Aalto University
MASTER'S THESIS 2019

From SNPs to Signals

Automatic Result Filtering and Novelty identification for Genome-Wide Association Studies

Arto Lehisto

Thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Technology. Espoo, 18.10.2019

Supervisor: Assistant Professor Pekka Marttinen
Advisor: Ph.D Mitja Kurki

Aalto University
School of Science
Master's Programme in Life Science Technologies

Author
Arto Lehisto

Title
From SNPs to Signals: Automatic Result Filtering and Novelty identification for Genome-Wide Association Studies

School School of Science

Master's programme Life Science Technologies

Major Bioinformatics **Code** SCI3058

Supervisor Assistant Professor Pekka Marttinen

Advisor Ph.D Mitja Kurki

Level Master's thesis **Date** 18.10.2019 **Pages** 52 **Language** English

Abstract

In recent years, genome-wide association studies (GWAS) have grown both in size and scope, with sample sizes growing to hundreds of thousands of samples and the focus of the efforts shifting to the amassing of phenome-wide, population-level data resources. These studies have brought with them an unprecedented amount of associations between genomic regions and phenotypic traits. Recently, the FinnGen project was started to create a population-level, phenome-wide GWAS recource of the Finnish population. The large amount of result data created by the FinnGen project creates a need for an automatic process of extracting significant results from the result data.

This thesis describes the automatic reporting tool, which was created for the needs of the FinnGen project. The tool extracts and annotates significant results from GWAS summary statistics and compares them to previously identified associations. The tool's motivation and function is described. A data analysis pipeline was created for the tool, and it was tested using a set of GWAS summary statistics. The results come in the form of identified signals per phenotype, as well as information about the novelty of the signals.The results of the experiment show the tool scales to the sizes necessary for the FinnGen project.

Keywords genome-wide association studies, data filtering, novelty identification, FinnGen, summary statistic

urn http://urn.fi/URN:NBN:fi:aalto-201912188201

Tekijä
Arto Lehisto

Työn nimi
Genominlaajuisten Assosiaationtutkimusten Tulosten Automaattinen Raportointi ja Vertaus

Korkeakoulu Perustieteiden korkeakoulu

Maisteriohjelma Life Science Technologies

Pääaine Bioinformatiikka **Koodi** SCI3058

Valvoja Apulaisprofessori Pekka Marttinen

Ohjaaja Tohtori Mitja Kurki

Työn laji Diplomityö **Päiväys** 18.10.2019 **Sivuja** 52 **Kieli** englanti

Tiivistelmä

Viimeaikaiset edistysaskeleet geenitutkimuksessa ovat mahdollistaneet genominlaajuisten assosiaatiotutkimusten (eng. genome-wide association study, GWAS) kasvamisen niin koossa kuin laajuudessa. Tutkimusten otoskoot ovat kasvaneet satoihin tuhansiin ja tutkimusten pääpaino on siirtynyt kohti koko fenotyyppikirjon sisältäviä, populaatiokohtaisia aineistoja. Näiden aineistojen ja niistä tehtyjen tutkimusten ansiosta genomin ja fyysisten ominaisuuksien välisten assosiaatioiden määrä on räjähtänyt. Vuonna 2017 alkanut FinnGen-projekti tähtää Suomen populaation kattavaan, koko suomalaisen tautikirjon sisältävään aineistoon. Valtavan datamäärän käsittelemiseksi työkalulle, joka erottelisi merkittävät tulokset projektin tuloksista, on syntynyt tarve.

Tämä diplomityö esittelee genominlaajuisten assosiaatiotutkimusten automaattisen raportointityökalun, joka luotiin FinnGen-projektin tarpeisiin. Raportointityökalu eristää merkittävät variantit GWAS-tiivistelmätilastoista, lisää niihin tunnetut geeniannotaatiot ja vertaa niitä jo löydettyihin assosiaatioihin. Diplomityössä kuvataan sekä työkalun tarkoitus että sen toiminta. Työkalun käyttämiseksi FinnGen-projektissa sille luotiin WDL-kieleen pohjautuva työnkulkuspesifikaatio, jota testattiin suorittamalla työkalun työnkulku joukolle GWAS-tiivistelmätilastoja. Työkalu tuottaa lopputuloksenaan joukon assosiaatiosignaaleja jokaiselle tiivistelmätilastolle. Näihin signaaleihin on lisätty tieto siitä, mitkä niistä on assosioitu aikaisemmin, ja mitkä ovat uusia assosiaatioita. Työkalun testauksen tulokset osoittavat, että työkalua voidaan käyttää myös FinnGen-projektin tarpeisiin.

Avainsanat genominlaajuinen assosiaatioanalyysi, datan suodatus, tiivistelmätilasto, FinnGen

urn http://urn.fi/URN:NBN:fi:aalto-201912188201

Acknowledgements

First of all, thank you to the FinnGen project for giving me the possibility of doing this thesis project. Thank you to my advisor Mitja Kurki for his continuing guidance during the thesis, as well as for the input he gave during its development. Without him this project would definitely not have been realized. Thanks to my supervisor Pekka Marttinen for the guidance during this project, as well as all of the help throughout it.

Thanks for the FinnGen data analysis team for providing me with the experimental data, and the countless times they helped me with any problems I stumbled upon. A special thanks to Padhraig Gormley, whose help in testing and thinking up features greatly helped in developing the thesis.

Finally, I'd like to thank my friends and family for supporting me during this thesis. A special thanks goes to my girlfriend Maija, whose support during the thesis proved invaluable.

Acronyms

BOLT-LMM BOLT-LMM. 12

CD/CV common disease/common variant. 5

DNA deoxyribonucleic acid. 8, 25

EFO Experimental Factor Ontology. 37

EMMA Efficient Mixed-Model Association. 12

EMR electronic medical record. 9

FDR false discovery rate. 13

FWER family-wise error rate. 12

GEMMA Genome-wide Efficient Mixed Model Association. 12

gnoMAD genome aggregation database. 3, 19, 20, 23, 24

GWA genome-wide association. 12, 13, 15, 19

GWAS genome-wide association study. 2, 4–17, 19, 21, 24, 26, 27

GWS genome-wide significant. 13, 19, 34, 36

HLA human leukocyte antigen. 16

LD linkage disequilibrium. 6–8, 13–15, 19, 21, 22, 25–27, 30

LDL low-density lipoprotein. 8

LoF Loss of Function. 4

MAF minor allele frequency. 15

NFE non-Finnish European. 15, 30

PC principal component. 10

PRS polygenic risk score. 13

SAIGE Scalable and Accurate Implementation of Generalized mixed model. 12

SNP single nucleotide polymorphism. 3, 4, 6, 7, 12, 13, 15

UKBB UK Biobank. 16

WDL Workflow Description Language. 18, 21, 26, 28

Glossary

biallelic Only having two alleles. 3, 25, 26

enrichment If an allele is more frequent in one population when compared to its frequency in another population, it is considered to be enriched compared to the other population. In the context of this thesis, enrichment means the enrichment of a variant in a Finnish population when compared to other European non-Finnish (NFE) populations.. 19

genetic architecture The underlying genetic structure resulting in a trait and how that trait is shown in a population[1]. 4–6, 24

locus Position in the genome. 4, 8

Contents

1. Introduction

1.1 Problem description

Modern genome-wide association studies (GWASs) produce summary statistics for many hundreds of phenotypes. At the sametime, the amount of measured variants can exceed one million. Recognizing novel phenome-genome associations from already identified associations is important for many applications, such as finding new cellular mechanisms behind the phenotype. Gathering the relevant associations from the statistics and comparing them to existing results manually is time-consuming. Therefore, there is a need to automate this process.

1.2 Approach

We propose to solve this problem by developing a tool that automatically selects the relevant associations in a GWAS summary statistic and compares these results to earlier studies, available in either online databases or specific summary statistics from chosen key studies.

1.3 Research scope and goals

The purpose of this master's thesis is to develop a tool for the needs of the FinnGen project. Specifically, the purpose of this thesis is to produce a tool for automatically filtering GWAS summary data, identifying significant variants from this data, as well as labelling these variants for further use. In order to test the suitability of this tool, an experiment will be performed to measure both its performance in identifying the significant results, as well as its runtime performance.

2. Background

2.1 Human Genome

2.1.1 Variation in the human genome

Research in human genetics has progressed tremendously during the last few decades. The Human Genome Project, starting in 1990 and ending in 2003, first determined the complete genetic sequence of humans[2][3]. Following that, multiple consortia have been established for quantifying the amount and types of variation in the human genome. Examples of these are the 1000 Genomes Project and the International HapMap Project[4][5]. Both of these projects span across multiple human populations, with the 1000 genomes project consisting of samples from 14 populations[4].

There are different types of structural variation in genomes, but they can be roughly divided into single nucleotide polymorphisms (SNPs) and structural variations[6]. The group of structural variation can be further divided into indels (insertions and deletions), block substitutions, inversion variants and copy number variants[6].

There is a large amount of variation in the human genome. The 1000 genomes project identified 38 million SNPs, 1.4 million indels with two alleles and 14 000 large deletions[4]. A subsequent analysis of structural variants using 2500 human genomes further raised the number of identified biallelic indels to 42 thousand, and identified more than 6000 biallelic duplications. Furthermore, they estimate that a median individual has approximately 2800 indels and 20 duplication sites in their genome. The latest development in the study of human genome variation is the genome aggregation database (gnoMAD), which studied 125 748 whole exome sequences and 15 708 whole genome sequences. Using this data, they identified 218 141 660 SNPs and 26 814 456 indels, as well as 366 412 structural variants.[7][8]

Alterations to a genetic sequence alter the function of that sequence. These mutations can be classified based on the effects that the mutation has on the amino acid sequence it codes for. A mutation that results in a different amino acid to be coded is called a missense mutation. A mutation that causes a codon to transform into a stop codon is called a nonsense mutation. An indel can change the reading frame of the genetic sequence, in which case the whole sequence after the mutation is altered. These mutations are called frameshift mutations. A mutation can also be classified based on whether it is located in an intragenic or intergenic region, with intragenic meaning inside a gene and intergenic between genes.

A type of mutations called Loss of Function (LoF) mutations are especially interesting in the context of GWASs, as the mutation causes a partial or complete loss of normal protein function in the protein it codes for. It should be noted that for the majority of SNPs, an apparent mutational effect can not be discerned.[9]

2.1.2 Genetics of human disease

The genetic basis for human diseases is a central point in genetic research[10][1].A vast amount of data implies that many diseases are heritable. Therefore, acquiring more information about the genetics of disease is crucial in transforming this knowledge into clinical information and eventually into treatments and medications. Here the genetic basis of human disease will be briefly introduced.[11]

The genetic basis for disease seems to differ greatly between diseases: Some diseases, such as cystic fibrosis, are affected by few small deleterious mutations with large effects in specific parts of the genome [12]. Other diseases, such as type 2 diabetes, have multiple associations with very weak effects across the genome, with no clear link between the associated variant and the disease mechanism[1]. This would suggest that the genetic architectures or the genetic basis of these diseases are different.

Generally speaking, the genetic architecture for traits can be classified into three observed categories: monogenic, polygenic and omnigenic architectures. In monogenic architectures, one or few variants affect the trait. In polygenic architectures, multiple or many variants contribute into the traits variability. In the omnigenic model many "non-core" genes affect the expression of the "core" genes that are directly responsible for the disease.[1][10]

The results from multiple studies on disease genetics seem to point that the genetic architecture behind diseases is complex and multifaceted. While there are many monogenic diseases, such as cystic fibrosis and Huntington's disease, research suggests that many complex diseases, such as type 2 diabetes and schizophrenia are highly polygenic and affected by multiple loci in the genome[12][1][13].

2.2 Genome-Wide Association Studies

In brief, this section describes the experimental design of GWASs, as well as the motivation behind using them. GWAS data and its analysis is also described.

2.2.1 Motivation

A fundamental question in genetics is how variation in the genome affects variation in the phenome, i.e. how traits are affected by genotype[12][10]. Historically, there have been assumptions that even complex traits, such as height or autism, might be determined by a

small number of genetic variants with relatively high effects.[10] However, growing knowledge of genetic variation and the challenges brought by complex traits have invalidated these assumptions[10]. In this section the rationale leading to genome-wide association studies, as well as the motivation behind them will be briefly described.

Some of the early successes in human genetics were identified by using methods such as linkage analysis. Linkage analysis was successfully used to uncover the genetic mutations behind rare genetic diseases such as the mutations in the gene *CFTR* that are responsible for cystic fibrosis[14]. In linkage analysis, families with the affected trait are genotyped using a set of genetic markers across the genome, and the way these markers are present in the families with the disease is examined. While linkage analysis proved to be successful in the case of rare genetic disorders, such as cystic fibrosis and Huntington disease, it has not been able to replicate these successes when applied to more complex traits, such as heart disease. This points to a a different genetic architecture behind these traits. The common disease/common variant (CD/CV) hypothesis has been more successful in explaining these traits.[12][15][11]

The CD/CV hypothesis states that diseases that are common in the population are affected by genetic variation that is also common in the population. This is in contrast to the disease model related to the aforementioned cystic fibrosis and Huntington disease, where the disease was largely due to few rare variants. The CD/CV hypothesis leads to two significant points: First, if diseases are governed by common variants, then the effect size of any one variant must be small, or else the phenotype would be very much correlated with the variant. Second, if the effect of any one variant is small, yet the diseases show heritability, then multiple alleles have to affect the prevalence of the disease. Through GWASs, the CD/CV hypothesis has been shown to be true for many complex diseases, and common variants do account for a part of the genetic makeup of common diseases[11]. However, rare variants with larger effects also make up part of the genetic variance[10].[12]

The small effect sizes in associations of complex traits require large sample sizes, which have been made possible by advances in chip-based genotyping arrays[12][15]. By reducing the price of genotyping, these arrays have made large sample sizes possible. The statistical power required for detecting associations depends on the experimental sample size, variant frequency, variant effect size and other factors[15][16]. In order to detect rarer variants or variants with a smaller effect on the trait, the sample size necessary to detect associations grows larger.

GWASs are able to answer to the need of large sample sizes. In recent years, GWASs have approached a scale of tens of thousands to hundreds of thousands of samples per study[17]. This makes them very useful for the search of common variants associated with common diseases, as they are highly replicable and have the statistical power to detect common variant associations [15]. Indeed, during the last 10 years that GWASs have been performed,

over 10 000 strong associations have been discovered, and association databases such as GWAS Catalog list over 100 000 genome-wide significant associations[15][18].

2.2.2 GWAS Study Design

GWASs search for associations between a phenotype and variants in a genome. In GWAS, the association search is performed over most of the genome. This differentiates a GWAS from other association studies, such as candidate gene association studies. Unlike many other types of studies, GWAS place no assumptions on the genetic architecture behind a trait, thus requiring no prior information about the gene function behind a trait .[19]

GWAS depend heavily on the fact that human genome has a strong linkage disequilibrium (LD) structure. Due to human genomes having a strong haplotype structure, an association with a SNP can point to one of two possibilities: The SNP in question can be the actual biological variant affecting the trait, which would make the association a direct association. Another possibility is that the variant affecting the trait is in high LD with the associated SNP, and thus the SNP gets associated indirectly, hence it being an indirect association. This is further demonstrated in figure 2.1. The concept of a region of high LD, or a LD block, is further explained in 2.2.[12]

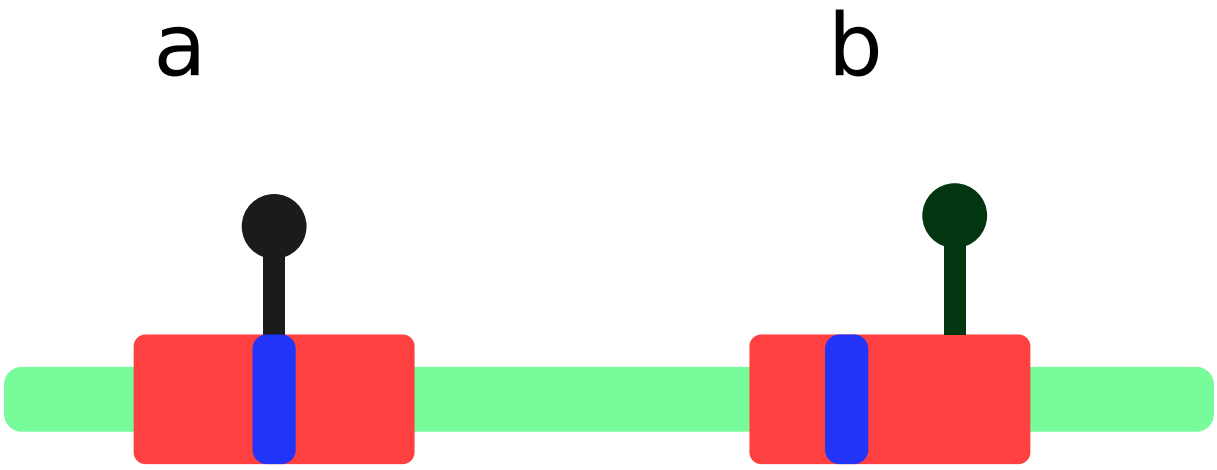

Figure 2.1. Direct and indirect association. A region of high linkage disequilibrium (red) is shown in the genome (light green). A causal mutation (dark blue) is present on the region of high LD, as well as a marker SNP (dark gray). **a)** The causal mutation is the marker SNP, and the GWAS therefore directly measures the causal genotype. **b)** The marker SNP is not the causal mutation, but due to the region of high LD, the causal effect of the mutation correlates with the marker SNP, therefore flagging it.

