Stockholms universitet

Examining the genetic overlap of rheumatoid arthritis and cardiovascular disease

Anton Öberg Sysojev

Masteruppsats i matematisk statistik
Master Thesis in Mathematical Statistics

Masteruppsats 2020:3
Matematisk statistik
Februari 2020

www.math.su.se

Matematisk statistik
Matematiska institutionen
Stockholms universitet
106 91 Stockholm

Stockholm University

Mathematical Statistics
Stockholm University
Master Thesis **2020:3**
http://www.math.su.se

Examining the genetic overlap of rheumatoid arthritis and cardiovascular disease

Anton Öberg Sysojev*

February 2020

Abstract

In this text we assess the amount of overlap in genetics of rheumatoid arthritis (RA) and cardiovascular disease (CVD). We use summary statistic data from genome-wide association studies for RA and acute myocardial infarction (AMI), using AMI to represent CVDs in general. Genetic overlap is measured as a correlation coefficient on the genetic part of the two traits, obtained through the implementation of linkage disequilibrium score regression. No significant genetic correlation was found between RA and AMI, indicating that little to no genetic overlap exists between the two traits, agreeing with a lone previous result. While power was not an issue in this study, a greater sample size for RA could hopefully shrink standard errors and give more precise estimates. Additionally, using only AMI as a proxy for CVDs in general might be naive and different results may be found for other CVD-phenotypes.

*Postal address: Mathematical Statistics, Stockholm University, SE-106 91, Sweden. E-mail: antonsysojev@gmail.com. Supervisor: Tom Britton.

Acknowledgements

I would like to offer my special thanks to my supervisor Helga Westerlind for the opportunity to write this thesis at the clinical epidemiology division of KI. Thank you for introducing me to a wonderfully interesting field and for all your guidance and direction during the writing of this thesis.

A big thank you also to the people over at KEP for welcoming me as one of them during this past semester.

And finally, I would like to thank my second supervisor Tom Britton for support and guidance in understanding and interpreting the mathematical theory behind my topic.

Contents

1 Introduction

1.1 Background

Rheumatoid arthritis (RA) is an autoimmune chronic disorder. Prevalence varies between countries but is estimated at around 1% in Sweden for the general population with a greater incidence in women than in men [1]. The etiology of RA is not fully understood but current consensus is that both environmental and genetic factors play a part in the development of the disease. Several risk factors have been established, of which smoking is considered the most influential environmental risk factor and certain genes in the human leukocyte antigen region have been demonstrated to collectively confer the greatest genetic risk [2].

Disorders of autoimmune type are characterized by the individual's immune system wrongly targeting functioning body parts, which in RA primarily manifests itself in the joints of the body. If left untreated, RA will lead to joint deterioration and physical disability. Neither direct prevention nor cure exists but treatment of the pain and symptoms is possible through medication and physical therapy. Beyond the burden of the disease itself, patients suffering from RA are exposed to a higher risk for various other diseases, which as a result, leads to an increased mortality rate for individuals diagnosed with RA. Such comorbidities include lymphomas, lung diseases and cardiovascular diseases [3].

Cardiovascular diseases (CVDs) cover a broad set of disorders characterized by primarily involving blood vessels and the heart. Such disorders include coronary artery disease, acute coronary syndrome, myocardial infarction (heart attack) and stroke among others. While genetic risk factors are currently poorly understood, many environmental risk factors such as diet, tobacco intake and physical activity are well known and established for CVDs in general [4, 5]. Despite this, CVDs are still the leading cause of death in humans globally, accounting for around 40% of deaths in Europe under the year of 2017 [6].

As stated above, individuals with RA suffer a greater risk for developing CVDs which prompted a recent article in which it was demonstrated that an increased risk could be found in siblings of patients too. This increased risk in turn implies an underlying shared susceptibility between RA and CVD. Whether this shared susceptibility is due to genetic similarities, environmental factors or both remains a question [7].

In this paper we look into this shared susceptibility by investigating the genetic overlap between RA and CVD through statistical examination of genetic data. We aim to study the overlap of RA and CVD by comparing results from genome-wide association studies (GWASs) for RA with GWAS results for acute myocardial infarction (AMI), using it as a proxy for CVDs in general. A previous study aiming to look at the genetic correlation of a broad range of traits found no significant correlation when comparing RA with the CVD of coronary artery disease [8]. We hope that this study can improve upon this estimate by using a more well defined type of CVD and by also studying the genetic overlap over different sub-types of RA.

