third cluster encompasses all tests that include regulatory variants.

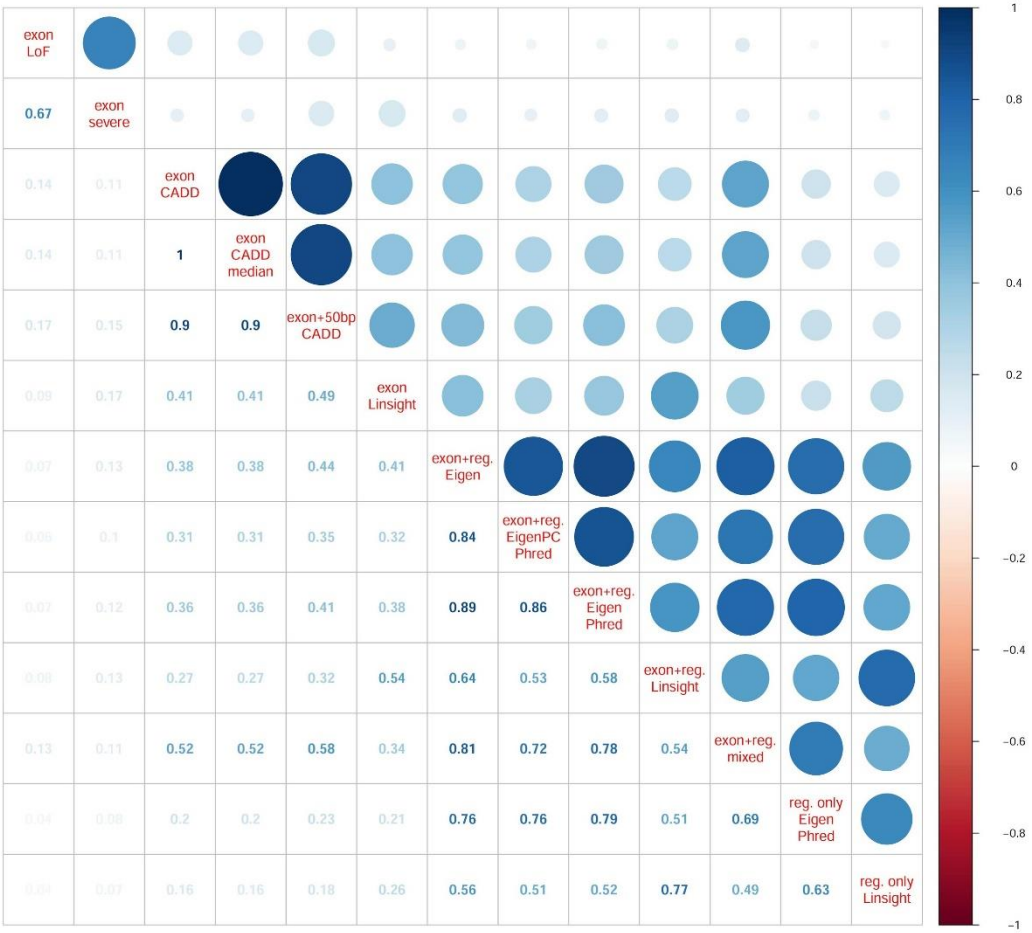



Figure 41: Correlogram of z-scores arising from all evaluated burden testing scenarios for the MANOLIS burden testing study.

A few runs seem to be less bound by this rule. Although the exon and regulatory, CADD+Eigen mixed score correlates most highly with fellow exon and regulatory runs, it has a higher correlation ($r^2$>0.5) with exonic runs than all others in its class. This is expected given how the score is constructed. After close examination of several signals arising from runs weighted using this score, we observe an unintended side-effect: some regulatory variants are weighted with the CADD score instead of the Eigen one if the regulatory region for a gene overlap the exon of another gene. Since CADD scores are likely to be very high for coding variants, the score of these regulatory variants is inflated, reflecting their consequences for a different gene than the one that is being studied. Due to this bias, we discard the CADD+Eigen mixed score from all further analyses.

Another set of runs that stand out are those that use the Linsight weighting system. In general, we observe that Linsight-weighted runs correlate more with each other than with other analyses run on similar regions. This can either reflect a biological fact, i.e. that the regions in which to select variants do not matter because regulatory burdens recapitulate signals found in exon-only analyses, or it can be due to a technical artefact, i.e. that Linsight-weighted scores behave like the mixed score, insufficiently accounting for the fundamental difference between regulatory and coding effects and the potential effect of a single variant on multiple genes.

A more detailed examination of burden testing results reveals several clues that suggest the latter option is more likely. We do still observe a large overlap between Linsight-weighted and other runs in their category. Mostly, when other runs have a low p-value for a trait-gene pair, Linsight-weighted runs appear significant as well. Conversely, we observe that a number of signals appear only in Linsight-weighted runs, across variant selection methods, suggesting that the increase in correlation within the Linsight group of analyses is driven at least in part by such signals exclusive to the category. Throughout our exploratory analyses, we found that some of these new signals were more likely to be found also by the flawed CADD+Eigen mixed score, which prompted us to not consider Linsight-weighted runs further in our downstream and prioritisation analyses.

### 6.3.2.Significance threshold

For burden testing, a strict Bonferroni correction would adjust for the number of traits, number of conditions and number of genes tested. We calculate $\alpha_{eff} = \frac{0.05}{N \times n_{cond} \times M_{eff}}$ where $N$ is the number of genes tested, $n_{cond}$ is the effective number of conditions tested and $M_{eff}$ is the effective number of traits. Since the $M = 57$ traits tested are highly correlated, we perform a PCA analysis, which indicates that 20 and 24 traits explain 90 and 95% of phenotype variance, respectively (Figure 42). Using the reduction method described in [138], we find a slightly higher $M_{eff} = 31$. We use $M_{eff} = 20$. For $n_{cond}$, we use the previous observation that the 10 analyses (4 were discarded as described above) clustered into three groups with very similar p-values, reducing the effective number of analyses to three. Further correlation is present, however, which means a value of $n_{cond}$ between one and three probably best represents the structure of the signal. Although $N = 18997$ protein-coding genes are available in GENCODE V25, not all genes are tested in every condition. For example, for many genes only one variant might pass inclusion criteria in a high-confidence loss-of-function run, thereby excluding those genes from the analysis (Table 7). On average, $N = 14{,}290$ genes are included in an analysis. Factoring in these elements, we obtain a significance threshold between $p = 1.7 \times 10^{-7}$ for $n_{cond} = 1$ and $p = 5.8 \times 10^{-8}$ for $n_{cond} = 3$.

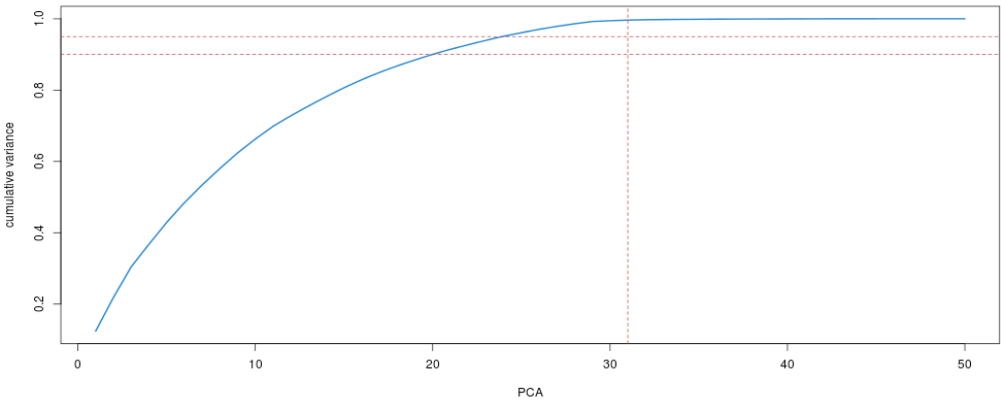



Figure 42 : PCA analysis for the 48 quantitative traits tested in the burden analysis. The two horizontal dashed lines indicate 90 and 95% variance explained, and the vertical dashed line indicates the estimate produced by the [138] method.

### 6.3.3.Pipelining script and burden visualisation software

Running MONSTER on a large number of traits using many different weighting systems and variant selection methods is a daunting task, especially when compounded by the relatively modern build 38 genomic coordinate system used for alignment and calling of the HELIC datasets, at a time where build 37 was still widely used for most annotations. We make available MUMMY (https://github.com/wtsi-team144/burden_testing), a flexible pipelining tool co-written by Daniel Suveges and myself, for running burden tests under various conditions using MONSTER. It has been extended by Emil Jorsboe for build 37 and imputed (dosage-based) genotypes.

Rare variant burden signals are harder to represent than their single-point counterparts. The extended LD structure at rare variant loci, as well as weights, allele frequencies, regulatory elements, variant consequences, exons and transcripts are all potentially useful pieces of information to display, which requires a flexibility that current regional association plotting tools do not provide. To address this issue, we write plotburden (https://github.com/wtsi-team144/plotburden), a python and bash script which makes use of the Bokeh library (https://bokeh.pydata.org) to create interactive plots representing burden associations. It represents both the burden p-value, the variants included in the burden with their weights, the genomic context (by displaying exons, introns and regulatory elements of nearby genes), the single-point p-values for all variants, every previous association at the locus as well as the LD structure. Because this information is quite dense when displayed at once, the plot is reactive and allows to display only certain types of information selectively and at will. Figure 45 presents a typical output, minus the interactive aspect. The plot is divided into two panels. The bottom one represents genomic context, with the analysed gene highlighted in pale red. Regulatory regions are added as tracks below the gene, in a shade of green corresponding to their type (hovering over any genomic feature displays a dialog containing relevant information such as the name of the feature and its Ensembl ID). The top panel contains association results, with the burden p-value displayed by a horizontal red line. P-values for all variants in the region are drawn as well in a LocusZoom-type layout. Variants not included in the burden are represented with small grey

dots, whereas variants included in the burden are larger, sized and coloured either by MAF (Orange-Purple scale) or weighting scheme (Viridis Green/Blue scale) depending on the user's wish. Optional vertical red lines present previous associations as provided by Ensembl (containing the GWAS Catalog, HGMD and OMIM). An interactive example can be found in the repository at https://rawgit.com/wtsi-team144/plotburden/master/example.html.

### 6.3.4.Burden testing results

We run MUMMY, the MONSTER wrapper, on the 18,997 genes in the Gencode V25 using the 14 region types and variant weights selected above, and although we run the full array of quantitative traits present in HELIC, we focus our discovery on non-haematological traits. Our main replication study is the first freeze of the INTERVAL WGS dataset, for which we describe preparation and QC in paragraph 6.3.5.

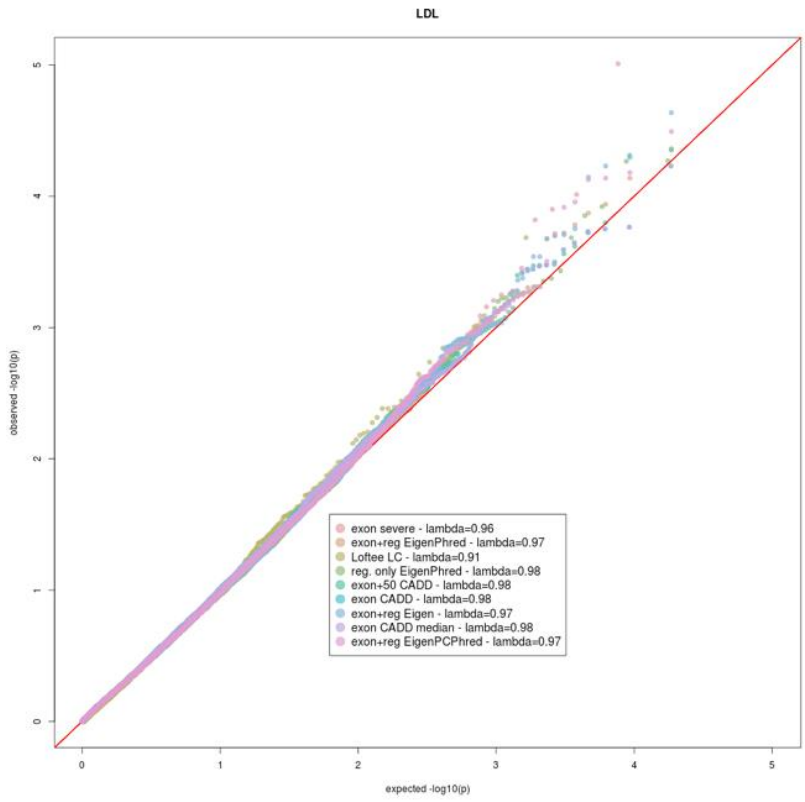



Figure 43 : Quantile-Quantile (QQ) plots for nine genome-wide runs using different testing conditions for the low-density lipoprotein (LDL) phenotype.

Overall, tests were well calibrated across all phenotypes and variant selection and weighting methods. Type-I error was well-controlled, and test statistics were slightly deflated only in loss-of-function runs, where variant selection criteria are most stringent ($\lambda_{GC} = 0.91$ for LDL).

In a typical GWAS setting, it is often assumed that the p-values arise from a test against the chi-square distribution, and $\lambda_{GC}$ is therefore calculated as $\lambda_{GC} = \frac{median_i\left(F^{-1}_{\chi^2_{(1)}}(p_i)\right)}{F^{-1}_{\chi^2_{(1)}}(0.5)}$, where $F^{-1}_{\chi^2_{(1)}}(p_i)$ is the inverse cumulative distribution function (or quantile function) of the chi-square distribution with 1 degree of freedom[139]. When evaluating burden tests, the underlying test statistic may not be chi-square distributed, and therefore we use the more general $\lambda_{GC} = \frac{median(-log_{10}(p))}{-log_{10}(0.5)}$, which only uses the weaker assumption that the p-value distribution is uniform.

#### *6.3.4.1.MANOLIS*

A total of 1,492 trait-gene pairs pass our suggestive significance threshold of $5x10^{-5}$ prior to conditional testing. 833 of these pairs involve non-haematological traits. After conditional analysis, these numbers drop to 574 and 368, respectively. Further restricting results to $p<1.7x10^{-7}$, we obtain 47 and 25 signals, respectively.

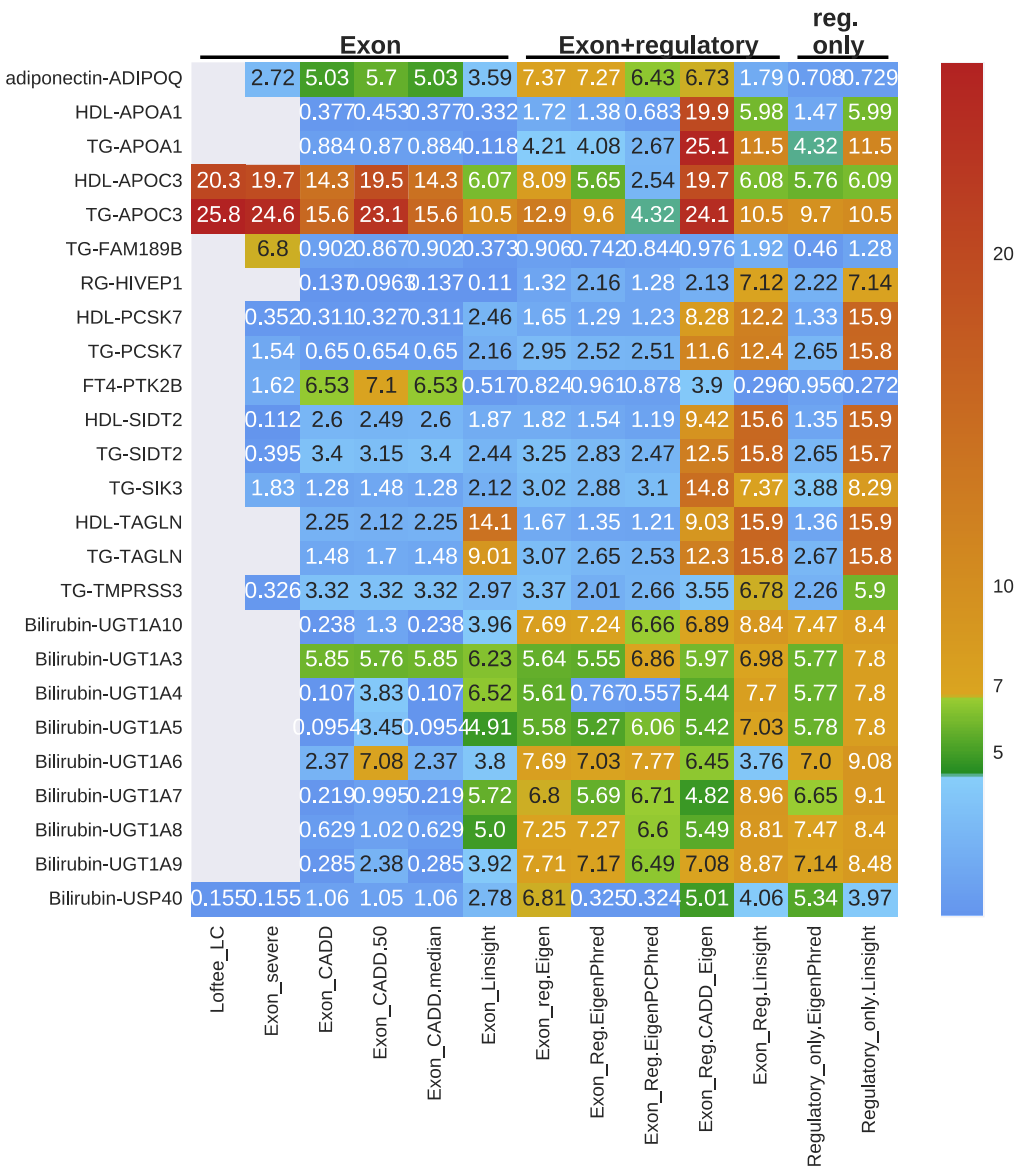



Figure 44: Signals passing conditional and study-wide significance in MANOLIS, all testing conditions, non-haematological traits (the numbers are the -$\log_{10}$(p) for the trait-gene pair in the specified testing condition).

High-level patterns are visible in the results. Some signals (*APOC3*) traverse all testing conditions whereas others are restricted to exonic variants (*PTK2B*) or regulatory ones (*UGT1Axx*). We observe the same distance between Linsight and all other weighting and variant selection methods that we had already observed in the correlogram. We disregard signals that are driven only by Linsight or by the combined Eigen/CADD runs (*TAGLN* for HDL and TG, *TMPRSS3* for TG, *HIVEP1* for RG). The remaining significant trait-gene pairs can be reduced to four independent signals (Figure 45, a-d):

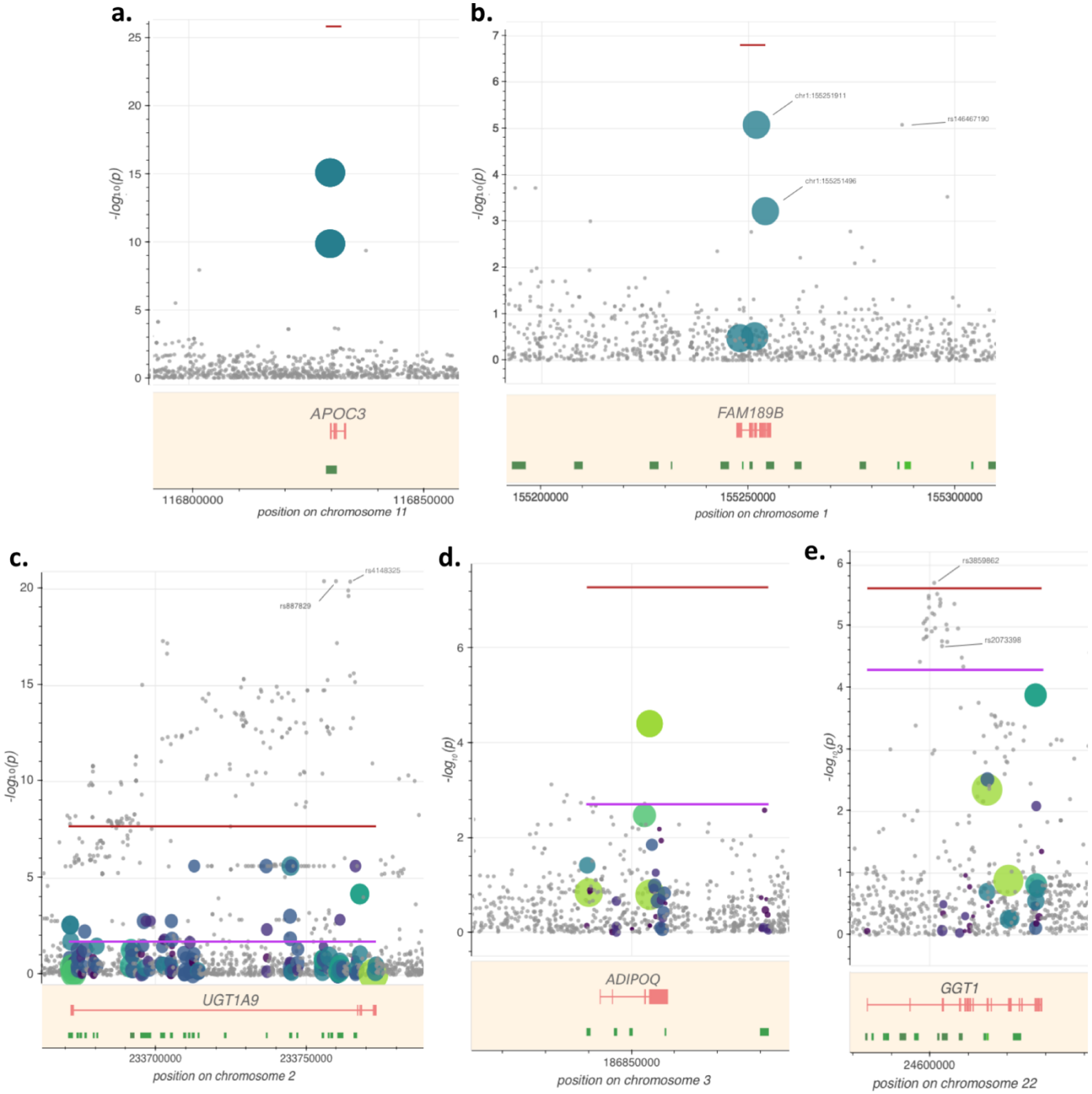
- A burden of loss-of-function variants with blood triglyceride and high-density lipoprotein levels in the *APOC3* gene, which has been previously documented, including by ourselves using the 1x data[111,140]. The strongest signal arises when only the splice-donor variant rs138326449 (MAC=38, MAF=0.013) and the stop-gained variant rs76353203 (MAC=62, MAF=0.022) are included in the analysis ($p=1.58\times10^{-26}$). Extended linkage disequilibrium (LD) in the region gives rise to regulatory variant-driven associations in the neighbouring *PCSK7, SIK3, SITD2, PAFAH1B2, TAGLN* and *APOA1* genes, all of which are conditionally dependent on rs138326449 or rs76353203.
- a novel association of triglyceride levels with rare variants in the *FAM189B* gene. The association ($p=1.56x10^{-7}$) is driven by two independent novel splice-region variants: chr1:155251911 G/A (human genome build 38, minor allele count (MAC)=3, $p=1.9\times10^{-6}$) and chr1:155254079 C/G (MAC=2, $p=1.9\times10^{-6}$). Both variants exhibit high variant quality scores (VQSLOD>19), high sequencing read depth (24x and 26.5x, respectively) and no missingness. A further novel splice region variant (chr1:155251496 T/C) and a stop gained variant (rs145265828), both singletons, were also included in the analysis; however their contribution to the burden was insignificant (with single-point P-values 0.30 and 0.33, respectively, and a burden $p=2.26x10^{-8}$ when excluding them). *FAM189B* has not been associated with blood lipid levels before. The gene overlaps with the assembly exception CHR_HSCHR1_2_CTG31, which contains the alternate sequence NW_003315906.1. This alternate sequence contains 15kb not present in the canonical reference for chromosome 1, with which it has 99.8% sequence homology. All discrepant bases overlap the *GBAP1* gene located 10kb upstream of *FAM189B*, and variant density is not lower in the region spanned by the exception, suggesting that it has little consequence in terms of integrity for the *FAM189B* signal.
- a low frequency and rare variant burden association with bilirubin levels in the *UGT1A9* gene ($p=1.9\times10^{-8}$), as well as to a lesser extent in *UGT1A1*, *UGT1A3*, *UGT1A4*, *UGT1A5*, *UGT1A6*, *UGT1A7*, *UGT1A8* and *UGT1A10*, all of which share exons and introns. This association arises from the analyses including exonic and regulatory variants, and in the analyses including regulatory variants only. A common variant in the first

intron of *UGT1A9* (rs887829, MAF=0.28, β=0.426, σ=4.27x10$^{-2}$, p=3.99×10$^{-21}$) has previously been associated with bilirubin levels [141,142]. As expected, correlation between genotypes at the common rs887829 and each of the low-frequency and rare variants included in the burden is low ($r_{max}^2$=0.1). The rs887829 signal is not attenuated when conditioning on carrier status for the two main drivers of the burden ($P_{conditional}$=4.45x10$^{-21}$), or when conditioning on the number of rare alleles carried per individual ($P_{conditional}$=3.99x10$^{-21}$). The evidence for association with the rare variant burden in *UGT1A9* drops to $P_{conditional}$=0.0146 when conditioned on rs887829 genotypes. Conversely, the two-variant burden signal is attenuated from $p$=1.4x10$^{-7}$ to $p_{conditional}$=7.0x10$^{-3}$ when conditioning on rs887829, indicating that it recapitulates part of a signal driven by a known common association in the region.

- an association of adiponectin levels with low-frequency and rare variants in the *ADIPOQ* gene ($p$=4.2×10$^{-8}$). The evidence for association is stronger for exonic and regulatory variants combined than in either the regulatory only ($p$=0.19) or exon-only ($p$=2.0x10$^{-6}$) analyses, suggesting a genuine contribution of both classes of variants to the burden. The variant with the lowest single-point *P*-value is the low-frequency missense rs62625753 (MAF=0.031, $p$=4.02×10$^{-5}$), which has previously been associated with type 2 diabetes [143]. The *P*-value of the burden is attenuated, but not completely, when conditioned on the genotypes of rs62625753 is $P_{conditional}$=8.9×10$^{-4}$. No common-variant single-point signal for adiponectin levels is present in this gene in our dataset. The strongest association for previously-reported single-point signals is for rs822387 (MAF=0.091, $p$=0.0018), and the burden remains significant upon conditioning on the genotypes of all variants with previous associations for adiponectin or type 2 diabetes that were polymorphic in our dataset. Among the 3 *ADIPOQ* variants found to be associated with adiponectin in a recent UK10K meta-analysis, only the low-frequency rs17366653 is included in the burden (MAF=0.01, single-point p=3.4x10$^{-3}$, conditional p=1.1x10$^{-6}$). The rare rs74577862 is present in our study, but is not included in the burden due to its intronic nature, and is not associated with adiponectin levels (p=0.145, conditional burden p=1.2x10$^{-7}$), whereas

the common variant rs201813484 is only moderately associated and also independent from our burden (single point p=1.48x10$^{-3}$, conditional p=3.1x10$^{-7}$).

Figure 45 : Burden signals in the *APOC3* (a), *FAM189B* (b), *UGT1A9* (c), *ADIPOQ* (d) and *GGT1* (d) genes.



A further signal reaches study-wide significance: variants in *PTK2B* are associated with thyroxine (FT4) ($p=7.9x10^{-8}$). However, upon inspection, this burden includes one variants whose single-point p-value is extremely close to the burden p-value. Despite this fact, the conditional p-value for that variant only increases the p-value by two orders of magnitude to $p_{conditional}=6.1x10^{-6}$. We further observe that the second most significantly associated variant included in the burden actually has a dampening effect: conditioning on it decreases the p-value to $1x10^{-9}$, which might explain why the p-value is so high despite having a driver variant with a low p-value. We are unable to replicate this burden due to the absence of thyroxine level measurements in INTERVAL, our main replication cohort (see 6.3.5), and we treat this signal as suspicious.

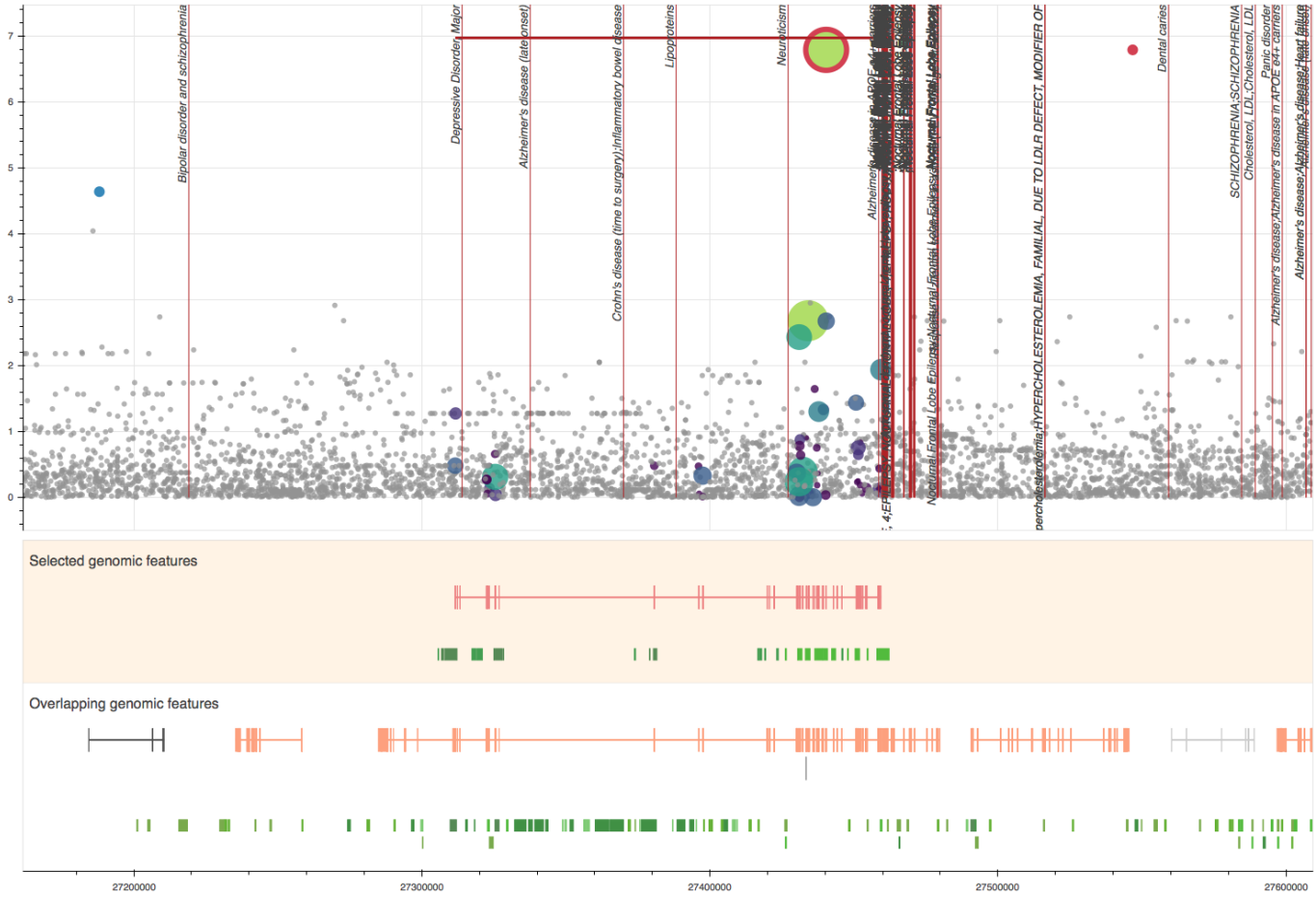



Figure 46: Burden of variants in *PTK2B* for thyroxine levels.

### Suggestively significant signals

353 signals pass conditional analysis and are suggestively significant without reaching study-wide significance. Removing Linsight-only signals and signals only arising in the CADD-Eigen mixed run produces 92 gene-trait pairs. Some signals are particularly worthy of interest. For example, gamma-glutamyltransferase levels are associated with low frequency and rare variants in the gamma-glutamyltransferase 1 (*GGT1*) gene ($p$=2.34×10$^{-6}$) in the exonic weighted analyses. A previously-reported, single-point common-variant association is also present in an intron of this gene (rs3859862, MAF=0.46, $p$=1.9×10$^{-6}$). LD as measured by $r^2$ is low between rs3859862 and all 34 variants included in the most significant burden analysis ($r_{max}^2$=0.02), whereas D', a measure of correlation more robust to heterogeneities in allele frequencies[18427557], is equal to 1 with 24 of these SNVs. The burden signal in *GGT1* is maintained when conditioning on rs3859862 ($p$=5.1×10$^{-5}$), suggesting that rare variants may be independently contributing to this established association. Similarly, the single-point association at rs3859862 conditioned on carrier status for all rare variants included in the burden is not attenuated ($p$=2.8×10$^{-5}$), a result recapitulated by conditioning the same variant on the number of rare alleles carried per individual ($p$=1.8×10$^{-5}$), strengthening the hypothesis of an independent rare variant signal at this locus. Gamma-glutamyltransferase levels are not available in the INTERVAL replication cohort.