In GWAS, it is sufficient to sequence a subset of known SNP variants to capture the majority of the variation[19]. This is due to the LD structure of the human genome, in which most of the SNPs in a genomic location correlate strongly with their neighbouring SNPs. In European populations, this can be done by using a subset of 500,000 variants

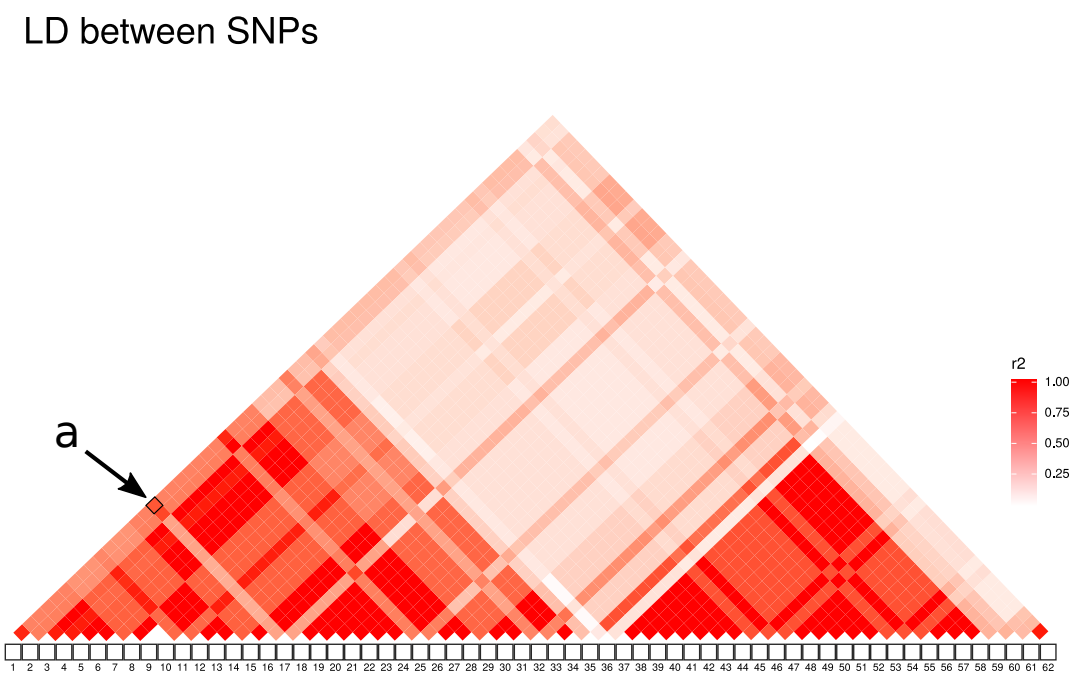

Figure 2.2. Pairwise LD matrix for a set of SNPs. Each square on in the matrix shows the squared correlation between the two SNPs on the SNPs line under the matrix. For example, the square marked with **a** shows the correlation between SNPs 1 and 17. A high value of r^2 means that the SNPs are correlated and are therefore in LD with each other. It can be seen from the figure that SNPs around indexes 37-58 are in high LD with each other, and therefore form an LD block. The LD matrix was formed by using data from the 1000 Genomes Project[4].

to capture upwards of 80% of variation in the genotype[12][20]. Further experience with GWASs has shown that using a SNP genotyping panel of half a million to a million SNPs is practical[12]. This amount of SNPs can further be augmented by imputing missing SNPs using additional LD information and tools such as IMPUTE2[21].

A GWAS can be performed for either quantitative or case/control traits. Quantitative traits have improved statistical power to detect effect alleles, in addition to them having easily understandable outcomes[12]. For example, in the case of height, the effect can be interpreted as unit of change per allele. In contrast, in case/control studies, there is more room for error in defining the trait and diagnosing it , and the association effects are presented as odds ratios[12]. While quantitative traits are preferred from a statistical standpoint, case/control GWASs have been successful in finding many significant associations[12][18]. It is notable that GWASs using quantitative traits are analysed using a different family of statistical models than studies using case/control traits[12].

In brief, GWASs search for associations between a trait and the whole genome. They utilise the haplotype structure and high LD present in the genome to efficiently cover the variation of whole genome with a much smaller subset of variants. GWASs can be used to find associations for both quantitative and case/control traits.

2.2.3 Data of GWAS

In this section, genotype and phenotype data used in GWASs will be described, as well as the procedures and problems related to data. The section will also touch on how the genotypes can be further imputed, so that a much larger subset of the genotype can be covered than what would be possible by using only directly genotyped data.

As mentioned in earlier sections, the genetic variation in GWASs is presented as the measured genotypes at different loci of the genome. These variants are measured using chip-based deoxyribonucleic acid (DNA) microarrays, in which short DNA sequences are placed on specific locations on a chip. The measured DNA binds to these locations differently based on the alleles that the sample has, and the genotype is measured by testing to which locations the sample DNA has bound to. A typical microarray chip is capable of measuring up to one million or more SNPs.[12]

Due to the haplotype structure present in the human genome, variation in the human genome appears as segments of highly correlated variants. This can be exploited to increase the amount of variants by genotype imputation, in which the LD structure of the genome and measured genotype information is used to infer genotypes for variants that were not directly genotyped[22]. This enables GWAS to use a larger set of variants than those that were measured using a genotyping chip, as well as imputing a common set of variants for meta-analysis from smaller GWAS that use disjoint sets of variants. IMPUTE2 is a tool that is commonly used for genotype imputation[21].[12]

As with other aspects of GWASs, great care must be used when measuring genotypes from samples, as well as in genotype imputation. In DNA chip genotyping, care must be taken in designing the DNA chip, so that it covers the study population's genomic variation properly. For example, differences in coverage for a single DNA microarray between European and African populations can be more than 20%[20]. Preferably the DNA microarray is designed based on the LD structure of the study population[19]. In genotype imputation, the LD information that is used to infer imputed genotypes has a large effect on the imputation quality. The LD information must be from the same population as the study sample for the imputation to succeed. LD information from another population can result in lower genotype imputation quality. The analysis methods used for the GWAS should also take the probabilistic nature of imputed genotypes into account.[12]

In short, the genotype data used in GWASs consists of measurements of the alleles in different loci of the genome. In GWASs, the measured genotypes consist of both directly measured genotypes and imputed genotypes.

The phenotype information used in GWAS can be either quantitative, such as height of the amount of low-density lipoprotein (LDL) cholesterol in blood, or qualitative, like in case/control traits. Quantitative traits are often based on physical measurements and are

therefore reliable to measure. Qualitative traits, such as disease diagnoses like multiple sclerosis, might not be directly measurable like quantitative traits. Instead, the diagnosis might be defined by considering the results of multiple physical measurements and by ruling out alternative diagnoses.

Standardised phenotype criteria are especially important to case/control traits. Especially if the data comes from multiple different entities, such as different hospitals, standardisation of the phenotype criteria prevents introducing entity-specific effects into the study. Another possible source of error is in how clinicians determining the phenotype interpret the phenotype criteria.[12]

Electronic medical records (EMRs) are also a possible source for phenotype information. Especially in the case of large studies or data resources that make use of national or regional biobanks, the usage of EMRs is possible.[12] Examples of data resources using EMRs include the FinnGen project[23] and the UK Biobank[24].

As a real-world example, in the UK Biobank resource, multiple different sources for phenotype information have been used. Questionnaires were presented to the participants about their health, lifestyle and social life. Several physical measures, such as blood pressure, lung capacity and grip strength were measured from all of the 500,000 participants. In addition to the direct measurements and questionnaires, electronic health records and follow-up questionnaires have been used to augment the phenotypic data records, to include updates to the participants' change in status like death or cancer diagnosis. The data resource is expected to be extended during the coming years with procedures such as medical imaging performed using magnetic resonance imaging, X-ray absorptiometry, and ultrasound imaging.[24]

2.2.4 GWAS statistical analysis

In this section, basic statistical analysis used in analysing single-marker associations will be explained. Basic analysis methods for quantitative and case-control traits are described. Modern methods for quantitative and case/control trait analysis are also examined briefly. The problem of multiple testing and how it is counteracted in GWASs is also explored.

Single marker association analysis

With continuous traits, one of the simplest ways to analyse the association of one variant is to do linear regression[25]. A simple formulation of the regression model is

$$y_i = b + x_i\beta + \epsilon_i, i = 1, 2...n, \epsilon \sim N(0, \sigma^2) \tag{2.1}$$

where y_i is the trait value for sample i, b is the intercept, x_i is the genotype for sample i, n is the sample amount, β is the genotype-specific effect size, and ϵ is an error term, assumed to

follow a normal distribution[26]. In GWAS statistical analysis, it is common to represent the genotype as the number of copies of one of the alleles[25]. In the case there were two alleles, a and A, and some genotypes were aa, aA and AA, the genotype representations would be 0, 1 and 2, respectively. This model takes into account the additive effect the variant may have, and is therefore called an additive model. While there are multiple ways in which the genotype can affect the trait, for example an additive, recessive or multiplicative effect, often only the additive model is explored. This is due to the fact that the additive model has reasonable statistical power to detect most of the other effects. Some studies use multiple genetic models and adjust the p-values for multiple testing in order to account for multiple possible genetic models.[12]

In proper GWAS, other covariates are often added to the model. These include population structure, age, sex and other clinical covariates. An especially important factor to take into account is population structure, which is often added to the model in the form of principal components (PCs). In case the genotypes have been analysed in batches, a batch indicator can also be added to regress out effects due to batch-specific differences.[12]

For the additive model, the null hypothesis H_0 is that the genotype has no effect on the trait, i.e. that $\beta = 0$. The alternate hypothesis H_1 is therefore $\beta \neq 0$. The previous formulation in equation 2.1 can be transformed to the form

$$y_i \sim N(b + x_i\beta, \sigma^2). \tag{2.2}$$

If the null hypothesis is in effect, then the test statistic

$$t = \frac{\hat{\beta}}{s/\sqrt{\Sigma_i(x_i - \bar{x})^2}} \tag{2.3}$$

follows the t-distribution with $n - 2$ degrees of freedom, where $\hat{\beta}$ is the least-squares estimator for β, s is the sample standard deviation, and $\bar{x}$ is the mean of x[26]. As p-value is defined as the probability of observing data at least as extreme as the observed data given the null hypothesis, the p-value in this case can be interpreted as the probability mass outside the observed value of the t statistic. The test is two-sided in this case, because β can be positive or negative. The interpretation can be seen in figure 2.3.

As with other instances of hypothesis testing, if the p-value is smaller than a pre-defined significance threshold, the null hypothesis H_0 is discarded and the alternate hypothesis H_1 is deemed significant[25]. In the case of association analysis, this means that the variant is thought of as being associated with the trait[12].

In the case of case/control traits, the formulation of the regression model is somewhat different. A logistic regression model can be used to determine whether there is an association. In a logistic regression model, the dependent variable y_i is given a value of 1 or 0,

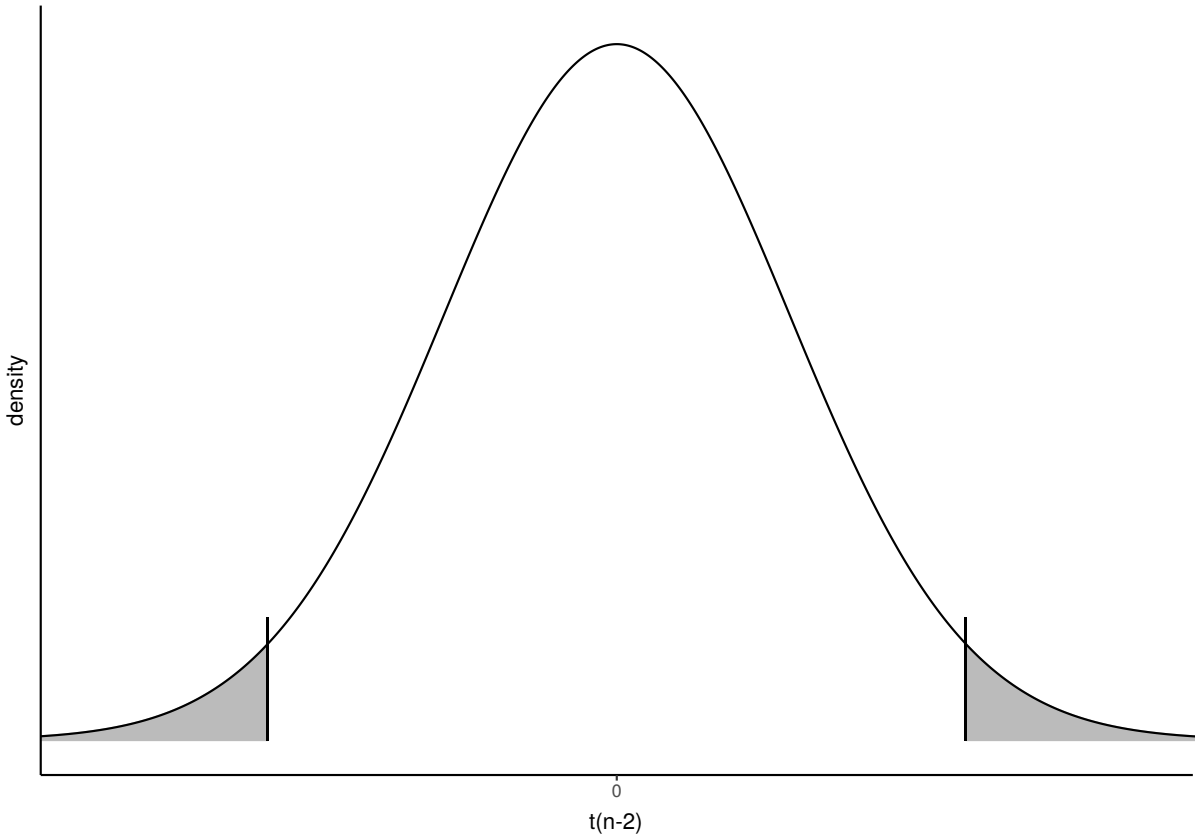

Figure 2.3. Visualization of the p-value. The t-test statistic distribution under the null hypothesis is plotted, and a sampled value of t is indicated using two vertical lines. Large positive or negative values of β make the observed t fall far from the center of the probability distribution. The grey area is the probability of observing a t-statistic at least this extreme under the null (i.e. the p-value). A large absolute value of the t-statistic implies that it is unlikely that this data was observed under the null.

depending on whether the patient is considered a case or a control. The logistic regression model is defined as

$$P(y_i = 1) = E(y_i) = p_i = \frac{e^{b+x_i\beta}}{1 + e^{b+x_i\beta}}, \tag{2.4}$$

which can be linearized as

$$ln(\frac{p_i}{1-p_i}) = b + x_i\beta. \tag{2.5}$$

As with the simple linear regression, hypothesis testing can be performed with logistic regression models as well. In the case of logistic regression, common tests for significance are Wald's test and the likelihood ratio test.[26][27]

Modern GWA analysis

While these basic regression models are sufficient for analysing associations between phenotypes and variants, they are susceptible to bias from a variety of sources, such as sample relatedness & case/control imbalance in the case of case/control studies[28]. Multiple tools for association analysis have been created to mitigate these problems, as well as to improve the efficiency of association analysis. During the last 10 years, sample size in GWASs has grown from thousands of samples to tens and hundreds of thousands of samples,

which has resulted in a need for more efficient computation[29][17].