The first part of this text gives a methodological background into assessing

the genetic overlap between two phenotypes through genome-wide data. We give a brief introduction to some of the more crucial concepts and continue with a thorough reading of the mathematical and statistical framework of linkage disequilibrium score regression. The second half of this text is then concerned with the implementation of linkage disequilibrium score regression, attempting to estimate the genetic overlap of RA and CVD through GWAS summary statistic data on RA and AMI.

Several minor lemmas, their proofs and a few concepts of which background knowledge is not mandatory to follow with in the details of this study have been placed in the opening section of the Appendix (Section A). Additionally, several acronyms are employed to simplify the reading of this study. Details on these acronyms can be found in the concluding glossary, placed towards the end of the Appendix but are also covered in the text as they are introduced.

1.2 Genome-wide association studies

The human genome can be represented as a linear sequence of nucleotides, commonly referred to as the human DNA sequence. While the majority of this DNA sequence is identical between human individuals, there are significant amounts of variation present among the three billion DNA base-pairs that it consists of. The most common type of genetic variation is in the form of single nucleotide polymorphisms (abbreviated SNPs; pronounced "snips"). A SNP is a base-pair change of a single nucleotide in the DNA sequence between different individuals, appearing in at least 1% of the population and they are by far the most abundant form of genetic variation in the human genome [9]. A visual illustration of a SNP is given in Figure 1.

The underlying hypothesis of the GWAS is that human traits (phenotypes) of interest vary as a function of the alleles (nucleotides) at these genetic variants of SNPs. An example of this would be an individual with a copy of an A allele at a given SNP being at higher risk for a type of cancer than a different individual with a copy of a T allele at the same SNP. The GWAS combs through the genome, testing a wide set of SNPs for association with trait of interest by regressing observed phenotypes against observed genotyped alleles for a set of individuals. It then outputs test statistics of each SNPs individual association with trait.

The GWAS was first developed in the early 2000's as a way to identify genomic regions (known as loci) in the form of SNPs, associated with given phenotypes of interest. Its main use was to identify loci that show strong association (either positively or negatively) with the phenotype of interest, as a way to aid researchers in better understanding the etiology and risk factors of complex traits [10]. It has since grown beyond that use, as researchers have developed ways to further study the genetics of phenotypes by using the GWAS output as genetic data. The core strengths of using GWAS data as opposed to genotyped data on an individual-level is twofold: firstly, a performed GWAS reduces data dimensionality immensely by compressing the previously $(N \times (M + 1))$ dimensional data into $(2 \times M)$, where N denotes sampled individuals and M

Figure 1: Visual illustration of a SNP in a DNA-sequence. When comparing the given DNA sequence between the six individuals, a SNP is revealed at the yellow position due to the difference in nucleotides between them. Source: https://www.genome.gov/genetics-glossary/Single-Nucleotide-Polymorphisms, taken on 6/12-2019.

denotes the number of SNPs genotyped, at the loss of very little information helping with reduction in computational complexity. Secondly, the sharing of individual-level genotyped data is difficult as privacy concerns are a big issue. However, the GWAS mitigates this problem by masking each person considered through summary statistics. This has led to the wide sharing of GWAS data between researchers and the establishment of several databases aimed at collecting large sets of meta-analyzed GWAS summary statistic data [11, 12].

In practice, the GWAS is performed as follows: a sample of N individuals are taken from the population with respect to the phenotype of interest. Both categorical phenotypes (disease status, educational attainment, existence of a particular anti-body) and continuous phenotypes (heart rate, BMI, height) are valid for GWASs. Genotyping these individuals for their observed alleles at M targeted SNPs leads to the individual-level genotyped data, dimension $(N \times (M+1))$ where the first column contains the recorded phenotype. These observed alleles are then tested for association with the phenotype through an appropriate regression model, i.e. by regressing the first column of observed phenotype on the observed genotyped alleles for each of the M remaining columns independently. Association significance is tested for through asymptotic Wald-tests, producing test statistics in the form of z-scores z_j for each SNP, $j = 1, ..., M$. Note that we may transform these to chi-squared statistics χ_j^2 by squaring the z-score estimates, i.e $z_j^2 = \chi_j^2$ for the same test hypothesis. The resulting GWAS output is then in the form of test statistics of each of the M SNPs individual associations with the phenotype of interest [9]. A visual representation of the output of a GWAS can be seen in Figure 3 (page 22) where

the results from our GWAS of RA-status is presented in a so called Manhattan plot.