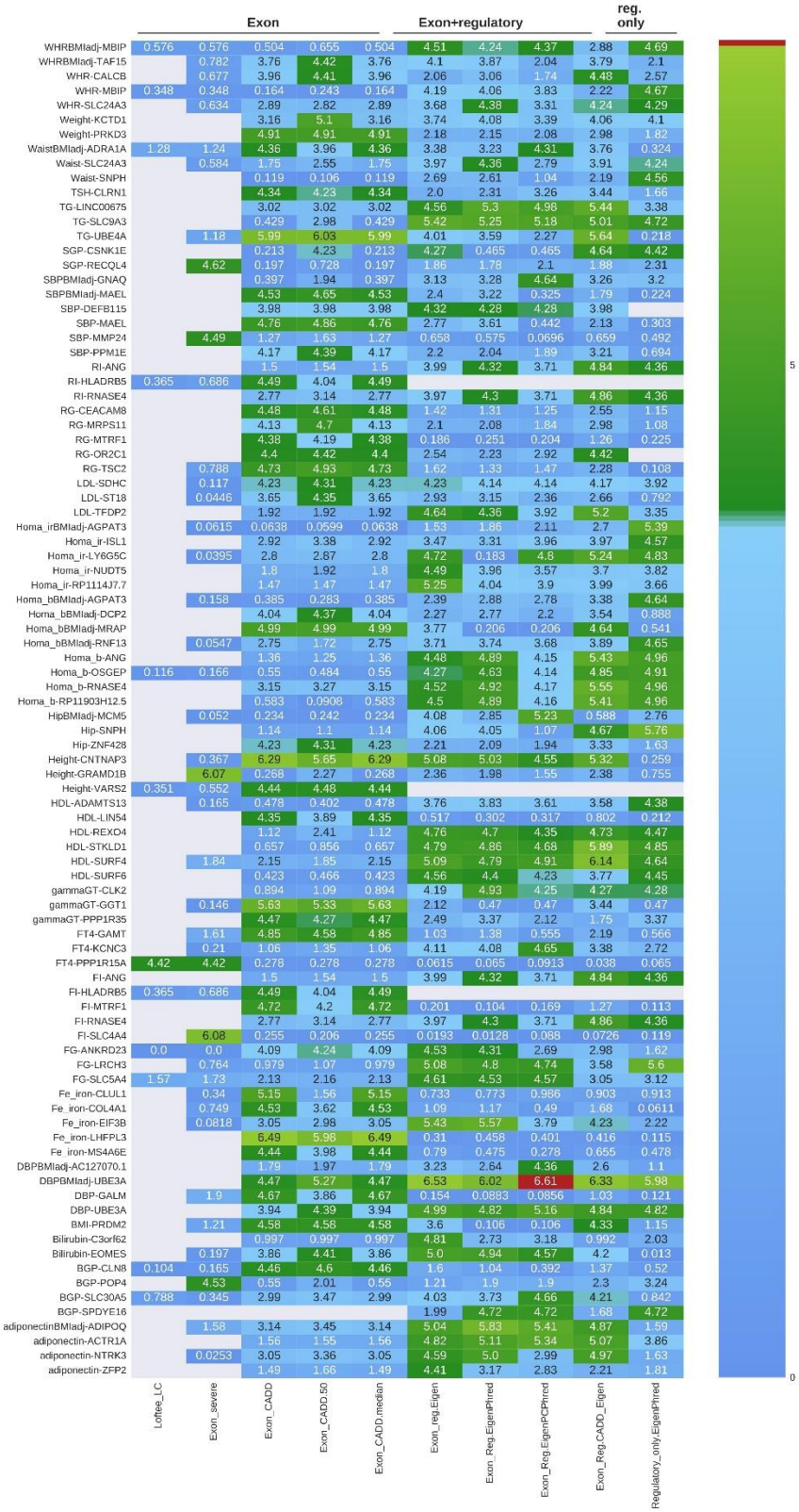



Figure 47: Suggestively significant non-haematological burden signals in MANOLIS.

Of particular interest is also an association between variants in *ISL1* and HOMA IR, a score calculated from insulin and glucose measurements that reflects insulin resistance. The burden is most significant in the regulatory-only analysis. *ISL1* activates a promoter of the insulin gene[144], is essential for pancreatic beta cell function[145], and has been associated with maturity-onset diabetes of the young and type-II diabetes. As such, it is a particularly good candidate for association with a proxy variable for insulin resistance such as HOMA ir. Conditional analysis attenuates the burden one order of magnitude to $1.6 \times 10^{-4}$, and no variant included in this burden has a p-value lower than $10^{-3}$, suggesting genuine contribution of several variants. It is most strongly associated in the regulatory-only run, and includes mainly UTR variants and up/downstream SNVs that overlap a promoter spanning the first exon and a transcription factor binding site intersecting the last exon. The burden is attenuated between one and two orders of magnitude in the exon and regulatory analysis, suggesting that this is a regulatory only signal that is dampened by exonic variants. No single point peak is present in the region. Due to the absence of insulin measurements and fasting information in INTERVAL, we are unable to replicate this signal.

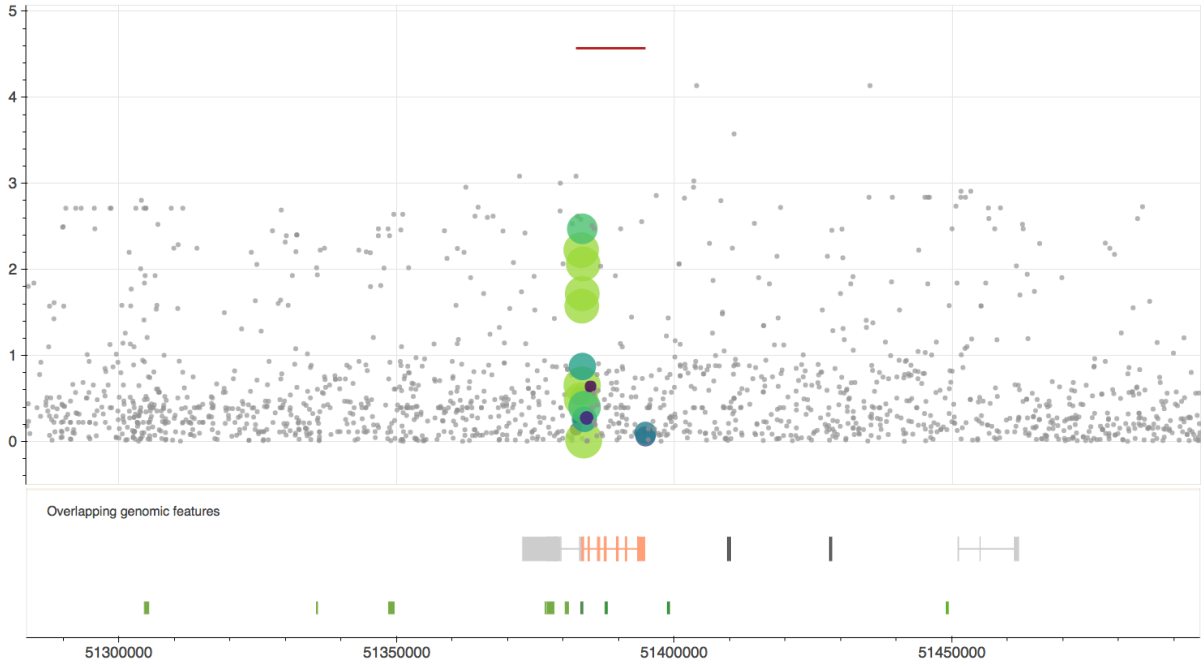



Figure 48: Burden of rare and low-frequency variants in the *ISL1* gene suggestively associated with HOMA insulin resistance.

We also note that similar to what was observed for the *USP40* association with bilirubin, which was driven by the nearby *UGT1Axx* signal, some signals can exhibit extended

association to nearby genes. One example is a signal in *RECQL4*, which is associated with levels of alanine aminotransferase (SGP/ALT). This signal, which arises in the exonic run filtered for severe variants, is driven by variants in linkage disequilibrium with a strong single-point signal in the nearby *GPT* gene, which codes for the SGP/ALT protein. Caution is therefore required when interpreting the results of burden tests, especially in gene-dense regions or when other signals are present nearby.

#### *6.3.4.2.Pomak*

Following the same protocol as for MANOLIS signals, only one non-haematological signal remains at study-wide significance ($p=1.9x10^{-7}$) after conditional testing, an association between variant in the *HBB* gene and bilirubin. The haemoglobin beta locus is one of the most well-documented loci influencing haemoglobin levels and the catabolic products of its degradation including bilirubin. The *HBB* locus is located in an extended LD region in the Pomak cohort, which harbours strong single-point associations between rare SNPs linked through long-range LD. We are currently deconvoluting the allelic architecture and LD structure at this complex locus, and this effort is not part of my thesis work.

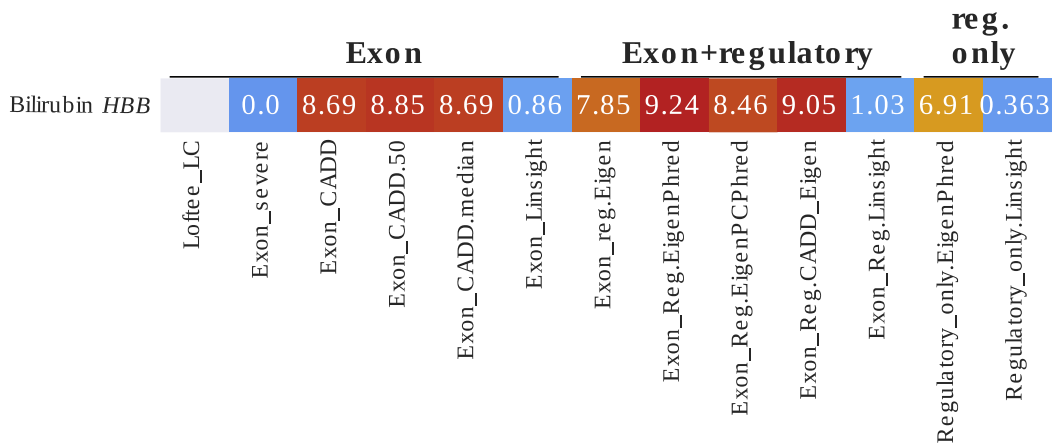



Figure 49: Study-wide significant burden(s) in Pomak for non-haematological traits.

### Suggestively significant signals

Similar to what was performed for MANOLIS, we also report 81 suggestively significant burdens that do not arise exclusively in Linsight or CADD/Eigen mixed runs for non-

haematological traits. Further excluding the *UGT1Axx* burden for Bilirubin yields 75 suggestively significant signals.

We highlight two remarkable associations:

- Variants in the *LTBP1* gene are associated with the HOMA variables for both beta cell function and insulin resistance, as well as insulin resistance adjusted for BMI. This analysis arises in the loss-of-function analysis only. *LTBP1* codes for the latent TGF-$\beta$1 binding protein, which has various interactions with TGF-beta, whose signalling pathway shows tight association with diabetes incidence and progression. In particular, *LTBP1* is thought to play a role in diabetic nephropathy, although the exact role and mechanism remain unclear.
- A burden of variants in the *ULK1* gene is associated with thyroid hormone levels (TSH) in the exonic weighted analyses ($p=4.46x10^{-5}$). This gene is involved in autophagy, which TSH can induce, and *ULK1* has been documented to play a role in TSH-induced autophagy. This gene encodes an enzyme, the Unc-51 like autophagy activating kinase, which is a key regulator of cell growth that senses nutrient status of the cell and determines whether it should be directed towards anabolic pathways and cell growth or catabolic pathways such as autophagy. Thyroid hormone induction of mitochondrial activity is coupled to mitophagy via the ROS-AMPK-ULK1 signaling, suggesting a potential feedback mechanism between levels of TSH and the ULK1 enzyme.

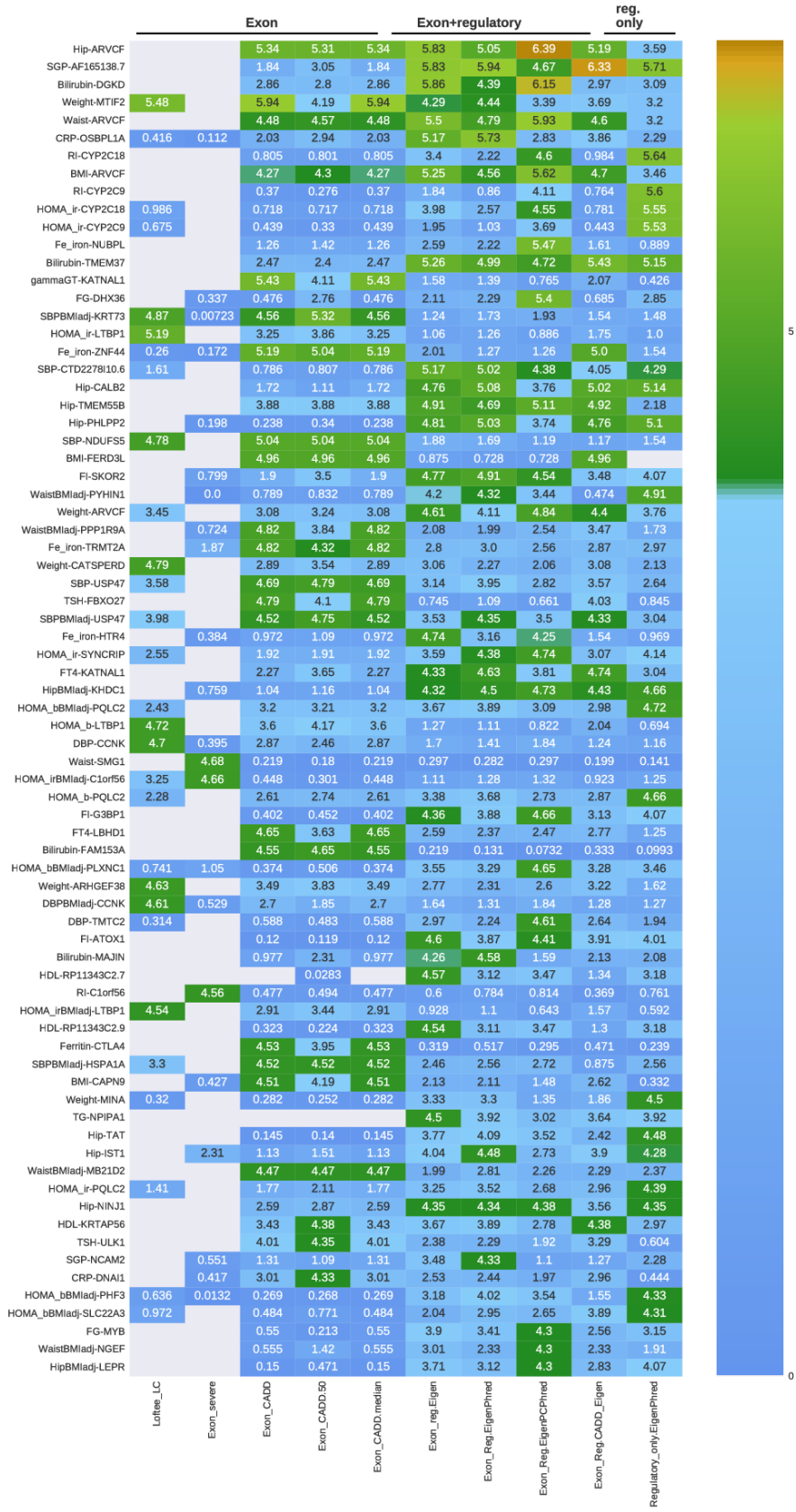



Figure 50: Suggestively significant burden signals in Pomak for non-haematological traits

#### *6.3.4.3.meta-analysis*

Meta-analysis of burden tests is an open issue in human genetics. Previous evidence including our own work on the *APOC3* burden in Chapter 4 suggests that the variants driving an association in one cohort may be different in another. In a meta-analysis context, this means that what appears as a regulatory variant burden might arise as an exonic burden in another cohort.

We therefore apply the approach of performing a p-value-based meta-analysis to the minimum p-value across all burden testing conditions. We apply this to the MANOLIS and Pomak cohort across 50 shared traits. We apply the traditional Stouffer's sum of z-scores:

$$p_{meta} = \Phi\left(\frac{\sum s_i \Phi^{-1}(1-p_i)}{\sqrt{\sum s_i^2}}\right)$$

where, in our case, the weights $s_i$ are the sample sizes in MANOLIS and Pomak, respectively.

We meta-analyse 928,610 trait-gene pairs and use an exploratory threshold of $1.7 \times 10^{-7}$, similar to the single-study threshold. We do not filter single-study results by conditional p-value at this stage, as this would require running conditional for all included trait-gene pairs, which is computationally intractable. We therefore run conditional as a last step, only for suggestively associated signals passing all other post-association criteria.

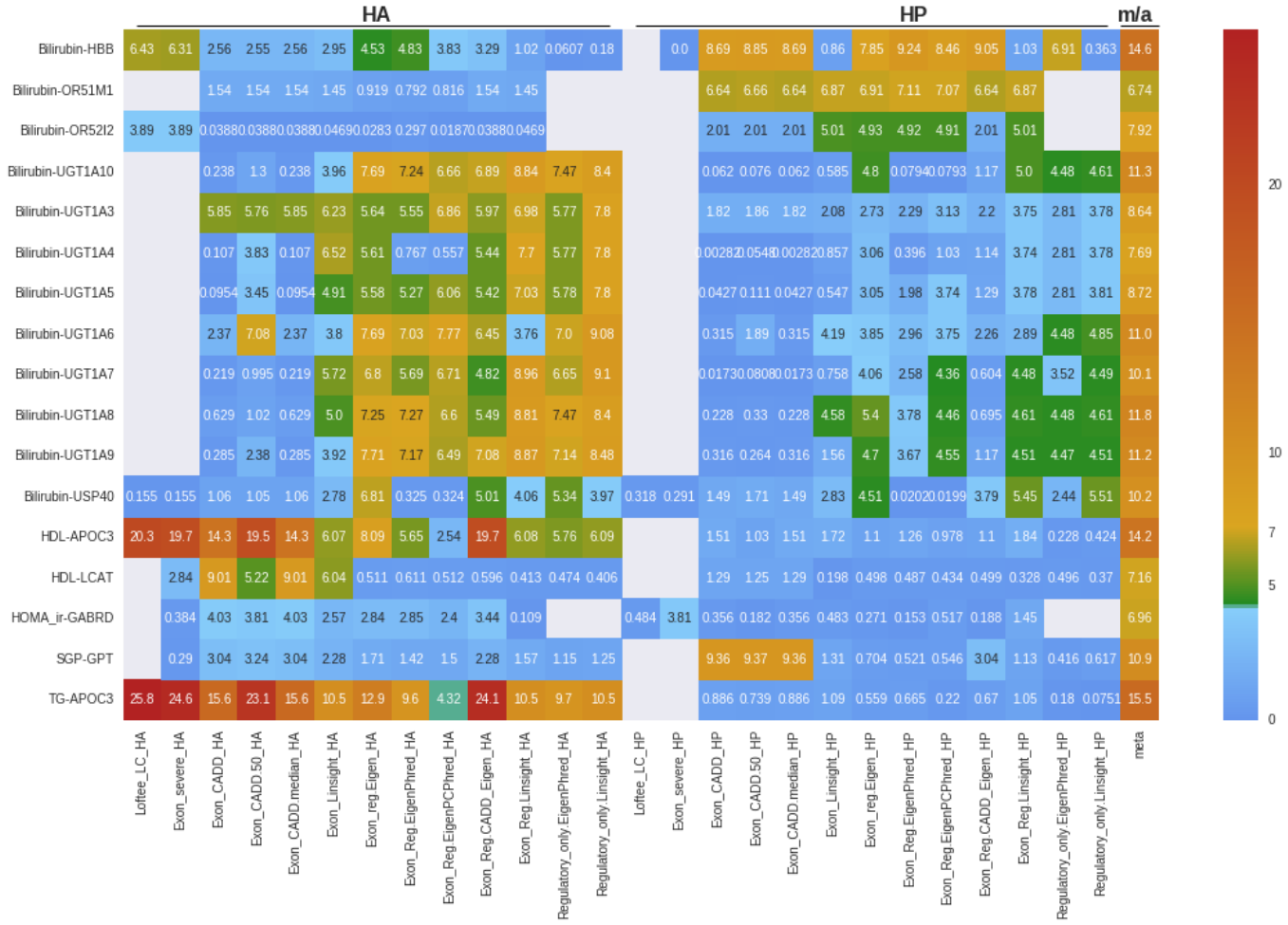



Figure 51: Results of the genome-wide burden meta-analysis, MANOLIS (HA) and Pomak (HP).

17 trait-gene pairs are study-wide significant after meta-analysis. First, a number of signals arise in known loci, some of which are driven by previously described signals arising in one of the two cohorts:

- We confirm the Bilirubin/*UGT1Axx* association previously evidenced in the MANOLIS cohort, this signal recapitulates a common variant association, and the signal extends into the *USP40* gene;
- We also confirm the MANOLIS-driven *APOC3* burden associated with triglycerides and high-density lipoprotein. The meta-analysis is attenuated compared to the single-cohort MANOLIS signal, due to the near-absence of signal in Pomak;

- Variants in *HBB* are associated with Bilirubin, a signal previously excluded from this report in MANOLIS due to it being driven by a single-point association;
- Alanine aminotransferase levels are associated with variants in *GPT,* however the SGP/*GPT* signal is driven by the Pomak cohort, in which it is caused by a single variant;
- The HDL/*LCAT* burden, which was driven by a single variant in MANOLIS, survives meta-analysis despite absence of signal in Pomak;
- Variants in *OR51M1* are associated with Bilirubin, however this association is driven by the Pomak cohort and attenuated by p-values in MANOLIS.

We then examine "true" meta-analysis signals, i.e. p-values that are more significant in the meta-analysis than in either cohort, and check whether they satisfy all previously documented inclusion criteria for burden signals. Two signals fall in this category:

- *OR52I2*, an olfactory receptor gene, is associated with Bilirubin levels. The burden arises for severe variants in MANOLIS and in exonic and regulatory ones in Pomak. *OR52I2* has not been associated with Bilirubin before, the only reported prior association is for response to serotonin reuptake inhibitor drugs in major depressive disorder. The association does not replicate in 2,722 individuals with a bilirubin measurement in INTERVAL, and conditional analysis fails for both MANOLIS and Pomak (p=1 and p=0.37, respectively);
- Variants in *GABRD* are associated with HOMA_ir, the homeostatic model assessment for insulin resistance. *GABRD* encodes the GABA-A receptor delta subunit, and has been associated with severe forms of epilepsy in humans, as well as with behavioural and psychiatric phenotypes in animal models. The signal fails conditional testing in MANOLIS ($p=4.5x10^{-2}$), and is present only in the exon severe run in Pomak where it also fails conditional analysis (p=0.5). The HOMA_ir phenotype is unavailable in INTERVAL due to the absence of insulin measurements.

From the above, it appears that no robust rare variant burden signal emerges from the meta-analysis beyond the ones that were already described in the individual cohorts.

### 6.3.5. Replication in the INTERVAL study

Whole-genome sequencing data of a similar or greater depth, as well as a large sample size, are required for replicating the rare variant burdens described above. We use whole-genome sequencing data in 3,724 samples from the interim release of the INTERVAL project, a cohort of healthy blood donors from the UK. Sequencing and quality control of the INTERVAL data is described in Chapter 2.

#### *6.3.5.1. Defining replication*

There is currently no gold standard method for replicating burdens of rare variants from whole-genome sequencing. Given that our meta-analysis has shown extensive allelic heterogeneity for associated loci in different cohorts, we consider replication across all possible testing conditions, not only the condition a particular signal arose in. We define replication quite liberally, as having the minimum p-value for all 14 runs (including Linsight-weighted ones) in INTERVAL below 0.05. We do not consider cases where only Linsight-based or CADD/Eigen mixed score runs are the only ones passing the threshold in the replication cohort.

#### *6.3.5.2. Study-wide significant signals*

**MANOLIS**

We replicate the association of a burden of rare coding *APOC3* variants with triglyceride levels in the UK-based INTERVAL cohort. We identify a burden of 25 exonic variants ($P = 3.1 \times 10^{-6}$) driven by rs138326449 and rs187628630, a rare 3' UTR variant (MAF=0.008, single-point $p = 8.9 \times 10^{-4}$, two-variant burden $p = 9.0 \times 10^{-7}$). rs138326449 is the only loss-of-function variant in *APOC3* present ($p = 1.3 \times 10^{-4}$) in this cohort, and is four times rarer ($MAF_{INTERVAL} = 0.003$ vs $MAF_{MANOLIS} = 0.013$).

We replicate evidence for a burden signal at *FAM189B* in the INTERVAL cohort ($p = 9.3 \times 10^{-3}$ in the exonic analysis). This burden includes two stop gained variants with one driving the association: chr1:155250417 (rs749626426, MAC=2, $p = 5.35 \times 10^{-3}$). The two novel splice-region

variants discovered in MANOLIS are not present in either the INTERVAL study or in a compendium of 123,136 exomes and 15,496 whole genomes assembled as part of the gnomAD project [118] (http://gnomad.broadinstitute.org accessed 6th February 2018).

We also find replicating evidence of a burden association in *UGT1A9* in the exon plus regulatory region burden analysis ($p$=1.7x10$^{-45}$), however, like in MANOLIS, this signal is fully recapitulated by a known association between the common rs887829 and bilirubin levels.

Adiponectin levels are not available in INTERVAL, preventing us from attempting to replicate the association we observe for rare variants in the *ADIPOQ* gene.

**Pomak**

The only study-wide significant hit was for Bilirubin in the *HBB* gene. It replicates in the INTERVAL cohort with a nominally significant p-value (p=0.015) in the exon and regulatory run weighted by EigenPC phred-scaled weights. No other run passes nominal significance.

**Meta-analysis**

We are only able to try and replicate the biilirubin/*OR52I2* association from the genome-wide burden meta-analysis, as it is the only signal arising as a genuine combination of effects from both cohorts with an available phenotype in the INTERVAL replication cohort. It does not reach our replication threshold, with a minimum p-value of 0.13 in the exon and regulatory run using EigenPhred weights.

### 6.3.6.Suggestively significant signals

#### *6.3.6.1.MANOLIS*

Among the 106 suggestively significant trait-gene pairs, 20 replicate at p=0.05 in INTERVAL. 6 signals further pass conditional analysis in MANOLIS:

- An exon and regulatory and regulatory-only burden arises for triglyceride levels in the *SLC9A3* gene (Figure 52). In INTERVAL, this signal also replicates at p=9x10$^{-3}$ in the exon and regulatory and regulatory analyses. The signal is complex, as the regulatory

regions extend far beyond the gene itself. The top associated variant in MANOLIS is in strong LD with rs73038962, a common intronic variant which overlaps the *AHRR* gene. This association is not strong enough to constitute a common variant peak. It is in high LD with a number of low-frequency variants driven by the synonymous rs6871053, which at 4.2% has a higher minor allele frequency in MANOLIS than any 1000 Genomes EUR (1.3%) and ExAC subpopulations except ones of African ascent (Non-Finnish European MAF=0.9%). Although there are no triglycerides or lipid signals in the region, variants included in the burden have been associated with ulcerative colitis, fat distribution in HIV, hereditary diarrhea in humans, and with low fat mass in mice.

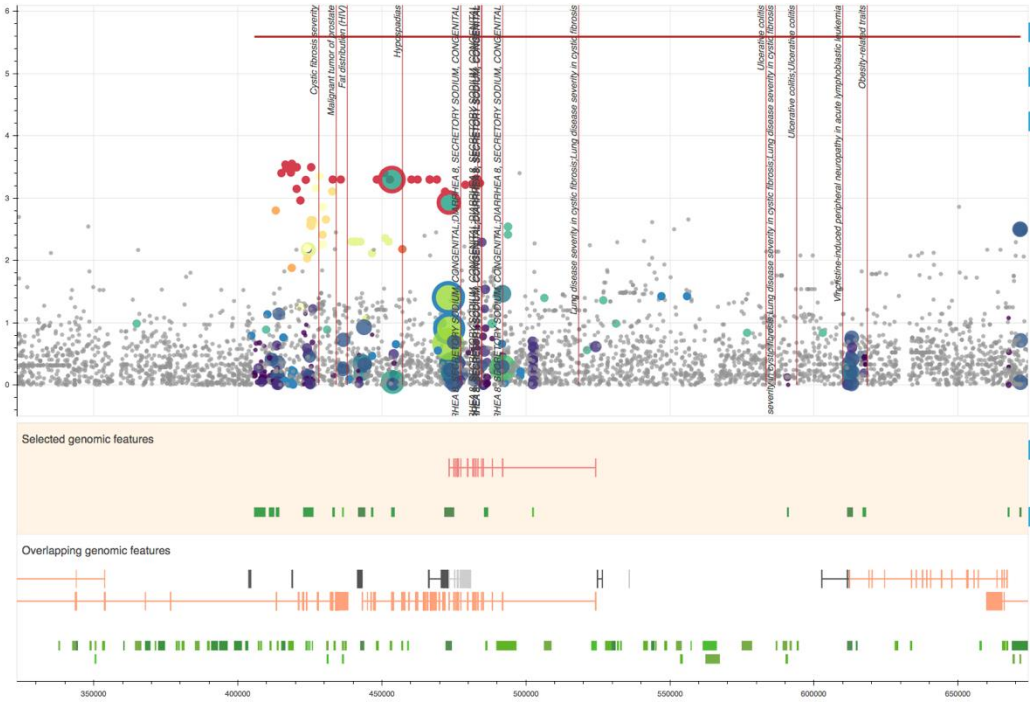



Figure 52 : Rare variant burden in SLC9A3 for triglycerides.

- variants in *UBE4A* are associated with triglyceride levels ($p=9.3x10^{-7}$ in the exonic-only extended run). This signal replicates weakly (p=0.021 and p=0.033 in the exon and regulatory, eigenPhred weighted and exon extended analyses, respectively) in INTERVAL. The burden is strongest in the exonic extended runs, and its main contributors are the novel rare missense variant chr11:118382699 (MAF=0.8%) and a

low-frequency 3'UTR variant rs554845199 (MAF=1.5%). rs554845199 is extremely rare in all 1000 Genomes Phase III populations (max MAF=0.1% in South Asians), as well as in gnomAD (max MAF=0.3% in the Ashkenazi Jewish subpopulation), suggesting that the strong association seen in MANOLIS may be isolate-specific. Both variants have a cardioprotective effect (β=-0.87 and β=-0.56). *UBE4A* codes for a protein involved in multiubiquitin chain assembly and plays a critical role in chromosome

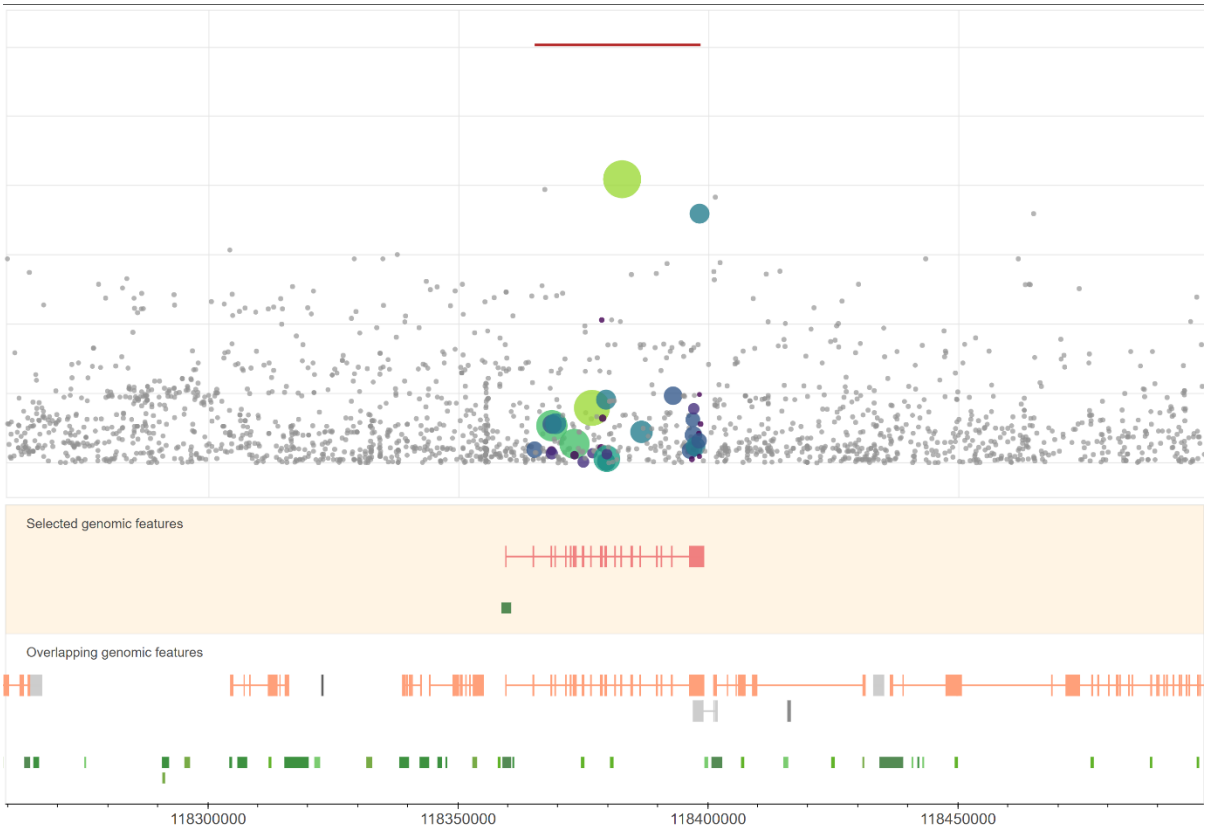



Figure 53 : Rare variant burden in UBE4A for triglycerides.

condensation and separation through the polyubiquitination of securin. Mutations in this gene are associated with various adenocarcinomas, Crohn's disease, fat distribution in HIV, and autosomal dominant deafness. The mechanism by which *UBE4A* could influence triglyceride levels is unknown.