While some of this need can be alleviated by the use of parallel computing and larger and larger computing clusters, more efficient and precise techniques make even more detailed analysis possible. For example, let us consider three algorithms of different complexities: an algorithm linear with respect to the sample amount, such as Scalable and Accurate Implementation of Generalized mixed model (SAIGE), an algorithm that is quadratic in computational complexity with respect to the amount of samples, such as Genome-wide Efficient Mixed Model Association (GEMMA), and an algorithm that is cubic with respect to the amount of samples, such as Efficient Mixed-Model Association (EMMA). In case it takes one hour to complete a task with n samples for each of these programs, a task with $10n$ samples would take 10 hours, 100 hours and 1000 hours for each of these algorithms, respectively. Considering that an increase of two magnitudes has already been observed in GWASs sample sizes, these algorithmic improvements are not only beneficial, they are necessary for the advancement of GWASs.[28][30][31]

By changing the model assumptions, modern techniques for genome-wide association (GWA) analysis can improve computational efficiency and reduce computational complexity significantly. For example, approaches such as BOLT-LMM and SAIGE can reduce the computational cost from quadratic ($O(MN^2)$, where M is the number of variants and N is the number of samples) to linear ($O(MN)$), and simultaneously account for case/control imbalance, improving statistical power.[32][28]

Multiple testing in GWA analysis

In single locus association analysis, the same test for significance is performed for each measured locus in the samples' genotypes[16][12]. In case no corrections for multiple testing are made, this would result in a large amount of false positives. For example, when using a significance threshold of 0.05, assuming the null hypothesis holds, with one test, the probability of the result being marked as significant is 5%. If there are 1,000,000 tests, the expected amount of tests that would be marked significant is 5% of that, or 50,000. Therefore, actions to reduce the burden of multiple testing are especially important in GWASs, where the amount of testable variants can grow to millions[16].

There are multiple methods for reducing false positives. One of the simplest methods for this is the Bonferroni correction, where the significance level is adjusted by the number of tests performed. Given an initial significance threshold α and n independent tests, the adjusted significance threshold would be $\alpha_a = \alpha/n$. This correction ensures that the probability of having at least one false positive result is α. This probability is also called the family-wise error rate (FWER)[33]. In case of genome-wide association (GWA) analysis, the correlated nature of SNPs violates the assumption of independent tests, and therefore the correction is considered highly conservative.[34]

Other methods for adjusting for multiple testing include determining the false discovery rate (FDR), such as the Benjamini-Hochberg procedure, or permutation testing. A somewhat different method of determining a significance threshold for GWAS is to estimate a genome-wide significance threshold. Since the LD blocks in the genome result in an "effective number of independent genomic regions"[12], this sets an upper limit to the amount of statistic tests that are performed during GWA analysis. This threshold differs between populations, and is estimated to be approximately 5×10^{-8} for European populations[34]. 5×10^{-8} has been widely accepted as the genome-wide significant (GWS) threshold for strong associations[15].[16][12]

2.2.5 GWAS Results and Their Interpretation

In this section, the nature of results that are acquired from GWASs is described, as well as some ways to use them. Finemapping will also be briefly explained, as well as how it can be used to find causal variants.

The main analysis method of GWASs, the single-locus association test, produces results for each of the tested variants with respect to the phenotype in question[12]. In the case of quantitative traits, these results consist of effect sizes and standard errors for each of the variants[35]. In the case of case/control traits, the effect size is replaced with odds ratio[35]. These measures can then be used to calculate p-values for each of the variants. These results, in which every variant is given a p-value/association test score and effect size, are usually compiled into a table and called summary association statistics, or summary statistics[35]. These results are often presented using Manhattan plots[13][36], which summarizes the whole dataset into a single, easily understandable format. Figure 2.4 shows an example of a Manhattan plot. It should be noted that GWAS associations should not be interpreted as the variant having a direct effect on the trait. Due to the fact that the variants are often chosen because they correlate with adjacent SNPs, they don't directly show a link between a trait and a genomic position[37]. In that case the causal variant has to be determined otherwise, e.g. by searching for possible missense or loss-of-function mutations near the marker SNP, or by finemapping. It is also common for the associations to reside in intergenic regions, which makes inferring the mechanism behind the observed effect difficult. What's more, often the desired end result is not a set of associations but a set of gene candidates that might give more information about the pathomechanisms of diseases, or act as drug targets.

Summary statistics are a result on their own, in that they can be understood as a list of the SNPs with the p-value assigned to each of the SNP portraying the significance of that association as well as the effect size showing the strength of the effect of that variant. However, they can be used in additional ways. For example, polygenic risk scores (PRSs) can be built from summary statistics when used in conjunction with LD information. The

LD information can be acquired by using the genotype-level data used to calculate the summary statistics, or by using a reference LD panel produced from the same population as the study sample has been drawn[35]. Other use cases for summary statistics include finding causal variants from the summary statistics by statistical finemapping and using summary statistics in conjunction with gene expression data to infer whether the same causal variants affect the GWAS and gene expression data signals[35].

In the context of GWASs, finemapping refers to the process of finding variants responsible for the phenotype[37]. Because the association scores of variants can get inflated due to LD between variants, a signal in GWASs often presents itself as a group of variants, out of which only some physically affect the trait. Since the amount of causal variants is generally not known, finemapping is usually done to identify a set or sets of variants, that with some predefined probability contain one or all of the causal variants. Several different finemapping strategies exist, for example Bayesian methods or heuristic methods[37].FINEMAP is one example of a widely used finemapping method[38].

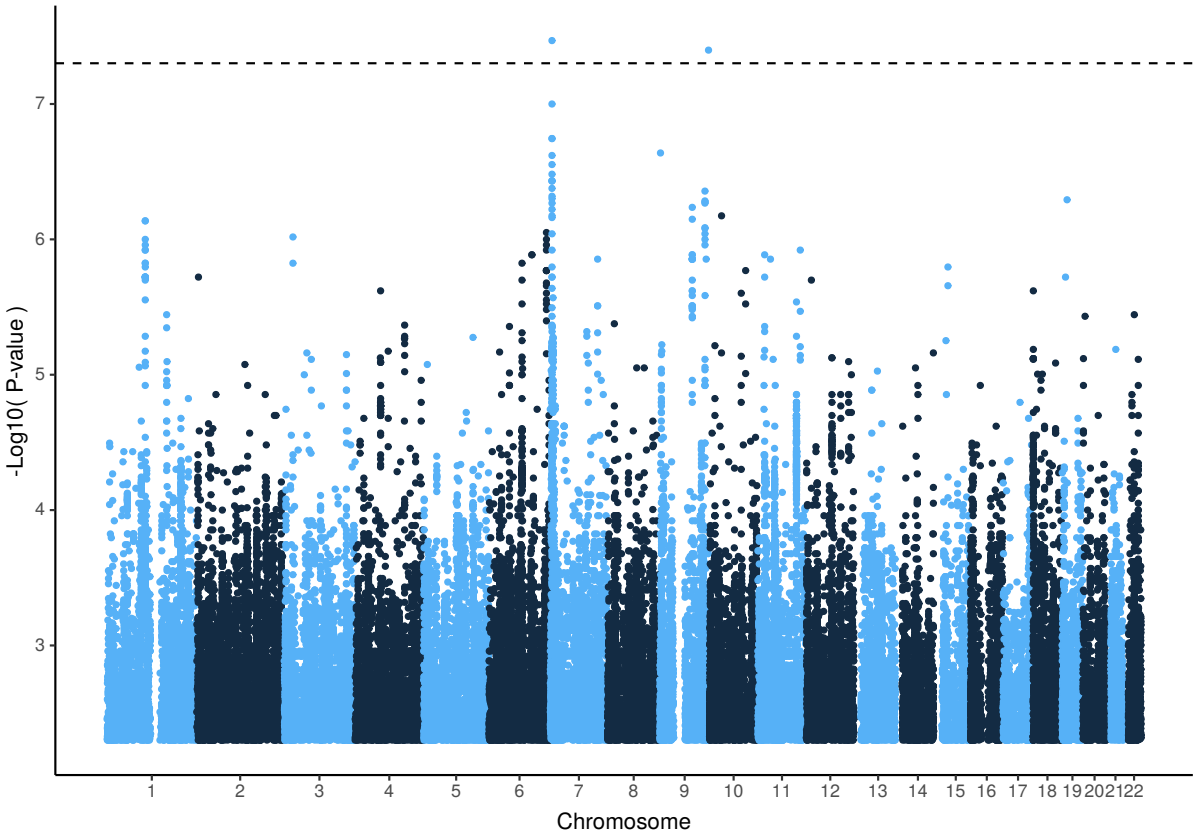

Figure 2.4. A Manhattan plot. Variants are plotted on the x-axis, and negative logarithm of the p-values on y-axis. The dashed line shows the genome-wide significance threshold, in this case $5e{-}8$. Adjacent chromosomes are coloured differently for easy visual separation. The data for this Manhattan plot was acquired from GWAS Catalog[18], and comes from [39]. Only variants with p-value smaller than 0.005 have been plotted.

2.3 Existing solutions

Bioinformatics toolkits

A wide variety of tools have been developed to analyse genome-wide association data. These include toolkits that cover a wide variety of analysis needs, such as PLINK[40], as well as tools with one particular focus, such as FINEMAP[38], tabix[41] and bgzip[42]. For example, PLINK provides extensive tools for everything related to analysing GWA data, providing options for e.g. association analysis, data wrangling, and for calculating linkage disequilibrium between variants[40].

Association Databases

Multiple association databases have been created for storing GWAS result associations. Some examples of these include the GWAS Catalog[18] and GWAS Central[43]. These curated results can then be compared to find out if the findings of a study are novel. GWAS Catalog also includes an API for programmatic access, which makes it possible to check for associations programmatically. Some of these databases, such as GWAS Catalog, harmonize the data to a certain reference genome, which makes comparing results substantially easier.[44] As a SNP can have a different position coordinate in different reference genomes, comparison without pre-harmonized results is not possible.

2.4 FinnGen study

2.4.1 The Finnish Genomic Landscape

The Finnish population is known to have gone through multiple bottlenecks that affect the Finnish gene pool today[45][46]. Due to the bottlenecks, the Finnish population shows evidence of a strong founder effect: There is less extremely rare variation (singletons, 1 in 3000 samples) in Finnish samples than in non-Finnish Europeans (NFEs)[46]. A low amount of variation in the Y chromosome has also been observed[47]. Finnish people also have much less variation in rare (minor allele frequency (MAF) < 0.5%) variants than the British population. However, the Finnish population shows a much higher proportion of low-frequency variants (2% < MAF < 5%)[45][46]. Therefore, while many variants did not get through the bottleneck event, those that did grew in frequency[46]. After the bottleneck effect, the Finnish population has grown significantly, with most of the growth happening during the last 1000 years. Under the last 250 years, the population grew from 0.5 million to over 5 million in size [45].

Variants with negative effects are under a negative selective pressure, as they reduce the

reproductive fitness of their carrier[45]. This causes variants with large negative effects to get weeded out from a population as the population grows older[45]. In relatively young populations, such as the Finnish population, these variants with large negative effects occur in with larger frequencies than in older populations, as there has not been enough time to drive the allele frequency down to the levels before the bottleneck effect[45].

From the point of view of genetic research, the Finnish population presents an interesting opportunity. The enriched rare variants add significantly to the statistical power needed to detect these variants, in some cases halving or even reducing the needed sample size to 10 % of the required sample size for detection compared to the British population[45]. Another advantage of this is that some variants that could not be studied in older populations due to too low allele frequency can have a high enough allele frequency to be studied when using a Finnish population.

In the study of common diseases, studying isolated populations has the advantage of having proportionally more deleterious variants compared to non-isolated populations. As the Finnish population is enriched in rare variants that may have an effect on the common diseases and other traits that are of interest, a population-level GWAS of the Finnish population has the potential of uncovering many associations that have not been shown in other populations due to lack of statistical power[45].[46]

2.4.2 FinnGen study

FinnGen is a project combining academic and industrial parties in order to add to information about the genetical basis behind diseases by analysing the genotypes and phenotypes of up to 500,000 Finnish individuals[48]. FinnGen was started in 2017, and is projected to take approximately 10 years[23][49]. The FinnGen study will combine approximately 200,000 existing samples from Finnish biobanks to 300,000 new samples, also collected by the Finnish biobanks.[23] The final data resource will include almost 10% of the whole Finnish population.[48]

Most of the large-scale genome-and biobank research projects have concentrated their efforts on gathering data from people that are of working age[23]. Examples of these projects include the UK Biobank (UKBB) project[17] and the Million Veterans Program. In contrast, a large proportion of the samples in FinnGen will come from hospital biobanks, and they represent a more hospitalised, older slice of the population[23].

The FinnGen project will combine extensive data from Finnish Health registries and the genetic data from samples from volunteers. The samples are processed using a custom genotyping array that in addition to normal GWAS markers combines Finnish enriched markers, markers specific to the human leukocyte antigen (HLA) region, pharmacogenomic markers and markers from the industrial partners. In total, the array contains 736,145 probes that can genotype 655,973 markers. The chip will be manufactured and the genotyping process is

performed by ThermoFisher. In addition to these directly genotyped variants, whole-genome sequences from 4000 Finns are used to further impute variants, which enables the study of 16 million variants.[49][23]

The phenotype information used in FinnGen is extracted mostly from national health registries[23]. These registries are listed in table 2.1. Special clinical expert groups, eight in total, with over 100 experts, will be involved in planning how to use the registry data and transform them into phenotypes[49]. With a large amount of phenotype endpoints, the FinnGen project will create a vast array of data for additional analysis in addition to its main data resource. This large amount of data creates a need for automated data processing pipelines.

Registry name
KELA
Statistics Finland
Register of Primary Health Care Visits
Care Register for Health Care
Finnish Cancer Registry
The Finnish Register of Visual Impairment
Population Register Centre
Finnish Registry for Kidney Diseases
Infectious Diseases Register
Medical Birth Register
Register of Social Assistance

Table 2.1. National registries used in FinnGen. Data from [49].

2.5 Challenges in bioinformatics

There are a lot of challenges in the case of analysing results from high-throughput biomedical research such as GWAS. Thanks to the decreasing price of genotyping, genome-wide studies have experienced an explosion in sample size. This creates a corresponding explosion in data amount, with genomic data taking terabytes of disk space or more. This creates a need for efficient data processing workflows. Simultaneously, GWASs are notoriously vulnerable for spurious associations due to factors such as batch effects and population stratification[12]. Controlling for these variables is important to ensure the correctness of results[12]. Another source for challenges is the fact that study replication is susceptible for errors due to changes in the data analysis workflow. The data analysis workflow itself can also act as a source for error, as programming errors and task failures during the data analysis can affect the end results.

These needs have prompted the creation of automatic workflow tools, which make it possible to automatically run a data analysis workflow. These include tools such as Toil and Cromwell[50][51]. These tools execute specifically designed workflow languages, such as

Workflow Description Language (WDL) and CWL[52][51]. These tools enable the replication and sharing of data analysis workflows, as well as offer the possibility of performing data analysis on a wide variety of computing environments, including single computers, machine clusters and cloud-based solutions[51][50]. This makes running large analyses without investment in a high-performance computing cluster possible, which reduces the initial cost of scientific computing significantly and removes the need for maintaining such a cluster, further reducing costs.

3. Tool implementation

3.1 Tool Overview

In this section, the tools function will be described.

The automatic reporting tool processes GWAS summary statistic files. The function of the tool can be divided into three distinct operations:

- Filter and group genome-wide significant variants from given summary statistic
- Annotate variants with gnoMAD and FinnGen variant annotation information
- Compare significant results to known associations

In the first step, a summary statistic file calculated from GWA analysis for one trait is filtered for GWS variants. These variants can then be grouped into possible signals, using either locus width or linkage disequilibrium as the grouping logic. The end result for this step is a list of genome-wide significant variants, grouped into possible signals.

In the second step, the previously filtered data gets annotated using gnoMAD and FinnGen-specific variant annotations. Information such as minimum allele frequency, allele frequencies of different populations, gene annotations, the most severe consequence of the alleles, as well as the calculated Finnish enrichment factor, are added to the data[7].

In the third step, the previously filtered and annotated variants are compared to current findings, either through using GWAS result databases like GWAS Catalog[18], or by supplying hand-picked summary statistics to the tool.

The input files are in the form of compressed tab-separated files, and the tool continues to use this tabular structure during its operation. An example of the input can be seen in table 3.1.

The tool and its inputs are further detailed in figure 3.1.

3.2 Technical details

The tool was created as a command line program. This decision was made because of the fact that the tool will be primarily run on a large batch of data. Since the tool is ran automatically on a large dataset, a command line interface makes it possible to use the tool

chrom	pos	ref	alt	beta	pval
1	11759478	A	G	0.11	0.45
2	83675651	T	C	0.24	1.27E-06
3	15476448	A	C	-0.45	8.00E-08
4	1005345	C	G	0.1	0.0025
5	987149	G	A	-0.12	0.00065
6	35954329	G	T	0.61	5.00E-120

Table 3.1. Tool data input. The input to the tool consists of a gzipped tab-separated table, in which the variants and their association analysis information are listed. The input has to contain at least the chromosome, position, reference and alternate alleles of the variant, as well as the p-value. Other columns can be present in the data, as long as they are not named with the same names as the aforementioned columns. While the effect size (beta) is listed in the input, it is not used in the tool.

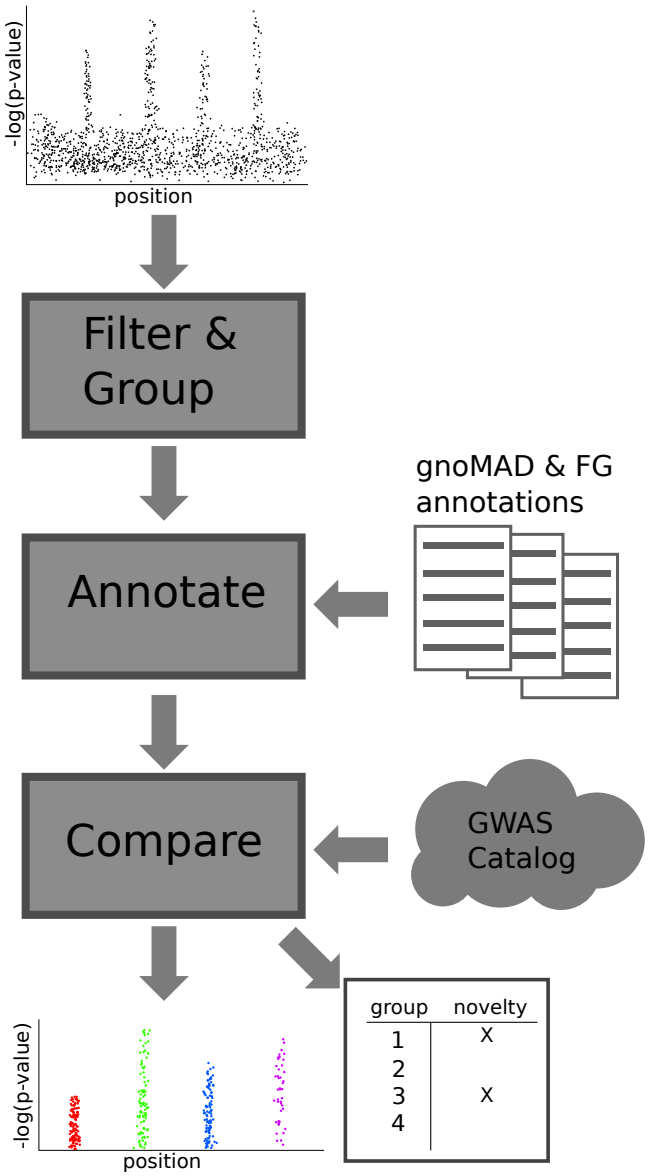

Figure 3.1. The automatic reporting tool. The data (presented as an uncolored manhattan plot) traverses through the tool. The tool consists of three distinct parts, each with its own function: In the first part, the variants are filtered and grouped into separate groups. In the second part, the variants are annotated using gnoMAD and FinnGen annotations. In the third and last part, the variants are compared against an outside source of associations, for example the online database GWAS Catalog. The output of the tool is the filtered and annotated variants. This is presented as a coloured manhattan plot, where colour separates the different signals. An additional novelty report is also outputted by the script, in which for each group their previous associations are enumerated.

in a variety of ways: On a single workstation, on larger computing clusters, or on cloud computing services by using data analysis pipelining tools. A WDL pipeline was developed in addition to the tool, to be run on on cloud computing services.

The tool is written using Python 3[53]. The data analysis package Pandas is used heavily throughout the script for operating on the variant data[54]. Multiple bioinformatics tools are used throughout the tool. These tools are Plink 1.9, tabix and LDstore[55][56][41].

The tool manipulates GWAS summary statistics, supplied to the tool as gzip-compressed tab-separated value files. The end result of the tool is set of tab-separated files, containing the filtered and grouped variants, the same variants annotated and with matching associations added, as well as a report containing the variant groups and all of the traits associated with those groups.

As the tool is implemented as a command-line program, its function is controlled with parameters supplied as command-line arguments. Using these parameters, it is possible to change key parameters of the tool, such as the used significance threshold for p-values, the type of grouping algorithm, or the database used for downloading known associations. A complete list of these parameters can be seen in table 1.1.

3.3 Variant filtering & grouping

In the first part of the tool, a summary statistics file is filtered down to include only genome-wide significant results, and the results are optionally grouped together. The grouping can be performed by using one of two algorithms: By ordering the variants by p-value, and adding a specified region, e.g. 1 megabase around the most significant variants into that variant's group, or by using LD information to determine which variants belong into which groups. The filtering and grouping concept is demonstrated in figure 3.2.

The variants are filtered according to their p-value, so that variants with a p-value lower than an user-set threshold are filtered in by the tool, and other variants are discarded.

After the variants are filtered, they are grouped. Due to LD between variants, the signal from a causal variant can be observed in the variants that are in high LD with the causal variant. Grouping variants with low p-values that are in high LD with each other and/or in close proximity to each other can therefore help in limiting the signals shown in the data to the plausible ones. A group of variants is indicated by the variants' group ID, called locus ID in the tool. The group IDs are given based on the most significant variant in the group.

In location-based grouping, the significant variants are ordered based on their p-value. Then, starting from the variant with the lowest p-value, all variants in a region around that variant, for example 1 megabase up-and downstream of the variant, are included into that variant's group. The variants that are part of a group can not form groups. This is repeated until there are no variants left to form groups. This algorithm is described in pseudocode in

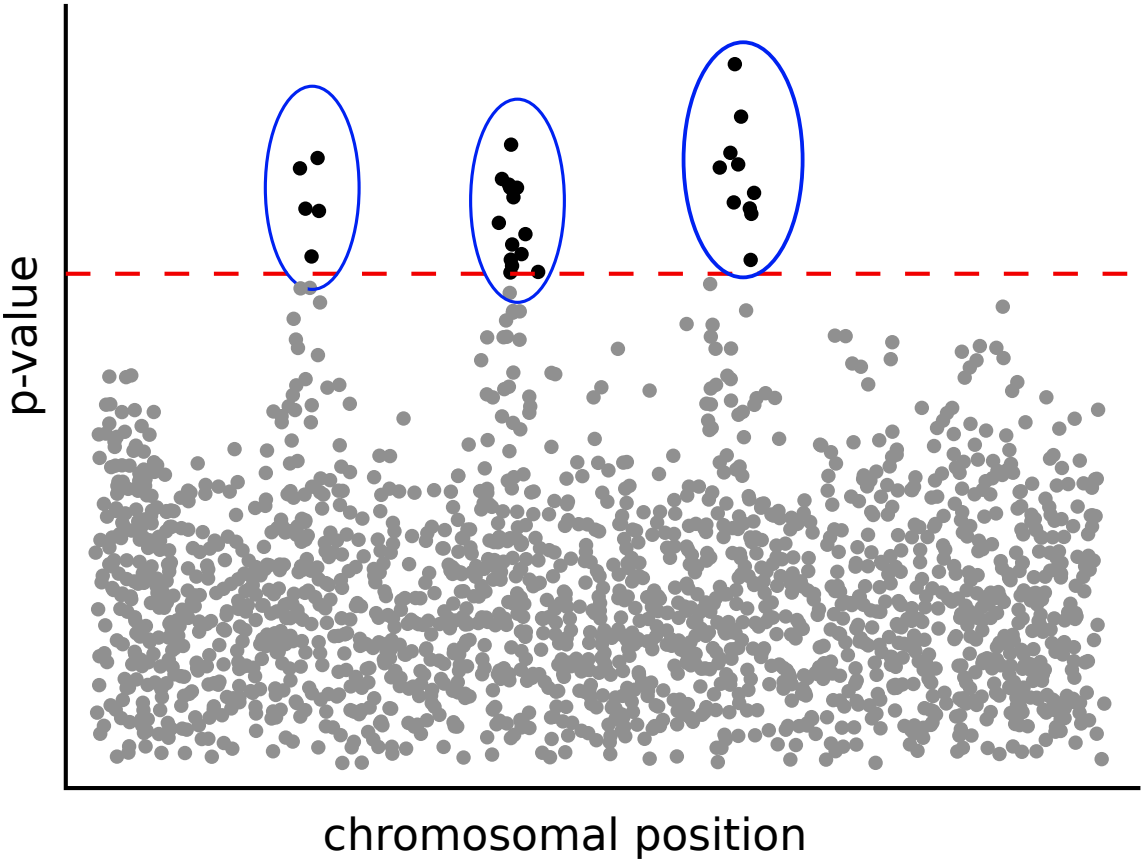

Figure 3.2. Filtering and grouping. The variants are represented as dots in the mock Manhattan plot. The variants are filtered by their p-value, such that only variants with p-value larger than a threshold (the dotted line) are filtered in. These filtered variants are then grouped (the blue circles around variants). Rest of the variants are discarded (greyed out dots).

listing A.1

In LD-based grouping, the groups are formed similarly, but the grouping algorithm takes into account the LD structure of the sample population. The variants are ordered based on their p-value, and all variants that 1) are in LD with the lead variant, 2) are closer to the lead variant than a set distance, and 3) are not yet in a group are added to that variant's group. This is continued until there are no variants left to form groups. This algorithm has been implemented in the software package PLINK[40]. The grouping method is called LD clumping. The tool itself does not implement this algorithm, and instead it calls PLINK using Python 3's *subprocess* library [53]. The thresholds for LD, p-value, and position range are configurable.

Additionally, in both of these grouping methods, the groups can be set to overlap, i.e. a variant can belong to more than one group. However, a variant already belonging to a group can not become a group's lead variant.

3.4 Variant Annotation

After the variants have been filtered and grouped, they are annotated using additional information from FinnGen variant annotation files and gnoMAD. These annotations serve as additional information that is relevant to the variants, such as allele frequencies in different populations, imputation quality scores, and the most severe consequence of the variant.

Information that is acquired from gnoMAD annotations includes allele frequencies for different populations, such as Finnish, Swedish and Estonian populations, as well as allele frequency enrichment in the Finnish population compared to the European population. The complete annotation list can be seen in tables 3.2 and 3.3. These annotations are used for interpreting significant associations. For example, if a variant is enriched in Finland, it can help explain why that variant was significant in FinnGen but not in other studies with similar sample sizes. The association might have been detected because the higher allele frequency in Finland gives more statistical power to detect it.

In addition to the gnoMAD annotations, annotations gathered during the FinnGen project are added to the variants. These annotations include batch-specific information about the imputation process, such as imputation quality and minor allele frequency. The most severe gene and most severe consequence are also added to the variants. Here, most severe consequence is defined as the consequence that is most severe to the functioning to that gene the variant resides in. For example, a nonsense mutation, where the mutation changes the codon it is coding to change into a stop codon, is considered more severe than a synonymous mutation, in which the amino acid the codon is coding to does not change. The most severe gene is then the gene that the most severe consequence affects.

The annotation information is then appended to the variant data as extra columns.

Annotation name	Description
FIN AF	Allele frequency of the effect allele in the Finnish population.
NFE AF	Allelle frequency of the effect allele in the non-Finnish European population.
EST AF	Allele frequency of the effect allele in the Estonian population.
NWE AF	Allele frequency of the effect allele in the North-Western European population.
ONF AF	Allele frequency of the effect allele people designated as other non-Finnish Europeans.
SEU AF	Allele frequency of the effect allele in the South-European population.
FIN enrichment vs NFE	The enrichment of the minor allele in the Finnish population compared to non-Finnish European population
FIN enrichment vs NFE, excl. EST.	The enrichment of the minor allele in the Finnish population compared to non-Finnish European population. Estonian population is excluded from the enrichment calculation.

Table 3.2. gnoMAD genome annotation information.

Annotation name	Description
FIN AF	Allele frequency of the effect allele in the Finnish population.
NFE AF	Allele frequency of the effect allele in the North-Western European population.
EST AF	Allele frequency of the effect allele in the Estonian population.
SWE AF	Allele frequency of the effect allele in the Swedish population.
NWE AF	Allele frequency of the effect allele in the North-Western European population.
ONF AF	Allele frequency of the effect allele people designated as other non-Finnish Europeans.
SEU AF	Allele frequency of the effect allele in the South-European population.
BGR AF	Allele frequency of the effect allele in the Bulgarian population.
FIN enrichment vs NFE	The enrichment of the minor allele in the Finnish population compared to non-Finnish European population.
FIN enrichment vs NFE, excl. EST	The enrichment of the minor allele in the Finnish population compared to non-Finnish European population. Estonian population is excluded from the enrichment calculation.
FIN enrichment vs NFE, excl. SWE	The enrichment of the minor allele in the Finnish population compared to non-Finnish European population. Swedish population is excluded from the enrichment calculation.
FIN enrichment vs NFE, excl. EST & SWE	The enrichment of the minor allele in the Finnish population compared to non-Finnish European population. Estonian & Swedish populations are excluded from the enrichment calculation.

Table 3.3. gnoMAD exome annotation information.

3.5 Comparing identified associations to literature

After the variants have been filtered, grouped and annotated, they are compared against associations that have been established earlier. There are two ways that the associations can be supplied to the tool: Either the associations can be manually curated beforehand, and supplied to the tool, or the associations can be searched for from an online database. It is also possible to use both of these options, using manually gathered association data in addition to the online association database. How the associations are used and what the tool produces as its end result will be described in the next paragraphs. Concerns related to association data preprocessing are also considered.

The automatic reporting tool seeks to automate finding significant associations from a GWAS, as well as to ascertain if the associations are novel. The comparison part of the tool seeks to find if the associations are novel. This is done by comparing the found associations to previous studies. An association is considered novel if a variant with the same chromosomal position and effect allele has not been identified as associated with the same trait before. Due to the fact that traits can have overlapping definitions, as well as pleiotropy due to the underlying genetic architecture, it is not uncommon for an associated variant to have multiple associations. Therefore it is beneficial to find out not only the novelty of a variant, but what other traits that variant is associated with, if any.

As previously mentioned, the associations can be supplied to the tool by either manually curating them or by using an online database to download them. In the case of a manual curation, the tool must be provided with a list of both the association files and the associated traits. The associations have to be reported against the correct reference genome. Only the human reference genome build 38 is supported by the tool.

When using the online association database, the tool downloads the associations belonging to the same genomic locations as the filtered and grouped variants. Two databases are available for use: The GWAS Catalog association database, as well as the GWAS Catalog summary statistic database. GWAS Catalog is an online database that contains more than 100,000 associations. The GWAS Catalog summary statistic database on the other hand contains complete summary statistics of studies, while GWAS Catalog association database contains reported associations of studies[18]. This means that the summary statistic database can contain thousands of times more association data per study than the association database. However, it only contains associations from studies that have released their complete summary statistics publicly, and therefore contains associations from fewer studies than the association database. The default option for the tool is the association database.[44]

Information concerning association chromosome, position, risk allele and trait are downloaded from the database. Due to the fact that the risk allele is not reported for all associations and the reference allele is not reported for any traits in the association database, the associated variant's allele information is downloaded from the Ensembl genome Browser, version 97 REST API[57]. The summary statistic database includes this information.

After the associations have been gathered, the tool compares them to the filtered and grouped variants. The association and variant have to have matching chromosome, position, and alleles in order to be considered a match. However, due to the fact that the variant can be measured from either strand of the DNA, and in the case of a biallelic variant, the risk allele can be either one of those, matching the alleles is not trivial. Table 3.4 shows all of the orientations of two different variants. Due to this ambiguity, the tool uses an algorithm to shift all of the variants into a unified presentation. The algorithm works by changing the strand of the variant into a strand that has an A allele if necessary, and ordering the alleles alphabetically. In this case, even variants such as 1) in table 3.4 are matched correctly, regardless of the orientation or strand they have been presented in.