- Variants with severe consequences were associated with height in the *GRAMD1B* gene ($p=8.5x10^{-7}$). This association replicated only in a single condition in the INTERVAL dataset ($p=1.2x10^{-3}$ in the exon and regulatory, EigenPhred weighted run).

It is driven in MANOLIS by two splice region variants, the low-frequency rs2306971 (MAF=4.4%) and the rare (MAF=0.005) rs141347127. rs2306971 is four times more common in MANOLIS than in the EUR subpopulation of 1000 Genomes and non-Finnish European cohort in gnomAD (MAF=0.007 and MAF=0.011, respectively), specifically this variant is absent in all European populations except the Tuscan and Spanish groups. The frequency observed in MANOLIS is closer to that in African, South Asian and East Asian populations. The mechanism by which *GRAMD1B* might influence height is unclear. Its protein product is involved in ectoderm differentiation, and the gene is ubiquitously expressed. rs11219210, a 3'UTR variant in that gene, was previously associated with body height, although not genome-wide significantly. That SNV is not included in the burden, and attenuates the burden p-value by only one order of magnitude. Other previous single-point associations in this gene include: bipolar disorder, autism spectrum disorder, schizophrenia, and several forms of leukemia.


Figure 54 : Rare variant burden in GRAMD1B for height.

- Exonic and regulatory variants were associated with bilirubin levels in *C3orf62,* replicating in INTERVAL regulatory analyses at a minimum p-value of 0.03. This association is particularly strange, as the exonic only and regulatory only runs are both flat, only the combination is suggestively significant. The only difference between the individual and combined runs is the weighting scheme applied to exonic variants (CADD for exonic only analyses, Eigen for exon plus regulatory runs). Given this, we suspect that this signal might be an artefact of the Eigen scoring method. Indeed, comparative analysis reveals that the three flat (p>0.1) missense variants rs150162031, rs61749312 and rs61749309 are upweighted by CADD and strongly downweighted by Eigen. This allows, in the latter case, the suggestively associated 5'UTR ($p=9x10^{-4}$) to drive the burden p-value down to suggestive significance.

- Exonic and regulatory variants in *EIF3B* were associated ($p=2.7x10^{-6}$) with iron levels. This association replicated in INTERVAL both in the exonic only (minimum p=0.005) and the exon and regulatory (weighted by raw Eigen scores, p=0.04). The signal is driven by the rare singleton 5'UTR variant rs995469072 and the low-frequency (MAF=0.022) synonymous rs144904326. The mechanism by which *EIF3B*, which


Figure 55 : Rare variant burden in *EIF3B* for iron levels.

codes for the eukaryotic translation initiation factor 3 (eIF-3) complex, which is required for several steps in the initiation of protein synthesis, could influence iron levels, is unclear. Prior single-point associations in the regions include height, hip circumference and BMI, but also multiple sclerosis and loneliness.

- Variants in the *STKLD1* gene are associated ($1.3x10^{-5}$) with high-density lipoprotein in the exon and regulatory and regulatory-only runs. It replicates weakly in INTERVAL (p=0.02) in the exonic severe run only. rs3758348, rs633862, rs495828, and rs651007 are associated with total cholesterol in the region, rs579459 and rs635634 are further associated with LDL cholesterol. These variants are located upstream of the *ABO* gene some 100kb downstream of *STKLD1*. No prior HDL association is present, and none of these variants are included in the burden. Conditioning the burden on


Figure 56 : Rare variant burden in *STKLD1* for high-density lipoprotein.

any of these variants does not attenuate the burden p-value. No variant in the gene itself were previously associated with any phenotype and the gene is poorly annotated, making a hypothetical mechanism of action on HDL hard to elucidate.

#### *6.3.6.2.Pomak*

In Pomak, 26 signals pass the replication threshold, of which 3 survive conditional analysis in the Pomak cohort:

- variants in the *ZNF44* gene are associated with iron levels in the exonic only analyses ($p=6.4x10^{-6}$). This signal replicates weakly both in the exonic severe (p=0.03) and the exon and regulatory runs (minimum p=0.02) in INTERVAL. The two main drivers of the burden are noticeable by their extreme allelic frequencies: rs758913897 is 163 times more common in MANOLIS (MAF=0.025) than in gnomAD European populations ($MAF=1.53x10^{-4}$), whereas rs747112581 is 233 times more frequent (MAF=0.007 vs $MAF=3x10^{-5}$). Both

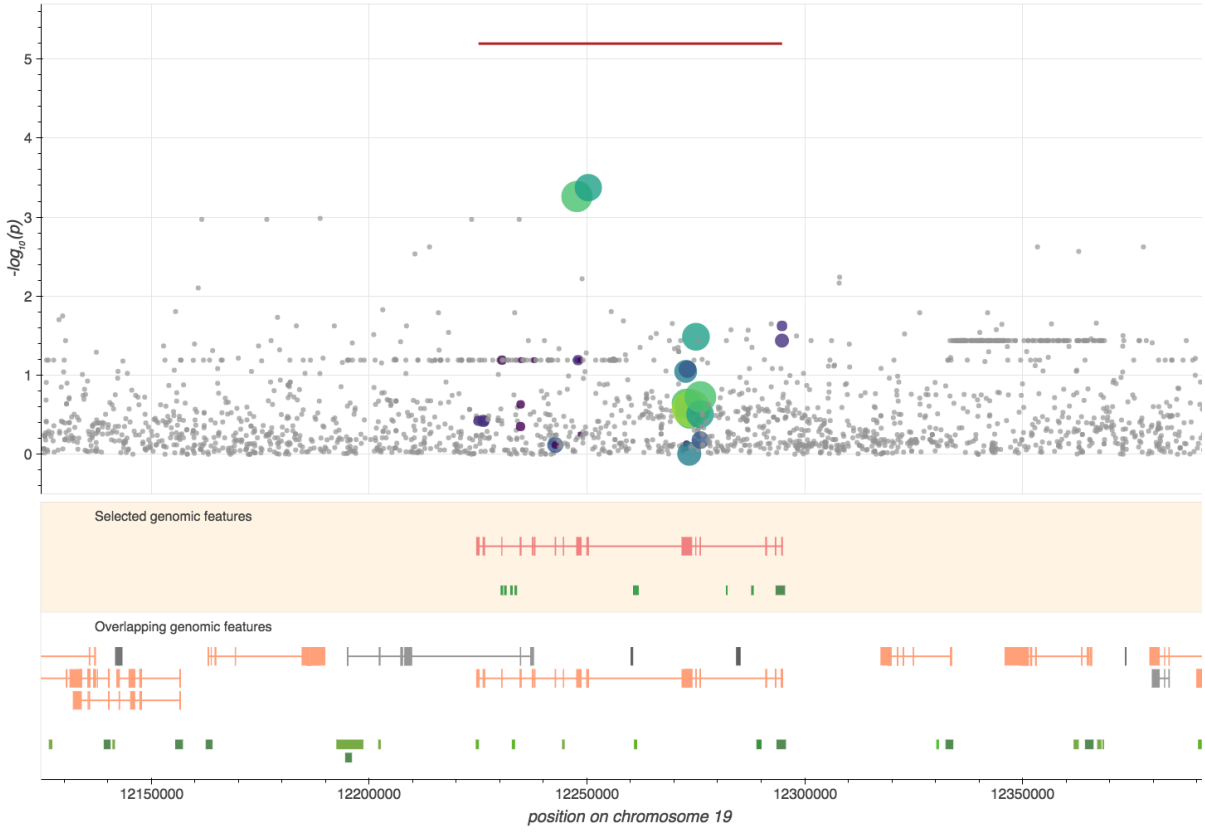



Figure 57 : Rare variant burden in *ZNF44* for iron levels.

variants are part of large LD blocks that extend in both directions into the other zinc-finger protein genes that populate the region. There are no previous associations with quantitative traits in the region.

- variants in the *DGKD* gene are associated with bilirubin in the exon and regulatory runs ($p=1.7x10^{-7}$). Some weak traces of signal ($p\sim10^{-2}$) are present in both the exonic only and regulatory only runs, suggesting either a genuine contribution of both types of variants or an artifact of the Eigen score, similar to what we observed for the *C3orf62* association in MANOLIS. rs1550532, rs6704644, rs1550532, rs6704644, rs1550532, rs6704644 are SNVs with known associations with bilirubin in the immediate vicinity of the gene, but the burden is conditionally independent on their genotypes. If we condition on rs887829 however, which is the lead common variant of the nearby *UGT1Axx* bilirubin signal, we fully attenuate the burden (p=0.09), which indicates non-independence from the common variant signal.

#### *6.3.6.3.Meta-analysis*

We consider meta-analysis signals for replication if the combined p-value is smaller than $5x10^{-5}$, the same threshold we used for the single-cohort analysis. We also discard trait-gene pairs whose meta-analysis p-value is larger than the minimum p-value in the individual cohorts, as this would be a sign of signal attenuation, rather than reinforcement. We also consider only signals for which the trait is available in the replication cohort.

Out of the 187 trait-gene pairs satisfying these conditions, 32 pass replication in INTERVAL at $5x10^{-2}$, of which 26 pass conditional at $5x10^{-2}$ in at least 1 cohort. 4 further signals arise only in runs weighted by Linsight or Cadd+Eigen mixed scores. The smallest replication p-value is 0.0045, which confirms that the generally weak replication evidence already reported for suggestively-significant signals in the single-cohort analyses transposes to meta-analysis.

In all 22 remaining cases, the three-way meta-analysis p-value between MANOLIS, Pomak and INTERVAL is larger than the suggestive threshold of $5x10^{-7}$ (the minimum is $1.1x10^{-6}$).

## 6.4.Genes associated with several traits

We examine whether burden signals are associated with more than one non-haematological trait among the suggestively significant signals in the two single-cohort analyses and the

meta-analysis. In the two single-cohort studies, we restrict analyses to signals passing conditional analyses.

| gene | *ADIPOQ* | *AGPAT3* | *ANG* | *APOA1* | *APOC3* | *HLADRB5* | *KCNC3* | *LY6G5C* | *MAEL* | *MBIP* | *MTRF1* | *RNASE4* | *SLC24A3* | *SNPH* | *TFDP2* | *UBE3A* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **traits** | adiponectin, adiponectinBMIadj | Homa_bBMIadj, Homa_irBMIadj | RI, FI | HDL, TG | HDL, TG | RI, FI | FT4, MID | RI, Homa_ir, FI | SBP, SBPBMIadj | WHR, WHRBMIadj | RG, FI | RI, FI | Waist, WHR | Hip, Waist | LDL, TC | DBP, DBPBMIadj |

Table 8: genes with pleiotropic burden association signals suggestively significant in MANOLIS

We add a suggestive association with adiponectin adjusted for BMI for rare variants in the *ADIPOQ* gene to the one we have already described for unadjusted levels. As expected, we see other cases where adjusted phenotypes are associated along with their unadjusted counterpart: systolic blood pressure in *MAEL,* diastolic blood pressure in *UBE3A*, waist-hip ratio in *MBIP.* LDL is a component of total cholesterol (TC), which may similarly suggest that the association of *TFDP2* with low-density lipoprotein and total cholesterol is a LDL-signal only, of a sufficient magnitude to influence TC levels. Most of the remaining genes in the table above are associated with phenotypes derived from one another: random and fasting insulin for *ANG, HLADRB5, RNASE4*, random and fasting insulin and HOMA insulin resistance for *LY6G5C,* Waist and Waist-Hip ratio for *SLC24A3.*

The association for both triglycerides and high-density lipoprotein in the *APOC3* region, (which is the only one reaching study-wide significance along with the *ADIPOQ*/adiponectin association) has been extensively described in this chapter and Chapter 4.

Five loci remain that are associated with traits not derived from each other: *AGPAT3* for both HOMA measurements adjusted for BMI, *KCNC3* for both free thyroxine and mixed cell count (a haematological measure), *MTRF1* for both fasting glucose and insulin and *SNPH* for hip and waist circumference.

| *gene* | *ADAM30* | *ARVCF* | *C1orf56* | *CCNK* | *CYP2C18* | *CYP2C9* | *DTX1* | *KATNAL1* | *LTBP1* | *PQLC2* | *R3HDM1* | *USP47* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| traits | DBP, SBP | BMI, Hip, Waist, Weight | RI, HOMA_irBMIadj | DBP, DBPBMIadj | RI, HOMA_ir | RI, HOMA_ir | SBP, SBPBMIadj | FT4, gammaGT | HOMA_b, HOMA_ir, HOMA_irBMIadj | HOMA_b, HOMA_bBMIadj, HOMA_ir | Hip, Weight | SBP, SBPBMIadj |

Table 9: genes with pleiotropic burden association signals suggestively significant in Pomak

In Pomak, applying the same broad-ranging criteria, we retain seven pleiotropic burden signals involving traits not derived from each other: *ADAM30* is associated with both systolic and diastolic blood pressure, *ARVCF* and *R3HDM1* with a range of anthropometric measurements, *KATNAL1* with both free thyroxine and gamma-glutamyltransferase, *LTBP1* and *PQLC2* with several adjusted and unadjusted HOMA measurements. Upon closer examination the burden in *ARVCF* only passes conditional in another run than the one in which it is most significant in, suggesting that the bulk of the observed strength of this signal is driven by one variant.

## 6.5.Discussion

Defining the genomic regions in which to select variants, as well as filtering strategies and weighting schemes are unresolved challenges in whole genome sequencing-based rare variant aggregation studies. In this analysis, association signal profiles of tests including regulatory region variants differred markedly from other scenarios, with some signals being driven solely by this variant class. Further, signal strength deviated substantially between analyses that include high-severity consequence exonic variants only, and those in which all exonic variants are weighted according to their predicted consequence. We find that, as a rule, variant and functional unit selection, rather than weighting scheme, plays the largest role in association testing.

We identify a role for rare regulatory variants in the allelic architecture of complex traits, confirming the added value conferred by whole-genome sequencing over exonic approaches. We observe congruent directions of effect among regulatory and coding rare variants in burden signals that combine both classes of variation, for example across eQTL and damaging missense variants in the ADIPOQ gene that are together associated with adiponectin levels.

We replicate the *FAM189B* association in an independent dataset with deep whole genome sequence data, in which the disruptive rare alleles are also associated with the same trait in the same direction. We further replicated all study-wide significant burden signals for which replication cohort trait measurements are available. In replicating burdens, we find that allelic heterogeneity is prevalent, partly due to the rare nature of the variants contributing to the burdens, and partly due to the distinct population genetics characteristics of the discovery and replication sets.

In general, we found replication p-values to be weak in the INTERVAL cohort, not allowing to push discovery and replication meta-analysis below study-wide significance, for any of the suggestively associated burdens studied, either in a single-cohort or meta-analysis setting. These findings have important consequences for defining replication in sequence-based studies of rare variants, and highlight the importance of defining replication at the locus level rather than the variant level for burden signals. In this proof-of-concept analysis, we used 0.05 as a replication p-value, however a more stringent threshold is likely needed, such as one that would correct for the effective number of replication analyses (0.05/3=0.016).

We demonstrate pervasive allelic heterogeneity at complex trait loci, and identify exonic and regulatory rare variant associations at established signals. We find multiple instances of burden signals that remain independent of colocalising common variant signals, and one instance of burden signal attenuation when conditioning on the established common variant association. Within the power constraints of the study, we do not find evidence for synthetic association at established signals, i.e. there is no evidence for multiple rare variants at a locus accounting for a common variant association. The presence of a large number of pleiotropic burden loci in the HELIC cohorts suggests that multi-trait rare variant association studies may be a relevant avenue to boost power in future sequencing-based studies.

# Chapter 7. Large structural variant detection using SNV calls from whole-genome sequencing data

## 7.1. Background

Copy number variants (CNVs) are large deletions or duplications at least 50 to 200 base pairs long. Up to 19.2% of the human genome is susceptible to this type of mutation, which can have a severe functional impact on gene function[146]. They have been implicated in several diseases, including epilepsy[147,148], Crohn's disease[149], schizophrenia[150,151], obesity[152], and autism[153] among others, as well as quantitative traits[154]. Although SNP array-based detection methods exist, whole-genome sequencing (WGS) at high depth has been the golden standard for detecting large polymorphisms. Several methods, including the study of anomalous read pairs, split-mapped reads, de-novo assembly or read depth have been developed[155], each able to optimally detect a certain subtype of variation. Recently, combinatorial methods that leverage two or more methods[156,157] have been proposed. Despite this wealth of methods available, calling structural variants genome-wide has been an ongoing challenge throughout the history of computational genetics. So far, even recent WGS-based structural variant studies are usually made in a limited number of samples or concentrated on targeted regions of the genome[146,158,159]. This is because detecting structural variants requires a different study design compared to association studies: for the latter, haplotype diversity and hence sample size are key[39,160], whereas for the former, high depth of sequencing is paramount, leading to prohibitive costs for population-wide studies. Structural variant detection also poses a computational challenge, since most algorithms use aligned reads or read pileups as a starting point for event detection. As these file formats describe the entire read pool, processing them genome-wide across an entire population is demanding both in terms of running time and memory. In contrast, detecting deletions and insertions from existing small variant callsets demands much less compute effort, as they only probe previously-called single-point variations for CNVs that overlap with them. Such methods were pioneered in the era of genotyping chips (PennCNV [161] and

PlatinumCNV [162]) and are still widely used [159,163], however no corresponding method exists for variant calls produced from whole-genome sequencing [164]. Read depth remains one of the clearest indicators of loss (through deletion) or gain (through duplication) of a genomic segment. Here, we evaluate the effect of copy number variants on sequencing depth in the HELIC data, and provide a proof-of-concept for calling these large variations in population-wide WGS datasets. We then test for association with quantitative traits of medical importance.

## 7.2.Methods

### 7.2.1.Influence of deletion events on read depth at marker sites

Large deletions (500 kbp or more in size) can influence chromosome-wide depth averages (Chapter 5). For example, in the Pomak 18x WGS data, a sample with unusually low average depth on chromosome 11 was first discarded as part of sample-level quality control, but in fact exhibited a genome-wide average depth value close to the population median. Closer examination showed a clear drop in sequencing depth for that sample in a 500kb region located 2Mb upstream of the centromere on chromosome 11 (Figure 58).

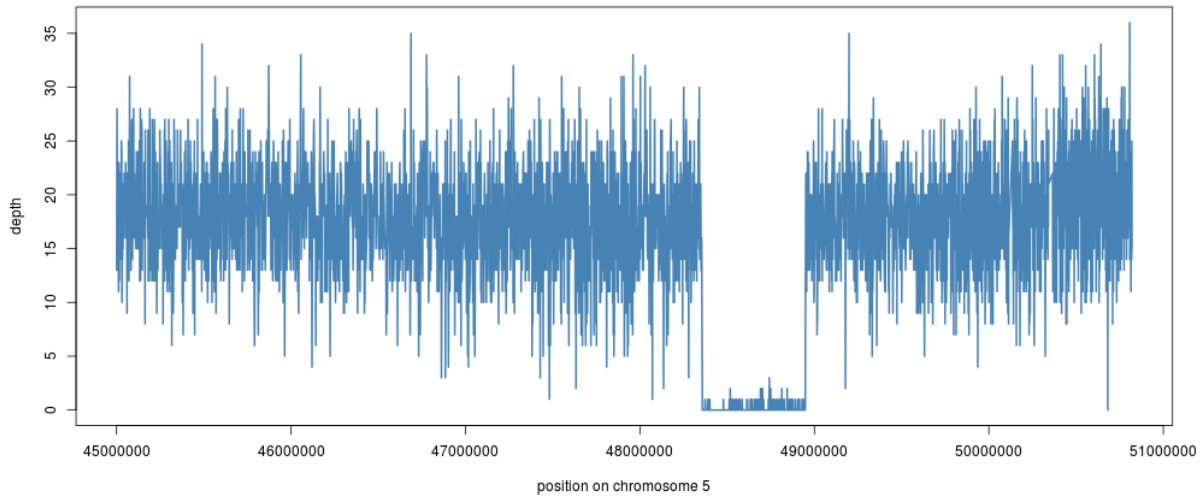



Figure 58 : Read depth at all variant sites for one individual in a 4Mbp region on chromosome 11.

When measured at every variant call, read depth exhibits wide variation even in regions where it is likely to be average. Using a rolling mean of 100 SNVs smoothens the signal by computing local mean depth at every position, at the cost of reduced resolution when pinpointing the exact boundaries of the depth abnormality (Figure 59).

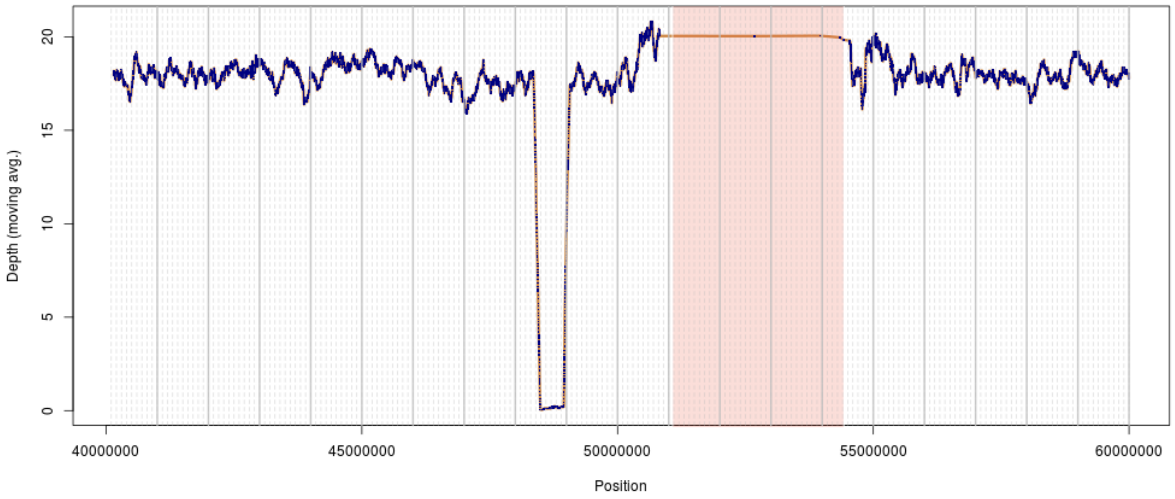



Figure 59: Rolling average depth (100 SNPs) for the same individual in a 20Mbp region on chromosome 11. Blue dots represent averages, with an orange line connecting them. The shaded red area represents the centromere, where only 1 variant is present.

Sudden sequencing drops can be caused by multiple factors, the most likely being a pipeline error. Since variant calling is performed on genomic chunks, a software fault might cause some regions to fail calling in some samples. If this error is not caught in the downstream process, the faulty chunk can be mistakenly incorporated in the full dataset. In the case of the sequencing drop on chromosome 11, the observed boundaries of the event did not correspond to a chunking interval, and this explanation is made even less likely by the presence of low, but non-zero depth at some markers in the region. We compute the chromosome-wide average depth for the affected sample as well as for 79 randomly selected individuals in the cohort, and plot relative depth in this region. If a technical artefact had been the cause of the observed drop, either only this sample (for a software error during calling) or the whole cohort (in case this region is genuinely hard to call, similarly to the centromere) should exhibit low sequencing depth. Instead, we observe a clustering of relative depth around both 1, 0.5 and 0 across the population (Figure 60). This signature is coherent with a copy number variation, with samples carrying a heterozygous deletion clustering around 0.5 and homozygous samples having lost all read coverage in the region.

We further examine whether other potential indicators of a deletion can be found at this locus in the Pomak cohort.

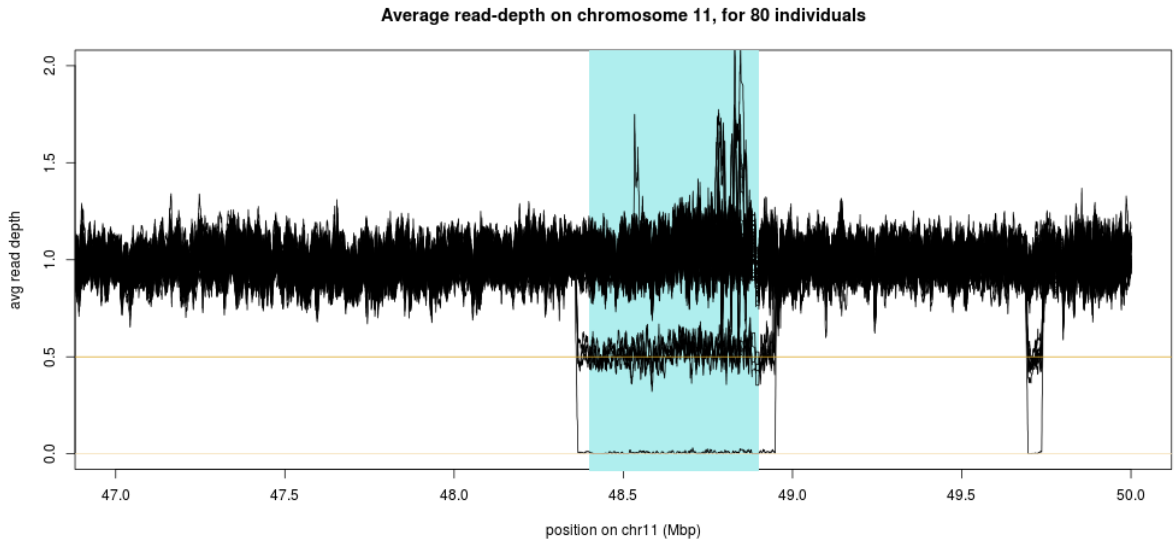



Figure 60: Rolling average read depth for 80 randomly selected individuals in the Pomak cohort. The shaded blue area represents the region of interest.

### 7.2.2.Influence of deletion events on heterozygosity rate and missingness at marker sites

If one copy of a genomic segment is deleted, sequencing depth at that locus will be halved and the reads mapping to that region will only support the remaining copy of the segment. Since single-nucleotide variant callers only marginally account for the genomic context around the variation being called, they should call SNVs as homozygotes for all heterozygote deletion carriers since all reads support a single allele at these variant positions. As a consequence, the number of heterozygote calls should be reduced in a deleted region, as the only source of heterozygosity will be calling error (which is also expected to rise in response to reduced read depth).

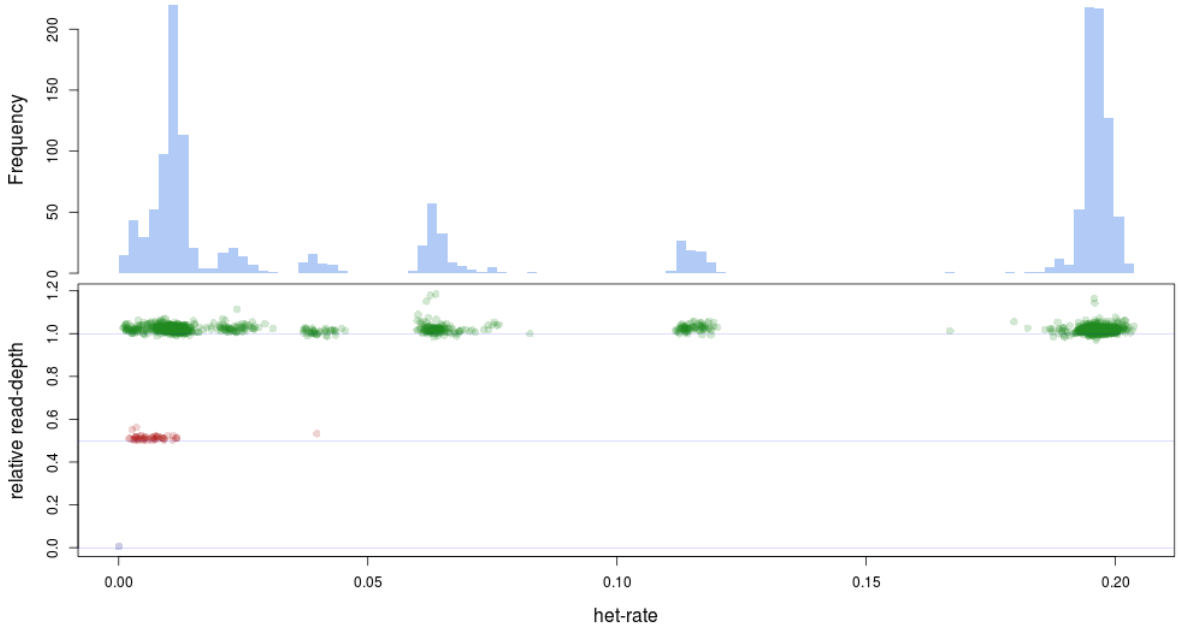



Figure 61 : Heterozygosity rate as a function of read depth(bottom), and as a histogram (top) in the region of interest. Each dot represents a sample.

At the chromosome 11 deletion locus, no carriers of the deletion also carry a high number of heterozygotes (heterozygosity rate >5%) across the 5,650 variants present in the region.

Five well-defined heterozygosity rate clusters are present in the cohort at this locus, which are likely due to reduced haplotype diversity. The individual exhibiting zero average depth in this region also has a null het rate. However, just as read depth was not uniformly constant around 0.5 and 0, heterozygosity rate is not strictly null for suspected heterozygous deletion carriers. In particular, although most values for carriers are within $[0,0.01]$, as much as 3.9% of all SNPs are heterozygotes in one particular sample. That sample exhibits an average read depth of 0.53. This suggests that although high heterozygosity rates may be indicative of low quality deletion calls, excluding potential deletions based on non-zero heterozygosity as part of a systematic QC procedure is probably too conservative.

Missingness rate, on the other hand, should be a good indicator of homozygous deletions: SNV callers are less likely to call a genotype at a locus for a sample if only a small fraction of the expected number of reads is present. This is verified in the case of the chromosome 11 locus, where the individual exhibiting close to zero depth also has a missingness rate of 91% across the 5,650 variants overlapping the sequencing drop. In the case of suspected heterozygous deletion carriers, where depth drops by half, we observe a higher average missingness across the same sites, although it remains in the expected range for samples without any depth anomaly (Figure 62). Interestingly, several of these samples had relatively high missingness (26%), whereas the single suspected heterozygote individual with unusually high heterozygosity rate had a missingness rate of 0.8%.

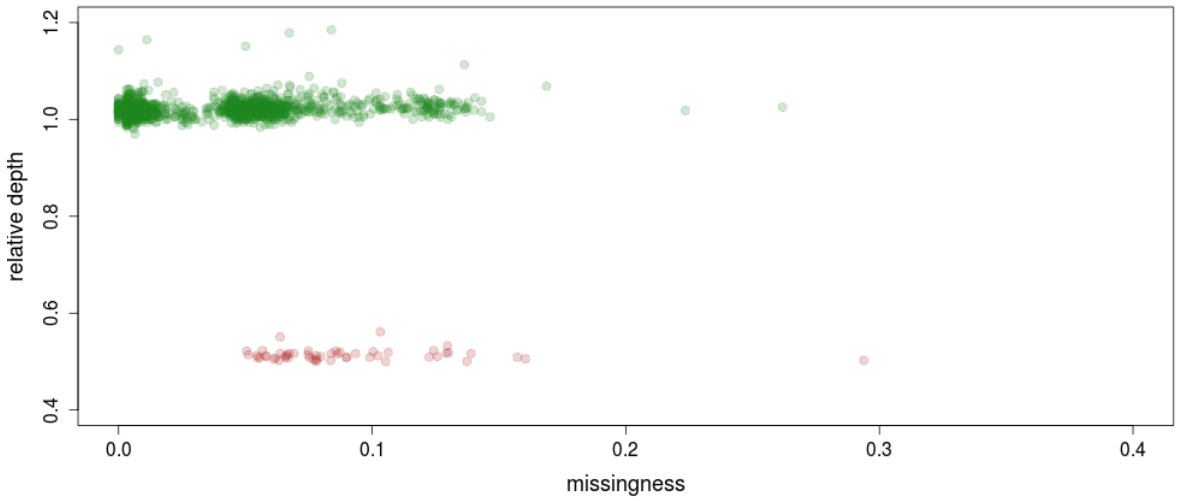



Figure 62 : Average missingness for all variants in the region of iterest, per sample, as a function of relative depth.