After the variants have been matched to associations, the associated traits are aggregated per group of variants, which produces a human-readable table for inspecting, which variant groups are novel and which have been identified in previous studies. This table can then be used to quickly get an understanding of which loci are novel, and which have ben associated to traits before. An example table is shown in table 3.5 .The LD between variants and previously identified associations is also inspected and reported as variant pairs. The

Strand	Effect allele	Alternate allele
1) Alleles A, C		
+	A	C
+	C	A
-	T	G
-	G	T
2) Alleles G, C		
+	G	C
+	C	G
-	C	G
-	G	C

Table 3.4. Possible allele orientations for two biallelic variants. **1)** The alleles A and C in all of the configurations. An association with these alleles can be present in summary statistics as A/C, C/A, T/G or G/T, depending on the strand and allele orientation. It is not possible to determine whether a variant with alleles A/C and a variant with any of the other allele pairs is the same association, or if the variant is multiallelic. **2)** The alleles C and G in all configurations. Because C and G pair with each other, a variant with C and G as its alleles is always presented with C and G, regardless of the strand or allele orientation.

rationale behind this is that if a variant identified by FinnGen is in LD with a previously identified association, then they might have the same causal variant. This process is clarified in figure 3.3

locus id	chrom	start	end	traits	other traits
chr1_101_A_C	1	13	2000	AF;T2D	
chr3_5034_T_TGA	3	2504	7890	T2D	asthma
chrX_2106_G_C	X	304	4605		

Table 3.5. Variant group table. The variant groups and associated traits are shown in the different columns. The first column is the group identification code, which is the same as the variant id of the group lead variant. The second column shows the chromosome the group of variants is in, and third and fourth columns show the group position range. the 'traits' and 'other traits'-columns show the traits associated with the group's variants. The difference between the columns is that 'traits'-column contains those traits that were specifically supplied to the tool as traits of interest. In case those columns are empty, the variant group did not have any matching associations, and therefore it might be novel. chrX_2106_G_C is an example of a possibly novel variant group.

3.6 Pipeline

In this section, the rationale for developing a data-analysis pipeline for the automatic reporting tool, as well as technical implementation details, will be described. Parallel processing of the tasks is also mentioned.

In order to process hundreds to thousands of GWASs, a WDL pipeline was written for the tool. The pipeline makes it possible to automatically process an arbitrary amount of GWAS summary statistics on a wide variety of computing hardware. Perhaps the most important advantage of the pipeline is that it makes it possible to run the tool on cloud computing resources, which makes it possible to scale the amount of hardware much more flexibly than what would be possible using traditional computing clusters.

In the case of this tool, the WDL pipeline is run on Google Cloud using Cromwell, a WDL pipelining server[51]. Cromwell executes pipelines by dividing the tasks defined in the WDL

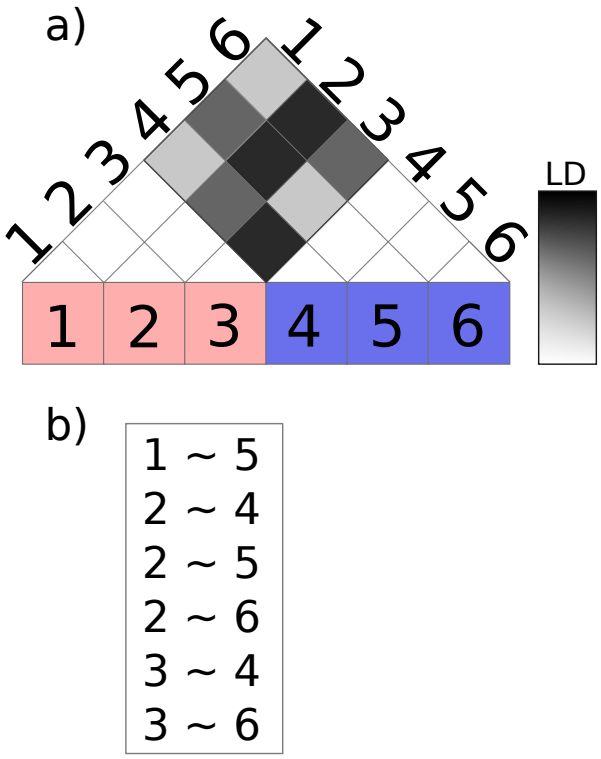

Figure 3.3. Association by LD. **a)** Consider a situation where variants 1, 2 and 3 have been identified as significant variants and grouped together. in addition to them, associations 4, 5 and 6 are located near their position on the genome, but they do not exactly match. The LD between variants and associations is drawn in a). LD between variants or LD between associations is not used. **b)** From a), we could see that variant 1 was in high LD with association 5, variant 2 was in high LD with associations 4,5, and 6, and variant 3 was in high LD with associations 4 and 6. The tool outputs these LD pairs as a separate output.

pipeline to virtual machines, each of which can be provided with appropriate resources to complete its task. Many of these virtual machines can be run at the same time, meaning that the real time required to complete a set of computationally demanding tasks can be as low as the time one task needs to be completed.

3.6.1 Parallel Processing

By its nature, much of bioinformatics is "embarrassingly parallel", i.e. a workload is easily separated into separate tasks that can be run relatively or completely independent of each other. Examples of this include processing RNA data or performing GWA analysis on multiple phenotypes, in which a single phenotype does not depend on any of the other phenotypes[50]. The workload can therefore be run on parallel on e.g. a cluster of multiple machines, which massively decreases the real time required to process the workload. In the context of this thesis, the workload of processing GWASs summary statistics with the tool can be trivially divided into independent tasks by considering each separate GWAS summary statistic as one task. Since these tasks do not depend on each other, they can be processed on different computers, e.g. on a dedicated cluster of computers, virtual machines,

or containers.

WDL implements parallel computing by its scatter call, which enables it to spin up a separate task for every input of a single call. In the case of Cromwell, these tasks are then assigned on Docker containers containing the environment necessary for that task.

Cromwell, the server implementing WDL, processes workflows by creating and setting up containers for individual tasks in that workflow, running them and copying the outputs to an output directory, which can be object storage in the cloud, a file system on a server etc. In the case of this tool, Cromwell is run on Google Cloud Platform, meaning that it will provision the containers on virtual machines provided by Google. This makes it possible to perform tasks of a wide range of scales, from small to massive, and to save up on workflow costs due to not needing to maintain computing hardware.

The computation performed by this tool is relatively fast, but the amount of data is quite large, on the order of tens of gigabytes. Because copying data from the data storage to containers takes time, it is beneficial to only partially separate the different phenotypes into separate tasks. That is implemented by dividing the list of phenotypes into a list of lists of phenotypes. A task is then started for each of those lists. Assuming that the time taken by a worker consists completely of worker intialization and task computation, the total compute time can be approximated using the formula

$$t_{Total} = t_{Phenotype} * N_{Phenotype} + t_{Initialization} * ceil(N_{Phenotype}/N_{Tasks}), \tag{3.1}$$

where t_{Total} is the total compute time, $t_{Phenotype}$ is the time to process one phenotype, $N_{Phenotype}$ is the number of phenotypes, $t_{Initialization}$ is the time to initialize a worker, and N_{Tasks} is the amount of tasks assigned to a single worker. The total real time can be calculated using the formula

$$t_{Real} = t_{Initialization} + N_{Tasks} * t_{Phenotype}, \tag{3.2}$$

where t_{Real} is the total real time. This has been further demonstrated in figure 3.4. The number of tasks per worker can be decided based on what the initialization and processing times are. For example, from figure 3.4 we can see that the decrease in compute time when switching from 3 to 4 tasks per worker is only approximately 4% of the maximum compute time. After that, the gain in decreased compute times keeps decreasing. Therefore, in a situation where the initialization time is roughly equal to the time taken processing one task, a worker should probably only have at most 2 or 3 tasks.

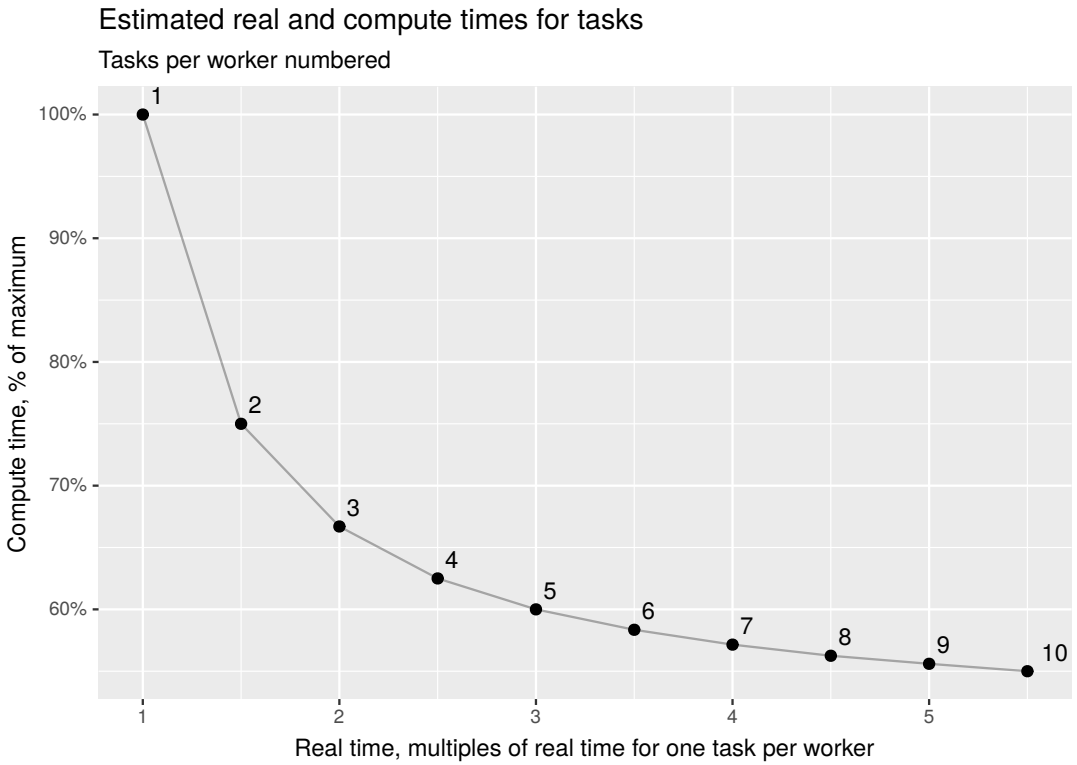

Figure 3.4. Compute & real time required when processing multiple tasks on a single worker. The x-axis shows the required real time as multiples of the time taken if only one task was assigned per worker. The initialization time and processing time of one task are assumed to be the same.

4. Methods and materials

An experiment was performed to confirm the tool works and to measure its performance.

4.1 Data

The experiment was performed using GWAS summary statistics from the Lee lab for Statistical Genetics and Genomics, which is part of the Department of Biostatistics at the University of Michigan. The GWAS was performed on the UK Biobank dataset using SAIGE[28]. The resulting summary statistics are openly available on the group's webpage. The summary statistics, being in human reference genome build 37, were lifted to build 38 using the liftOver tool by the FinnGen project data analysis team[58]. The total dataset consists of 22 GWAS summary statistics, corresponding to 22 different traits. The complete trait information is shown in table 4.3.[59]

To calculate LD between variants in the analysis, a reference genotype panel was used. The panel consisted of the NFE populations (CEU, TSI, GBR, IBS) in the 1000 Genomes Project[8]. In total, 195 samples were included in the genotype panel. The original panel data was acquired from the 1000 Genomes Project, and is openly available in [60]. The original data was converted into PLINK .bed-format and lifted to human genome build 38 using liftOver by the FinnGen project analysis team. The panel was then restricted to only the european populations using PLINK.

4.2 Experiment

The summary statistics were processed using the automatic reporting tool. To run the tool, a WDL pipeline was run with Cromwell in a Google Cloud computing environment. The resources given to individual machines on which the tool ran can be seen in table 4.1. The tool was used to filter the significant variants from the summary statistics, group the variants using the LD clumping option in PLINK, and compare the variants to the GWAS Catalog. The variants were not further annotated, as no British annotation resources were obtained. The complete parameter list for the pipeline can be seen in table 4.2.

Property	value	Description
Container Image	Ubuntu 18.04	The operating system of the container. Additional software, i.e. PLINK 1.9, tabix, LDStore and the autoreporting tool were installed onto the container image.
CPUs	4	Number of logical CPUs.
Memory	16 GB	Amount of RAM the worker had access to.
Disk space	300 GB	The amount of disk space given to the workers.

Table 4.1. Worker specification.

Parameter description	Value
# of phenotypes assigned per worker	3
significance threshold	5×10^8
alternate significance threshold for grouping	1×10^5
Grouping locus width	1500
LD clumping r^2 threshold	0.1
PLINK memory amount	13500 MB
GWAS Catalog significance threshold	5×10^8
# of logical CPUs	4
# of threads for the GWAS Catalog	8
Toggle grouping variants on or off	On
Are groups allowed to overlap	False
The method of grouping: LD clumping or position-based grouping	LD clumping
Online catalog to use for variant matching	GWAS Catalog

Table 4.2. Chosen parameters for the experiment.

Phenotype	Description	Phenotype Category	Cases	Controls
153.2	Colon cancer	neoplasms	3051	382756
165.1	Cancer of bronchus; lung	neoplasms	2101	406226
172	Skin cancer	neoplasms	13752	395071
174.11	Malignant neoplasm of female breast	neoplasms	11874	388549
185	Cancer of Prostate	neoplasms	6743	169185
250.1	Type 1 diabetes	endocrine/ metabolic	2660	388756
250.2	Type 2 diabetes	endocrine/ metabolic	18945	388756
290.11	Alzheimer's disease	mental disorders	404	402383
332	Parkinson disease	neurological	1127	395209
335	Multiple sclerosis	neurological	1356	395209
362.2	Degeneration of macula and posterior pole of retina	sense organs	2191	396859
365.11	Primary open angle glaucoma	sense organs	1037	397761
366.2	Senile cataract	sense organs	8369	388609
411.2	Myocardial infarction	circulatory system	11703	377103
427.2	Atrial fibrillation and flutter	circulatory system	14820	380919
495	Asthma	respiratory	26332	375505
496	Chronic airway obstruction	respiratory	10502	375505
555	Inflammatory bowel disease and other gastroenteritis and colitis	digestive	4528	334783
696.4	Psoriasis	dermatologic	2237	398199
714.1	Rheumatoid arthritis	musculoskeletal	4412	365085
715.2	Ankylosing spondylitis	musculoskeletal	620	365085
939	Atopic/contact dermatitis due to other or unspecified	dermatologic	2110	404817

Table 4.3. Phenotypes for the pipeline.

5. Results

The tool took 2 hours 14 minutes to process all of the 22 traits of the dataset. The work was divided into 8 tasks, with 3 phenotypes to process for each worker. The processing took 25 minutes of real time to complete. The average time for each worker to complete was 17 minutes. The complete timing information can be seen in table 5.1. Pipeline function-specific timing information can be seen in table 5.2, as well as in figure 5.1. No workers were pre-empted during the pipeline.

Based on Google pricing for custom pre-emptible machines with 4 virtual CPUs and 16 GB of RAM, the whole pipeline cost was approximately 0.11 USD or approximately 0.10 EUR [61][62]. This extrapolates to approximately 0.45 EUR /100 phenotypes. The complete pricing calculation can be seen in table 5.3.

Worker	# of phenotypes	Start	End	Duration
1	3	13:37:38.101	13:55:32.962	17:54.86
2	3	13:37:38.101	13:53:47.962	16:09.86
3	3	13:37:38.101	13:52:35.973	14:57.87
4	3	13:37:38.101	13:56:50.963	19:12.86
5	3	13:37:38.101	13:52:14.972	14:36.87
6	3	13:37:38.101	13:56:35.972	18:57.87
7	3	13:37:38.101	14:02:05.972	24:27.87
8	1	13:37:38.101	13:45:17.971	07:39.87

Table 5.1. The timing data for the pipeline. The time taken for processing is similar for the workers, except for the last worker. This can be explained by the fact that the worker only had 1 phenotype to process.

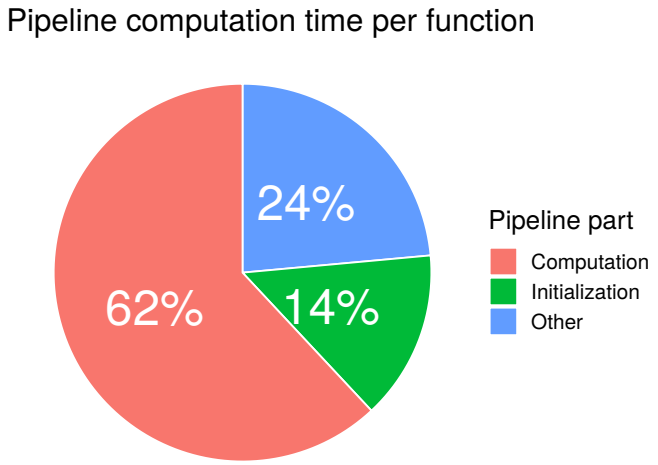

Figure 5.1. Time taken per different functions of the pipeline. 'Computation' refers to the actual computation of the automatic reporting tool, 'Initialization' refers to the downloading of the data files to each worker, and 'Other' is all of the other time spent on the pipeline not belonging to either of those categories.

A list of the amounts of identified variants as well as the amount of groups with and without association hits is available in table 5.4. In total, 21 phenotypes contained significant

Worker	# Phenos	Total	Init	Comp	Comp/Pheno	Other
0	3	17:54.86	02:23.96	10:01.47	03:20.49	05:29.44
1	3	16:09.86	02:52.62	11:16.75	03:45.58	02:00.49
2	3	14:57.87	02:24.46	09:08.01	03:02.67	03:25.40
3	3	19:12.86	02:27.83	10:21.99	03:27.33	06:23.04
4	3	14:36.87	02:34.99	09:29.10	03:09.70	02:32.78
5	3	18:57.87	02:33.96	13:15.48	04:25.16	03:08.43
6	3	24:27.87	02:07.19	17:24.89	05:48.30	04:55.80
7	1	07:39.87	02:04.66	02:02.86	02:02.86	03:32.35

Table 5.2. Pipeline timing data, separated by function. We can see that the while the computation time taken by the different workers differs between workers, the initialization time stays somewhat constant. The time taken by miscellaneous actions, shown in the 'Other' column, varies greatly between workers. From the table it can be seen that the average computation time per phenotype is quite small, ranging from 2 minutes to slightly less than 6 minutes.

Attribute	Price/h (pre-emptible)	Amount	Price
Virtual CPU	0.00768 USD	4	0.0686 USD
Memory, 1 GB	0.00103 USD	16	0.0368 USD
Total	0.0472 USD	-	0.105 USD

Table 5.3. Price calculations for the pipeline. The price takes into account the duration of the computation.

results. An example of the tool's output after filtering and grouping the GWS variants in phenotype 366.2 can be seen in table 5.5. An example of the tool's final outputs, the complete variant report and the group report, for phenotype 366.2 can be seen in tables 5.6 and 5.7, respectively.

As we can see, The amount of groups and significant associations differs by a great margin between phenotypes. For example, in the case of phenotype 366.2, Senile cataract, the variants are grouped into separate groups, with the exception of two variants, which contain no other variants in their groups. This can be seen in figure 5.2. The amount of groups and the amount of associations is relatively small. In some other phenotypes, such as 335, multiple sclerosis, there is a large amount of large groups in chromosome 6, which makes interpreting this region difficult. The amount of groups and associations is relatively large. In addition, some variants did not group with other variants. This can be seen in figure 5.3.

Phenotype	Description	Variants	# gr.	# Novel gr.	# Assoc. gr.
153.2	Colon cancer	59	3	1	2
165.1	Cancer of bronchus; lung	313	5	3	2
172	Skin cancer	6491	194	129	65
174.11	Malignant neoplasm of female breast	857	32	15	17
185	Cancer of Prostate	923	171	149	22
250.1	Type 1 diabetes	11680	528	466	62
250.2	Type 2 diabetes	6914	281	207	74
290.11	Alzheimer's disease	306	6	0	6
332	Parkinson disease	0	0	0	0
335	Multiple sclerosis	6619	370	337	33
362.2	Degeneration of macula and posterior pole of retina	249	4	2	2
365.11	Primary open angle glaucoma	152	5	2	3
366.2	Senile cataract	273	9	5	4
411.2	Myocardial infarction	2615	37	13	24
427.2	Atrial fibrillation and flutter	4497	118	63	55
495	Asthma	12036	465	371	94
496	Chronic airway obstruction	971	10	5	5
555	Inflammatory bowel disease and other gastroenteritis and colitis	2426	82	57	25
696.4	Psoriasis	11162	353	303	50
714.1	Rheumatoid arthritis	7653	471	423	48
715.2	Ankylosing spondylitis	6078	216	183	33
939	Atopic/contact dermatitis due to other or unspecified	14	3	2	1

Table 5.4. Reporting tool result summary. The tool found significant results in 21 of the 22 phenotypes, with no significant results found in phenotype 332, Parkinson's disease. All of the results were grouped. In some phenotypes, such as 250.1, Type 1 diabetes, there is a large amount of single variants that were not grouped with any other variants. This is due to LD information not being available for them, for example because the variant is not biallelic, or because of inconsistencies between the LD panel and the summary statistic used.

#chrom	pos	ref	alt	beta	sebeta	pval	#variant	locus_id	pos_rmax	pos_rmin
11	69140554	G	A	0.209	0.0345	1.29E-09	chr11_69140554_G_A	chr11_69172840_G_C	69200682	69140554
11	69141754	T	C	0.21	0.0345	1.19E-09	chr11_69141754_T_C	chr11_69172840_G_C	69200682	69140554
11	69141980	A	C	0.21	0.0345	1.08E-09	chr11_69141980_A_C	chr11_69172840_G_C	69200682	69140554
11	69146940	C	T	0.209	0.0345	1.41E-09	chr11_69146940_C_T	chr11_69172840_G_C	69200682	69140554
11	69147244	C	T	0.209	0.0345	1.26E-09	chr11_69147244_C_T	chr11_69172840_G_C	69200682	69140554
11	69148063	C	T	0.208	0.0345	1.52E-09	chr11_69148063_C_T	chr11_69172840_G_C	69200682	69140554
11	69148200	A	G	0.208	0.0345	1.51E-09	chr11_69148200_A_G	chr11_69172840_G_C	69200682	69140554
11	69148203	C	T	0.209	0.0345	1.41E-09	chr11_69148203_C_T	chr11_69172840_G_C	69200682	69140554
11	69148569	C	T	0.208	0.0345	1.52E-09	chr11_69148569_C_T	chr11_69172840_G_C	69200682	69140554
11	69148979	A	G	0.208	0.0345	1.52E-09	chr11_69148979_A_G	chr11_69172840_G_C	69200682	69140554
11	69149111	T	C	0.208	0.0345	1.51E-09	chr11_69149111_T_C	chr11_69172840_G_C	69200682	69140554
11	69149614	A	G	0.208	0.0345	1.64E-09	chr11_69149614_A_G	chr11_69172840_G_C	69200682	69140554
11	69149883	T	C	0.208	0.0344	1.62E-09	chr11_69149883_T_C	chr11_69172840_G_C	69200682	69140554
11	69150704	T	C	0.208	0.0344	1.64E-09	chr11_69150704_T_C	chr11_69172840_G_C	69200682	69140554
11	69151517	A	C	0.206	0.0344	2.22E-09	chr11_69151517_A_C	chr11_69172840_G_C	69200682	69140554
11	69151839	T	C	0.208	0.0344	1.56E-09	chr11_69151839_T_C	chr11_69172840_G_C	69200682	69140554
11	69152181	A	G	0.208	0.0344	1.51E-09	chr11_69152181_A_G	chr11_69172840_G_C	69200682	69140554
11	69153920	A	G	0.209	0.0345	1.27E-09	chr11_69153920_A_G	chr11_69172840_G_C	69200682	69140554
11	69154651	G	A	0.211	0.0345	9.32E-10	chr11_69154651_G_A	chr11_69172840_G_C	69200682	69140554
11	69155266	C	T	0.21	0.0345	1.05E-09	chr11_69155266_C_T	chr11_69172840_G_C	69200682	69140554
11	69155745	C	T	0.21	0.0345	1.11E-09	chr11_69155745_C_T	chr11_69172840_G_C	69200682	69140554
11	69158599	C	T	0.186	0.0328	1.42E-08	chr11_69158599_C_T	chr11_69172840_G_C	69200682	69140554
11	69159890	A	G	0.197	0.0267	1.67E-13	chr11_69159890_A_G	chr11_69172840_G_C	69200682	69140554
11	69171308	G	A	0.216	0.0276	4.7E-15	chr11_69171308_G_A	chr11_69172840_G_C	69200682	69140554
11	69171980	A	G	0.19	0.0267	9.87E-13	chr11_69171980_A_G	chr11_69172840_G_C	69200682	69140554
11	69172840	G	C	0.221	0.0279	2.85E-15	chr11_69172840_G_C	chr11_69172840_G_C	69200682	69140554
11	69173724	T	G	0.22	0.028	3.49E-15	chr11_69173724_T_G	chr11_69172840_G_C	69200682	69140554
11	69174306	T	C	0.22	0.028	3.27E-15	chr11_69174306_T_C	chr11_69172840_G_C	69200682	69140554
11	69174319	G	A	0.221	0.028	2.91E-15	chr11_69174319_G_A	chr11_69172840_G_C	69200682	69140554

Table 5.5. Some variants from the filtering & grouping output of the tool for phenotype 366.2, Senile cataract. The tool has filtered the variants into GWS variants and grouped them based on their LD structure.The variants shown belong to 5 distinct groups, The group lead variant is shown in the locus_id column. The pos_rmax and pos_rmin columns show the position coordinates of the last and first variant in the group.

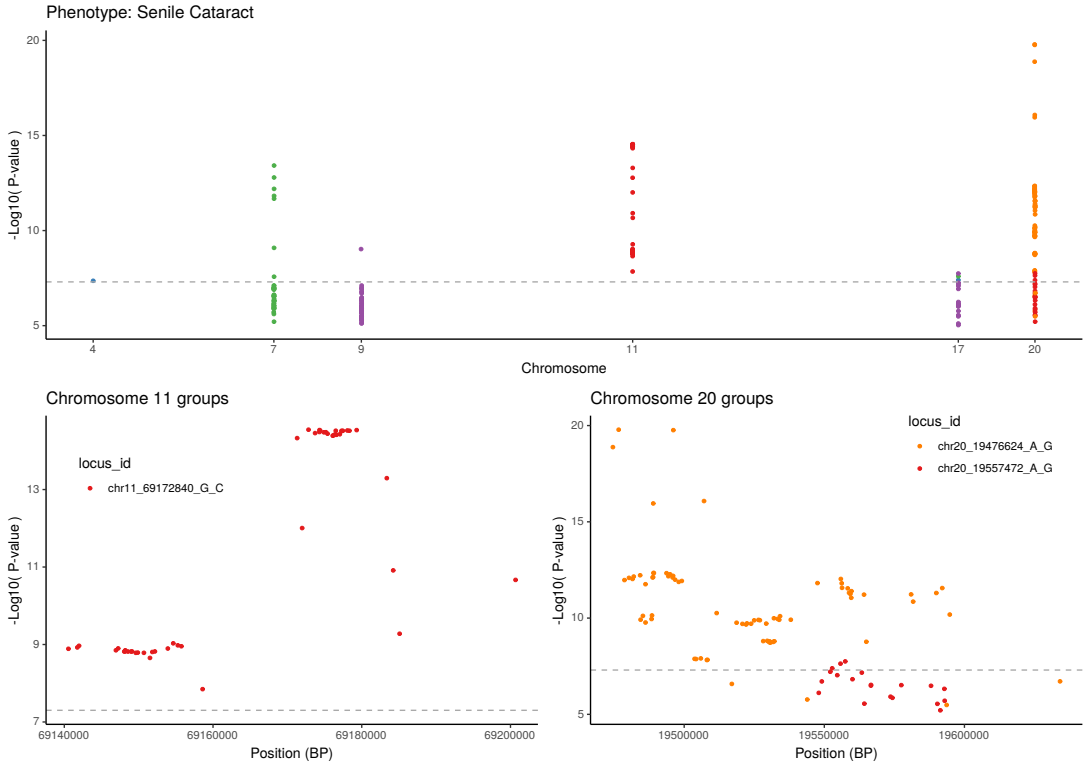

Figure 5.2. A Manhattan plot of the phenotype 366.2, Senile cataract, significant results, as well as closer views of the separate groups in chromosomes 11 and 20. In chromosome 11, all of the significant variants are neatly grouped into one group. In chromosome 20, the variants are grouped into two overlapping groups that would be difficult if not impossible to separate without the LD-based grouping. Note that variants with p-values larger than the significance threshold have been included into the groups, due to the LD between the group lead variant and those variants.

#chrom	pos	ref	alt	beta	sebeta	pval	#variant	locus_id	pos_rmax	pos_rmin	pval_trait	#variant_hit	trait	trait_name
11	69178340	T	C	0.221	0.028	3.03E-15	chr11_69178340_T_C	chr11_69172840_G_C	69200682	69140554				
11	69179313	C	T	0.221	0.028	2.92E-15	chr11_69179313_C_T	chr11_69172840_G_C	69200682	69140554				
11	69183367	C	T	0.212	0.0282	5.08E-14	chr11_69183367_C_T	chr11_69172840_G_C	69200682	69140554				
11	69184229	G	A	0.163	0.0241	1.22E-11	chr11_69184229_G_A	chr11_69172840_G_C	69200682	69140554				
11	69185090	G	A	0.229	0.0369	5.27E-10	chr11_69185090_G_A	chr11_69172840_G_C	69200682	69140554	4E-08	chr11_69185090_A_G	EFO_0004339	body height
11	69200682	C	A	0.168	0.0251	2.15E-11	chr11_69200682_C_A	chr11_69172840_G_C	69200682	69140554				
17	83094678	G	A	0.0928	0.0169	3.97E-08	chr17_83094678_G_A	chr17_83094678_G_A	83094678	83094678				
17	83100564	T	C	0.0945	0.017	2.53E-08	chr17_83100564_T_C	chr17_83100564_T_C	83100564	83100564				
17	83104098	A	G	0.0957	0.017	1.81E-08	chr17_83104098_A_G	chr17_83104098_A_G	83112886	83104098	3E-18	chr17_83104098_G_A	EFO_0005842	colorectal cancer
17	83104098	A	G	0.0957	0.017	1.81E-08	chr17_83104098_A_G	chr17_83104098_A_G	83112886	83104098	3E-18	chr17_83104098_G_A	EFO_0005406	colorectal adenoma
17	83104230	T	C	0.0877	0.0179	8.97E-07	chr17_83104230_T_C	chr17_83104098_A_G	83112886	83104098				
17	83104291	A	G	0.0876	0.0178	9.17E-07	chr17_83104291_A_G	chr17_83104098_A_G	83112886	83104098				
17	83104747	C	T	0.0873	0.0178	9.88E-07	chr17_83104747_C_T	chr17_83104098_A_G	83112886	83104098				
17	83105016	A	G	0.0882	0.0178	7.62E-07	chr17_83105016_A_G	chr17_83104098_A_G	83112886	83104098				
17	83106035	A	G	0.0942	0.0176	8.13E-08	chr17_83106035_A_G	chr17_83104098_A_G	83112886	83104098				
17	83106294	G	T	0.0952	0.0175	5.49E-08	chr17_83106294_G_T	chr17_83104098_A_G	83112886	83104098				
17	83106373	G	A	0.0952	0.0176	6.46E-08	chr17_83106373_G_A	chr17_83104098_A_G	83112886	83104098				
17	83106725	G	A	0.0866	0.0186	3.25E-06	chr17_83106725_G_A	chr17_83104098_A_G	83112886	83104098				
17	83107253	G	A	0.0932	0.0176	1.17E-07	chr17_83107253_G_A	chr17_83104098_A_G	83112886	83104098				
17	83107668	C	T	0.0772	0.0173	7.73E-06	chr17_83107668_C_T	chr17_83104098_A_G	83112886	83104098				
17	83107711	C	T	0.0766	0.0173	9.34E-06	chr17_83107711_C_T	chr17_83104098_A_G	83112886	83104098				
17	83108165	G	A	0.0873	0.0186	2.79E-06	chr17_83108165_G_A	chr17_83104098_A_G	83112886	83104098				
17	83109824	T	C	0.0871	0.0176	7.14E-07	chr17_83109824_T_C	chr17_83104098_A_G	83112886	83104098				
17	83112879	C	G	0.0968	0.0194	5.69E-07	chr17_83112879_C_G	chr17_83104098_A_G	83112886	83104098				
17	83112886	G	C	0.093	0.0194	1.7E-06	chr17_83112886_G_C	chr17_83104098_A_G	83112886	83104098				
20	19474559	G	A	0.215	0.0238	1.32E-19	chr20_19474559_G_A	chr20_19476624_A_G	19634107	19474559				
20	19476624	A	G	0.208	0.0224	1.64E-20	chr20_19476624_A_G	chr20_19476624_A_G	19634107	19474559				
20	19478684	A	C	0.132	0.0185	1.06E-12	chr20_19478684_A_C	chr20_19476624_A_G	19634107	19474559				
20	19480307	G	C	0.133	0.0185	8.15E-13	chr20_19480307_G_C	chr20_19476624_A_G	19634107	19474559				
20	19481563	A	C	0.132	0.0185	9.29E-13	chr20_19481563_A_C	chr20_19476624_A_G	19634107	19474559				
20	19481992	G	A	0.133	0.0185	6.97E-13	chr20_19481992_G_A	chr20_19476624_A_G	19634107	19474559				
20	19484282	C	T	0.134	0.0186	5.97E-13	chr20_19484282_C_T	chr20_19476624_A_G	19634107	19474559				

Table 5.6. The variants from 5.5 after the pipeline. Four additional columns have been added to the results: the matched variant ids from GWAS Catalog, shown in column #variant_hit, the p-value of the reported association in pval_trait column, the Experimental Factor Ontology (EFO) code for the associated trait in trait column, and the trait name in trait_name column.

locus_id	chr	start	end	enrichment	lead_pval	matching_pheno_gwas_catalog_hits	other_gwas_hits
chr11_69172840_G_C	11	69140554	69200682		2.85E-15		body height
chr17_83094678_G_A	17	83094678	83094678		3.97E-08		
chr17_83100564_T_C	17	83100564	83100564		2.53E-08		
chr17_83104098_A_G	17	83104098	83112886		1.81E-08		colorectal cancer; colorectal adenoma
chr20_19476624_A_G	20	19474559	19634107		1.64E-20		migraine disorder; pulse pressure measurement; diastolic blood pressure; FEV/FEC ratio
chr20_19557472_A_G	20	19548023	19592951		1.78E-08		
chr4_173978109_G_A	4	173978109	173978109		4.38E-08		
chr7_46174655_A_G	7	46083961	46174655		3.81E-14		
chr9_22206988_C_T	9	22206988	22373458		9.41E-10		chronic lymphocytic leukemia

Table 5.7. The group report for phenotype 366.2, Senile cataract. The associations from GWAS Catalog can be seen in the other_gwas_hits column.