In summary, although both missingness and heterozygosity rates do follow trends that are expected if the observed drop in sequencing depth was caused by a deletion, normalised read counts paint a much clearer picture of such events. Similarly, it is unclear how robustly missingness or heterozygosity rates could be used as a quality metric for deletions called using sequencing depth, especially in the case of heterozygous deletion carriers. Based on these observations at the chromosome 11 locus in the Pomak population, I generalise this approach by writing UN-CNVc, the Unimaginatively-Named Copy Number Variant caller, which is based on normalised sequencing depth calculated at a finite number of SNVs.

### 7.2.3.Systematic copy number variation calling using marker-level depth data

The depth signal exhibits high amounts of noise (Figure 58), and rolling averages can be used to smoothen it. Even after normalisation and averaging however, assigning deletion carrier status to each individual based on depth is non-trivial as the signal continues to exhibit high variance (Figure 60). Systematic deletion calling and genotyping therefore requires, first, to smoothen the read depth signal even further, for each individual separately; second, to collate these traces across the cohort to call regions of interest; and finally, to genotype each individual on the basis of read depth.

#### *7.2.3.1.Downsampling the read depth signal*

For a single sample, read depth across the genome can be modelled as a sum of a random noise function and several indicator functions:

$$\hat{d}(x) = \sum_{k} k.\mathbb{1}_{d(x)=k}(x) \ + \epsilon$$

where $d(x) = 0.5n$ is the ideal relative depth at position $x$, $n$ is the copy number at this position, $\mathbb{1}_{d(x)=k}(x) = \begin{cases} 1 \ if \ d(x) = k \\ 0 \ otherwise \end{cases}$ is the indicator function for copy number $k$ genome-wide and $\epsilon \sim \mathcal{N}(0, \sigma)$ is the error in estimating true read counts.

This model will not account for variations in depth caused by variations in GC or repeat content. Average depth is also expected to vary dramatically within regions where the reference sequence is uncertain or masked altogether, such as assembly exceptions and centromeric or telomeric regions. Apart from these cases however, read depth should remain approximatively constant across long stretches of the reference sequence.

The problem of finding a strong enough downsampling of the depth signal to get a clean view of the deletions is difficult both to formalise and to solve. In contrast, fitting a piecewise constant function over an arbitrarily long sequence is a well-studied problem, and falls within

the scope of piecewise constant regression, which is commonly achieved using regression trees.

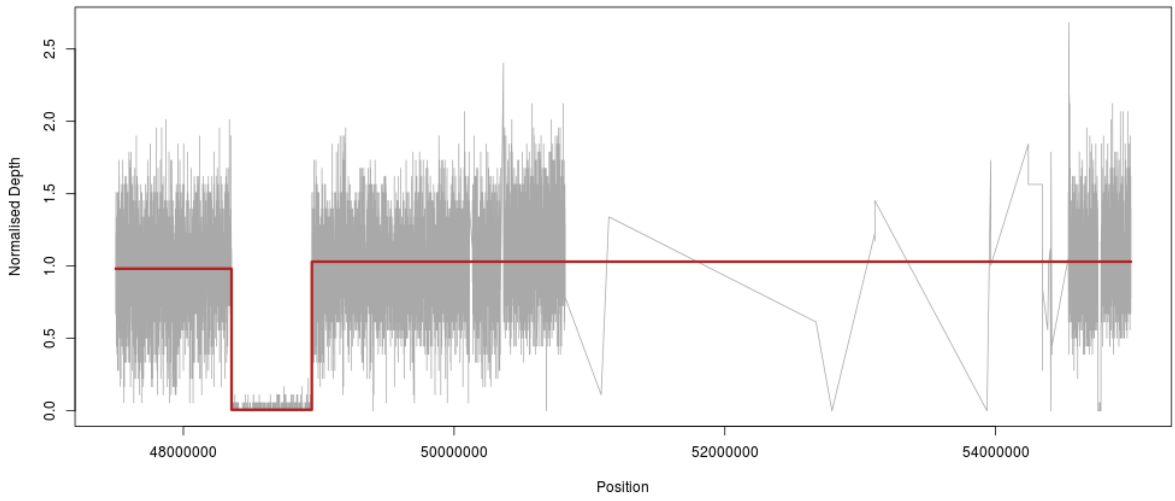



Figure 63 : Raw depth signal (grey) and superimposed piecewise constant regression (red) for the individual carrying the signature of a homozygous deletion.

It is important to note that the model provided by piecewise constant regression is slightly different from the one we previously described, as it does not incorporate our knowledge that $d(x)$ and $k$ take values that are $0.5n$. Instead it tries to find the set of indicator functions $\mathbb{1}_{d(x)=z}$ that best describe the data, where $z$ is an arbitrary real number, and where the number of fitted indicator functions is controlled by a bandwidth or complexity parameter *c*. In practice, applying a regression tree over entire chromosomes is impractical, we therefore choose a window (10Mb) in which to compute relative depth and apply the regression tree. The sufficient size of this window will vary depending on the size of the analysed cohort: since this is used to calculate average depth, it should be sufficiently large to allow an accurate estimation of the mean, even if copy number variants are present. Then, for each sample, we apply a regression tree using the popular `rpart` R library and we use values predicted by this model instead of the actual depth.

We use the recommended values of 0.01 for the complexity parameter of the regression tree (the overall goodness-of-fit $r^2$ of the model must increase of at least this value at each

iteration) and 6 for the minimum leaf size of the regression tree. At the variant densities expected in cohort-wide WGS data, these parameters are very restrictive, i.e. they will only fit a model that follows very broad variations of the data, which is appropriate for this use case (Figure 63).

#### *7.2.3.2.calling copy number variants*

Assembling the constant segments of depth across the entire set of samples provides a global picture of broad depth changes in each 10Mb window (Figure 64). Despite an apparent wide diversity of observed depths, the regressed segments cluster around multiples of 0.5 relative depth, as expected if these anomalies indeed corresponded to copy number variants (Figure 65). Hence, sequencing depth paints an accurate, yet complex picture of

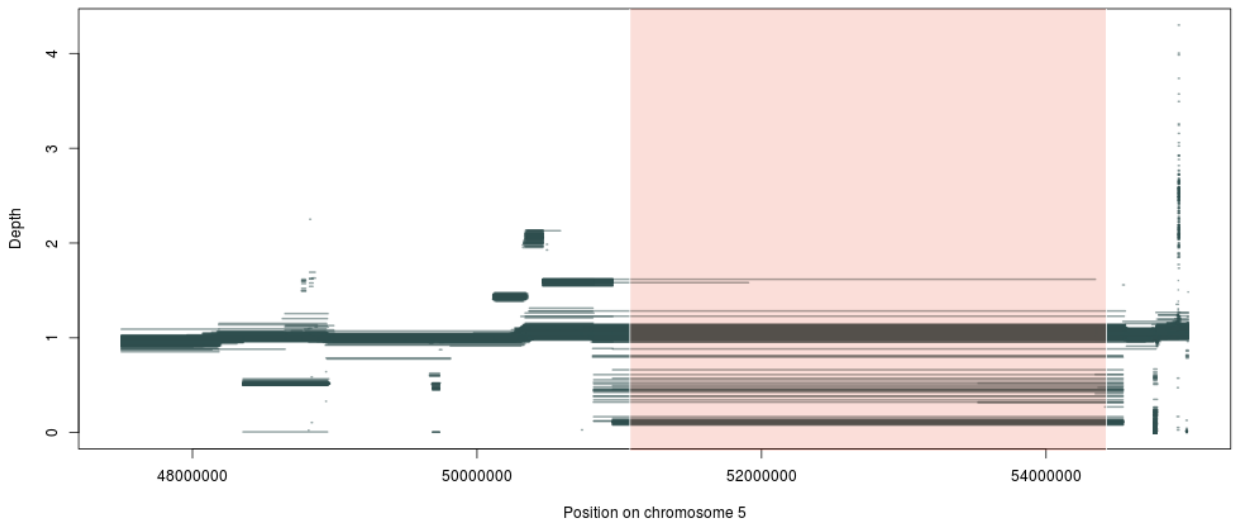



Figure 64 : regressed segments across the Pomak cohort in the centromeric region of chromosome 11.

copy number variants: in particular, some putative deletions (such as the one observed around 48.5Mb on chromosome 11 which led to this work) overlap with smaller duplications of various sizes.

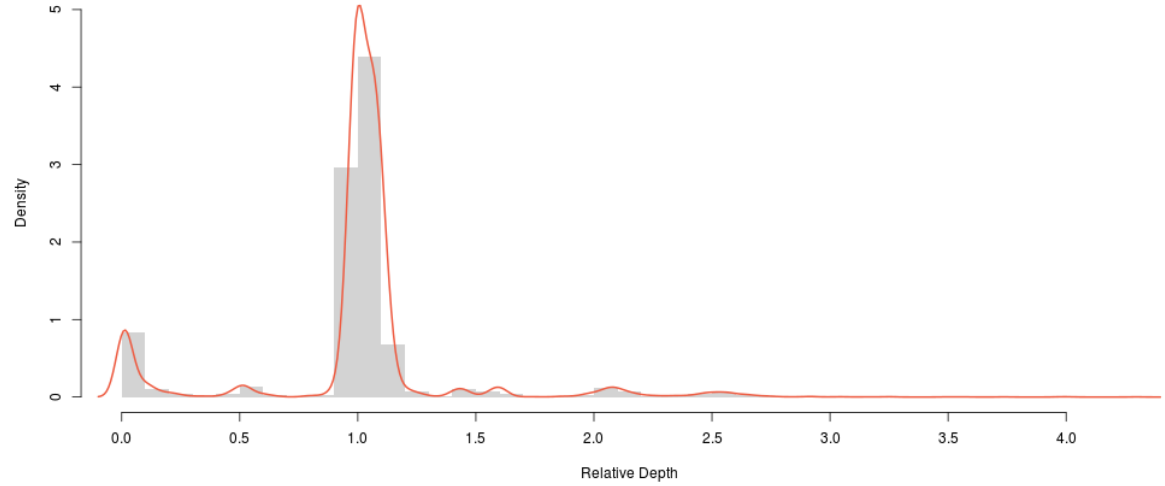



Figure 65 : Histogram of relative depths of regressed segments in the pericentromeric region of chromosome 11.

The algorithm should be able to differentiate between events whose depth signatures correspond to a likely genetic event (the modes of the depth distribution are multiples of 0.5) from those that might indicate an inexact reference sequence.

### Modelling sequencing depth

Read depth at a single base-pair position is often modelled as following a Poisson distribution. In practice, read depths are often over-dispersed compared to a Poisson distribution (the observed variance is larger than the mean), which has led some authors to use the negative binomial distribution instead, mostly in very high depth applications such as RNA-seq data. For moderate depths such as the ones seen in WGS experiments, the Poisson model remains suitable. For an ideal sample of individuals sequenced using a strictly identical protocol and having the same underlying depth, the observed depths at a given site can be seen as realisations of independent and identically distributed Poisson variables with mean and variance $\lambda$.

Here, we are interested in modelling the distribution of segments of relative depth. Segments can be seen as averages of depth measurements over a set of partially correlated SNPs. These variants have different genomic contexts, and variables influencing depths, such as GC content or accessibility, may vary slightly between them.

A scaled Poisson distribution, while obviously not Poisson-distributed, follows a Poisson-like distribution with a PDF of $p(x) = e^{-\lambda}\frac{\lambda^{\frac{x}{c}}}{(\frac{x}{c})!}$.

A sum of independent Poisson variables $X_1, \ldots, X_n$ with parameters $\lambda_1, \ldots, \lambda_n$ is Poisson distributed with parameter $\sum_n \lambda_i$. The average of these variables follows the scaled Poisson described above, where $c = \frac{1}{n}$. Poisson distributions can be approximated by the normal when λ is very large. Although the normal approximation is likely invalid for any of the $X_1, \ldots, X_n$ at usual WGS depths such as the 22x average in MANOLIS, the sum or average, with a parameter $\sum_n \lambda_i$, is closer to the normal approximation.

We hypothesize that, at variant sites, segment depths will cluster around multiples of 0.5 following Gaussian distributions with the same parameters, i.e. that $\sigma$, the variances of the error component of our model, are equal throughout the genome. This is likely to be inaccurate, as the variance of the approximate Gaussian will be proportional to a value close to the average depth in the region. This assumption therefore underestimates dispersion for duplications and overestimates it for deletions.

For each window, we fit a Gaussian mixture model, with means constrained to multiples of 0.5 within the observed depth range at that region, and standard errors constrained to equality across all components of the mixture. The main information source for this model will be the majority of depth segments clustered around a relative depth of 1. For each fitted Gaussian, a p-value can be computed for a depth segment, that represents the probability that this segment originated from each Gaussian component.

### Variant Calling

We then call deletions by discretizing the window in 5kb chunks, and counting for each chunk the number of boundaries of depth segments supporting any depth anomaly. We define a depth segment as supporting a depth anomaly if the component of the fitted Gaussian mixture that produces the maximum two-sided p-value does not have a mean of 1.

To call a region as variable, we use two indicators: the raw count of high-confidence segments supporting a depth anomaly, and the ratio of high- versus low-confidence segments supporting a depth anomaly. The first indicator must be greater than 0, and the second greater than 1, for the region to be called as variable using each calling method, respectively. The first method is much more permissive and is designed to maximise sensitivity, as it just requires one segment to fall within a high-confidence band. For this reason, UN-CNVc uses the second calling method by default. In practice, we found that these two methods disagreed in only a handful of cases. Regions in which these indicators are greater than their trigger limit are called using run-length encoding (RLE), which converts a discrete sequence of values to intervals of constant values. This causes the algorithm to extend regions of interest as much as possible when region boundaries do not exactly overlap, making it more likely to overestimate the length of called CNVs.

*7.2.3.3.genotyping*

**Segment-based genotyping**

Because copy number events can be complex, it is common for a sample to have several segments, and hence several assigned depths per called CNV interval. To produce a single genotype per individual, we compute, for each region of interest, the mean of the assigned depths weighted by the length of each segment, which is rounded to the next multiple of 0.5. Similarly we produce an aggregate score summarising the average quality of the regressed segments for that sample. This allows for the easy application of a quality control (QC) step, whereby genotypes with too high a number of segments, or too low an aggregate quality can be set to missing.

**Means-based genotyping**

The ability of the regression tree to correctly detect drops or increases in depth depends on the number of markers spanned by a CNV, as well as on the complexity parameter: for a constant complexity, smaller events are harder to distinguish from noise, hence harder to detect. At the limit of detection, it is therefore possible that not every carrier sample exhibits abnormal depth segments, leading to correct calling of the presence of a CNV, but false

negative errors in genotyping. To address this issue, we implement means-based genotyping, where each sample gets assigned the multiple of 0.5 that is closest to the average depth across all markers spanning the CNVs called by the regression step. The quality score is the distance between the average and assigned depths. This genotyping method is particularly sensitive to inaccurate boundaries, but it can perform well on smaller events where segment-based genotyping is inaccurate. We therefore also implement a manual genotyper, which applies means-based genotyping on genomic coordinates specified by the user.

We implement a deletion-only mode, where calls are restricted to the deletion part of a CNV even if duplications overlap with a deleted interval, and a duplication mode, which includes duplication genotypes. The deletion-only mode generates PLINK output genotypes which are suitable for association analysis, whereas the duplication mode generates files in the PLINK-specific "cnv" format, which is not currently supported by any association methods.

The method was written in R and bash and is freely available as part of the UN-CNVc package (https://github.com/agilly/un-cnvc).

#### 7.2.4.Quality control

Quality control (QC) is required after calling with UN-CNVc, and the software provides plots and flat files containing quality metrics for manual or automated screening. First, the software produces a plot for each 10Mb region, containing all regressed segments and the variant regions called by the algorithm (Figure 66).

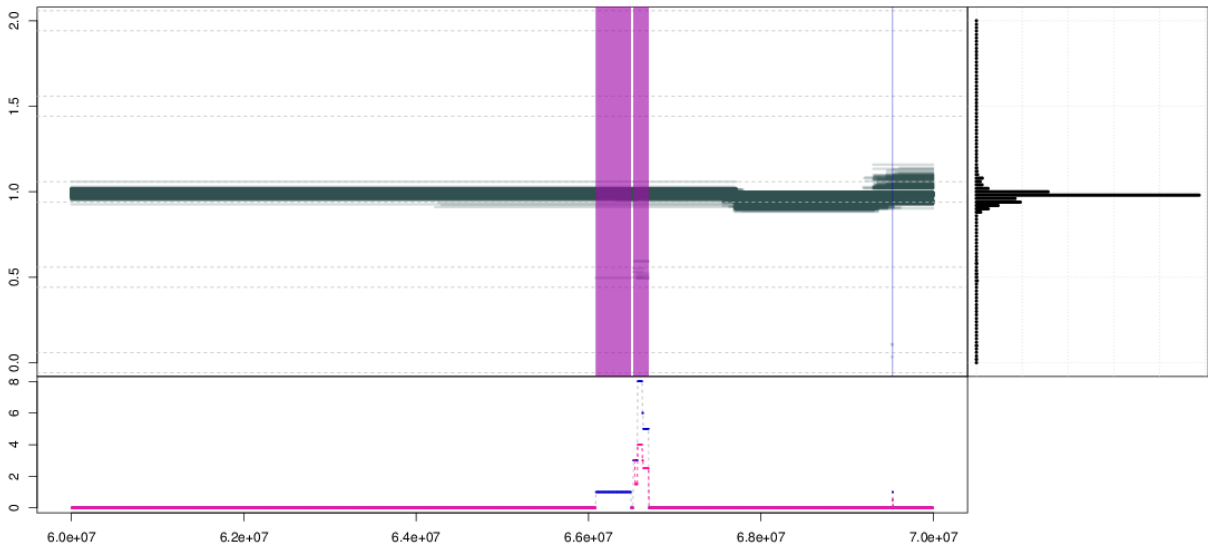



Figure 66 : Segment QC plot generated by UN-CNVc for region chr10:60000000-70000000 in the MANOLIS cohort. The top left panel represents the piecewise constant relative depth intervals. There will be a cluster around 1, representing the normal depth. Horizontal dashed intervals represent expected locations of CNV segments (heterozygote/homozygote deletion or duplication). Vertical highlighted regions are regions called as variants by UN-CNVc. Top right panel is a histogram of the observed segment depths. Bottom left panel represents the RLE statistics used to call variable regions, with the blue and purple lines representing the first and second genotyping method, respectively.

This can be used to diagnose the robustness of a call for variable regions. In Figure 66, the first deletion contains a single heterozygote with extremely high confidence, as the single segment's depth value is almost exactly 0.5. The second deletion is more common and is also solely carried by heterozygotes. However, it is less trustworthy, as some segments are outside the confidence band around 0.5, and the RLE statistics show multiple breakpoints, suggesting a complex region. The last event, in blue, is only called by the high-sensitivity calling method: although it does contain one high-confidence homozygote segment, two are present outside the confidence band, yielding a 2/3 ratio, which is smaller than the required threshold of 1. Another automatically generated plot is used to perform quality control on genotyping accuracy (Figure 67). It provides a more detailed picture of the regressed segments for each event, as well as a rolling average depth, allowing to visualise depth variation at single-SNP resolution. It also compares the two genotyping methods, and

is particularly useful when deciding whether to manually redefine boundaries for small or otherwise hard to call events.

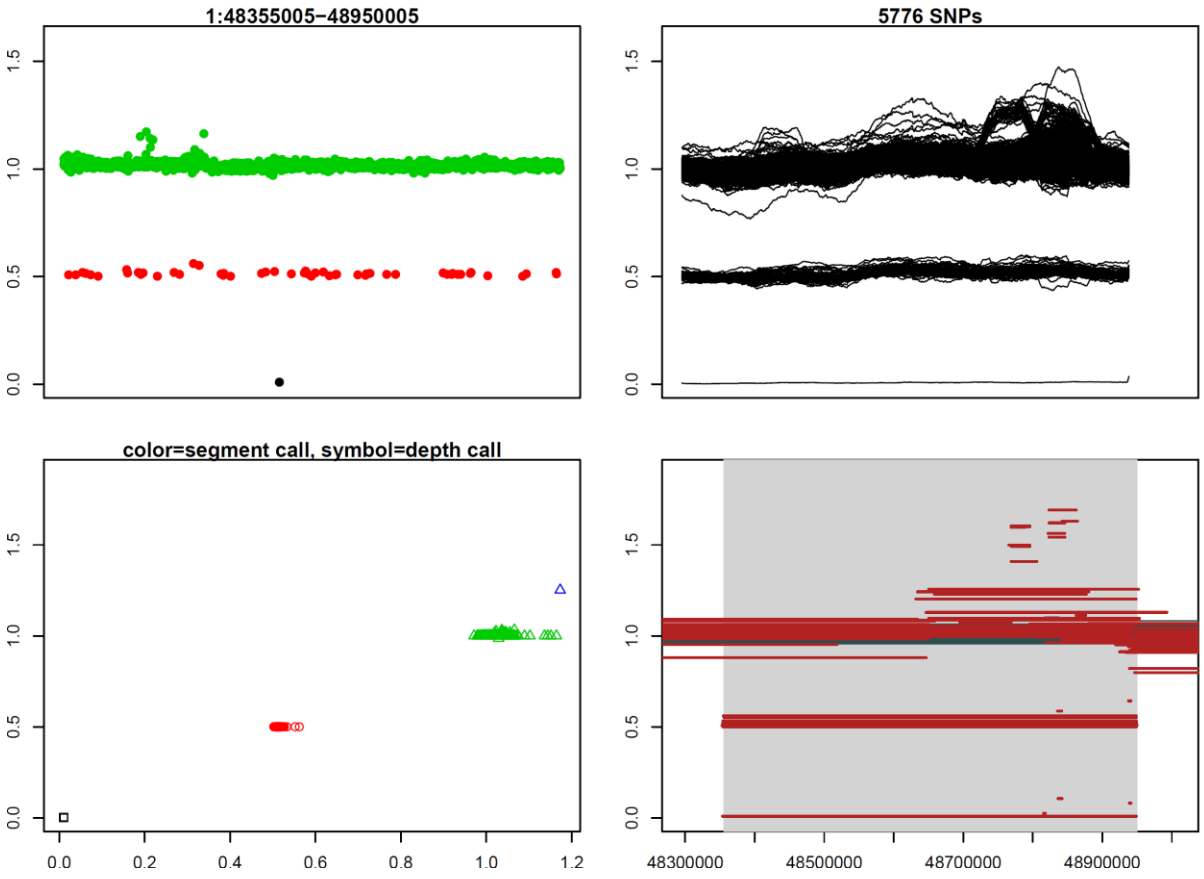



Figure 67 : Genotyping QC plot generated by UN-CNVc. The top left panel represents average depth calls in the region, with depth calls shown as colors. A good graph has clearly defined clusters around multiples of 0.5 that each mostly contain 1 colour. The top right panel shows rolling mean depths, per sample, in the region. Lines should be mostly horizontal (no slopes), and should be clustered around multiples of 0.5. The bottom right panel represents piecewise constant intervals, with samples containing multiple intervals in the region coloured in red. In good quality events, most non-REF intervals are about the same length and largely overlap. The bottom left panel is a diagnostics plot comparing genotyping results from both methods. Ideally, results will be similar, meaning that n clusters of mostly identical colours and shapes should appear on the diagonal of the plot.

In addition to this plot, a file containing various quality metrics, such as the number of segments for each sample in each variable region, the depth of segments, weighted by their length, and the p-value of the call being true. These metrics allow for automatic QC to pre-filter results when a large amount of events is present.

## 7.3. Results

### 7.3.1.Genome-wide calling of deletions in four European population-based cohorts

We ran UN-CNVc on 6,898 samples across four studies: 1,457 and 1,617 samples from both HELIC cohorts (mean depth 22.5x and 18.6x), as well as 100 TEENAGE samples (mean depth 32x) and 3,724 samples from the INTERVAL cohort (18.4x depth). Both these cosmopolitan cohorts were previously used for replication and comparison, and their QC is described in Chapter 2.

QC of the variants was carried out based on the plots and statistics files generated by UN-CNVc. Variants called within the centromeric and telomeric regions were first removed due to their low assembly quality and repeat-rich nature. Then, interval QC flagged variants with multiple breaks within the call regions or vastly different segment boundaries (Figure 68).

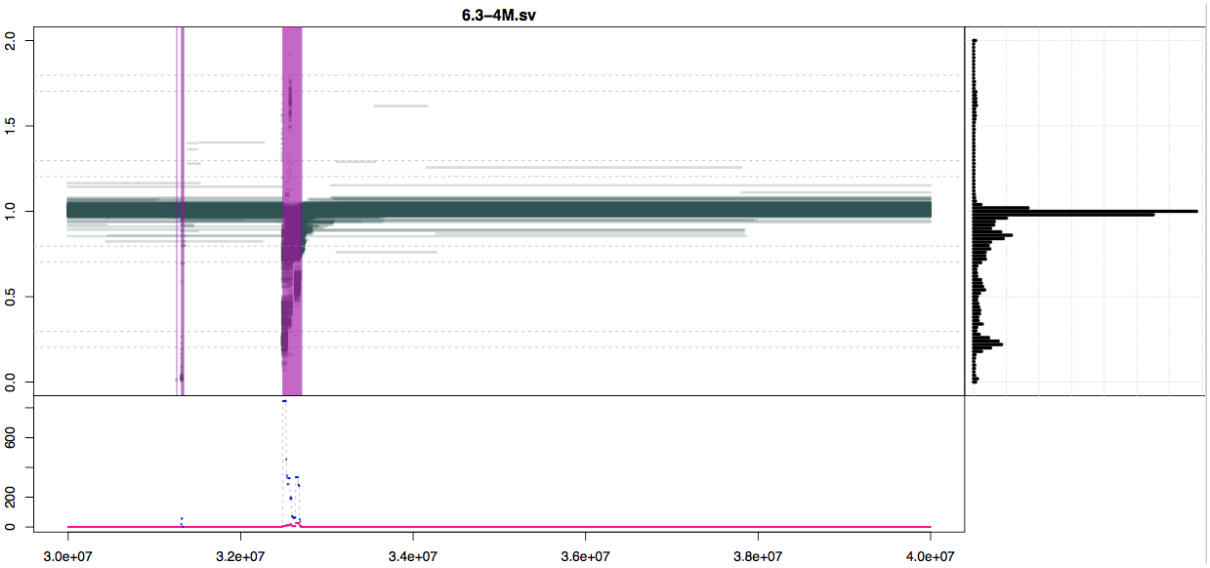



Figure 68 : Example (rightmost on the plot) of a pericentromeric event exhibiting both a complex structure (at least 3 depetions and two duplications are visible) and regressed depths that cluster insufficiently around the expected multiples of 0.5.

Second, genotype QC was performed using the genotype diagnostics plots. For complex events with multiple breakpoints flagged by the previous step, or small events with incorrect genotypes, boundaries were adjusted using the manual genotyper. Interval-based and

genotype quality control for MANOLIS, Pomak, TEENAGE and INTERVAL was undertaken by Grace Png as part of her MPhil dissertation, under my day-to-day supervision.

Using boundary calls from UN-CNVc, we generate a map of the large deletion landscape comprising 1,170 CNVs in four European populations (Figure 69).

This map confirms previous accounts that CNVs are not located uniformly in the genome[146], with some regions being relatively CNV-poor, like the long arm of chromosome 9. A large proportion of CNVs excluded as part of QC were located in pericentromeric and telomeric regions, with depth patterns not indicative of discrete copy number changes. The fact that these depth anomalies largely overlap between all four cohorts is coherent with the hypothesis of imperfectly assembled regions. In general, we found that regions rich in low-complexity sequences, such as segmental duplications and assembly exceptions and patches, harboured a large number of CNVs. In many cases it was possible to disentangle the structure of these regions despite their apparent complexity; in fact the *CCL3* and *NOMO* loci discussed below are both located in regions where an alternative reference is present.

101 (8.6%) of the high-quality CNVs were shared between two or more cohorts, of which 12 CNVs were shared between all four cohorts and 37 between at least three cohorts. The largest overlap was between Pomak and INTERVAL, which share 54 CNVs, followed by MANOLIS and INTERVAL, with 42 CNVs (Figure 70). To make comparisons more meaningful, we applied a 80% reciprocal overlap criterion, which avoids counting as overlapping cases where a large event spans a much smaller one in another cohort.

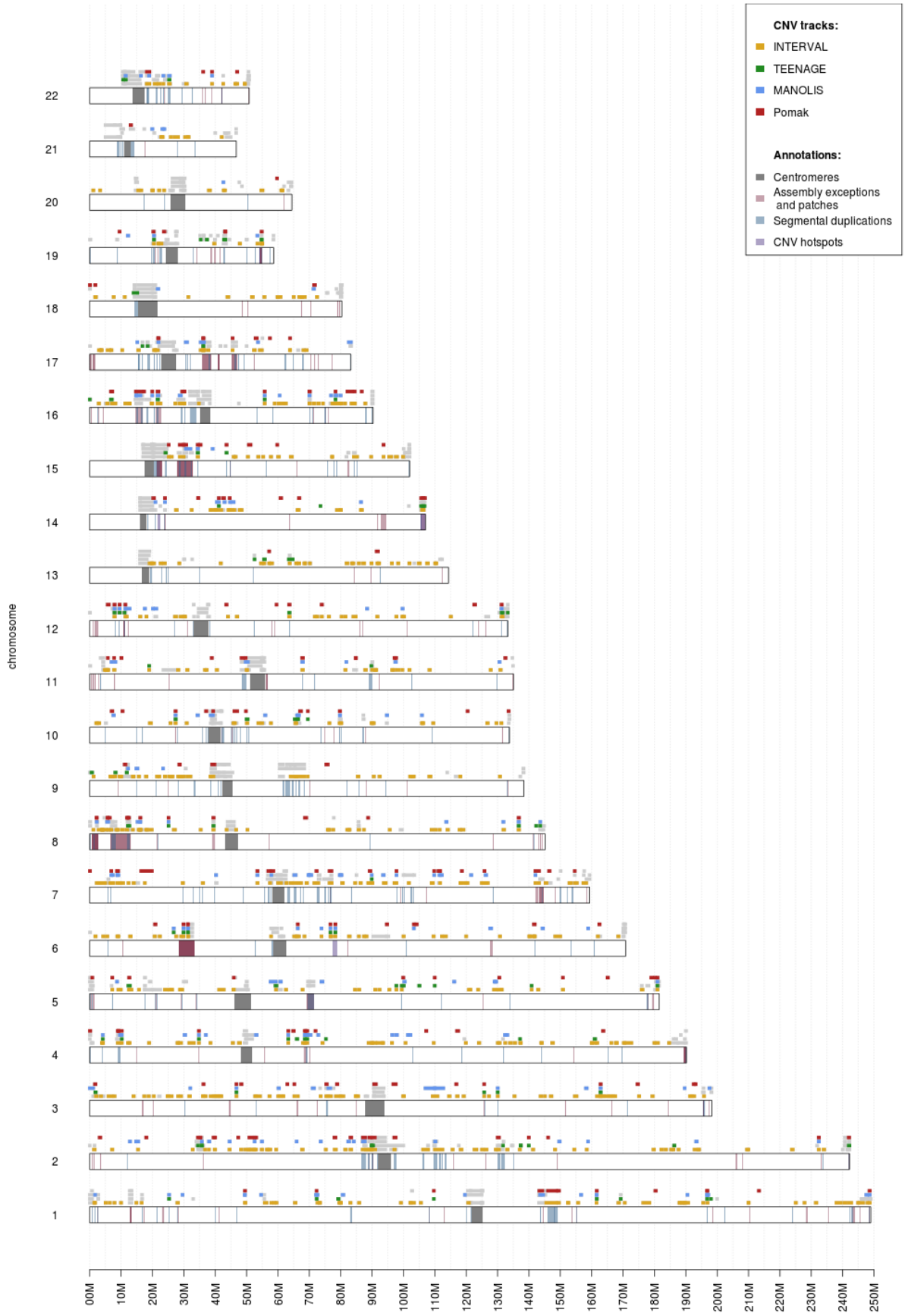



Figure 69 : Genome-wide map of events called using UN-CNVc. Regions in grey in CNV tracks above chromosomes were QC-ed out. Regions marked in pink are assembly exceptions and patches, taken from the GRC data for GRCh38.p12, regions in blue are segmental duplications (from UCSC), and regions in purple are "CNV hotspots", which are known, highly variable regions comprising an intergenic region on chr6q14.1, an olfactory receptor gene cluster (*OR4C11-OR5L2*) on chr11q11, a leukocyte immunoglobulin gene cluster (*LILRB3-LILRB5*) on chr19q13.42, the immunoglobulin $\kappa$, $\lambda$, and heavy chain loci (*IGKC, IGLC1, IGH*), and the T cell receptor alpha locus (*TRA*).