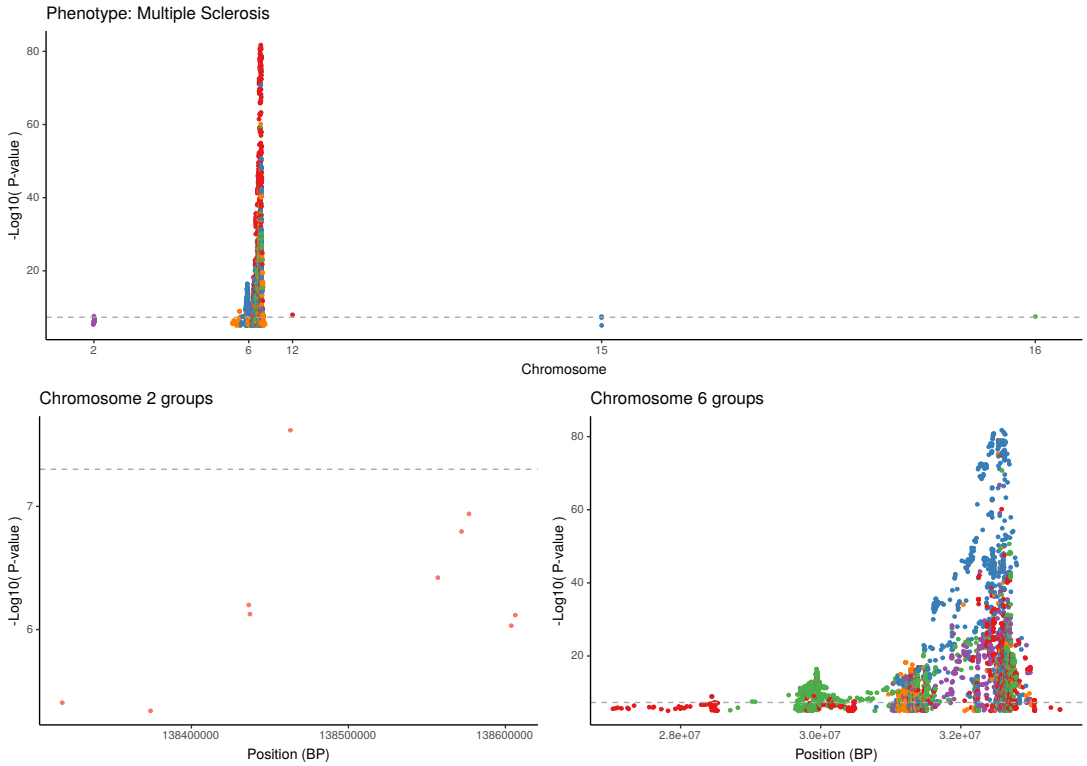

Figure 5.3. A Manhattan plot of the phenotype 335, Multiple sclerosis, significant results, as well as closer views of the chromosomes 2 and 6. A very large majority of the significant results are located in chromosome 6, around the HLA region. The LD structure around the HLA region seems to be quite complex, with many overlapping groups of variants. In addition to them, there is a large amount of variants that did not group with other variants. This is again most likely due to inconsistencies between the LD panel and summary statistic.

6. Discussion

The tool processed all of the phenotypes in a relatively small amount of time and with relatively small computational and monetary expenses. The processing of the phenotypes can be easily scaled up to handle hundreds or thousands of phenotypes. Considering the costs of 0.10 EUR per 22 phenotypes, processing 1000 phenotypes could perhaps be processed with less than 10 EUR in computing costs.

An interesting question that arises from the results is how big are the computation time savings acquired from giving multiple tasks for a single worker vs giving only one task per worker.Based on the timing information in 5.2, we can estimate the overhead per worker and computation time per phenotype. The calculations are presented in table 6.1, and the estimated savings based on the experiment timing data can be seen in figure 6.1. The different plots in figure 6.1 show the proportional differences in compute times with different assumed worker overheads.

Figure 6.1 shows that the estimated savings in computing time are significant, being either approximately 40% or 25% in the case of 3 tasks per worker, depending on assumptions. If all time other than computation was assumed to be worker-specific overhead, then the time savings were larger than if the worker-specific overhead was assumed to only consist of initialization (data transfer).

One aspect that the timing estimation does not take into account is the possibility of a worker to be pre-empted. Presumably, a worker taking more time to complete the work it has been given is under a larger risk of being pre-empted during its work. Therefore assigning more tasks for a worker might affect the expected computation time and expected real time taken more that this simple model estimates. Another point that the timing estimation does not take into account is that the amount of summary statistics moved to the workers depends on the amount of tasks given to a worker. However, because the summary statistics are much smaller in size than the other input data that does not depend on task amount, this effect is small.

Pipeline part	Mean
Computation/Pheno	03:37.76
Initialization	02:26.21
Other	03:55.96
Overhead	06:22.17

Table 6.1. Calculated timing information & their means. Overhead is the sum of the times for initialization and other time taken by the worker.

The tool was also successfully ran on the FinnGen project release 3 data. Due to the

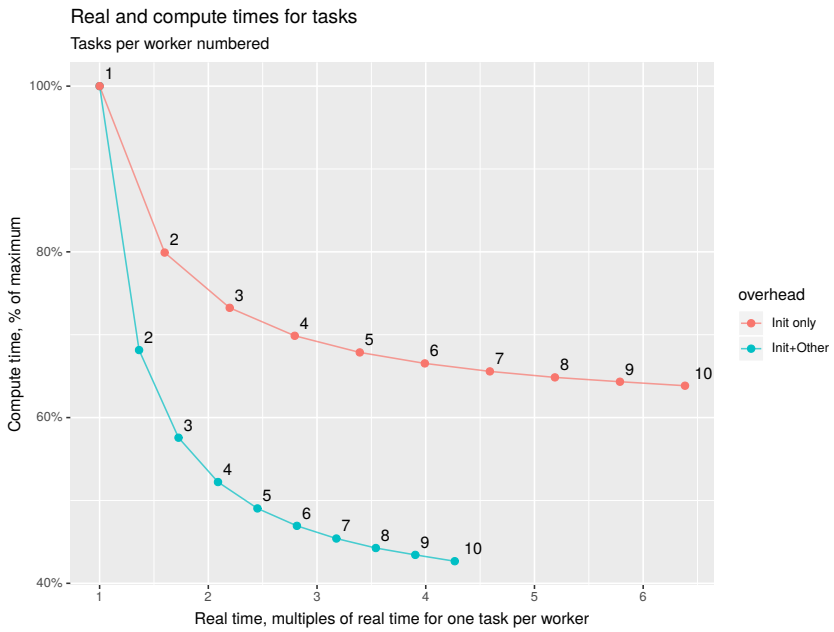

Figure 6.1. Compute & real time required when processing multiple tasks on a single worker. The x-axis shows the required real time as multiples of the time taken if only one task was assigned per worker. The processing time per phenotype and the initialization/overhead time per worker were estimated from the experiment timing data. Values from table 6.1 were used.

proprietary nature of the results, they cannot be presented here. A summary of the performance is therefore presented. The processing took approximately 95 hours 53 minutes of computing time per 1000 phenotypes. Out of that, 3% of total computing time was spent on pre-empted computation, i.e. the workers were cancelled before completion due to other users of Google Cloud requesting those resources, and new workers were tasked with the same computations. The computation cost, based on google VM pricing, was approximately 5.72 USD per 1000 phenotypes. With 4 phenotypes per worker, approximately 46% of the worker time was spent on actual computation. The time taken for everything else by the worker, i.e. the processing overhead, was 54%. The differences to the experiment can be partially explained with the differences in input data. For this analysis, a much larger LD panel was used, which increases the time taken for initializing each worker. A picture of the estimated compute & real times for different amounts of phenotypes can be seen in figure 6.2.

While the tool did complete the analysis for all of the phenotypes and the filtering, comparing and annotation worked well, the grouping of variants did not perform flawlessly with all variants. With some phenotypes, such as type 1 diabetes, some variants were grouped into groups with only a lead variant. This cluttered the group report output with many groups of only one variant and made it more difficult to interpret, diminishing its value. This can be seen in figures 6.3 and 6.4. One explanation for this problem is the fact that the grouping algorithm uses LD information acquired from an LD panel, which in the case of this experiment comes from a different dataset compared to the summary statistic. Those datasets don't completely overlap, which means that some of the variants in the

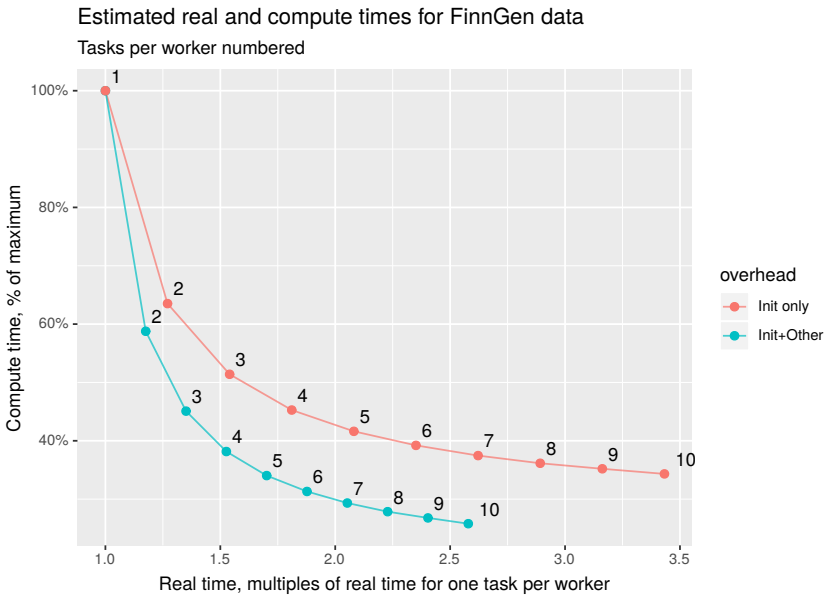

Figure 6.2. Compute & real time estimates for different amounts of tasks per worker. The x-axis shows the required real time as multiples of the time taken if only one task was assigned per worker. The processing time per phenotype and the initialization/overhead time per worker were estimated from the timing data acquired from running the autoreporting tool for FinnGen release 3 data.

summary statistics were not available in the LD panel. Variants missing LD information can not be grouped, and therefore they end up as single variant groups. For the intended purpose of this tool, i.e. processing summary statistics of the FinnGen project, this is not a problem. This is because the LD panel used for those summary statistics includes all of the variants in the summary statistics.

A possible improvement, in addition to using summary statistics and LD information that come from a same source, is not to include the single variants in the group summary report. These variants would still be in the complete variant list, and they would not clutter the group report.

The tool is useful, and it does fulfill the need it was created for, but further improvement can be done. One specific improvement that could be done is extending the association information in the group-specific report with associations that are in linkage disequilibrium with the group's variants. This could be done by e.g. selecting all of the associations in a certain range from the group, calculating LD between the group and the associations, and by reporting all of those associations that are in higher LD with the group's variants than a set threshold. This procedure would make it possible to identify associations that are not exact matches with the significant variants, but are in LD with the group of significant variants. This would be somewhat akin to the LD clumping procedure, in that variants in LD with the signal would get included to the signal. Another improvement would be highlighting finemapped variants in the reports. Considering that the finemapped variants are in a sense more likely to be the true causal signals compared to just looking at high p-values, this could make it easier to identify interesting signals in the reports.

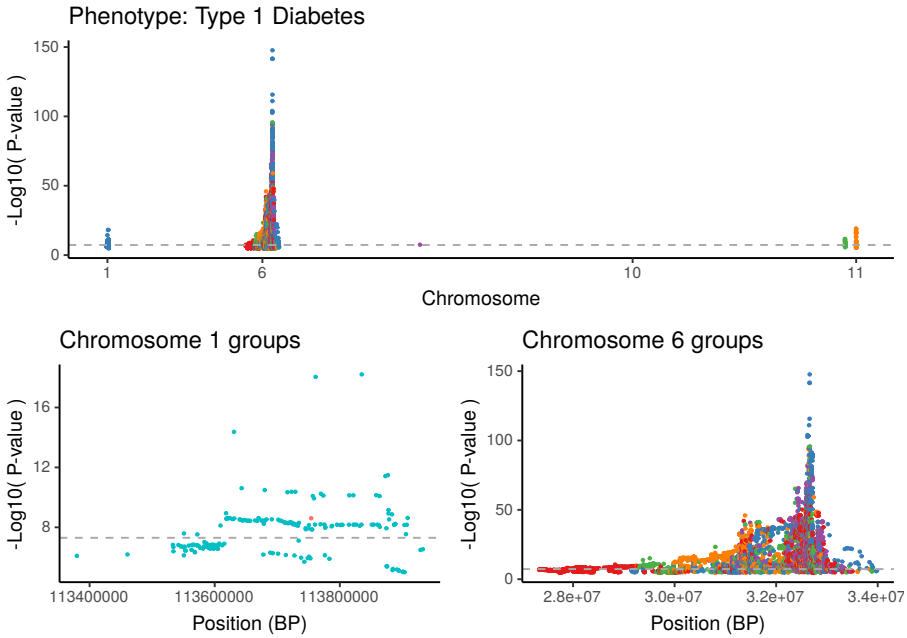

Figure 6.3. A Manhattan plot of the phenotype 250.1, Type 1 Diabetes, significant results, as well as closer views of the chromosomes 1 and 6. A very large majority of the significant results are located in chromosome 6, around the HLA region. The LD structure around the HLA region seems to be quite complex, with many overlapping groups of variants.

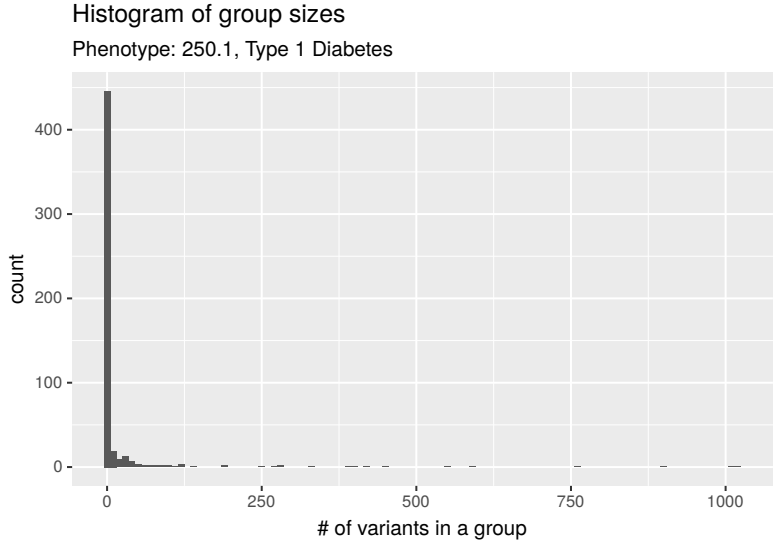

Figure 6.4. Histogram of group sizes for phenotype 250.1, Type 1 Diabetes. There is a large amount (>400) of groups with only a single variant in them.

7. Conclusion

Genome-wide association studies produce a great amount of data, and recent projects such as FinnGen and UK Biobank have increased the amount of analysed traits to hundreds or thousands of different traits, with a GWA analysis done for each trait. The automatic reporting tool successfully filters and annotates this information, as well as compares it to the GWAS Catalog, an association database. The resulting information provides researchers an automatic way of distilling the summary statistics based on their needs. The WDL pipeline for the tool makes it possible to easily perform the pipeline to data of any size, and makes it possible to replicate the operation with minimal effort. This keeps the results comparable and removes a source for error.

Using the tool, 22 GWA analysis summary statistics were filtered and compared against GWAS Catalog. The tool performed well, and the resource usage data and cost analysis indicates that the tool scales well for tens or hundreds of times larger amounts of traits, while still being comparatively inexpensive. The results showed that the tool performed well, and despite a few problem areas, mainly regarding the LD clumping of variants and a small amount of incompatibility between the LD panel and the summary statistics, the tool succeeded in is goals. Some improvements to the tool were identified.

Bibliography

[1] Nicholas J Timpson et al. “Genetic architecture: the shape of the genetic contribution to human traits and disease”. In: *Nature Reviews Genetics* 19.2 (2018), p. 110.

[2] Francis S Collins, Michael Morgan, and Aristides Patrinos. “The Human Genome Project: lessons from large-scale biology”. In: *Science* 300.5617 (2003), pp. 286–290.

[3] *Human Genome Project, NHGRI, webpage*. https://www.genome.gov/human-genome-project/. Accessed: 2019-09-30.

[4] 1000 Genomes Project Consortium et al. “An integrated map of genetic variation from 1,092 human genomes”. In: *Nature* 491.7422 (2012), p. 56.

[5] International HapMap Consortium et al. “The international HapMap project”. In: *Nature* 426.6968 (2003), p. 789.

[6] Kelly A Frazer et al. “Human genetic variation and its contribution to complex traits”. In: *Nature Reviews Genetics* 10.4 (2009), p. 241.

[7] Konrad J Karczewski et al. “Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes”. In: *BioRxiv* (2019), p. 531210.

[8] Peter H Sudmant et al. “An integrated map of structural variation in 2,504 human genomes”. In: *Nature* 526.7571 (2015), p. 75.

[9] B.A. Pierce. *Genetics: A Conceptual Approach*. W. H. Freeman, 2016. ISBN: 9781319127121.

[10] Evan A Boyle, Yang I Li, and Jonathan K Pritchard. “An expanded view of complex traits: from polygenic to omnigenic”. In: *Cell* 169.7 (2017), pp. 1177–1186. URL: https://doi.org/10.1016/j.cell.2017.05.038.

[11] Guy Sella and Nicholas H Barton. “Thinking About the Evolution of Complex Traits in the Era of Genome-Wide Association Studies”. In: *Annual Review of Genomics and Human Genetics* 20 (2017).

[12] William S Bush and Jason H Moore. “Genome-wide association studies”. In: *PLoS computational biology* 8.12 (2012), e1002822. URL: https://doi.org/10.1371/journal.pcbi.1002822.

[13] Stephan Ripke et al. “Biological insights from 108 schizophrenia-associated genetic loci”. In: *Nature* 511.7510 (2014), p. 421.

[14] Bat-sheva Kerem et al. “Identification of the cystic fibrosis gene: genetic analysis”. In: *Science* 245.4922 (1989), pp. 1073–1080.

[15] Peter M Visscher et al. “10 years of GWAS discovery: biology, function, and translation”. In: *The American Journal of Human Genetics* 101.1 (2017), pp. 5–22. URL: https://doi.org/10.1016/j.ajhg.2017.06.005.

[16] Pak C. Sham and Shaun M. Purcell. “Statistical power and significance testing in large-scale genetic studies”. In: *Nature Reviews Genetics* 15 (2014). Review Article, 335 EP –. URL: https://doi.org/10.1038/nrg3706.

[17] Cathie Sudlow et al. “UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age”. In: *PLoS medicine* 12.3 (2015), e1001779.

[18] Annalisa Buniello et al. “The NHGRI-EBI GWAS Catalog of published genome-wide association studies, targeted arrays and summary statistics 2019”. In: *Nucleic acids research* 47.D1 (2018), pp. D1005–D1012.

[19] Joel N Hirschhorn and Mark J Daly. “Genome-wide association studies for common diseases and complex traits”. In: *Nature reviews genetics* 6.2 (2005), p. 95.

[20] Mingyao Li, Chun Li, and Weihua Guan. “Evaluation of coverage variation of SNP chips for genome-wide association studies”. In: *European Journal of Human Genetics* 16.5 (2008), p. 635.

[21] Bryan N Howie, Peter Donnelly, and Jonathan Marchini. “A flexible and accurate genotype imputation method for the next generation of genome-wide association studies”. In: *PLoS genetics* 5.6 (2009), e1000529.

[22] Yun Li et al. “Genotype imputation”. In: *Annual review of genomics and human genetics* 10 (2009), pp. 387–406.

[23] Aarno Palotie et al. “FinnGen-tutkimuksen lupaukset”. In: *Duodecim* (2019).

[24] Clare Bycroft et al. “The UK Biobank resource with deep phenotyping and genomic data”. In: *Nature* 562.7726 (2018), p. 203.

[25] Ben Hayes. “Overview of statistical methods for genome-wide association studies (GWAS)”. In: *Genome-wide association studies and genomic prediction*. Springer, 2013, pp. 149–169.

[26] Alvin C Rencher and G Bruce Schaalje. *Linear models in statistics*. John Wiley & Sons, 2008.

[27] Ludwig Fahrmeir et al. *Regression: models, methods and applications*. Springer Science & Business Media, 2013.

[28] Wei Zhou et al. “Efficiently controlling for case-control imbalance and sample relatedness in large-scale genetic association studies”. In: *Nature genetics* 50.9 (2018), p. 1335.

[29] Gilles Thomas et al. “Multiple loci identified in a genome-wide association study of prostate cancer”. In: *Nature genetics* 40.3 (2008), p. 310.

[30] Hyun Min Kang et al. “Efficient control of population structure in model organism association mapping”. In: *Genetics* 178.3 (2008), pp. 1709–1723.

[31] Xiang Zhou and Matthew Stephens. “Genome-wide efficient mixed-model analysis for association studies”. In: *Nature genetics* 44.7 (2012), p. 821.

[32] Po-Ru Loh et al. “Efficient Bayesian mixed-model analysis increases association power in large cohorts”. In: *Nature genetics* 47.3 (2015), p. 284.

[33] Erich Leo Lehmann and Joseph P Romano. “Generalizations of the familywise error rate”. In: *Selected Works of EL Lehmann*. Springer, 2012, pp. 719–735.

[34] International HapMap Consortium et al. “A haplotype map of the human genome”. In: *Nature* 437.7063 (2005), p. 1299.

[35] Bogdan Pasaniuc and Alkes L Price. “Dissecting the genetics of complex traits using summary association statistics”. In: *Nature Reviews Genetics* 18.2 (2017), p. 117.

[36] Padhraig Gormley et al. “Meta-analysis of 375,000 individuals identifies 38 susceptibility loci for migraine”. In: *Nature genetics* 48.8 (2016), p. 856.

[37] Daniel J Schaid, Wenan Chen, and Nicholas B Larson. “From genome-wide associations to candidate causal variants by statistical fine-mapping”. In: *Nature Reviews Genetics* 19.8 (2018), p. 491.

[38] Christian Benner et al. “FINEMAP: efficient variable selection using summary data from genome-wide association studies”. In: *Bioinformatics* 32.10 (2016), pp. 1493–1501.

[39] Luke C Pilling et al. “Human longevity is influenced by many genetic variants: evidence from 75,000 UK Biobank participants”. In: *Aging (Albany NY)* 8.3 (2016), p. 547.

[40] Shaun M Purcell and Christopher C Chang. “PLINK 1.9”. In: (). URL: https://www.cog-genomics.org/plink/1.9/.

[41] Heng Li. “Tabix: fast retrieval of sequence features from generic TAB-delimited files”. In: *Bioinformatics* 27.5 (Jan. 2011), pp. 718–719. ISSN: 1367-4803. DOI: 10.1093/bioinformatics/btq671. eprint: http://oup.prod.sis.lan/bioinformatics/article-pdf/27/5/718/5504485/btq671.pdf. URL: https://doi.org/10.1093/bioinformatics/btq671.

[42] *bgzip, webpage*. http://www.htslib.org/doc/bgzip.html. Accessed: 2019-08-20.

[43] Tim Beck et al. “GWAS Central: a comprehensive resource for the comparison and interrogation of genome-wide association studies”. In: *European journal of human genetics* 22.7 (2014), p. 949.

[44] *GWAS Catalog, webpage*. https://www.ebi.ac.uk/gwas/. Accessed: 2019-09-04.

[45] Himanshu Chheda et al. “Whole-genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom”. In: *European Journal of Human Genetics* 25.4 (2017), p. 477.

[46] Elaine T Lim et al. “Distribution and medical impact of loss-of-function variants in the Finnish founder population”. In: *PLoS genetics* 10.7 (2014), e1004494.

[47] Juha Kere. “Human population genetics: lessons from Finland”. In: *Annual review of genomics and human genetics* 2.1 (2001), pp. 103–128.

[48] Aarno Palotie et al. “FinnGen-tutkimus luo perustaa genomitiedon hyödyntämiseksi terveydenhuollossa”. In: *Duodecim* (2018).

[49] *FinnGen research project, webpage*. https://www.finngen.fi/. Accessed: 2019-08-27.

[50] John Vivian et al. “Toil enables reproducible, open source, big biomedical data analyses”. In: *Nature biotechnology* 35.4 (2017), p. 314.

[51] *WDL workflow language, webpage*. https://software.broadinstitute.org/wdl/. Accessed: 2019-08-19.

[52] *OpenWDL homepage*. http://www.openwdl.org/. Accessed: 2019-07-23.

[53] *Python programming language, webpage*. https://www.python.org/. Accessed: 2019-09-03.

[54] Wes McKinney. “Data Structures for Statistical Computing in Python”. In: (2010). Ed. by Stéfan van der Walt and Jarrod Millman, pp. 51 –56.

[55] Christian Benner et al. “Prospects of fine-mapping trait-associated genomic regions by using summary statistics from genome-wide association studies”. In: *The American Journal of Human Genetics* 101.4 (2017), pp. 539–551.

[56] Christopher C Chang et al. “Second-generation PLINK: rising to the challenge of larger and richer datasets”. In: *GigaScience* 4.1 (Feb. 2015). ISSN: 2047-217X. DOI: 10.1186/s13742-015-0047-8. eprint: http://oup.prod.sis.lan/gigascience/article-pdf/4/1/s13742-015-0047-8/25512027/13742_2015_article_47.pdf. URL: https://doi.org/10.1186/s13742-015-0047-8.

[57] Yuan Chen et al. “Ensembl variation resources”. In: *BMC genomics* 11.1 (2010), p. 293.

[58] *liftOver tool, webpage*. https://genome.ucsc.edu/cgi-bin/hgLiftOver. Accessed: 2019-09-06.

[59] *Lee lab for Statistical Genetics and Genomics, webpage*. https://www.leelabsg.org. Accessed: 2019-09-06.

[60] *The International Genome Sample Resource data portal, webpage*. https://www.internationalgenome.org/data/. Accessed: 2019-09-06.

[61] *Google Cloud VM Instance Pricing, website*. https://cloud.google.com/compute/vm-instance-pricing. Accessed: 2019-09-07.

[62] *Bank of Finland, webpage*. https://www.suomenpankki.fi/. Accessed: 2019-09-07.

A. Appendix

A.1 Automatic reporting tool parameter table

Argument	Description
–sign-threshold	Signifigance threshold for variants. Variants with a p-value larger than the threshold are filtered out from the results.
–prefix	A prefix for all of the output and temporary files. Useful in cases where there might be confusion between processes running in the same folder. A dot is inserted after the prefix if it is passed.
–fetch-out	Output file path for filtered and/or grouped variants.
–group	Supplying this flag results in the variants being grouped into groups using either grouping by distance from a lead variant or by LD.
–grouping-method	Grouping method used if –group flag is supplied. options are 'simple', i.e. grouping based on distance from lead variant, or 'ld', i.e. grouping using PLINK's LD clumping.
–locus-width-kb	Group widths in kilobases. In case of ld clumping, the value is supplied to plink's –clump-kb flag.
–alt-sign-threshold	Optional alternate signifigance threshold for including less significant variants into groups.
–ld-panel-path	Path to ld panel. Ld panel must be in plink's .bed format, as a single file. Accompanying .bim and .fam files must be in the same directory.
–ld-r2	Plink clump-r2 argument. Variants that are under –locus-width-kb distance away from the lead variant, have a lower p-value than the alternate significance threshold, and have have a squared correlation coefficient larger than this value with the lead variant are included in the group.
–plink-memory	Plink –memory argument. Defines the PLINK maximum allowed memory amount in megabytes.
–overlap	If this flag is supplied, the groups of gws variants are allowed to overlap, i.e. a single variant can appear multiple times in different groups. Variants that are already grouped can not function as lead variants.

Argument	Description
–ignore-region	The region supplied to this flag is ignored in the tool, i.e. the variants in this region will not be included in the output.
–gnomad-genome-path	Path to gnomad genome annotation file. Must be tabixed. Required for annotation.
–gnomad-exome-path	Path to gnomad exome annotation file. Must be tabixed. Required for annotation.
–include-batch-freq	Include batch frequencies from finngen annotation file
–finngen-path	Path to finngen annotation file. Required for annotation.
–annotate-out	annotation output file.
–compare-style	Whether to use gwascatalog and/or additional summary statistics to compare findings to literature. Use values 'file', 'gwascatalog' or 'both'.
–summary-fpath	path to a file containing external summary statistic file paths. List one summary file per line.
–endpoint-fpath	path to a file containing endpoints for summary statistic files. List one endpoint per line. The endpoints should be in the same order as the summary files in –summary-fpath file
–check-for-ld	When supplied, gws variants and summary statistics (from file or gwascatalog) are tested for ld using LDstore.
–report-out	Comparison output file.
–ld-report-out	Output file containing the LD between GWS variants and associations.
–gwascatalog-pval	Associations with p-value smaller than this are included in the GWAS Catalog results.
–gwascatalog-width-kb	The region from which associations are downloaded form GWAS Catalog is incremented up-and downstream with this many kilobases.
–gwascatalog-threads	Number of concurrent queries to gwasgatalog API. Default 4. Increase to speed up gwascatalog comparison.
–ldstore-threads	Number of threads to use with LDstore. At most the number of logical cores in the processor.
–ld-threshold	LD threshold for LDstore. Associations in higher LD with our variants are included.
–cache-gwas	Save GWAScatalog results into a file, so they do not need to be downloaded again. Useful for testing.
–column-labels	One can supply custom input file column names with this (chrom, pos, ref, alt, pval only)
–top-report-out	Filename of top-level report.

Argument	Description
–efo-traits	specific traits that you want to concentrate on the top level locus report. Other found traits will be reported on a separate column from these. Uses Experimental Factor Oncology codes.
–local-gwascatalog	File path to a local copy of GWAS Catalog.
–db	Choose which comparison database to use, gwas catalog proper, gwas catalog's summary statistic api, or a local copy of gwas catalog. With local copy, you need to supply the –local-gwascatalog parameter.
gws_path	Path to the tabixed and gzipped summary statistic that is going to be filtered, annotated and compared. Required argument.

Table 1.1. Parameter list to automatic reporting tool.

A.2 Position grouping pseudocode

```
let Vs = set(genome-wide significant variants)
let Gs = set()
while Vs is not empty:
  let v = min(Vs)
  let g = v' in Vs where distance(v',v) < locus width
  Gs = Gs + g + v
  Vs = Vs - v - g
return Gs
```

Listing A.1. Simple grouping of variants. In this context, the minimum is calculated based on p-value.