We call an average of 70 deletion alleles per sample in MANOLIS, Pomak and TEENAGE. This number varied massively however, with Pomak having barely a third (n=46) as many deleted segments as TEENAGE (n=134). It is unknown whether this is due to sample size discrepancies influencing the software's sensitivity, or whether it reflects a true population heterogeneity. The latter is more likely, given that a factor of 2 differentiates average CNV copies in MANOLIS (n=92) and Pomak, despite these cohorts having very similar sample sizes(n=1,457 and n=1,618 respectively). Sequencing depth is higher in TEENAGE than in both HELIC cohorts, which may have caused increased sensitivity, as the caller is better powered to differentiate depths when average read coverage is high.

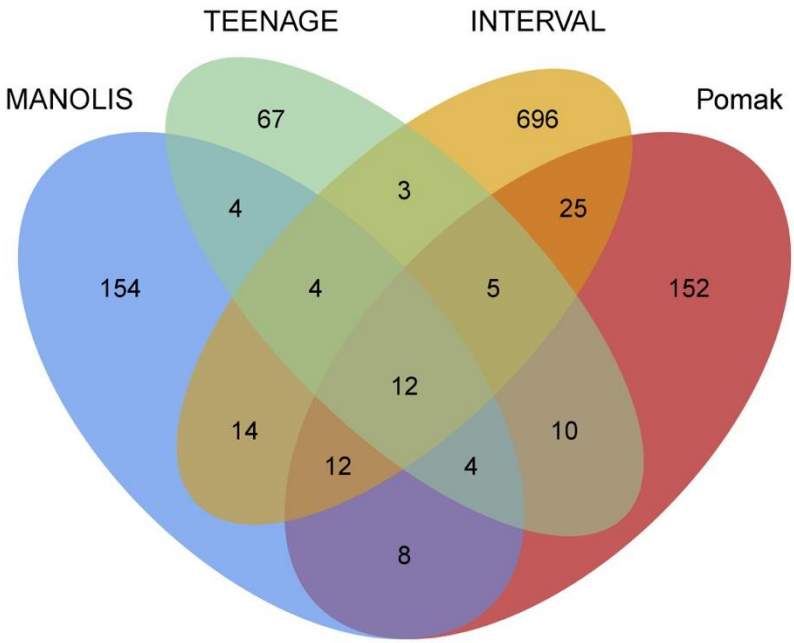



Figure 70 : Overlap between the four cohorts after quality control.

Overlap calculations performed by Grace Png using intersectBed.

### 7.3.2.Performance and comparison with other callers

For all cohorts, the genome was divided into 332 equal-sized 10 Mbp chunks, which were run in parallel, with some chunks empty due to overlap with centromeric regions. UN-CNVc's run time had a power dependency to sample size, between linear and quadratic (Figure 71 a) with the linear model giving 2.4 seconds/sample (the best fit was for a $n^{1.5}$ dependency). On a

cluster with at least 332 nodes, this means UN-CNVC can call CNVs genome-wide on a 1,000-sample cohort in 40 minutes. Peak RAM usage was between a square and a cubic function of the sample number, with approximately 10Gb required for 3,000 samples (Figure 71 b).

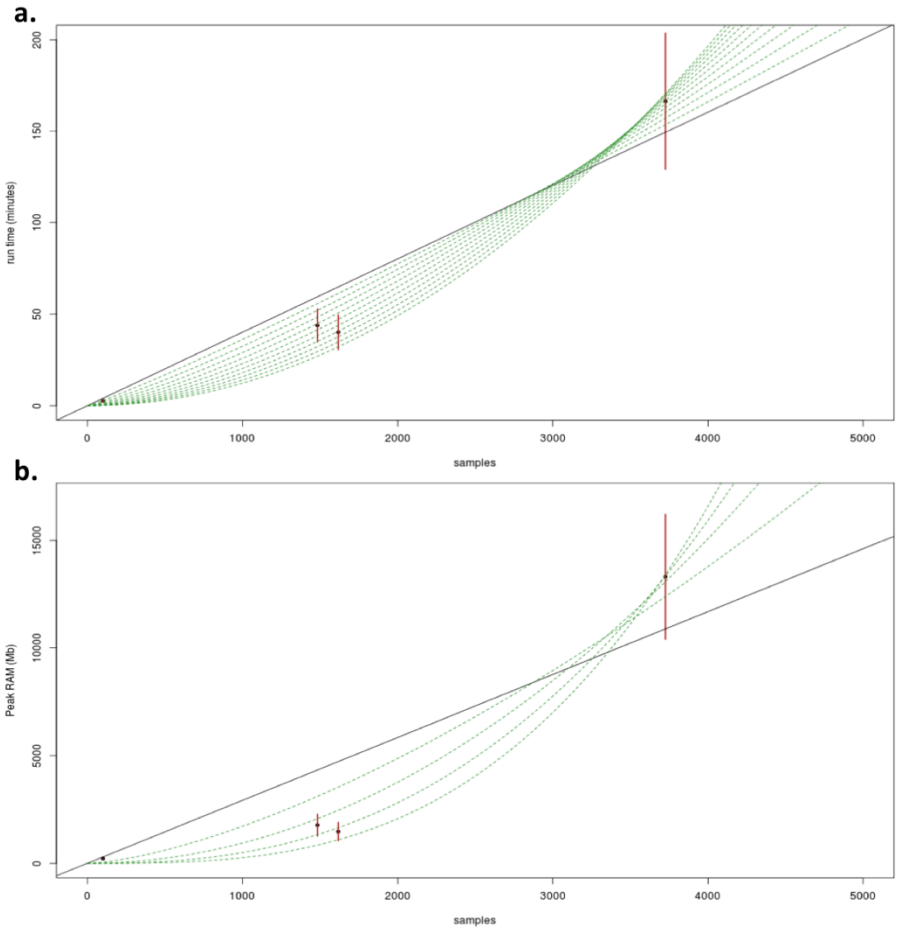



Figure 71 : a. Median runtime as a function of sample size for all four cohorts analysed. The black line is a linear model with null intercept, dashed lines represent the ten consecutive $y = x^{1+0.1n}$ power function models between the linear and quadratic models. b. Peak RAM usage. The dashed lines represent the four $y = x^{1+0.5n}$ between the linear and cubic models. Red intervals represent the mean absolute deviation.

UN-CNVc shares characteristics of both array-based and WGS read-depth based CNV calling methods. We therefore compare UN-CNVc's genome-wide calls with both PennCNV, a commonly used array-based tool, and GenomeStrip, the gold-standard method for WGS-based CNV calling. This comparison is run on 211 MANOLIS samples with both WGS and

CoreExome array data. On this subset, PennCNV took 2 hours to run with 586Mb peak RAM use, and GenomeStrip took 14.5 hours with peak RAM use of 3Gb, compared to 16 minutes and 798Mb for UN-CNVc. For PennCNV, these statistics did not include preprocessing, as it was necessary to recall the array data from intensities in order to gather the necessary b-allele frequencies and log ratios. This was done separately, using the GenomeStudio software from Illumina. PennCNV itself was run by Grace Png. UN-CNVc calls 253 CNVs in deletion-only mode, whereas PennCNV and GenomeStrip call 1,404 and 10,660 deletions, respectively. As expected from its design, UN-CNVc method called on average larger CNVs than the other two methods (Figure 72). Interestingly, the distribution of GenomeStrip CNVs seemed to have both a hard lower bound of 1kb, similar to UN-CNVc, and an upper bound of around 100kb. This is unexpected, given that unlike UN-CNVc, it analyses raw read counts and does not use a discretisation window that might limit its resolution. PennCNV has the widest distribution of sizes, and calls events as small as 3 base pair long which would usually be classified as indels.

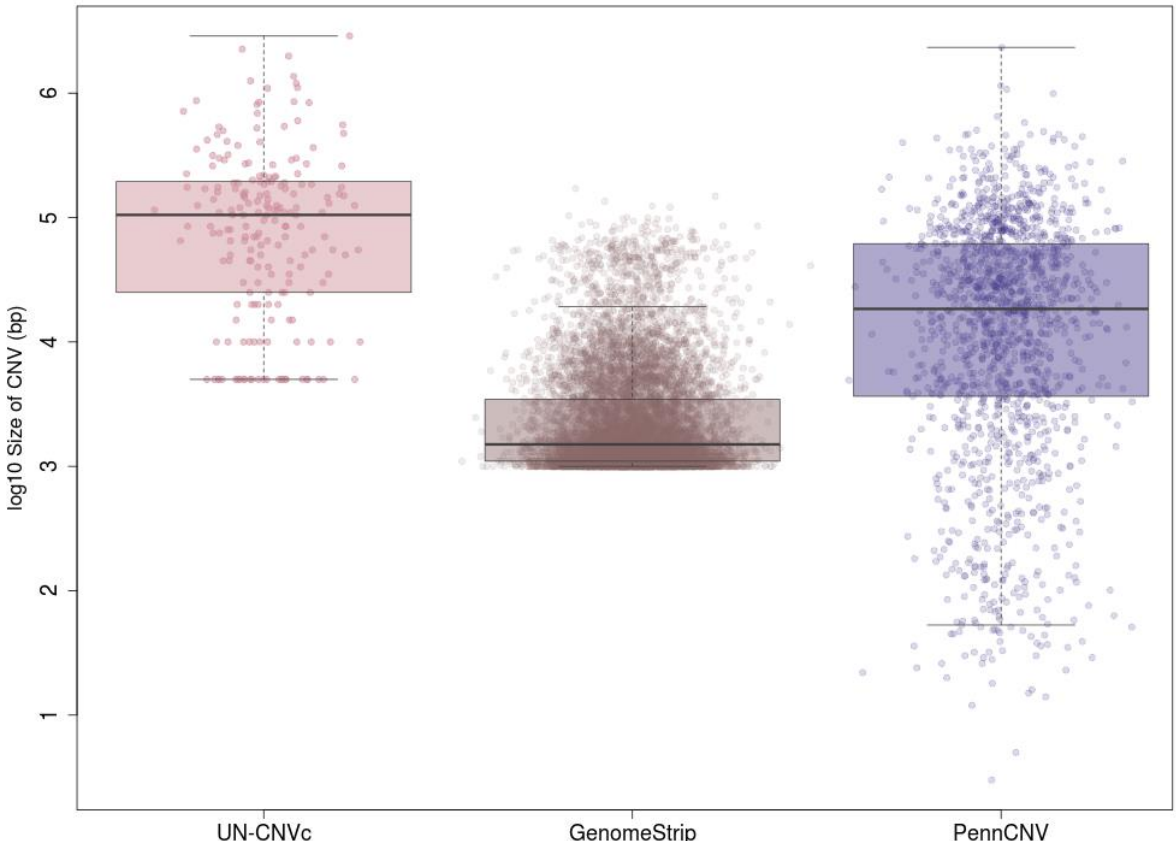



Figure 72 : Comparison of event sizes from CNV callsets in 211 MANOLIS samples using three different methods.

Computing overlap between callsets is generally harder with CNVs than with SNVs, as it is unclear which proportion of shared bases constitutes true concordant calling of the same event. Only 7% of deletions called by UN-CNVc overlapped with the GenomeStrip set when using 80% reciprocal overlap, however, this increases to 55% when assessing whether UN-CNVc completely contained one or more GenomeStrip CNVs. For PennCNV, these values were 9% and 28%, respectively.

#### 7.3.3.Systematic evaluation of sharing with DGV

Compared to the other two methods, a higher percentage of CNVs detected by UN-CNVc had 80% reciprocal overlap with known CNVs in the Database of Genomic Variants (DGV; build 38, May 2016), although PennCNV and GenomeStrip each called more known CNVs. For UN-CNVc, 110 (55%) CNVs overlapped with known variants in DGV, versus 437 (16%) and 2,615 (25%) for PennCNV and GenomeStrip, respectively, indicating that UN-CNVc offers higher specificity and lower sensitivity than the other methods. Notably, the well-known complete deletion of the *RHD* gene was detected only by UN-CNVC in the 211 MANOLIS samples. For the array-based PennCNV, this was likely due to the lack of tagging SNPs within the region. Only 5 tagging SNPs in the CoreExome array were within the *RHD* gene coordinates, compared to 141 SNPs from the WGS data used by UN-CNVC, demonstrating the advantage of using WGS data for CNV calling. For GenomeStrip, the deletion was split into six smaller CNVs with an average size of 11kb. This example, where the whole gene is known to be deleted, indicates that in some cases GenomeStrip may be "tiling" large CNVs by cutting them up into smaller events.

#### 7.3.4.Notable deletions

A main question for any new CNV method is its sensitivity, which can be approached by looking at how well it recalls known, common deletions.

##### *7.3.4.1.RHD deletion in MANOLIS, Pomak and TEENAGE*

UN-CNVc detected a deletion spanning the *RHD* gene in all cohorts (after manual genotyping fo Pomak, INTERVAL and TEENAGE). This gene codes for the rhesus D antigen (RhD), which determines an individual's rhesus blood group. In Europeans, this well-documented deletion of *RHD* is the most common determinant of rhesus negative status[165]. Among the carriers across MANOLIS and Pomak cohorts, 166 individuals had rhesus blood group among their phenotypic information, 32 of whom reported rhesus negative. We confirm this by genotyping in only 15 cases (the remainder carrying at least one copy of *RHD*), however there is great uncertainty regarding the quality of this phenotype, as self-reported blood groups are often inaccurate. This makes it hard to use this particular event to evaluate UN-CNVC's calling accuracy, however interesting observations can be made in terms of Rh- frequencies. The frequency of the RHD homozygous deletion was 8% in MANOLIS, 14% in Pomak, 15% in TEENAGE and 24% in INTERVAL. European frequency of Rh- is estimated at 15%, coherent with observations in Pomak and TEENAGE. The increased frequency in INTERVAL may be due to a bias in selecting samples, as it is a blood donor cohort. Conversely, the lower proportion in MANOLIS may be due to either a founder effect or population history (populations with a history of African admixture, such as African Americans, exhibit Rh- frequencies of around 7%).

#### *7.3.4.2.GSTM1 deletion in Pomak*

UN-CNVc detects a deletion of the entire *GTSM1* gene and several exons of the *GTSM2* gene (chr1:109680723-109700723) in the Pomak cohort. At 20kb, it is close to UN-CNVc's limit of detection, and we manually genotype it in the other cohorts. This variation overlaps two 1000 Genomes Project variants from the DGV database of structural variants: esv3587155 (gain/loss chr1:109687453-109698625) and esv3587154 (deletion, chr1:109682397-109702658). Both variants have a loss frequency of 0.72 in European subpopulations, and we confirm frequencies of 0.73, 0.77, 0.72, and 0.73 for MANOLIS, Pomak, TEENAGE, and INTERVAL, respectively, using UN-CNVc genotypes. The *GSTM* family of genes encodes the mu class of Glutathione S-transferases, responsible for the detoxification of electrophilic compounds, including carcinogens, environmental toxins and products of oxidative stress. As a consequence, loss of function of this gene has been associated with multiple

phenotypes [166-169], however we do not find this event associated with any of our quantitative trait in our association study (see below).

#### *7.3.4.3.ADAMTS19 deletion*

A large deletion spanning the 8 first exons of the *ADAMTS19* gene was found by F. Wünemann's lab at the Cardiovascular unit at CHU Sainte Justine Research Center, Montreal, in four individuals from two consanguineous families with a rare form of cardiac valve disease[170]. Using a mouse model, they showed that the homozygous *ADAMTS19* $^{-/-}$ genotype is linked to abnormal development of the aortic valve, a phenotype coherent with the observed disease phenotypes in humans. This deletion is extremely rare, having been found in only 1 family of Haitian ancestry in the exAC cohort before Dr. Wünemann's study. exAC also reports two duplications in the region, one with exactly the same boundaries as the deletion, and one extending much farther upstream (chr5:128988598-129113006), and present only in one non-Finnish European sample. The deletion was found in MANOLIS with coordinates 5:129370075-129590075, where it was carried by four distinct heterozygous samples. This represents a near-doubling of the known worldwide carrier count for this variant. The clinical phenotype data for the carriers did not reveal any disease information related to cardiac phenotypes, however, this might be caused by the largely silent nature of the disease, which can go unnoticed for many years, as well as the uncertainty surrounding the functional consequences of a heterozygous genotype. There were three high-consequence variants in *ADAMTS19* in the MANOLIS cohort: rs142924298 (inframe insertion MAF=8.1%), rs30645(splice region variant MAF=45.5%), and rs73246875, a splice region variant with MAF=0.37%. We determined that no sample carried both the deletion and one of the three high-consequence alleles for *ADAMTS19*, ruling out the possibility of disease-causing compound heterozygotes. The variant is absent from the Pomak, INTERVAL and TEENAGE cohorts. In INTERVAL, two events with different boundaries are present in the region: a duplication that extends further downstream and spans one exon less than the MANOLIS deletion, and a 500kb deletion further upstream, which marginally overlaps the MANOLIS deletion (Figure 73). The duplication has the same right boundary as the singleton

duplication in exAC, however the left boundaries of the two duplications are not concordant. This suggests that the *ADAMTS19* region is subject to complex rearrangements.

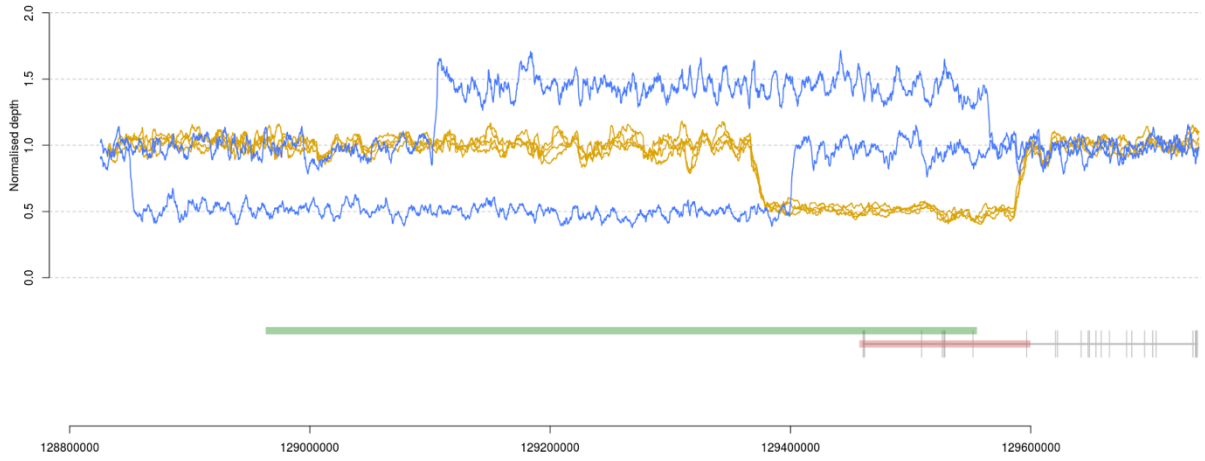



Figure 73: Structure of the ADAMTS19 region in MANOLIS and INTERVAL. The top panel represents rolling mean depth averaged over 100 SNPs in the region, coloured by cohort (blue=INTERVAL, yellow=MANOLIS) for the 6 samples that exhibited anomalous depths at this locus. The bottom panel represents the exons of *ADAMTS19* (in grey) and the two documented CNVs in exAC in this region. In red, a deletion/duplication event (exAC $n_{del}$=1, $n_{dup}$=6), and in green a duplication-only event ($n_{dup}$=1).

### 7.3.5.Exon-deleting variants

The aforementioned events are gene-deleting, hence more likely to be functional. We examine gene-deletion overlap genome-wide using Ensembl REST API, and find that 57% (673/1170) of deletions called by UN-CNVc in all cohorts overlap with protein-coding genes, 42% (489/1170) delete at least one exon and 26% (305/1170) delete entire genes. These proportions are surprisingly high, given that genic regions make up only a small percentage of the human genome. 89 exon-deleting events affect olfactory receptor genes or genes without an HGNC name, however the majority of affected genes do not belong to a specific class and are not clustered in hypervariable regions. This overrepresentation of exon-deleting events may be due to a bias in UN-CNVc's calling algorithm, however the only known specificity of this method is its restriction to large variants. One explanation may be the evolutionarily-constrained high GC content found in gene-rich regions[171], which in turn could

lead to depth decreases and spurious calls by UN-CNVc. However, we find that most variations in GC content are probably too weak to influence UN-CNVc's calling accuracy (see Discussion). A more likely explanation is found in the size distribution of the deletions, whose median is three orders of magnitude higher than the size of most exons (100-400bp) [172,173]. Under the assumption that genes are spatially randomly distributed, this size imbalance could make it more likely that large deletions such as the ones called by UN-CNVc would overlap coding segments of the genome.

#### 7.3.6.Frequency comparison

We compare the population deletion allele frequencies between any event that was present in at least two cohorts, adjusting for the number of comparisons performed ($p < \frac{0.05}{211} = 2.37 \times 10^{-4}$ for the two-proportion chi-squared test). We find that 40.5% (41/101) of all shared deletions exhibit significant allelic frequency differences(Table 10). Most differences were observed between Pomak and INTERVAL (22 significant differences) and MANOLIS and INTERVAL (20 significant differences), whereas no differences were significant between MANOLIS and TEENAGE. This broadly reflects the number of shared deletions between pairs of cohorts. No direction of frequency was significantly overrepresented in any of the compared sets.

The most significant difference ($AF_{MANOLIS}(0.24) > AF_{INTERVAL}(0.099)$, $p=1.68 \times 10^{-80}$) arises for the known deletion esv3608493, which overlaps four HLA pseudogenes and is significantly differentiated, albeit to a lower extent, in all other cohort pairs.

Table 10 : Significant allelic frequency differences between overlapping CNVs across the four analysed cohorts. In the test column, cohort codes are used for brevity: HA: MANOLIS, HP: Pomak, TEEN: TEENAGE, INT: INTERVAL.

| | | | | | deletion frequencies | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| chr | start | end | p | test | MANOLIS | Pomak | TEENAGE | INTERVAL | exons deleted | present in DGV |
| 6 | 29881000 | 29939000 | 1.68E-80 | HA > INT | 0.2426 | 0.1252 | 0.2250 | 0.0987 | | y |
| | | | 1.09E-32 | HA > HP | | | | | | |
| | | | 1.15E-08 | TEEN > INT | | | | | | |
| | | | 5.15E-05 | HP > INT | | | | | | |
| | | | 7.83E-05 | TEEN > HP | | | | | | |
| 2 | 52525285 | 52560078 | 1.52E-30 | HP > HA | 0.2450 | 0.3819 | | 0.3087 | | y |
| | | | 1.76E-13 | HP > INT | | | | | | |
| | | | 1.65E-10 | INT > HA | | | | | | |
| 4 | 34780149 | 34828486 | 2.67E-15 | INT > HP | 0.2526 | 0.2171 | 0.3350 | 0.2911 | | y |
| | | | 9.62E-05 | INT > HA | | | | | | |
| | | | 1.48E-04 | TEEN > HP | | | | | | |
| 7 | 8791473 | 8830064 | 1.08E-49 | HA > INT | 0.0463 | 0.0145 | | 0.0048 | | y |
| | | | 3.62E-13 | HA > HP | | | | | | |
| | | | 2.96E-07 | HP > INT | | | | | | |
| 3 | 46755055 | 46810039 | 1.36E-29 | HP > INT | 0.0645 | 0.0764 | | 0.0281 | *PRSS50, AC109583.3* | y |
| | | | 5.88E-18 | HA > INT | | | | | | |
| 4 | 63270000 | 63290000 | 6.29E-05 | HP > INT | 0.1201 | 0.1153 | | 0.0901 | | y |
| | | | 4.81E-06 | HA > INT | | | | | | |
| 5 | 180950060 | 181005000 | 5.92E-06 | INT > HP | 0.2859 | 0.2832 | | 0.3277 | *BTNL8, BTNL3* | y |
| | | | 4.15E-05 | INT > HA | | | | | | |
| 6 | 78260010 | 78330010 | 7.32E-12 | INT > HP | 0.2471 | 0.2038 | | 0.2662 | | y |
| | | | 5.47E-05 | HA > HP | | | | | | |
| 7 | 6865064 | 6885064 | 6.22E-17 | HA > HP | 0.0906 | 0.0383 | | 0.0509 | | y |
| | | | 7.00E-14 | HA > INT | | | | | | |
| 7 | 53395127 | 53525000 | 7.49E-39 | HA > INT | 0.0275 | 0.0009 | | 0.0011 | | y |
| | | | 6.20E-19 | HA > HP | | | | | | |
| 8 | 39375116 | 39535116 | 3.85E-06 | INT > HP | | 0.3667 | 0.5250 | 0.4146 | | y |
| | | | 1.04E-05 | TEEN > HP | | | | | | |
| 8 | 136670241 | 136850069 | 1.85E-09 | HP > HA | 0.0100 | 0.0328 | | 0.0247 | | y |
| | | | 2.85E-06 | INT > HA | | | | | | |
| 10 | 27315163 | 27420045 | 1.20E-29 | HA > INT | 0.0261 | 0.0136 | | 0.0024 | | y |
| | | | 7.10E-12 | HP > INT | | | | | | |
| 15 | 34400292 | 34560000 | 3.43E-15 | HA > HP | 0.1095 | 0.0544 | | 0.0871 | *GOLGA8A, GOLGA8B* | y |
| | | | 7.47E-09 | INT > HP | | | | | | |
| 16 | 19935007 | 19960007 | 5.24E-17 | INT > HA | 0.0782 | 0.1024 | | 0.1383 | | y |
| | | | 3.64E-07 | INT > HP | | | | | | |
| 16 | 55765071 | 55805071 | 1.76E-15 | HA > INT | 0.2244 | 0.1753 | | 0.1579 | *CES1* | y |
| | | | 1.69E-06 | HA > HP | | | | | | |
| 1 | 1693835 | 1743835 | 6.00E-09 | INT > HA | 0.0233 | | | 0.0490 | *CDK11A, SLC35E2A* | y |
| 1 | 72300362 | 72305013 | 1.29E-07 | HA > HP | 0.7450 | 0.6837 | | | | n |
| 1 | 248545020 | 248650000 | 4.82E-13 | INT > HP | | 0.0903 | | 0.1410 | *OR* cluster | y |
| 4 | 4125167 | 4151933 | 8.74E-05 | HA > INT | 0.3301 | | | 0.2905 | | y |
| 4 | 10210028 | 10233028 | 7.01E-08 | INT > HA | 0.6301 | | | 0.6857 | | y |
| 4 | 68505250 | 68570000 | 6.68E-07 | INT > HA | 0.2828 | | | 0.3336 | *UGT2B17* | y |
| 4 | 68575000 | 68626250 | 1.69E-08 | INT > HA | 0.2749 | | | 0.3324 | | y |
| 4 | 69254955 | 69365288 | 4.82E-08 | HP > INT | | 0.1685 | | 0.1284 | *UGT2B28* | y |
| 5 | 7171982 | 7201982 | 3.07E-17 | HP > INT | | 0.0529 | | 0.0218 | | y |
| 6 | 73885313 | 73892010 | 1.38E-08 | INT > HA | 0.2821 | | | 0.3404 | | n |
| 6 | 76730618 | 76750010 | 7.02E-07 | HP > INT | | 0.3086 | | 0.2617 | | y |
| 7 | 97770098 | 97773348 | 3.45E-07 | INT > HP | | 0.1200 | | 0.1582 | | n |
| 7 | 142070034 | 142095025 | 4.55E-08 | HP > HA | 0.1599 | 0.2149 | | | *MGAM* | y |
| 12 | 73915320 | 74035023 | 1.15E-08 | HP > INT | | 0.0077 | | 0.0009 | | y |
| 13 | 57180063 | 57215063 | 1.18E-11 | HP > INT | | 0.1534 | | 0.1066 | | y |
| 14 | 23975051 | 24020051 | 8.44E-06 | HP > HA | 0.0251 | 0.0467 | | | *DHRS4L2* | y |
| 14 | 41140051 | 41205051 | 1.26E-43 | HP > HA | 0.1472 | 0.2953 | | | | y |
| 14 | 44710051 | 44760051 | 1.20E-07 | HA > INT | 0.0202 | | | 0.0078 | | y |
| 14 | 106420191 | 106485042 | 4.22E-06 | HP > INT | | 0.2106 | | 0.1728 | | y |
| 15 | 25177531 | 25185870 | 6.09E-10 | INT > HP | | 0.0111 | | 0.0320 | | y |
| 16 | 78340096 | 78351000 | 7.46E-10 | HP > HA | 0.4616 | 0.5405 | | | | y |
| 17 | 36202500 | 36225000 | 1.46E-05 | TEEN > INT | | | 0.2300 | 0.1241 | *CCL4L2* | n |
| 22 | 38965107 | 38995009 | 2.83E-08 | HP > INT | | 0.1039 | | 0.0717 | *APOBEC3A, APOBEC3B* | y |

### 7.3.7.Genome-wide association study

We ran genetic association testing using GEMMA in 214 MANOLIS and 228 Pomak deletions, using the same relatedness matrix and phenotype preparation as described in Chapter 5 and Chapter 6. In addition to the quantitative traits described in previous chapters, we also analysed 273 quantitative proteomic traits across the CVDII, CVDIII and META OLINK platforms in MANOLIS. The phenotype data for these proteomics panels were prepared by Young-Chan Park as part of his MPhil thesis under my day-to-day supervision. In MANOLIS, study-wide significance was set to $\frac{0.05}{M_{eff}.N} = \frac{0.05}{165\times212} = 1.4 \times 10^{-6}$ where $M_{eff}$ is determined using the eigenvalues of the phenome-wide correlation matrix[138] across both proteomic and non-proteomic traits, and $N_{eff}$ is the average number of deletions in the two cohorts. In Pomak, only non-proteomic traits are available, yielding a study-wide significance of $\frac{0.05}{24\times228} = 9.1 \times 10^{-6}$.

In Pomak, a single common (MAF=0.06) deletion on chromosome 11 (11:7790388-7815388) passes study-wide significance. It is associated with increased mean corpuscular haemoglobin concentration ($\beta$=0.5, $\sigma$=0.075, p=9.42x10$^{-11}$) and decreased red cell distribution width ($\beta$=-0.35, $\sigma$=0.076, p=4.62x10$^{-6}$). These coordinates overlap with those of the 1000 Genomes variation esv3625301, which has a deletion frequency of 0.03 and 0.25 in European and African samples, respectively. The deletion spans the *OR5P2* olfactory receptor gene, and the region is devoid of any regulatory features that might link the deletion to a distal gene. Furthermore, it overlaps with the GRCh38 assembly exception CHR_HSCHR11_1_CTG1_2:7779133-7982684, which corresponds to issue HG-2109 in the Genome Reference Consortium data. This exception provides the alternate contig AC243812.1, which represents 10kbp of sequence not present in GRCh38. The alternate segment overlaps the downstream end of esv3625301, which might indicate that samples identified as carriers of the deletion are carrying the haplotype described by AC243812.1. The HELIC WGS datasets were aligned to a version of the reference that contained decoy sequences but not alternate contigs, and full characterisation of this locus would therefore require either reference-based reassembly or realignment to the alternate sequences, which are outside the scope of this thesis.

In MANOLIS, three association signals involving protein level measurements pass study-wide significance (Table 11). All three deletions overlap assembly exceptions, regions of the reference sequence that are flagged by Ensembl as containing either uncertainties, errors or multiple possible assemblies. Intense structural variation activity at a locus decreases assembly accuracy and contig length, so a co-localisation of CNV events and assembly exceptions is not unexpected.

Table 11 : Study-wide significant deletion associations in the MANOLIS cohort.

| coordinates | protein | Effect size | p-value | Overlapping genes | SVid, if known | Frequency in MANOLIS | Frequency in EUR | Assembly exceptions |
|---|---|---|---|---|---|---|---|---|
| 17:36195241-36196130 | CCL3 | -0.36 | $2.28x10^{-10}$ | *CCL3L3* | novel | 0.15 | - | HSCHR17_7_CTG4 |
| 6:29881000-29939000 | TFF3 | 0.25 | $2.37x10^{-7}$ | HLA-{H,T,K,U} pseudogenes | esv3608493 | 0.241 | 0.11 | MHC alternate haplotypes |
| 16:14833681-14896160 | NOMO1 | -0.71 | $4.34x10^{-7}$ | *NOMO1* | novel | 0.022 | - | HSCHR16_1_CTG1 |

#### *7.3.7.1. CCL3L3 deletion and CCL3 protein levels*

The most significant association links a deletion spanning the *CCL3L3* gene and CCL3 protein levels (Figure 74). The deletion was called manually as part of the quality control process, as the region is structurally complex. The deletion is not restricted to *CCL3L3* but extends far downstream, across a SNP-poor region. The duplication of *CCL3L3* by contrast is a much smaller event, which we used for determining the boundaries of the region to genotype.

Rolling mean coverage across the region reflects its structural complexity (Figure 74, third panel from top). Coverage is relatively uniform upstream of *CCL4* (Figure 74, region A), followed by a small region devoid of all genes, which is observed in up to 5 copies (region B). A copy-number variable region containing *TBC1D3B*, *CCL3L3*, *CCL4L2* is seen from 0 to 4 times (region C), with a sub-segment spanning CCL3L3 can be duplicated up to 7 copies (region D, which was used for genotyping the deletion). Region E is a long stretch of sequence with consistently lower coverage across the entire MANOLIS sample set, a strong indication of a poorly assembled sequence. Inspection of the gEval database (https://geval.sanger.ac.uk), a resource for the diagnostic of reference sequence quality, confirms that this stretch of the GRCh38 reference suffers from numerous tiling problems.

The junction between segment C, which contains *CCL3L3*, and both its neighbours B and E exhibits very low coverage for all samples, which indicates that the evidence anchoring this segment at this particular location in the genome may be weak.

Although two full copies (*CCL3L3* and *CCL3L1*) and one truncated copy (*CCL3L2*) of the same gene have been identified in the human genome, there is confusion in online resources as to the location and copy number of the *CCL3L3* gene. Ensembl places *CCL3L1* on chromosome 17 and *CCL3L3* on an alternate reference sequence. *CCL3L3* is referred by OMIM as "a centromeric copy" of *CCL3L1,* which seems to indicate it is located on an unanchorable sequence. NCBI Gene however reports that *CCL3L3* is on the main sequence of chromosome 17, and *CCL3L1* is located on the NT_187661.1 alternate reference. All resources confirm that the two genes have 100% sequence homology. We use the NCBI gene convention and use *CCL3L3* as the gene on the canonical reference sequence of chromosome 17, against which our data were aligned.

Although no known copy number variants are present in publicly accessible databases at the *CCL3L3* locus, direct experimental evaluation of copy-number variation of *CCL3L3* have been extensively performed. In addition to levels of their protein product [174], they have been shown to be associated with rheumatoid arthritis[175,176], immune reconstitution following HIV therapy [177], and protection against malaria [178]. *CCL3L1* and *CCL3* have highly similar (96%) nucleotide and protein sequences, and the former is thought to have evolved through duplication and subsequent divergence from the latter. Both genes encode isoforms of the same MIP-1$\alpha$ pro-inflammatory cytokine which acts as a ligand for CCR5, the co-receptor used by HIV-1 virus for cell entry, and *CCL3L3* exhibits increased affinity to the receptor compared to *CCL3*. No commercially available antibody that can distinguish the two protein products[179] exists, and the provider of the OLINK proteomics assay has confirmed that measurement of CCL3 protein levels may not be specific, as the antibodies used to measure abundance may not be able to differentiate between the two ligands. It is therefore likely that our assay measures abundance of the protein product of the copy variant *CCL3L3* gene against a background level of CCL3 protein. Up to 14 copies of *CCL3L3* have been validated in some genomes[180], with 1 to 6 being the most common copy numbers in the general

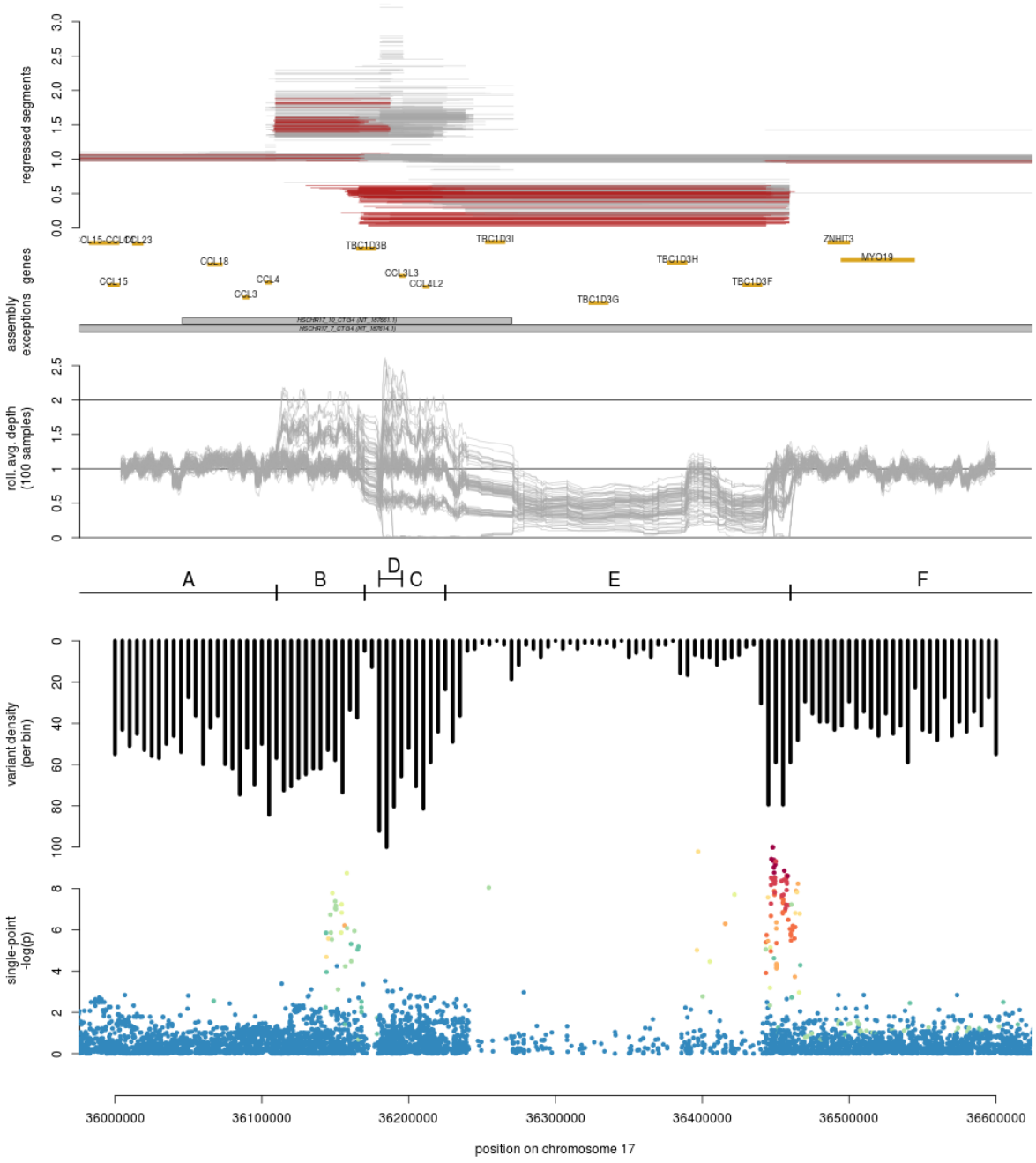



Figure 74 : CCL3 region, read depth, structure and quality metrics. From top to bottom: Segments called by UN-CNVc, with carriers of the deletion in red; gene (from Ensembl) and assembly exception (from USCS) annotations; rolling average depth for 100 randomly selected samples in MANOLIS; subdivision of the region in segments of interest; variant counts in fixed bins; single-point association statistics, with markers coloured by LD with the lead variant.

population [181]. We confirm up to 7 copies in the MANOLIS cohort. Complete absence of

*CCL3L3* is reported in 2% of the British population, close to the MANOLIS frequency of 2.6%. It has been hypothesised that increased copy number of this gene resulted in higher levels of expression of its protein product, however in our study, including copy numbers greater than 2 in the model did not strengthen the association signal compared to a deletion-only model (Figure 75). This phenomenon has been previously attributed to post-transcriptional

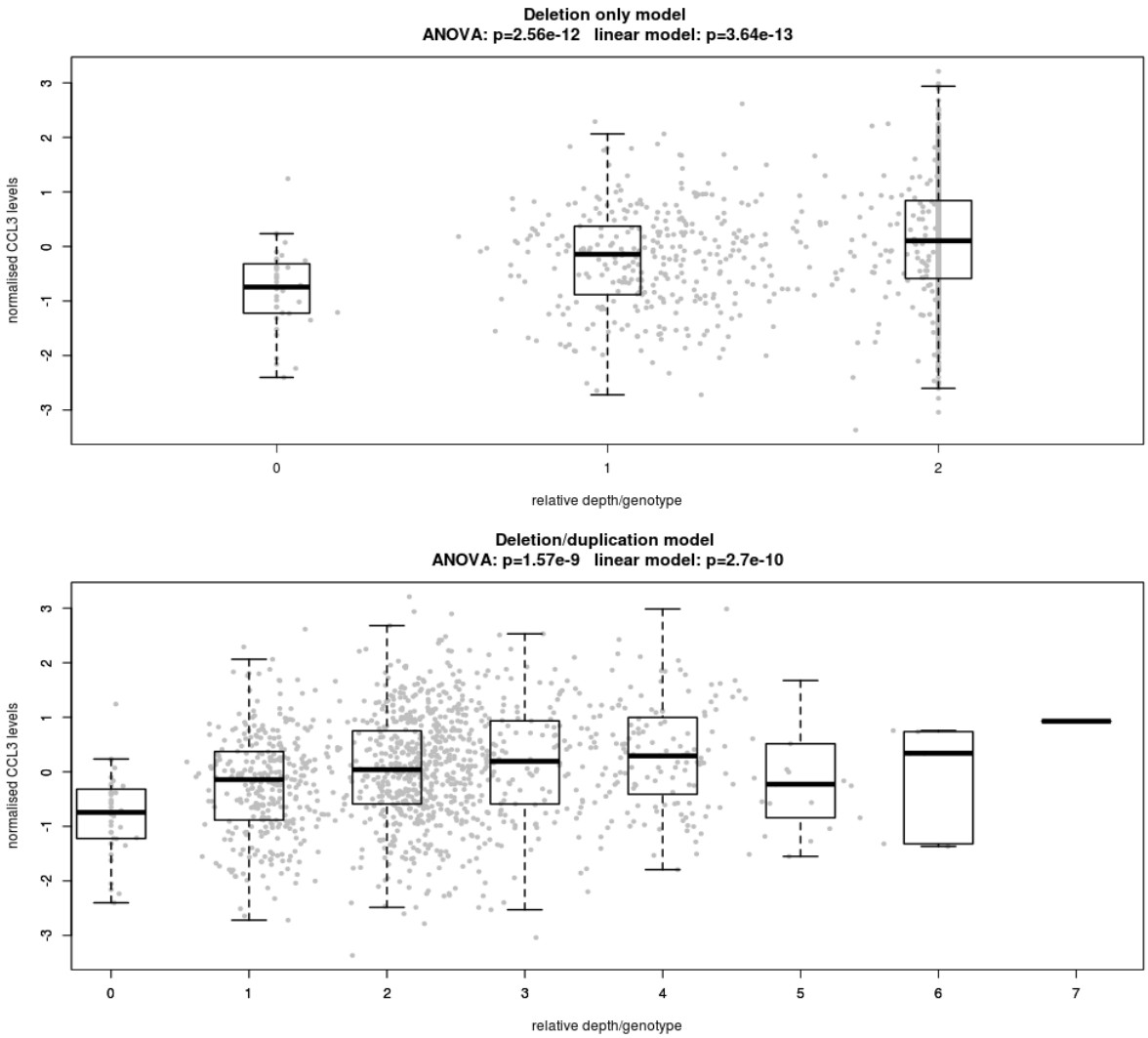



Figure 75 : Association of CCL3 protein levels with CCL3L3 copy number in the deletion (top) and deletion/duplication(bottom) models

regulation of MIP-1$\alpha$, although prior to this work, no study has been able to link even homozygous *CCL3L3* deletion to any variation of the combined CCL3/CCL3L3 protein product[179].

A single-point association signal for CCL3 levels is apparent in sections B and F. Its peculiar structure, with two distinct, co-dependent peaks either side of the SNP-poor region E, further attests to the complex nature and potential mis-assembly of the region. The signal is driven by a common novel variant located at chr17:36448038 (MAF=0.14, $p=2.78x10^{-11}$), and is in strong LD with two variants previously associated with CCL3 levels ($r^2=0.72$ with rs113877493[182], and in $r^2=0.51$ with rs4796217[183]). The first study to report association between rs4796217 and CCL3 levels hypothesised a potential linkage disequilibrium between that variant and *CCL4L1* copy number. In MANOLIS, the lead variant is in weak LD ($r^2=0.37$) with the secondary peak in section B, which culminates at rs7210781 ($p=5.8x10^{-10}$). This variant in turn is in linkage with the genotyped *CCL3* deletion ($r^2=0.51$). The association of the *CCL3L3* deletion with CCL3 levels is attenuated to $3.6x10^{-4}$ when conditioned on the lead variant and to $4.7x10^{-3}$ when conditioned on rs7210781, which suggests both variants may be partially tagging the deletion. Due to this complex linkage structure, the respective contributions of the associated variants and the deletion are difficult to disentangle.

#### *7.3.7.2. HLA pseudogenes and TFF3 protein levels*

We genotype the previously described esv3608493 copy number variant deleting four HLA pseudogenes, and find it associated with TFF3 protein levels ($\beta=0.25$, $p=2.37x10^{-7}$). The deletion allele frequency was double in MANOLIS compared to the European 1000 Genomes subpopulation and INTERVAL. TFF3, the trefoil factor 3 protein, acts as a motogen for epithelial cells lining the gastrointestinal tract, promoting repair of wounded mucosal lining in the intestines. As such, it has been poised as a potential biomarker for mucosal healing in inflammatory bowel disease (IBD) [184]. Although the deletion event is located in the IBD3 chr6p21.1-23 locus, which has previously been associated with both Crohn's disease and IBD [185], this region is located within the wider hypervariable MHC locus, and multiple alternate haplotypes are present. Mapping and variant calling are generally thought to be poor in the region, and reassembly or genotyping using dedicated tools are necessary to accurately assess the structure of the HLA region [186]. The association between TFF3 levels and an HLA gene deletion, while coherent with prior results in autoimmune disease research, would therefore need to be confirmed by HLA typing in the MANOLIS population.

#### *7.3.7.3. NOMO1 protein levels and the NOMO gene cluster*

A cis-deletion of the *NOMO1* gene is associated with NOMO1 protein levels (MAF=0.022, β=-0.71, $p=4.34x10^{-7}$) in a complex region harbouring an assembly exception and numerous repeats. The assembly exception reflects the presence of an alternate contig containing an inversion of two regions, one of which contains *NOMO1*(Figure 76). Although this inversion is frequent enough to have been observed using the optical maps used to build the reference assembly, read depth information is not informative as to the presence of inversions, and we are therefore unable to quantify its frequency in the MANOLIS cohort. A number of repeated sequences and a concentration of repeats of unknown lengths further complicate the analysis of this region. Wide fluctuations in depth, even in locations where agreement

between the alternate and canonical reference sequences is high, attest to the

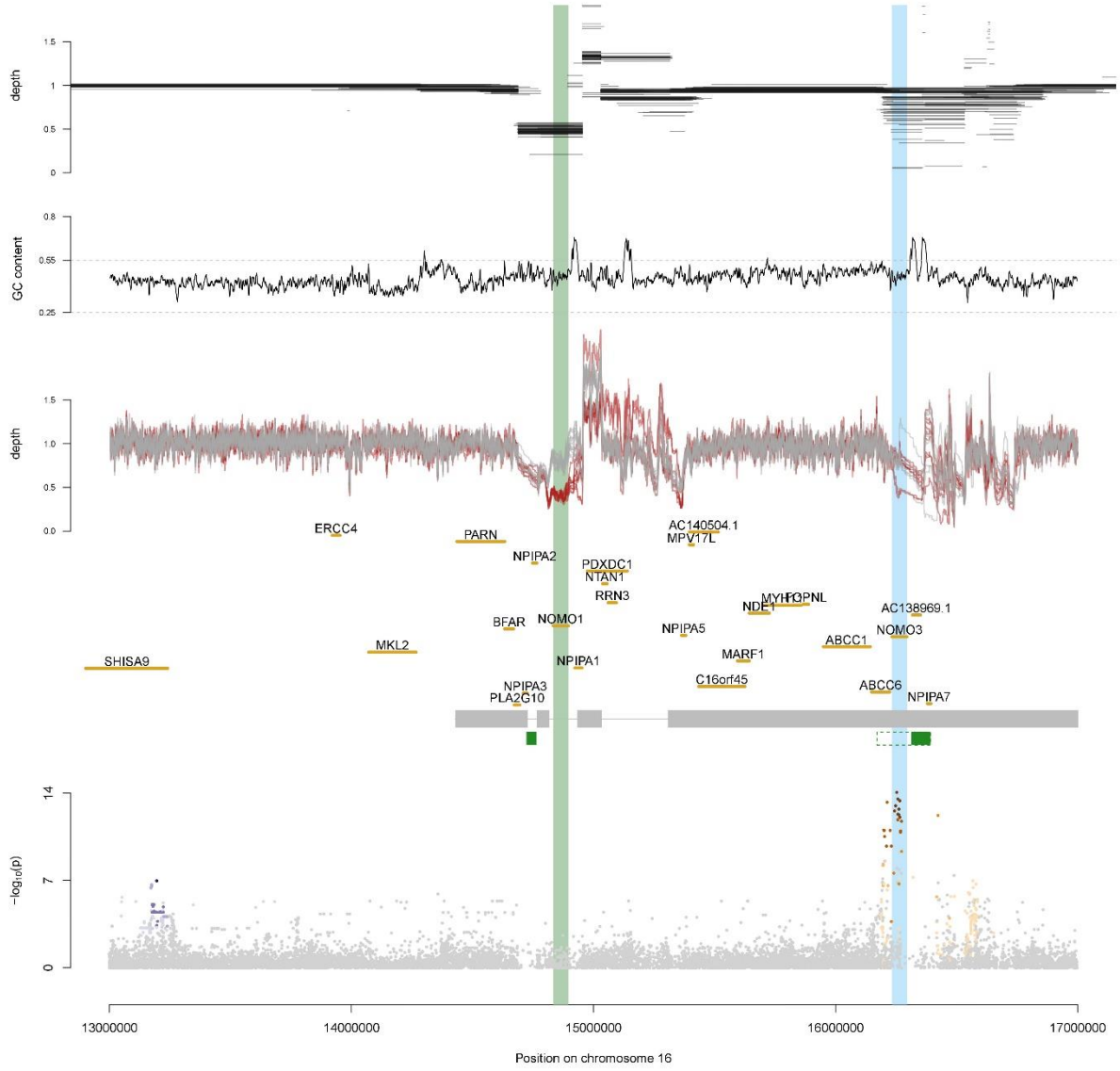



Figure 76 : Structure of the cis-region for the NOMO1 protein. The green and blue vertical bands represent the genotyped deletions of *NOMO1* and *NOMO3*. Genes are annotated in yellow. The gray gapped bar below the gene track is the NT_187607.1 alternative contig, where the second and third gap are inverted compared to the reference sequence of chromosome 16. Green horizontal rectangles represent large tandem repeats (full) and uncertain regions containing unknown numbers of dinucleotide repeats (dashed). The bottom track represents the single-point association of NOMO1 protein levels, with the two cis associations indicated in shades of purple and brown, respectively, according to LD with the lead SNVs.

hypervariability of this region. Although no single-point signal is present in the *NOMO1* gene itself, two independent intronic variants are associated with NOMO1 levels in the *SHISA9* (rs200517050 , MAF=0.059, β=-0.462, $p=1.01x10^{-7}$) and *NOMO3* (rs3891245, MAF=0.257, β=-0.337, $p=8.34x10^{-15}$) genes, respectively. We also manually genotype a deletion of the *NOMO3* gene, however it does not show any sign of association with NOMO1 protein levels (p=0.41). All four variations are independent of each other ($r^2<0.1$), suggesting that *NOMO1* $^{(del)}$, rs200517050 and rs3891245 independently contribute to decreasing NOMO1 protein levels.

*NOMO1*, *NOMO2,* and *NOMO3* are closely located genes with very high sequence similarity (99.4% and 99.5% homology (BLAST)), and cannot be distinguished by the polyclonal antibody used in the Olink proteomics assay. It is therefore possible that rs3891245 acts directly on the function of *NOMO3* by reducing the combined NOMO1 and NOMO3 protein product measured by our assay. In this case however, it is unclear why *NOMO3* $^{(del)}$ would not have the same effect. According to OMIM, the high similarity in nucleotide sequence between *NOMO1* and *NOMO3* makes it impossible to pinpoint variation specifically to either copy of the gene, making it possible, in theory, that the rs3891245-driven signal is actually located within *NOMO1*. Finally, it is also possible that the reduced depth in some samples is not indicative of a deletion but of reduced mapping accuracy due to a poor quality reference sequence, or the presence of an undocumented alternate haplotype. This hypothesis is reinforced by the fact that *NOMO3* is located in a repeat-rich region, where the variance in sequencing depth is abnormally high.

We observe a single homoozygote carrier of *NOMO1* $^{(del)}$ (whose depth is not represented on Figure 76), however this sample does not exhibit extremely low levels of the gene's protein product. In fact, the sample with the lowest levels of NOMO1 is homozygote for the common *NOMO3* rs3891245 variant and does not carry the deletion.

## 7.4.Discussion

We present UN-CNVc, a method for fast calling of large copy number variants from WGS variant calls, and apply it to a large European population sample with high-depth WGS. The

method uses piecewise constant regression achieved using regression trees, which we demonstrate is a viable and fast alternative to the circular binary segmentation algorithm used in many array-based CNV detection methods[187,188]. We run UN-CNVc on 6,898 high-depth whole-genome sequences of European ancestry and establish a chromosome-wide map of large copy number variation. We carry out a genome-wide association study with over 300 quantitative traits and protein biomarkers in the two HELIC cohorts, and report two *cis* deletions influencing NOMO1 and CCL3 protein levels.

UN-CNVC's reliance on marker-level data confers it a substantial speed improvement over existing methods, with run times over 50 times shorter than commonly used tools such as GenomeSTRiP. This makes its application to large population-based WGS studies computationally tractable, allowing processing of the INTERVAL cohort (n=3,724 samples with 15x WGS) in under three and a half hours.

Another advantage of using marker-level data is that the input dataset has often already undergone stringent read- and variant-level QC, as is standard for sequencing-based association studies.

Despite providing a certain level of automation, UN-CNVC still requires post-run manual QC, in the same way as array-based genotypes require inspection of cluster plots. The software generates extensive diagnostic tables and plots to make this task easier for the user. In the calling step, segments are aggregated together to construct variant intervals in a greedy way, causing the algorithm to expand the boundaries of events as much as it can. This can cause events to be called as larger than they really are, however inspection of the diagnostics plots as part of the standard QC pipeline will reveal if this is the case for any given CNV. Boundaries can then be adjusted using the manual genotyper provided as part of the UN-CNVC package.

Although piecewise constant regression can accurately model WGS depth in a single individual, UN-CNVC leverages large sample sizes (n>100) to differentiate signal from noise, which makes it applicable only to large studies. Furthermore, since the software performs clustering on depth averages, a high enough depth (greater than 15x) is required to ensure

proper cluster separation. Finally, using marker-level depth puts limits on the precision of the boundaries as well as the sizes of detected CNVs. The maximum precision achievable by a method such as UN-CNVC is the distance between two consecutive SNVs, in practice it is limited to around 10kb by the greedy segment aggregation algorithm and the discretization interval. For smaller events, it relies on at least one of the carriers to be correctly genotyped using piecewise constant regression, which makes small, rare CNVs difficult to call.

UN-CNVc expects all analysed samples to exhibit similar sequencing depths. This was not strictly true in the HELIC cohorts, where some samples were re-sequenced and had twice the expected sequencing depth. Depths are normalised internally at the first stage of the algorithm, however the higher depth observed in some samples may increase the depth variance and cause overdispersion of the segments. However, only calling, not genotyping, relies solely on modelling the segment distribution. In consequence, UN-CNVc may have a slightly lower recall rate for rare CNVs arising solely in these very high depth samples.

A common problem with CNV callsets from whole-genome sequencing is the presence of high amounts of noise caused by technical artefacts arising as part of the sequencing protocol [147]. The piecewise constant model, as well as the UN-CNVC's tendency to call larger events can protect it from such artefacts, as any chimeric depth variation would need to persist over stretches of DNA much longer than UN-CNVC's 10kb detection limit. In practice, we do observe abnormal depth segments, notably in centromeric and telomeric regions, but they are mostly due to low-quality reference data, with most called segments exhibiting remarkable resistance to local effects. One of the main local variables that can bias CNV callers is sequence GC content, which is a combined effect of bias in several steps of the sequencing pipeline, notably PCR amplification[189]. Experimental evaluation has shown that 25 to 55% of GC content ensure 90% median coverage in human WGS [190]. Previous studies have shown that although the choice of a window size to compute read counts and means does not affect the presence of this bias, it does change the variance in observed coverage. UN-CNVc does not normalise read counts by GC content, hence we investigated the effect of GC content on depth in the complex *CCL3L3* region using a publically available script (https://github.com/DamienFr/GC-content-in-sliding-window-) with a window size of 100

base pairs. We found that GC content in the entire 10Mbp chunk varied from 0.4 to 0.54, well within the optimal boundaries guaranteeing optimal coverage. In the *NOMO* region, we found that although the effect of high GC content on depth can be clearly seen, notably at the *MLK2* locus and downstream of both *NOMO* deletions, it only marginally contributes to the large variations in depth, which are likely caused by a low-quality reference. In fact, the variation around *MLK2* and the *NOMO* loci are averaged out by the regression algorithm into wider regressed segments, suggesting that overall, GC content has a limited effect on UN-CNVC's calling accuracy, although it may impair its ability to correctly genotype called events. In general, population-wide approaches such as the one implemented here, which both normalise depth across a population and average it in finite genomic windows, are able to correct through averaging a significant fraction of sample- and locus-specific technical biases[147].

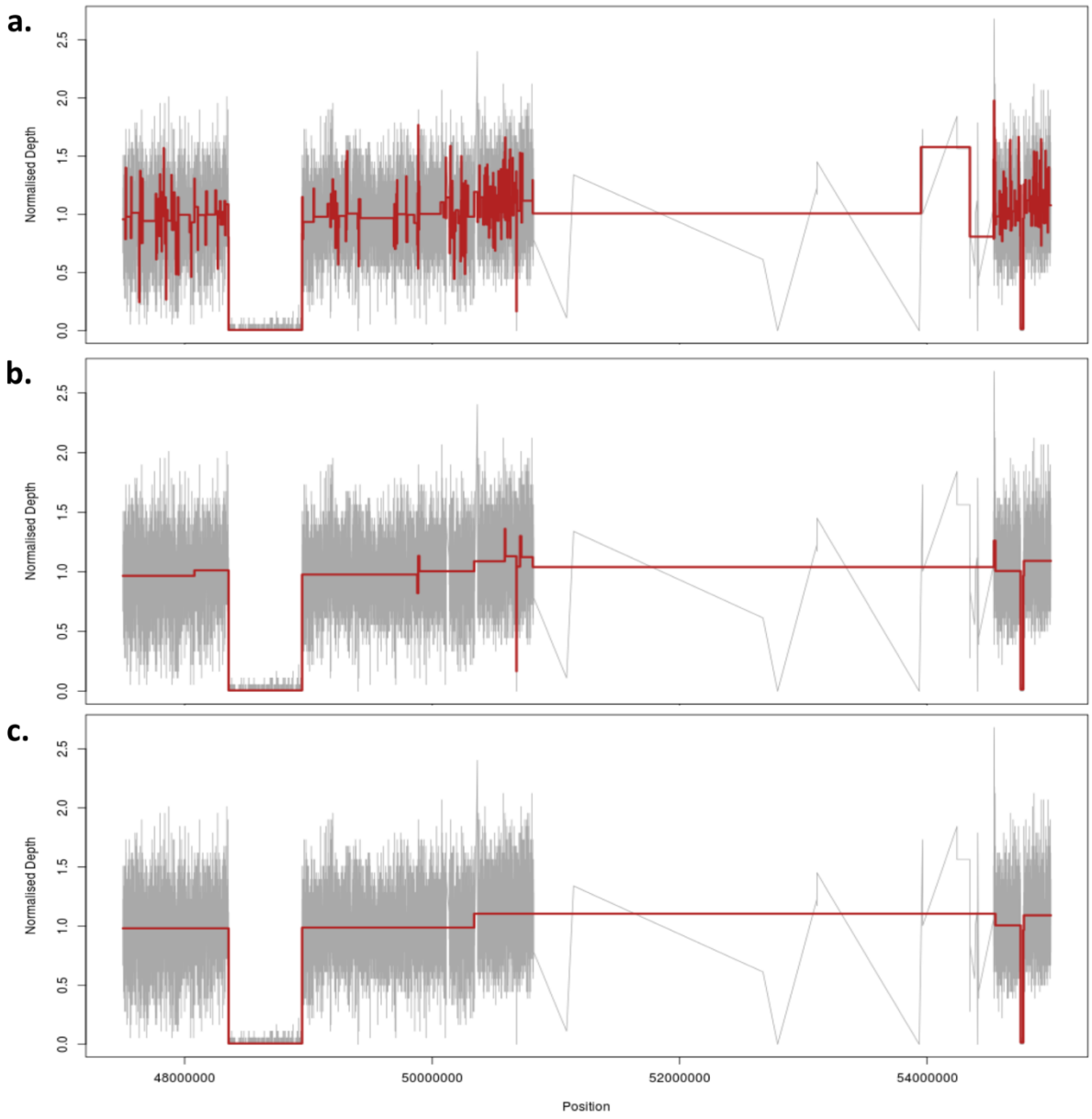



Figure 77 : Influence of the complexity parameter *c* on depth regression. Values of *c* represented are $1x10^{-4}$ for a, $3x10^{-4}$ for b, and $5x10^{-5}$ for c, in a single sample from the Pomak cohort over a pericentromeric region of chromosome 11. The default used by UN-CNVc reflects R's default value of 0.01, which over a 10Mb windows results in low sensitivity.

By design, UN-CNVc averages out local variations of depth at the cost of failing to call smaller events. This bandwidth effect is controlled by the regression tree's complexity parameter *c,* which limits the tree's branching to splits that increase the overall model fit by at least *c*. A lower value of this parameter causes the regression to follow local depth variations more closely, increasing sensitivity but also false positives (Figure 77).

Other parameters of the algorithm include the length of the window in which CNVs will be called, and the trigger limit above which to call a region as variable using the two internal indicator functions. All three parameters indirectly control bandwidth (for example, a smaller window will provide less data points, loosening the constraints on the algorithm even for a constant value of *c*), such that adjusting them should provide the same result as controlling the *c* parameter. One case in which adjusting the window size may be necessary is when analysing small samples. In the case of the smaller TEENAGE cohort (n=100), UN-CNVc underestimated mean depth in windows that spanned regions of relatively low coverage, causing events to either go undetected or be genotyped incorrectly. In such cases, larger windows should be used, so as to ensure that average mean depth for most segments are reasonably close to chromosome-wide averages.

We compare UN-CNVc's called deletions with those produced by two methods, one array-based (PennCNV) and one sequencing-based (GenomeStrip). The small percentage of overlap in both cases confirms the high caller specificity generally observed with structural variant callers. We showed that UN-CNVc exhibits a higher specificity than either method when compared to known CNVs from the Database of Genomic Variants (DGV). However, overlap with a database of known variants (like dbSNP for SNVs) only provides limited insight into the true performance of any calling method. Ultimately, the absence of a dependable truth set, combined with the high structural heterogeneity at copy number variable loci and the high costs involved with experimental validation of large variants will continue to be important obstacles to any thorough comparison of CNV caller performance.

Despite these limitations, we show that UN-CNVc recapitulates well-known deletions, notably at the *HLA, GTSM1, RHD* and *ADAMTS19* loci, and we identify two *cis* loci where gene deletions decrease circulating protein levels. Given the small size of the callsets it produces and its increased specificity, this proof of concept study establishes UN-CNVc as an efficient method for quick examination of the large CNV landscape in WGS population cohorts where SNV variant calls are already available.

# Chapter 8. Discussion

In recent years, the multiple successes of GWAS, and increasingly, whole-exome and whole-genome sequencing studies have identified thousands of loci affecting complex traits, shedding insights into their genetic architecture. Most GWAS studies, despite their ever-increasing scale, have discovered phenotype-associated variants with relatively small effect sizes that collectively explain a small fraction of the trait's genetic heritability. For example, despite heritability estimates ranging from 22% to 59% for a disorder such as insomnia[191], a recent study of over a million samples explained 2.6% of the phenotype variance[192]. The typical stringent thresholds used to declare genome wide significance might mask many common variants with small effect sizes[193]. This, combined with the observation that SNPs associated with a given trait tend to be distributed evenly across the genome has led some in the field to move from a "few genes-few variants" model inherited from Mendelian disorders to an omnigenic approach, where both polygenic and pleiotropic effects are spread ubiquitously across the genome[194]. In such a model, a significant portion of common genetic variants, as well as rare and low-frequency variants with potentially larger effects, affect most quantitative traits including risk of diseases through a complex interplay of feedback loops and tissue-specific regulatory networks. Polygenic scores, which can aggregate single-variant effects at increasingly lenient significance thresholds, are now a commonly used tool in large association studies.

Even in such a model, a reduced set of "core genes" of unknown cardinality is assumed to directly affect the aetiology of complex traits. Equally, widespread pleiotropy is not necessarily an obstacle to interpretation, as it can be leveraged in the context of association studies to boost power in detecting such core genes through the use of multivariate association methods[195-197]. In both single- and multi-trait frameworks, a mixture of common and rare variants are likely to contribute at these loci, even though it has been postulated that genes implicated through rare variant and common variant effects could explain fundamentally different aspects of trait aetiology[198]. A sizeable proportion of common variants with modest effect sizes affecting well-studied complex traits and diseases are likely to have already been implicated through large GWAS meta analyses. Larger biobanks

and consortial efforts will enable the detection of common variants with lower effect sizes, and will start to better elucidate the role of low-frequency variants. Although accurate imputation down to MAF=0.25% has been reported[199], rarer variants still remain mostly inaccessible through array-based studies even at very large sample sizes. There is ongoing debate as to whether whole-exome sequencing or deep whole-genome sequencing is the most relevant approach to unlock the lower end of the allelic spectrum in both clinical and basic research settings [200-203]. Both approaches remain costly to deploy at biobank scales despite a continued fall in sequencing prices.

However, there are alternative approaches to sample size increases for boosting power in association studies. In particular, the study of isolated populations remains relevant in this context, as it can enable the detection of phenotype-associated variants using relatively small sample sizes[204] through their observed enrichment in low-frequency functional variants[205,206]. Comparatively smaller population sizes allow the application of sequencing to empower cohort-wide rare variant studies. The second angle is methodological, as moving beyond the tried and tested single-variant effect model can reveal subtler effects, for example through the evaluation of effects arising from rare variant burdens or structural variants.

## 8.1.Summary

In this thesis, I explored multiple ways in which whole-genome sequencing can be used to empower quantitative trait association studies in population isolates, by leveraging the sequencing datasets available in HELIC.

### 8.1.1.Meta-analysis of summary statistics accounting for unobserved levels of sample relatedness and overlap

Due to the allelic and haplotypic landscape of isolated populations, their study can benefit from the creation of a custom reference panel for imputation of genotype array data. This paradigm has been successfully applied in large population-scale studies[117], as well as in the HELIC cohorts, where genotyping and imputation using 249 MANOLIS low-depth WGS

samples enabled the discovery of population-specific trait-associated variants[72]. This approach required meta-analysis of subsets of MANOLIS and Pomak samples typed using different chips, as well as the joint analysis of two isolates. Not accounting for relatedness or sample overlap across meta-analysed strata may lead to inflation of test statistics and to an excess of false positives. This issue is particularly relevant to large-scale GWAS meta-analyses, in which genotype level data are typically not available for sharing, and in which sample overlap or relatedness across cohorts needs to be adjusted for. To address this challenge, I developed METACARPA, a C++ implementation of a p-value and effect-size based meta-analysis that uses the tetrachoric correlation coefficient to account for unknown sample relatedness or overlap (Chapter 3). Due to its underlying assumptions, the tetrachoric coefficient is more robust to the presence of associated SNVs not under the null, and is hence advantageous in the presence of a highly polygenic signal. This makes it especially relevant as meta-analysis studies of increasing sample sizes continue to give an ever more detailed view of the genetic factors influencing complex traits. The METACARPA method was published, is available on github, and has to my knowledge been used by several ongoing meta-analysis efforts to date.

### 8.1.2.Very low-depth sequencing

One option to access low-frequency and rare variants at reduced costs compared to deep WGS is very low depth sequencing. As a proof of concept, I examined the accuracy of genotypes obtained through imputation of 1x WGS genotypes in the HELIC datasets[207](Chapter 4). When compared to high-depth sequencing, 1x achieves 90% concordance and sensitivity down to minor allele frequencies of 1%. These figures increase to 95% and 99%, respectively, when compared to the less dense OmniExpress and ExomeChip data.

Single-point analysis using the 1x data captured well-known common and low-frequency variant association loci such as *APOC3* for triglycerides and HDL, *UGT1A10* for bilirubin, *CETP* for blood lipid levels and *HBB* for haematological traits. In general, 1x genotypes recapitulated 96% of all associations found using imputed GWAS data, while discovering up to twice as

many signals that were also suggestively significant in a high-depth WGS based study in the same samples. Most of these additional signals were driven by rare or low-frequency variants. This study exemplifies the potential added value of very low depth sequencing for association studies, as 62% of low-frequency variants unique to the 1x dataset were true positives, which corresponded to 140,844 true additional low-frequency variants in MANOLIS. In MANOLIS, this allowed the discovery of an externally-replicating cardioprotective burden driven by two independent, low-frequency variants in the *APOC3* gene[111].

However, this increase in sensitivity comes with a high false positive rate, which is most evident in the rare variant category. 97% of variants called with the 1x data, but missed by GWAS and imputation, are rare, and 96% of these are false positives. This low specificity further underscores the need for any signal-driving variants to undergo thorough genotype validation, coupled with replication in external cohorts.

#### 8.1.3.Rare variant aggregation tests

This gene-based approach can be extended genome-wide using high-depth WGS, which allows the accurate description of the whole allelic spectrum. After establishing a quality control and association pipeline for this type of data (Chapter 5), I used high-depth WGS in HELIC (22x in MANOLIS, 18x in Pomak), to perform rare and low-frequency variant aggregation testing genome-wide using a range of variant selection and weighting methods[90], across the wealth of phenotype measurements available in both HELIC cohorts(Chapter 6). Tests were performed on a gene-by-gene basis in both exonic regions, regulatory regions linked to genes by eQTL overlap, and combined exonic and regulatory regions. Overall, the definition of this region of interest, rather than the functionality score used to assess a variant's pathogenicity, had the largest effect on burden testing p-values genome-wide.

Variants in four different genes (including *APOC3*) drove burden associations that passed quality control at study-wide significance, all of them in the MANOLIS cohort. Two of these loci (*ADIPOQ* for adiponectin, *UGT1A10* for bilirubin) had already been implicated in the

aetiology of the studied traits through common variant studies. In the case of *UGT1A10*, the burden was not independent of the strong common variant signal present at the locus, whereas for *ADIPOQ,* the previously documented signal was not present in MANOLIS, and the signal was driven by a combined effect of rare coding and regulatory variants. Further rare variant burden associations at suggestive significance in previously documented association loci (such as *GGT1* for gamma-glutamyltransferase and *ISL1* for insulin resistance) provide additional evidence for an effect of rare variants at common-variant loci. However, this is not always the case, as only one (*APOC3*) of 21 lipid loci that were first selected as potential positive controls for benchmarking the pipeline were found to harbour an aggregated rare variant signal, within the power limits of the current study. Conversely, the burden of rare variants private to MANOLIS in *FAM198B*, associated with increased triglyceride levels, was replicated in the larger, cosmopolitan INTERVAL cohort, in which it was driven by a different set of variants of a different functional class. Together, these results provide evidence for a role of low-frequency and rare, regulatory and coding variants in complex traits, and highlight the complex nature of locus- and population-specific allelic architectures at established and newly emerging signals.

#### 8.1.4.Structural variant calling in four European populations

Despite their increased density compared to 1x WGS and imputed GWAS, the high-depth WGS based variant callsets were restricted to single-nucleotide variants or short indels up to a few hundred basepairs long. CNVs, and in particular deletions, are an understudied class of structural variation that has been shown to play a role in the aetiology of complex traits[146], and that requires high sequencing depth to be accurately detected and genotyped. Most CNV detection algorithms require re-analysis of the entire read pool, however structural variant analyses are not yet part of standard variant calling pipelines. This means most existing WGS datasets might only have short variant calls available. I wrote UN-CNVc, a fast copy-number variant caller that detects and genotypes large (>5kb) CNVs based on SNV calls only, with increased specificity compared to existing methods (Chapter 7). Applied in the two HELIC cohorts as well as the TEENAGE and INTERVAL studies, it allowed the characterisation of the large CNV landscape in over 6,500 high-depth WGS samples. The

software correctly typed known structural variations, such as those spanning the *RHD* and *GTSM1* genes, and accurately detected very rare events, such as the recessive pathogenic deletion of the first 8 exons of the *ADAMTS19* gene in 4 MANOLIS samples.

A genome-wide association of over 50 quantitative traits available in HELIC, as well as 273 quantitative proteomics measurements, allowed further characterisation of the contribution of large CNVs to complex traits. Deletions of the *CCL3L3* and *NOMO1* genes were found to significantly decrease CCL3 and NOMO1 protein levels, respectively, in regions where single-point associations were present. In the case of *NOMO1*, two SNVs acted independently of each other and of the deletion to exert a combined protein level lowering effect. In contrast, for *CCL3L3,* a novel variant chr17:36448038 was in partial linkage disequilibrium with both the deletion and variants previously associated with CCL3. These results show that characterising structural variation can further improve our understanding of well-studied traits through the detection rate of novel quantitative trait loci, as well as shed light into structural variation underpinning known single-variant associations.

## 8.2.Future directions

### 8.2.1.Isolate studies

Isolated populations have proven their value in empowering single-variant association studies through the discovery of association signals driven by variants that were either unique to, or present at higher frequency in the founder population sample[72,117,204]. Similarly, the discovery of rare variant burden associations with the modest sample sizes found in HELIC has been made possible due to the special population genetics characteristics and reduced environmental variability of the cohorts under study. Rare variant signals, such as the ones discovered in *APOC3* and *FAM189B* in MANOLIS, are driven by variants with severe consequences that are rarer or absent in cosmopolitan populations. This demonstrates that the well-rehearsed power gains conferred by isolated cohorts in genome-wide association studies[204] extend to whole genome sequence-based rare variant association designs, and

that such studies are likely to remain relevant in the future. This is especially true since analysis pipelines employed for studying founder populations are largely transposable from those used in cosmopolitan cohorts. They differ mainly at the quality control stage, where consequences of isolatedness can affect quality indicators, and by the need to account for relatedness when associating genotypes to phenotypes.

### 8.2.2.Low and very low depth sequencing

As shown in Chapter 4, the reduced financial cost incurred at the sequencing stage for 1x studies is offset by a comparatively long pipeline and additional steps compared to imputed GWAS designs. This is mainly due to the computational cost of imputation, as well as the necessary experimental validation and QC of associated rare variants. The main advantage of very low depth sequencing is therefore the increased access to variants that it provides due to whole-genome coverage. However, as publicly-available, sequencing-based cosmopolitan reference panels increase in size[14], advantages of both custom reference panels and very low-depth may diminish, at least in populations represented in reference panels. Such approaches will however remain relevant alternatives to GWAS and imputation in diverse populations that are underrepresented in common reference panels. For low-frequency variants in particular, very low depth sequencing and imputation not only exhibits high genotyping accuracy, but is able to capture variants that are private to the population. As methods specific to low depth data continue to be developed in areas such as imputation and prediction of genotype quality, well-designed studies based on very low depth WGS should apply stringent MAF thresholds to focus on well-imputed low-frequency variants. This will streamline analysis pipelines, and enable them to exhaustively survey a sparsely described portion of the allelic spectrum potentially rich in phenotype-associated variants, at a fraction of the cost of deep WGS.

### 8.2.3.Rare variant aggregation tests

There is a general lack of best practice in the application of rare variant aggregation tests in genome-wide studies.

This thesis offers answers to some of the challenges associated with designing burden testing approaches, but more efforts are required to develop both the theoretical and practical aspects. Chapter 6 proposes a variant selection and testing framework centered around four types of allelic architectures that may affect complex traits: loss-of-function variants, exonic variants with an effect proportional to their predicted consequence, combinations of exonic and regulatory variants, and regulatory-only burdens.

Applying this framework in both HELIC isolates evidenced a role of rare regulatory variation in complex traits, both in combination with coding sequence variants and as part of purely regulatory burdens.

This work also highlighted the importance of the selection of variants and regions of interest. The analyses performed in HELIC were limited to a gene-centric approach, where only variants were included in the test that could be linked, either through disruption of the coding sequence or regulatory effects, to the function of a gene. This linking of regulatory regions to genes was performed in a tissue-agnostic way, which may decrease power by including regulatory variants not active in a tissue relevant to the trait under study. However, in many of the traits studied, the relevant tissues are not known. Weighting variants by tissue-specific functionality scores[208] can circumvent this problem, however this would require performing one test per tissue when the relevant cell type for a trait is unknown, with consequences on significance thresholds. Tests that allow the integration of multiple scores[209] may be helpful in this respect, but such methods have not yet been widely tested on real data.

Testing rare variants across biological pathways or genes contributing to protein complexes may be another relevant extension of the approach described here, although loss of power and computational intractability will ensue if the number of variants included is too high. Moreover, only the small fraction of intronic variants that were less than 50 base pairs away from exons, as well as those overlapping with regulatory elements, were considered for inclusion, making results highly dependent on the quality of the regulatory build. One possible solution could be the inclusion of all intronic variants, or even the expansion of the regions of interest beyond genes through window-based approaches. In both cases the

dependency would simply be shifted towards weighting methods, and power concerns remain. Non coding rare variant selection for inclusion in burden testing is an area under active development, and new methods both to define optimal regions of interest[210] and to assess functionality outside of genes[129,137] have been proposed. Recent developments, such as semi-supervised approaches that can incorporate tissue-specific data from massively parallel reporter assays[211] are a welcome step forward, but a thorough review of variant prediction methods is direly needed. Ultimately, the interpretation of both single point and rare variant aggregation test results in noncoding regions will require functional assays to link regions to causal pathways influencing measurable traits.

#### 8.2.4.Computing infrastructure

Another key area of development for enabling the next generation of WGS-based genetic association studies is the hardware and software architecture on which these analyses are run. Methods that allow automated parallelization on the software side, and flexible cloud computing on the hardware side, have been standard tools in many private industries for a number of years[212]. Innovative technological solutions for genetics research are a field under active development, with specialized field programmable gate array (FPGA) chips having recently been proposed for the rapid analysis of whole genomes[213]. Adoption of these technologies has been slow in the healthcare sector[214] and in research settings in particular. Their speed and efficiency has conferred a competitive advantage to early adopters in the early dissemination of research outcomes[215,216]. Concerns over costs and data privacy are often cited by researchers as reasons to favour traditional in-house computing facilities, but lack of familiarity with the methodological aspects of these technologies may also play a role[217]. In addition, the institutions more likely to have the funds to embrace decentralised computing may also be the ones with the least incentive to do so: the inevitable implementation costs and incompatibility issues with existing infrastructures will seem to validate past sizeable investments in in-house compute farms. A more pertinent issue is that of visibility and control over the storage, algorithms, hardware and platforms used in next-generation computing tools such as the Google Genomics platform, Amazon S3 storage or the Edico Dragen microchip. Despite the fact that the community itself has started to

develop powerful, scalable alternative cloud-compatible analysis frameworks[173,218], diminishing user insight into the inner workings of a piece of software has always been linked to its maturation as well as increasing abstraction and ease of use. In the future, institutions will need to find a suitable balance between data security and administrative burden for cloud-enabled research, so as not to discourage potential users. Equally, increasing awareness and training for both researchers and support staff will be key to driving up adoption, and ultimately, ameliorating the efficiency and pace of bioinformatics research.

#### 8.2.5.Genetic diversity and genome structure

Human geneticists performing association studies on SNVs have usually viewed genetic diversity in terms of different allele frequencies at fixed single-base-pair loci across population strata. The study of copy number variation presented in Chapter 7 shows that CNVs are omnipresent, and exhibit wide spatial heterogeneities across populations. Indeed, the presence of multiple overlapping deletions and duplications in the *ADAMTS19* and *CCL3* loci confirms that although several individuals in a sample usually exhibit deletions and duplications with the same approximate boundaries, variable regions usually exhibit more than one event with entirely different boundaries, as well as more complex composite events. This complex structure has been previously evidenced by studies of multiallelic CNVs(mCNV)[219], structural events that present with a different number of copies of a segment on each chromosomal copy. These specificities may explain why it has been so difficult to create a unified map of large structural variation in humans akin to dbSNP for point mutations. This heterogeneity also complicates the study of potential functional consequences for these events. As exemplified by the case of *CCL3*, taking the classical approach of studying the impact of a region's copy number, irrespective of the structural events underlying sequence abundance, can reduce statistical power in detecting phenotype associations instead of, for example, focusing on copy loss.

Calling of CNVs in silico from WGS data remains challenging, as computational methods generally lack in specificity and sensitivity[220], and comparison with databases of established

variants may overestimate the detection of most methods due to the population-specific structural heterogeneity mentioned above. Although WES has been widely used to describe the CNV landscape in humans[221], restricting in-silico analyses to exons does not seem to make the challenge of calling this class of variation any easier[222-224]. Furthermore, WES confers an imperfect coverage of the structural diversity even in genic regions, as evidenced by the case of *ADAMTS19* described in Chapter 7. Both of the two events that span exons of the gene are reported to span a much shorter region in exAC compared to WGS-based calls in our datasets.

Integrated methods that combine information from any mixture of array, WES or WGS data have recently been developed[225], however the gold standard for validation remains experimental, through methods such as array comparative genome hybridization (array CGH)[226,227] or droplet digital PCR (ddPCR)[219,228,229]. More recently, long read sequencing technologies such as graphene-based nanopore [230] or zero-mode waveguides[231] approaches, have been used to validate or detect large structural variants[232-234]. These technologies allow long range phasing through production of reads up to 800kb long[235], however they currently do not scale well to more than a handful of genomes.

### 8.2.6.Improving reference sequence, genome representation and annotation

The CNV work described in Chapter 7 highlights the fact that the human reference sequence is subject to the effects of large and complex rearrangements that yield inaccuracies and approximations. All study-wide significant phenotype-associated deletions identified in this work were located in assembly exceptions, confirming previous reports that CNVs are enriched in regions of low mappability[236]. More generally, any WGS association study will only ever be as good as the reference it uses, and improvements in reference genome assembly, aided notably through promising long read sequencing technologies[237,238], are vital to building a better reference and increasing the coverage of WGS studies to the true whole genome. In the long term, as CNVs and complex structural variants begin to be studied and described better, the limits of a linear sequence representation adopted since the early days of sequencing will be more evident. More fluid representations, such as graph-based

approaches, will be able to fully render the structural diversity of the genome[239-241]. In the meantime, large-scale WGS based studies should embrace the study of complex areas of the genome. The common practice of masking repeat-rich regions of the genome, assembly exceptions or segments particularly prone to structural variation potentially suppresses the discovery of medically-relevant results, such as the associations at the *CCL3L3* and *NOMO1* loci described in Chapter 7 and the rare variant burden in *FAM189B* (Chapter 6). UN-CNVc, through marker-level depth data, can clarify the high-level structure at complex loci, and powerful computational tools have been developed to accurately type the most convoluted and variation-prone regions of the genome, such as the MHC complex[242].

This dependence of future WGS studies on the quality of public resources is not restricted to functional prediction scores or the reference sequence. Between the time the HELIC burden work was performed and the writing of this thesis, a new version of the Ensembl regulatory build was released. Overlap between the two builds was poor, and although in some regulatory region boundaries simply changed by a few base pairs, some features were added and others removed entirely. Using the new build, the set of linked features for the set of genes suggestively significant in the burden analyses changed dramatically, indicating that results using the current (and future) builds would significantly change the signal landscape. This reflects the current state of research in regulatory annotation, and while some changes are expected as datasets go through iterations in curation, further crystallisation will be needed in order to improve replicability and robustness of results.

## 8.3.Conclusion

Larger-scale, cohort-wide, deep whole genome sequencing initiatives are poised to substantially enhance our understanding of the genetic underpinning of complex traits. Increasing sample sizes in cohort studies remains a main avenue for enabling the discovery subtler genetic effects through GWAS studies. However, methodological work in modelling and interpreting the effect of genetic variants on complex traits is urgently needed. In this thesis, I have shown that even at established quantitative trait loci, rare variant burdens, as well as CNVs can affect quantitative traits synergistically and independently of well-

established common variant signals. Isolated population studies can empower such discoveries even for rare variants, using comparatively small sample sizes. Genome-wide in-silico and experiment-based assembly, annotation and scoring methods are in their infancy despite considerable recent advances in the field. Stringent curation and collation of these resources will be critical to enable future studies that aim to move past the single-SNV, linear effects model of complex trait aetiology.

# References


1 Zhou, X. & Stephens, M. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet* 44, 821-824, doi:10.1038/ng.2310 (2012).

2 Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* 38, 904-909, doi:10.1038/ng1847 (2006).

3 McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 20, 1297-1303, doi:10.1101/gr.107524.110 (2010).

4 Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* 526, 68-74, doi:10.1038/nature15393 (2015).

5 Gamazon, E. R., Cox, N. J. & Davis, L. K. Structural architecture of SNP effects on complex traits. *Am J Hum Genet* 95, 477-489, doi:10.1016/j.ajhg.2014.09.009 (2014).

6 Mace, A. *et al.* CNV-association meta-analysis in 191,161 European adults reveals new loci associated with anthropometric traits. *Nat Commun* 8, 744, doi:10.1038/s41467-017-00556-x (2017).

7 Escaramis, G., Docampo, E. & Rabionet, R. A decade of structural variants: description, history and methods to detect structural variation. *Brief Funct Genomics* 14, 305-314, doi:10.1093/bfgp/elv014 (2015).

8 Ott, J., Wang, J. & Leal, S. M. Genetic linkage analysis in the age of whole-genome sequencing. *Nat Rev Genet* 16, 275-284, doi:10.1038/nrg3908 (2015).

9 Siontis, K. C., Patsopoulos, N. A. & Ioannidis, J. P. Replication of past candidate loci for common diseases and phenotypes in 100 genome-wide association studies. *Eur J Hum Genet* 18, 832-837, doi:10.1038/ejhg.2010.26 (2010).

10 Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* 322, 881-888, doi:10.1126/science.1156409 (2008).

11 Karlsson, E. K., Kwiatkowski, D. P. & Sabeti, P. C. Natural selection and infectious disease in human populations. *Nat Rev Genet* 15, 379-393, doi:10.1038/nrg3734 (2014).

12 Welter, D. *et al.* The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res* 42, D1001-1006, doi:10.1093/nar/gkt1229 (2014).

13 Marigorta, U. M., Rodriguez, J. A., Gibson, G. & Navarro, A. Replicability and Prediction: Lessons and Challenges from GWAS. *Trends Genet* 34, 504-517, doi:10.1016/j.tig.2018.03.005 (2018).

14 McCarthy, S. *et al.* A reference panel of 64,976 haplotypes for genotype imputation. *Nat Genet* 48, 1279-1283, doi:10.1038/ng.3643 (2016).

15 Yang, J. *et al.* Genetic variance estimation with imputed variants finds negligible missing heritability for human height and body mass index. *Nat Genet* 47, 1114-1120, doi:10.1038/ng.3390 (2015).

16 Charlesworth, B. Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. *Nat Rev Genet* 10, 195-205, doi:10.1038/nrg2526 (2009).

17 Varilo, T. & Peltonen, L. Isolates and their potential use in complex gene mapping efforts. *Curr Opin Genet Dev* 14, 316-323, doi:10.1016/j.gde.2004.04.008 (2004).

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

19 Arcos-Burgos, M. & Muenke, M. Genetics of population isolates. *Clin Genet* 61, 233-247 (2002).

20 Norio, R. Finnish Disease Heritage I: characteristics, causes, background. *Hum Genet* 112, 441-456, doi:10.1007/s00439-002-0875-3 (2003).

21 Charrow, J. Ashkenazi Jewish genetic disorders. *Fam Cancer* 3, 201-206, doi:10.1007/s10689-004-9545-z (2004).

22 Otto, S. P. & Whitlock, M. C. The probability of fixation in populations of changing size. *Genetics* 146, 723-733 (1997).

23 Holm, H. *et al.* A rare variant in MYH6 is associated with high risk of sick sinus syndrome. *Nat Genet* 43, 316-320, doi:10.1038/ng.781 (2011).

24 Sulem, P. *et al.* Identification of low-frequency variants associated with gout and serum uric acid levels. *Nat Genet* 43, 1127-1130, doi:10.1038/ng.972 (2011).

25 Jonsson, T. *et al.* A mutation in APP protects against Alzheimer's disease and age-related cognitive decline. *Nature* 488, 96-99, doi:10.1038/nature11283 (2012).

26 Gudmundsson, J. *et al.* A study based on whole-genome sequencing yields a rare variant at 8q24 associated with prostate cancer. *Nat Genet* 44, 1326-1329, doi:10.1038/ng.2437 (2012).

27 Kurki, M. I. *et al.* High risk population isolate reveals low frequency variants predisposing to intracranial aneurysms. *PLoS Genet* 10, e1004134, doi:10.1371/journal.pgen.1004134 (2014).

28 Lencz, T. *et al.* Genome-wide association study implicates NDST3 in schizophrenia and bipolar disorder. *Nat Commun* 4, 2739, doi:10.1038/ncomms3739 (2013).

29 Tachmazidou, I. *et al.* A rare functional cardioprotective APOC3 variant has risen in frequency in distinct population isolates. *Nat Commun* 4, 2872, doi:10.1038/ncomms3872 (2013).

30 Pollin, T. I. *et al.* A null mutation in human APOC3 confers a favorable plasma lipid profile and apparent cardioprotection. *Science* 322, 1702-1705, doi:10.1126/science.1161524 (2008).

31 Kristiansson, K., Naukkarinen, J. & Peltonen, L. Isolated populations and complex disease gene identification. *Genome Biol*/9, 109, doi:10.1186/gb-2008-9-8-109 (2008).

32 Terwilliger, J. D., Zollner, S., Laan, M. & Paabo, S. Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum Hered* 48, 138-154 (1998).

33 Stephens, J. C. & Bamshad, M. Population choice as a consideration for genetic analysis study design. *Cold Spring Harb Protoc* 2011, 917-922, doi:10.1101/pdb.top122 (2011).

34 Zuk, O. *et al.* Searching for missing heritability: designing rare variant association studies. *Proc Natl Acad Sci U S A* 111, E455-464, doi:10.1073/pnas.1322563111 (2014).

35 Wright, A. F., Carothers, A. D. & Pirastu, M. Population choice in mapping genes for complex diseases. *Nat Genet* 23, 397-404, doi:10.1038/70501 (1999).

36 Huyghe, J. R. *et al.* Exome array analysis identifies new loci and low-frequency variants influencing insulin processing and secretion. *Nat Genet* 45, 197-201, doi:10.1038/ng.2507 (2013).

37 Kiezun, A. *et al.* Exome sequencing and the genetic basis of complex traits. *Nat Genet* 44, 623-630, doi:10.1038/ng.2303 (2012).

38 Do, R., Kathiresan, S. & Abecasis, G. R. Exome sequencing and complex disease: practical aspects of rare variant association studies. *Hum Mol Genet* 21, R1-9, doi:10.1093/hmg/dds387 (2012).

39 Le, S. Q. & Durbin, R. SNP detection and genotyping from low-coverage sequencing data on multiple diploid samples. *Genome Res* 21, 952-960, doi:10.1101/gr.113084.110 (2011).

40 Li, Y., Sidore, C., Kang, H. M., Boehnke, M. & Abecasis, G. R. Low-coverage sequencing: implications for design of complex trait association studies. *Genome Res* 21, 940-951, doi:10.1101/gr.117259.110 (2011).

41 Zeggini, E. Next-generation association studies for complex traits. *Nat Genet* 43, 287-288, doi:10.1038/ng0411-287 (2011).

42 Kang, H. M. *et al.* Efficient control of population structure in model organism association mapping. *Genetics* 178, 1709-1723, doi:10.1534/genetics.107.080101 (2008).

43 Aulchenko, Y. S., Ripke, S., Isaacs, A. & van Duijn, C. M. GenABEL: an R library for genome-wide association analysis. *Bioinformatics* 23, 1294-1296, doi:10.1093/bioinformatics/btm108 (2007).

44 Zhang, Z. *et al.* Mixed linear model approach adapted for genome-wide association studies. *Nat Genet* 42, 355-360, doi:10.1038/ng.546 (2010).

45 Kang, H. M. *et al.* Variance component model to account for sample structure in genome-wide association studies. *Nat Genet* 42, 348-354, doi:10.1038/ng.548 (2010).

46 Lippert, C. *et al.* FaST linear mixed models for genome-wide association studies. *Nat Methods* 8, 833-835, doi:10.1038/nmeth.1681 (2011).

47 Chen, H. *et al.* Comparison of statistical approaches to rare variant analysis for quantitative traits. *BMC Proc* 5 Suppl 9, S113, doi:10.1186/1753-6561-5-S9-S113 (2011).

48 Asimit, J. & Zeggini, E. Rare variant association analysis methods for complex traits. *Annu Rev Genet* 44, 293-308, doi:10.1146/annurev-genet-102209-163421 (2010).

49 Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nat Rev Genet* 11, 773-785, doi:10.1038/nrg2867 (2010).

50 Basu, S. & Pan, W. Comparison of statistical tests for disease association with rare variants. *Genet Epidemiol* 35, 606-619, doi:10.1002/gepi.20609 (2011).

51 Chen, H., Meigs, J. B. & Dupuis, J. Sequence kernel association test for quantitative traits in family samples. *Genet Epidemiol* 37, 196-204, doi:10.1002/gepi.21703 (2013).

52 Jiang, D. & McPeek, M. S. Robust rare variant association testing for quantitative traits in samples with related individuals. *Genet Epidemiol* 38, 10-20, doi:10.1002/gepi.21775 (2014).

53 Wang, X., Lee, S., Zhu, X., Redline, S. & Lin, X. GEE-based SNP set association test for continuous and discrete traits in family-based association studies. *Genet Epidemiol* 37, 778-786, doi:10.1002/gepi.21763 (2013).

54 Listgarten, J. *et al.* A powerful and efficient set test for genetic markers that handles confounders. *Bioinformatics* 29, 1526-1533, doi:10.1093/bioinformatics/btt177 (2013).

55 Oualkacha, K. *et al.* Adjusted sequence kernel association test for rare variants controlling for cryptic and family relatedness. *Genet Epidemiol* 37, 366-376, doi:10.1002/gepi.21725 (2013).

56 Carmi, S. *et al.* The variance of identity-by-descent sharing in the Wright-Fisher model. *Genetics* 193, 911-928, doi:10.1534/genetics.112.147215 (2013).

57 Powell, J. E., Visscher, P. M. & Goddard, M. E. Reconciling the analysis of IBD and IBS in complex trait studies. *Nat Rev Genet* 11, 800-805, doi:10.1038/nrg2865 (2010).

58 Zeggini, E. & Ioannidis, J. P. Meta-analysis in genome-wide association studies. *Pharmacogenomics* 10, 191-201, doi:10.2217/14622416.10.2.191 (2009).

59 Asimit, J. *et al.* An evaluation of different meta-analysis approaches in the presence of allelic heterogeneity. *Eur J Hum Genet* 20, 709-712, doi:10.1038/ejhg.2011.274 (2012).

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

61 Asimakopoulou, F. *The Muslim Minority of Thrace and the Greco-Turkish Relations*. 211-339 (Livanis, 2002).

62 Weekes, R. *Muslim peoples. A world ethnographic survey*. (Greenwood, 1984).

63 Ghodsee, K. R. *Muslim Lives in Eastern Europe: Gender, Ethnicity, and the Transformation of Islam in Postsocialist Bulgaria*. (Princeton University Press, 2009).

64 Christopoulos, K. T. D. *Minorities in Greece*. (Kritiki, 1997).

65 Southam, L. *et al.* Whole genome sequencing and imputation in two Greek isolated populations identifies associations with complex traits of medical importance. *Nat Comms* in review (2017).

66 Moayyeri, A., Hammond, C. J., Hart, D. J. & Spector, T. D. The UK Adult Twin Registry (TwinsUK Resource). *Twin Res Hum Genet* 16, 144-149, doi:10.1017/thg.2012.89 (2013).

67 Golding, J., Pembrey, M., Jones, R. & Team, A. S. ALSPAC--the Avon Longitudinal Study of Parents and Children. I. Study methodology. *Paediatric and perinatal epidemiology* 15, 74-87 (2001).

68 Moore, C. *et al.* The INTERVAL trial to determine whether intervals between blood donations can be safely and acceptably decreased to optimise blood supply: study protocol for a randomised controlled trial. *Trials* 15, 363, doi:10.1186/1745-6215-15-363 (2014).

69 Province, M. A. & Borecki, I. B. A correlated meta-analysis strategy for data mining "OMIC" scans. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, 236-246 (2013).

70 Lin, D. Y. & Sullivan, P. F. Meta-analysis of genome-wide association studies with overlapping subjects. *Am J Hum Genet* 85, 862-872, doi:10.1016/j.ajhg.2009.11.001 (2009).

71 Panoutsopoulou, K. *et al.* Genetic characterization of Greek population isolates reveals strong genetic drift at missense and trait-associated variants. *Nat Commun* 5, 5345, doi:10.1038/ncomms6345 (2014).

72 Southam, L. *et al.* Whole genome sequencing and imputation in isolated populations identify genetic associations with medically-relevant complex traits. *Nat Commun* 8, 15606, doi:10.1038/ncomms15606 (2017).

73 Zhou, X. & Stephens, M. Efficient multivariate linear mixed model algorithms for genome-wide association studies. *Nature methods* 11, 407-409, doi:10.1038/nmeth.2848 (2014).

74 Teslovich, T. M. *et al.* Biological, clinical and population relevance of 95 loci for blood lipids. *Nature* 466, 707-713, doi:10.1038/nature09270 (2010).

75 Magi, R. & Morris, A. P. GWAMA: software for genome-wide association meta-analysis. *BMC Bioinformatics* 11, 288, doi:10.1186/1471-2105-11-288 (2010).

76 Consortium, U. K. *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 526, 82-90, doi:10.1038/nature14962 (2015).

77 Astle, W. J. *et al.* The Allelic Landscape of Human Blood Cell Trait Variation and Links to Common Complex Disease. *Cell* 167, 1415-1429 e1419, doi:10.1016/j.cell.2016.10.042 (2016).

78 Pasaniuc, B. *et al.* Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat Genet* 44, 631-635, doi:10.1038/ng.2283 (2012).

79 consortium, C. Sparse whole-genome sequencing identifies two loci for major depressive disorder. *Nature* 523, 588-591, doi:10.1038/nature14659 (2015).

80 Luo, Y. *et al.* Exploring the genetic architecture of inflammatory bowel disease by whole-genome sequencing identifies association at ADCY7. *Nat Genet* 49, 186-192, doi:10.1038/ng.3761 (2017).

81 Delaneau, O., Howie, B., Cox, A. J., Zagury, J. F. & Marchini, J. Haplotype estimation using sequencing reads. *Am J Hum Genet* 93, 687-696, doi:10.1016/j.ajhg.2013.09.002 (2013).

82 Browning, S. R. & Browning, B. L. Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 81, 1084-1097, doi:10.1086/521987 (2007).

83 UK10K Consortium *et al.* The UK10K project identifies rare variants in health and disease. *Nature* 526, 82-90, doi:10.1038/nature14962 (2015).

84 Browning, B. L. Private communication. (2014).

85 Arthur, R., Schulz-Trieglaff, O., Cox, A. J. & O'Connell, J. AKT: ancestry and kinship toolkit. *Bioinformatics* 33, 142-144, doi:10.1093/bioinformatics/btw576 (2017).

86 Hao, K., Li, C., Rosenow, C. & Wong, W. H. Detect and adjust for population stratification in population-based association study using genomic control markers: an application of Affymetrix Genechip Human Mapping 10K array. *Eur J Hum Genet* 12, 1001-1006, doi:10.1038/sj.ejhg.5201273 (2004).

87 Li, M. X., Yeung, J. M., Cherny, S. S. & Sham, P. C. Evaluating the effective numbers of independent tests and significant p-value thresholds in commercial genotyping arrays and public imputation reference datasets. *Hum Genet* 131, 747-756, doi:10.1007/s00439-011-1118-2 (2012).

88 Xu, C. *et al.* Estimating genome-wide significance for whole-genome sequencing studies. *Genet Epidemiol* 38, 281-290, doi:10.1002/gepi.21797 (2014).

89 Chang, C. C. *et al.* Second-generation PLINK: rising to the challenge of larger and richer datasets. *Gigascience* 4, 7, doi:10.1186/s13742-015-0047-8 (2015).

90 Gilly, A. *et al.* Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. *bioRxiv* (2018).

91 Gurdasani, D. *et al.* The African Genome Variation Project shapes medical genetics in Africa. *Nature* 517, 327-332, doi:10.1038/nature13997 (2015).

92 Weiss, E. *et al.* Farming, Foreign Holidays, and Vitamin D in Orkney. *PLoS One* 11, e0155633, doi:10.1371/journal.pone.0155633 (2016).

93 Laakso, M. *et al.* The Metabolic Syndrome in Men study: a resource for studies of metabolic and cardiovascular diseases. *J Lipid Res* 58, 481-493, doi:10.1194/jlr.O072629 (2017).

94 Danjou, F. *et al.* Genome-wide association analyses based on whole-genome sequencing in Sardinia provide insights into regulation of hemoglobin levels. *Nat Genet* 47, 1264-1271, doi:10.1038/ng.3307 (2015).

95 Esko, T. *et al.* Genetic characterization of northeastern Italian population isolates in the context of broader European genetic diversity. *Eur J Hum Genet* 21, 659-665, doi:10.1038/ejhg.2012.229 (2013).

96 Granot-Hershkovitz, E. *et al.* A study of Kibbutzim in Israel reveals risk factors for cardiometabolic traits and subtle population structure. *Eur J Hum Genet*, doi:10.1038/s41431-018-0230-3 (2018).

97 Wu, M. C. *et al.* Rare-variant association testing for sequencing data with the sequence kernel association test. *Am J Hum Genet* 89, 82-93, doi:10.1016/j.ajhg.2011.05.029 (2011).

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

99 Abecasis, G. R. *et al.* An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56-65, doi:10.1038/nature11632 (2012).

100 Ballman, E. S., Rugman-Jones, P. F., Stouthamer, R. & Hoddle, M. S. Genetic structure of Graphocephala atropunctata (Hemiptera: Cicadellidae) populations across its natural range in California reveals isolation by distance. *J Econ Entomol* 104, 279-287 (2011).

101 http://exac.broadinstitute.com. Exome Aggregation Consortium (ExAC), Cambridge, MA. *accessed april 2015*.

102 McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* 26, 2069-2070, doi:10.1093/bioinformatics/btq330 (2010).

103 Tg and Hdl Working Group of the Exome Sequencing Project, N. H. L. B. I. *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *The New England journal of medicine* 371, 22-31, doi:10.1056/NEJMoa1307095 (2014).

104 Li, A. H. *et al.* Analysis of loss-of-function variants and 20 risk factor phenotypes in 8,554 individuals identifies loci influencing chronic disease. *Nat Genet* 47, 640-642, doi:10.1038/ng.3270 (2015).

105 Browning, B. L. & Browning, S. R. Genotype Imputation with Millions of Reference Samples. *Am J Hum Genet* 98, 116-126, doi:10.1016/j.ajhg.2015.11.020 (2016).

106 Bycroft, C. *et al.* Genome-wide genetic data on ~500,000 UK Biobank participants. *bioRxiv* (2017).

107 Davies, R. W., Flint, J., Myers, S. & Mott, R. Rapid genotype imputation from sequence without reference panels. *Nat Genet* 48, 965-969, doi:10.1038/ng.3594 (2016).

108 ←https://gencove.com→ (

109 Livne, O. E. *et al.* PRIMAL: Fast and accurate pedigree-based imputation from sequence data in a founder population. *PLoS Comput Biol* 11, e1004139, doi:10.1371/journal.pcbi.1004139 (2015).

110 Herzig, A. F. *et al.* Strategies for phasing and imputation in a population isolate. *Genet Epidemiol* 42, 201-213, doi:10.1002/gepi.22109 (2018).

111 Gilly, A. *et al.* Very low-depth sequencing in a founder population identifies a cardioprotective APOC3 signal missed by genome-wide imputation. *Hum Mol Genet* 25, 2360-2365, doi:10.1093/hmg/ddw088 (2016).

112 Zheng-Bradley, X. *et al.* Alignment of 1000 Genomes Project reads to reference assembly GRCh38. *Gigascience* 6, 1-8, doi:10.1093/gigascience/gix038 (2017).

113 Arthur, R., O'Connell, J., Schulz-Trieglaff, O. & Cox, A. J. Rapid genotype refinement for whole-genome sequencing data using multi-variate normal distributions. *Bioinformatics* 32, 2306-2312, doi:10.1093/bioinformatics/btw097 (2016).

114 Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am J Hum Genet* 91, 839-848, doi:10.1016/j.ajhg.2012.09.004 (2012).

115 Gilly, A. *et al.* Cohort-wide deep whole genome sequencing and the allelic architecture of complex traits. *Nat Commun* 9, 4674, doi:10.1038/s41467-018-07070-8 (2018).

116 MacArthur, D. G. *et al.* A systematic survey of loss-of-function variants in human protein-coding genes. *Science* 335, 823-828, doi:10.1126/science.1215040 (2012).

117 Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nat Genet* 47, 435-444, doi:10.1038/ng.3247 (2015).

118 Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* 536, 285-291, doi:10.1038/nature19057 (2016).

119 Zerbino, D. R. *et al.* Ensembl 2018. *Nucleic Acids Res* 46, D754-D761, doi:10.1093/nar/gkx1098 (2018).

120 Tanaka, T. *et al.* A genome-wide association analysis of serum iron concentrations. *Blood* 115, 94-96, doi:10.1182/blood-2009-07-232496 (2010).

121 Chambers, J. C. *et al.* Genome-wide association study identifies loci influencing concentrations of liver enzymes in plasma. *Nat Genet* 43, 1131-1138, doi:10.1038/ng.970 (2011).

122 Porcu, E. *et al.* A meta-analysis of thyroid-related traits reveals novel loci and gender-specific differences in the regulation of thyroid function. *PLoS Genet* 9, e1003266, doi:10.1371/journal.pgen.1003266 (2013).

123 Elliott, P. *et al.* Genetic Loci associated with C-reactive protein levels and risk of coronary heart disease. *JAMA* 302, 37-48, doi:10.1001/jama.2009.954 (2009).

124 Reiner, A. P. *et al.* Genome-wide association and population genetic analysis of C-reactive protein in African American and Hispanic American women. *Am J Hum Genet* 91, 502-512, doi:10.1016/j.ajhg.2012.07.023 (2012).

125 Moore, C. B. *et al.* Phenome-wide Association Study Relating Pretreatment Laboratory Parameters With Human Genetic Variants in AIDS Clinical Trials

Group Protocols. *Open Forum Infect Dis* 2, ofu113, doi:10.1093/ofid/ofu113 (2015).

126 Brody, J. A. *et al.* Analysis commons, a team approach to discovery in a big-data environment for genetic epidemiology. *Nat Genet* 49, 1560-1563, doi:10.1038/ng.3968 (2017).

127 Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. An application and empirical comparison of statistical analysis methods for associating rare variants to a complex phenotype. *Pac Symp Biocomput*, 76-87 (2011).

128 Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet* 46, 310-315, doi:10.1038/ng.2892 (2014).

129 Ionita-Laza, I., McCallum, K., Xu, B. & Buxbaum, J. D. A spectral approach integrating functional genomic annotations for coding and noncoding variants. *Nat Genet* 48, 214-220, doi:10.1038/ng.3477 (2016).

130 Morris, A. P. & Zeggini, E. An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34, 188-193, doi:10.1002/gepi.20450 (2010).

131 Lee, S. *et al.* Optimal unified approach for rare-variant association testing with application to small-sample case-control whole-exome sequencing studies. *Am J Hum Genet* 91, 224-237, doi:10.1016/j.ajhg.2012.06.007 (2012).

132 Svishcheva, G. R., Belonogova, N. M. & Axenovich, T. I. FFBSKAT: fast family-based sequence kernel association test. *PLoS One* 9, e99407, doi:10.1371/journal.pone.0099407 (2014).

133 Eu-Ahsunthornwattana, J. *et al.* Comparison of methods to account for relatedness in genome-wide association studies with family-based data. *PLoS Genet* 10, e1004445, doi:10.1371/journal.pgen.1004445 (2014).

134 Manichaikul, A. *et al.* Robust relationship inference in genome-wide association studies. *Bioinformatics* 26, 2867-2873, doi:10.1093/bioinformatics/btq559 (2010).

135 McLaren, W. *et al.* The Ensembl Variant Effect Predictor. *Genome Biol* 17, 122, doi:10.1186/s13059-016-0974-4 (2016).

136 GTEx Consortium. Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science* 348, 648-660, doi:10.1126/science.1262110 (2015).

137 Huang, Y. F., Gulko, B. & Siepel, A. Fast, scalable prediction of deleterious noncoding variants from functional and population genomic data. *Nat Genet* 49, 618-624, doi:10.1038/ng.3810 (2017).

138 Li, J. & Ji, L. Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity (Edinb)* 95, 221-227, doi:10.1038/sj.hdy.6800717 (2005).

139 Hinrichs, A. L., Larkin, E. K. & Suarez, B. K. Population stratification and patterns of linkage disequilibrium. *Genet Epidemiol* 33 Suppl 1, S88-92, doi:10.1002/gepi.20478 (2009).

140 TG and HDL Working Group of the Exome Sequencing Project, N. H. L. *et al.* Loss-of-function mutations in APOC3, triglycerides, and coronary disease. *N Engl J Med* 371, 22-31, doi:10.1056/NEJMoa1307095 (2014).

141 van Es, H. H. *et al.* Assignment of the human UDP glucuronosyltransferase gene (UGT1A1) to chromosome region 2q37. *Cytogenet Cell Genet* 63, 114-116 (1993).

142 Sanna, S. *et al.* Common variants in the SLCO1B3 locus are associated with bilirubin levels and unconjugated hyperbilirubinemia. *Hum Mol Genet* 18, 2711-2718, doi:10.1093/hmg/ddp203 (2009).

143 Vasseur, F. *et al.* Single-nucleotide polymorphism haplotypes in the both proximal promoter and exon 3 of the APM1 gene modulate adipocyte-secreted adiponectin hormone levels and contribute to the genetic risk for type 2 diabetes in French Caucasians. *Hum Mol Genet* 11, 2607-2614 (2002).

144 Zhang, H. *et al.* The LIM-homeodomain protein ISL1 activates insulin gene promoter directly through synergy with BETA2. *J Mol Biol* 392, 566-577, doi:10.1016/j.jmb.2009.07.036 (2009).

145 Ediger, B. N. *et al.* Islet-1 Is essential for pancreatic beta-cell function. *Diabetes* 63, 4206-4217, doi:10.2337/db14-0096 (2014).

146 Zarrei, M., MacDonald, J. R., Merico, D. & Scherer, S. W. A copy number variation map of the human genome. *Nat Rev Genet* 16, 172-183, doi:10.1038/nrg3871 (2015).

147 Monlong, J. *et al.* Global characterization of copy number variants in epilepsy patients from whole genome sequencing. *PLoS Genet* 14, e1007285, doi:10.1371/journal.pgen.1007285 (2018).

148 Striano, P. *et al.* Clinical significance of rare copy number variations in epilepsy: a case-control survey using microarray-based comparative genomic hybridization. *Arch Neurol* 69, 322-330, doi:10.1001/archneurol.2011.1999 (2012).

149 McCarroll, S. A. *et al.* Deletion polymorphism upstream of IRGM associated with altered IRGM expression and Crohn's disease. *Nat Genet* 40, 1107-1112, doi:10.1038/ng.215 (2008).

150 International Schizophrenia, C. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* 455, 237-241, doi:10.1038/nature07239 (2008).

151 Hosak, L., Silhan, P. & Hosakova, J. Genomic copy number variations: A breakthrough in our knowledge on schizophrenia etiology? *Neuro Endocrinol Lett* 33, 183-190 (2012).

152 Bochukova, E. G. *et al.* Large, rare chromosomal deletions associated with severe early-onset obesity. *Nature* 463, 666-670, doi:10.1038/nature08689 (2010).

153 Stefansson, H. *et al.* CNVs conferring risk of autism or schizophrenia affect cognition in controls. *Nature* 505, 361-366, doi:10.1038/nature12818 (2014).

154 Zekavat, S. M. *et al.* Deep coverage whole genome sequences and plasma lipoprotein(a) in individuals of European and African ancestries. *Nat Commun* 9, 2606, doi:10.1038/s41467-018-04668-w (2018).

155 Zhao, M., Wang, Q., Wang, Q., Jia, P. & Zhao, Z. Computational tools for copy number variation (CNV) detection using next-generation sequencing data: features and perspectives. *BMC Bioinformatics* 14 Suppl 11, S1, doi:10.1186/1471-2105-14-S11-S1 (2013).

156 Hwang, M. Y. *et al.* Combinatorial approach to estimate copy number genotype using whole-exome sequencing data. *Genomics* 105, 145-149, doi:10.1016/j.ygeno.2014.12.003 (2015).

157 Pirooznia, M., Goes, F. S. & Zandi, P. P. Whole-genome CNV analysis: advances in computational approaches. *Front Genet* 6, 138, doi:10.3389/fgene.2015.00138 (2015).

158 Lu, J. *et al.* Assessing genome-wide copy number variation in the Han Chinese population. *J Med Genet* 54, 685-692, doi:10.1136/jmedgenet-2017-104613 (2017).

159 Kayser, K. *et al.* Copy number variation analysis and targeted NGS in 77 families with suspected Lynch syndrome reveals novel potential causative genes. *Int J Cancer*, doi:10.1002/ijc.31725 (2018).

160 Alex Buerkle, C. & Gompert, Z. Population genomics based on low coverage sequencing: how low should we go? *Mol Ecol* 22, 3028-3035, doi:10.1111/mec.12105 (2013).

161 Wang, K. *et al.* PennCNV: an integrated hidden Markov model designed for high-resolution copy number variation detection in whole-genome SNP genotyping data. *Genome Res* 17, 1665-1674, doi:10.1101/gr.6861907 (2007).

162 Kumasaka, N. *et al.* PlatinumCNV: a Bayesian Gaussian mixture model for genotyping copy number polymorphisms using SNP array signal intensity data. *Genet Epidemiol* 35, 831-844, doi:10.1002/gepi.20633 (2011).

163 Selvanayagam, T. *et al.* Genome-wide copy number variation analysis identifies novel candidate loci associated with pediatric obesity. *Eur J Hum Genet*, doi:10.1038/s41431-018-0189-0 (2018).

164 do Nascimento, F. & Guimaraes, K. S. Copy Number Variations Detection: Unravelling the Problem in Tangible Aspects. *IEEE/ACM Trans Comput Biol Bioinform* 14, 1237-1250, doi:10.1109/TCBB.2016.2576441 (2017).

165 Flegel, W. A. The genetics of the Rhesus blood group system. *Blood Transfus* 5, 50-57, doi:10.2450/2007.0011-07 (2007).

166 Petrovic, D. & Peterlin, B. GSTM1-null and GSTT1-null genotypes are associated with essential arterial hypertension in patients with type 2 diabetes. *Clin Biochem* 47, 574-577, doi:10.1016/j.clinbiochem.2014.03.012 (2014).

167 Grubisa, I. *et al.* Combined GSTM1 and GSTT1 null genotypes are strong risk factors for atherogenesis in a Serbian population. *Genet Mol Biol* 41, 35-40, doi:10.1590/1678-4685-GMB-2017-0034 (2018).

168 Wang, M., Li, Y., Lin, L., Song, G. & Deng, T. GSTM1 Null Genotype and GSTP1 Ile105Val Polymorphism Are Associated with Alzheimer's Disease: a Meta-Analysis. *Mol Neurobiol* 53, 1355-1364, doi:10.1007/s12035-015-9092-7 (2016).

169 Nath, S., Das, S., Bhowmik, A., Ghosh, S. K. & Choudhury, Y. The GSTM1 and GSTT1 null genotypes increase the risk for Type 2 diabetes mellitus and the subsequent development of diabetic complications: A meta-analysis. *Curr Diabetes Rev*, doi:10.2174/1573399814666171215120228 (2017).

170 Wünnemann, F. *et al.* in *American Society of Human Genetics Meeting.*

171 Galtier, N., Piganeau, G., Mouchiroud, D. & Duret, L. GC-content evolution in mammalian genomes: the biased gene conversion hypothesis. *Genetics* 159, 907-911 (2001).

172 Sakharkar, M. K., Chow, V. T. & Kangueane, P. Distributions of exons and introns in the human genome. *In Silico Biol* 4, 387-393 (2004).

173 Ivashchenko, A. T., Khailenko, V. A. & Atambaeva Sh, A. [Variations of the length of exons and introns in human genome genes]. *Genetika* 45, 22-29 (2009).

174 Townson, J. R., Barcellos, L. F. & Nibbs, R. J. Gene copy number regulates the production of the human chemokine CCL3-L1. *Eur J Immunol* 32, 3016-3026, doi:10.1002/1521-4141(2002010)32:10←3016::AID-IMMU3016→3.0.CO;2-D (2002).

175 Ben Kilani, M. S. *et al.* Characterization of copy number variants for CCL3L1 gene in rheumatoid arthritis for French trio families and Tunisian cases and controls. *Clin Rheumatol* 35, 1917-1922, doi:10.1007/s10067-015-3156-y (2016).

176 Nordang, G. B. *et al.* Association analysis of the CCL3L1 copy number locus by paralogue ratio test in Norwegian rheumatoid arthritis patients and healthy controls. *Genes Immun* 13, 579-582, doi:10.1038/gene.2012.30 (2012).

177 Aklillu, E. *et al.* CCL3L1 copy number, HIV load, and immune reconstitution in sub-Saharan Africans. *BMC Infect Dis* 13, 536, doi:10.1186/1471-2334-13-536 (2013).

178 Carpenter, D., Farnert, A., Rooth, I., Armour, J. A. & Shaw, M. A. CCL3L1 copy number and susceptibility to malaria. *Infect Genet Evol* 12, 1147-1154, doi:10.1016/j.meegid.2012.03.021 (2012).

179 Carpenter, D., McIntosh, R. S., Pleass, R. J. & Armour, J. A. Functional effects of CCL3L1 copy number. *Genes Immun* 13, 374-379, doi:10.1038/gene.2012.5 (2012).

180 Sudmant, P. H. *et al.* Diversity of human copy number variation and multicopy genes. *Science* 330, 641-646, doi:10.1126/science.1197005 (2010).

181 Rimoin, D. L., Pyeritz, R. E. & Korf, B. R. Emery and Rimoin's principles and practice of medical genetics. (2013).

182 Ahola-Olli, A. V. *et al.* Genome-wide Association Study Identifies 27 Loci Influencing Concentrations of Circulating Cytokines and Growth Factors. *Am J Hum Genet* 100, 40-50, doi:10.1016/j.ajhg.2016.11.007 (2017).

183 Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet* 4, e1000072, doi:10.1371/journal.pgen.1000072 (2008).

184 Srivastava, S. *et al.* Serum human trefoil factor 3 is a biomarker for mucosal healing in ulcerative colitis patients with minimal disease activity. *J Crohns Colitis* 9, 575-579, doi:10.1093/ecco-jcc/jjv075 (2015).

185 Muro, M., Lopez-Hernandez, R. & Mrowiec, A. Immunogenetic biomarkers in inflammatory bowel diseases: role of the IBD3 region. *World J Gastroenterol* 20, 15037-15048, doi:10.3748/wjg.v20.i41.15037 (2014).

186 Ka, S. *et al.* HLAscan: genotyping of the HLA region using next-generation sequencing data. *BMC Bioinformatics* 18, 258, doi:10.1186/s12859-017-1671-3 (2017).

187 Olshen, A. B., Venkatraman, E. S., Lucito, R. & Wigler, M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* 5, 557-572, doi:10.1093/biostatistics/kxh008 (2004).

188 Hsu, F. H. *et al.* A model-based circular binary segmentation algorithm for the analysis of array CGH data. *BMC Res Notes* 4, 394, doi:10.1186/1756-0500-4-394 (2011).

189 Benjamini, Y. & Speed, T. P. Summarizing and correcting the GC content bias in high-throughput sequencing. *Nucleic Acids Res* 40, e72, doi:10.1093/nar/gks001 (2012).

190 xu, F., Wang, W. & Wang, J. *Assessment of Mapping and SNP-Detection Algorithms for Next-Generation Sequencing Data in Cancer Genomics*. Vol. 1 (2013).

191 Lind, M. J. & Gehrman, P. R. Genetic Pathways to Insomnia. *Brain Sci* 6, doi:10.3390/brainsci6040064 (2016).

192 Jansen, P. R. *et al.* Genome-wide Analysis of Insomnia (N=1,331,010) Identifies Novel Loci and Functional Pathways. *bioRxiv* (2018).

193 Shi, H., Kichaev, G. & Pasaniuc, B. Contrasting the Genetic Architecture of 30 Complex Traits from Summary Association Data. *Am J Hum Genet* 99, 139-153, doi:10.1016/j.ajhg.2016.05.013 (2016).

194 Boyle, E. A., Li, Y. I. & Pritchard, J. K. An Expanded View of Complex Traits: From Polygenic to Omnigenic. *Cell* 169, 1177-1186, doi:10.1016/j.cell.2017.05.038 (2017).

195 Hackinger, S. & Zeggini, E. Statistical methods to detect pleiotropy in human complex traits. *Open Biol* 7, doi:10.1098/rsob.170125 (2017).

196 Solovieff, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* 14, 483-495, doi:10.1038/nrg3461 (2013).

197 Galesloot, T. E., van Steen, K., Kiemeney, L. A., Janss, L. L. & Vermeulen, S. H. A comparison of multivariate genome-wide association methods. *PLoS One* 9, e95923, doi:10.1371/journal.pone.0095923 (2014).

198 Wray, N. R., Wijmenga, C., Sullivan, P. F., Yang, J. & Visscher, P. M. Common Disease Is More Complex Than Implied by the Core Gene Omnigenic Model. *Cell* 173, 1573-1580, doi:10.1016/j.cell.2018.05.051 (2018).

199 Mahajan, A. *et al.* Fine-mapping of an expanded set of type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *bioRxiv* (2018).

200 Belkadi, A. *et al.* Whole-genome sequencing is more powerful than whole-exome sequencing for detecting exome variants. *Proc Natl Acad Sci U S A* 112, 5473-5478, doi:10.1073/pnas.1418631112 (2015).

201 Lelieveld, S. H., Spielmann, M., Mundlos, S., Veltman, J. A. & Gilissen, C. Comparison of Exome and Genome Sequencing Technologies for the Complete Capture of Protein-Coding Regions. *Hum Mutat* 36, 815-822, doi:10.1002/humu.22813 (2015).

202 Meienberg, J., Bruggmann, R., Oexle, K. & Matyas, G. Clinical sequencing: is WGS the better WES? *Hum Genet* 135, 359-362, doi:10.1007/s00439-015-1631-9 (2016).

203 Maroti, Z., Boldogkoi, Z., Tombacz, D., Snyder, M. & Kalmar, T. Evaluation of Whole Exome Sequencing as an Alternative of BeadChip and Whole Genome Sequencing in Human Population Genetic Analysis. *bioRxiv* (2018).

204 Hatzikotoulas, K., Gilly, A. & Zeggini, E. Using population isolates in genetic association studies. *Brief Funct Genomics* 13, 371-377, doi:10.1093/bfgp/elu022 (2014).

205 Xue, Y. *et al.* Enrichment of low-frequency functional variants revealed by whole-genome sequencing of multiple isolated European populations. *Nat Commun* 8, 15927, doi:10.1038/ncomms15927 (2017).

206 Chheda, H. *et al.* Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *Eur J Hum Genet* 25, 477-484, doi:10.1038/ejhg.2016.205 (2017).

207 Gilly, A. *et al.* Very low depth whole genome sequencing in complex trait association studies. *bioRxiv* (2017).

208 Backenroth, D. *et al.* FUN-LDA: A Latent Dirichlet Allocation Model for Predicting Tissue-Specific Functional Effects of Noncoding Variation: Methods and Applications. *Am J Hum Genet* 102, 920-942, doi:10.1016/j.ajhg.2018.03.026 (2018).

209 He, Z., Xu, B., Lee, S. & Ionita-Laza, I. Unified Sequence-Based Association Tests Allowing for Multiple Functional Annotations and Meta-analysis of Noncoding Variation in Metabochip Data. *Am J Hum Genet* 101, 340-352, doi:10.1016/j.ajhg.2017.07.011 (2017).

210 Loehlein Fier, H. *et al.* On the association analysis of genome-sequencing data: A spatial clustering approach for partitioning the entire genome into nonoverlapping windows. *Genet Epidemiol* 41, 332-340, doi:10.1002/gepi.22040 (2017).

211 Ionita-Laza, I. in *Statistical Society of Canada* (2018).

212 Morgan, L. & Conboy, K. in *21st European Conference on Information Systems 2013 (ECIS)* 1-12 (Utrecht University, 2013).

213 Miller, N. A. *et al.* A 26-hour system of highly sensitive whole genome sequencing for emergency management of genetic diseases. *Genome Med* 7, 100, doi:10.1186/s13073-015-0221-8 (2015).

214 Wang, Y., Kung, L. & Byrd, T. A. Big data analytics: Understanding its capabilities and potential benefits for healthcare organizations. *Technological Forecasting and Social Change* 126, 3-13 (2018).

215 McInnes, G. *et al.* Global Biobank Engine: enabling genotype-phenotype browsing for biobank summary statistics. *bioRxiv* (2018).

216 Neale, B. M. http://www.nealelab.is/uk-biobank/, 2018).

217 Charlebois, K., Palmour, N. & Knoppers, B. M. The Adoption of Cloud Computing in the Field of Genomics Research: The Influence of Ethical and Legal Issues. *PLoS One* 11, e0164347, doi:10.1371/journal.pone.0164347 (2016).

218 Wang, L., Lu, Z., Van Buren, P. & Ware, D. SciApps: A cloud-based platform for reproducible bioinformatics workflows. *Bioinformatics*, doi:10.1093/bioinformatics/bty439 (2018).

219 Handsaker, R. E. *et al.* Large multiallelic copy number variations in humans. *Nat Genet* 47, 296-303, doi:10.1038/ng.3200 (2015).

220 Alkan, C., Coe, B. P. & Eichler, E. E. Genome structural variation discovery and genotyping. *Nat Rev Genet* 12, 363-376, doi:10.1038/nrg2958 (2011).

221 Ruderfer, D. M. *et al.* Patterns of genic intolerance of rare copy number variation in 59,898 human exomes. *Nat Genet* 48, 1107-1111, doi:10.1038/ng.3638 (2016).

222 Kadalayil, L. *et al.* Exome sequence read depth methods for identifying copy number changes. *Brief Bioinform* 16, 380-392, doi:10.1093/bib/bbu027 (2015).

223 Hong, C. S., Singh, L. N., Mullikin, J. C. & Biesecker, L. G. Assessing the reproducibility of exome copy number variations predictions. *Genome Med* 8, 82, doi:10.1186/s13073-016-0336-6 (2016).

224 Yao, R. *et al.* Evaluation of three read-depth based CNV detection tools using whole-exome sequencing data. *Mol Cytogenet* 10, 30, doi:10.1186/s13039-017-0333-5 (2017).

225 Zhou, Z., Wang, W., Wang, L. S. & Zhang, N. R. Integrative DNA copy number detection and genotyping from sequencing and array-based platforms. *Bioinformatics* 34, 2349-2355, doi:10.1093/bioinformatics/bty104 (2018).

226 Guo, Y. *et al.* Comparative study of exome copy number variation estimation tools using array comparative genomic hybridization as control. *Biomed Res Int* 2013, 915636, doi:10.1155/2013/915636 (2013).

227 Retterer, K. *et al.* Assessing copy number from exome sequencing and exome array CGH based on CNV spectrum in a large clinical cohort. *Genet Med* 17, 623-629, doi:10.1038/gim.2014.160 (2015).

228 Harmala, S. K., Butcher, R. & Roberts, C. H. Copy Number Variation Analysis by Droplet Digital PCR. *Methods Mol Biol* 1654, 135-149, doi:10.1007/978-1-4939-7231-9_9 (2017).

229 Eisfeldt, J., Nilsson, D., Andersson-Assarsson, J. C. & Lindstrand, A. AMYCNE: Confident copy number assessment using whole genome sequencing data. *PLoS One* 13, e0189710, doi:10.1371/journal.pone.0189710 (2018).

230 Wasfi, A., Awwad, F. & Ayesh, A. I. Graphene-based nanopore approaches for DNA sequencing: A literature review. *Biosens Bioelectron* 119, 191-203, doi:10.1016/j.bios.2018.07.072 (2018).

231 Levene, M. J. *et al.* Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* 299, 682-686, doi:10.1126/science.1079700 (2003).

232 Huddleston, J. *et al.* Discovery and genotyping of structural variation from long-read haploid genome sequence data. *Genome Res* 27, 677-685, doi:10.1101/gr.214007.116 (2017).

233 Merker, J. D. *et al.* Long-read genome sequencing identifies causal structural variation in a Mendelian disease. *Genet Med* 20, 159-163, doi:10.1038/gim.2017.86 (2018).

234 Couldrey, C. *et al.* Detection and assessment of copy number variation using PacBio long-read and Illumina sequencing in New Zealand dairy cattle. *J Dairy Sci* 100, 5472-5478, doi:10.3168/jds.2016-12199 (2017).

235 Pollard, M. O., Gurdasani, D., Mentzer, A. J., Porter, T. & Sandhu, M. S. Long reads: their purpose and place. *Hum Mol Genet* 27, R234-R241, doi:10.1093/hmg/ddy177 (2018).

236 Monlong, J. *et al.* Human copy number variants are enriched in regions of low mappability. *Nucleic Acids Res* 46, 7236-7249, doi:10.1093/nar/gky538 (2018).

237 Jain, M. *et al.* Nanopore sequencing and assembly of a human genome with ultra-long reads. *Nat Biotechnol* 36, 338-345, doi:10.1038/nbt.4060 (2018).

238 Shi, L. *et al.* Long-read sequencing and de novo assembly of a Chinese genome. *Nat Commun* 7, 12065, doi:10.1038/ncomms12065 (2016).

239 Paten, B., Novak, A. M., Eizenga, J. M. & Garrison, E. Genome graphs and the evolution of genome inference. *Genome Res* 27, 665-676, doi:10.1101/gr.214155.116 (2017).

240 Novak, A. M. *et al.* Genome Graphs. *bioRxiv* (2017).

241 Rand, K. D. *et al.* Coordinates and intervals in graph-based reference genomes. *BMC Bioinformatics* 18, 263, doi:10.1186/s12859-017-1678-9 (2017).

242 Lee, H. & Kingsford, C. Accurate Assembly and Typing of HLA using a Graph-Guided Assembler Kourami. *Methods Mol Biol* 1802, 235-247, doi:10.1007/978-1-4939-8546-3_17 (2018).