There is however one core issue with this approach. The number M of genotyped SNPs is often large, ranging from several hundred thousands up to millions, with the actual number depending on the genotyping chip utilized on the sample. As each SNP is tested individually against phenotype for association, the GWAS performs an enormous amount of tests, meaning that using the common significance level of $\alpha = 0.05$ will lead to an alarming amount of false positives. Several approaches exist to mitigate this problem of multiple testing in a GWAS but the, in the literature, standard solution is to use an adjusted p-value of $p = 5 \cdot 10^{-8}$ as the threshold for significant SNP-association. This threshold is commonly referred to as the genome-wide significance level and is based on a Bonferroni correction of the significance level for $\alpha = 0.05$ when running a million tests [13]. While Bonferroni corrections are notoriously conservative, this type of threshold is simplistic in nature and has come to be established as an in-field standard to facilitate replication and consensus among researchers without requiring tedious and intricate computations or computer intensive algorithms.

1.3 Linkage disequilibrium

A key factor in why the GWAS can be performed efficiently is due to linkage disequilibrium (LD). LD refers to the non-random association of alleles at different loci in the genome and can be seen as a kind of correlation structure between the alleles at these different regions [10, 14]. Due to this correlation structure, we can reduce the number of genotyped SNPs needed to get an accurate representation of the variation in the human genome, as the information contained in several SNPs is already represented through their correlation with targeted markers [10]. Mathematically we define LD between two genomic regions as follows.

Definition 1. *Consider two biallelic loci on the genome with the first loci having alleles A and a and the second having alleles B and b. The population frequencies are denoted by π_A and π_B with $\pi_a = 1 - \pi_A$ and $\pi_b = 1 - \pi_B$. Letting π_{AB} denote the population frequency of the inherited AB haplotype, we get a measure of the amount of LD between the two loci as*

$$D = \pi_{AB} - \pi_A \pi_B. \tag{1}$$

The measure D can be both positive and negative, the sign depending on the (arbitrary) labeling of the alleles. To this end we note that the magnitude of D is independent of the alleles considered, i.e.

$$\begin{aligned} D &= \pi_{AB} - \pi_A \pi_B = \pi_{ab} - \pi_a \pi_b \\ -D &= \pi_{Ab} - \pi_A \pi_b = \pi_{aB} - \pi_a \pi_b \end{aligned}$$

The magnitude of D in turn depends on the population frequencies of the alleles. For instance, if $\pi_A = \pi_B = \pi_a = \pi_b = 0.5$, the maximum of D is obtained at $D = 0.25$ which happens when $\pi_{AB} = 0.5$. If $D = 0$ we say that the two genomic regions are in linkage equilibrium [14].

The LD measure of D is not always optimal due to its dependence on the population frequencies of the alleles it is referring to. As a result, various different measures and estimates of LD exist in the literature although most of them are simply variations of Definition 1 [14]. A common such measure that we will make frequent reference to in our discussion of linkage disequilibrium score regression (Section 2) is the squared correlation coefficient of r^2.

Definition 2. *Consider two biallelic loci on the genome with the first loci having alleles A and a and the second having alleles B and b. The population frequencies are denoted by π_A and π_B with $\pi_a = 1 - \pi_A$ and $\pi_b = \pi_B$. Let D be as given in Definition 1, then*

$$r^2 = \frac{D^2}{\pi_A(1-\pi_A)\pi_B(1-\pi_B)}. \tag{2}$$

As r^2 is the squared correlation coefficient of alleles at two different loci we note that it inherits all of its nice properties, including being normalized to the range of $[0, 1]$ [14]. It follows that for genomic regions in linkage equilibrium, we get $r^2 = 0$. Furthermore, if two sites are correlated in such a way that $r^2 = 1$, we say the two regions are in complete disequilibrium.

1.4 Genetic overlap and heritability

Studying the genetic overlap of phenotypic traits can lead to newfound knowledge and a greater understanding of the genetic etiology, susceptibility and risk of phenotype development in a population. Plenty of different methods exist for studying genetic association, overlap and correlation but to our knowledge, no single standard approach exists in the literature. In Section B of the Appendix we give a brief introduction to some of the available methods, reviewing their statistical framework and covering some of their strengths and problems.

A common quantity to many of these methods that we will make frequent reference to, is the genetic correlation coefficient, r_g. The quantity r_g is given as a correlation coefficient on the genetic part of the phenotypic trait and is a key feature in methods based on parametric models.

Definition 3. *Let G_i denote the genetic part of any phenotypic trait. The genetic correlation coefficient r_g is then defined as*

$$r_g = \frac{Cov(G_1, G_2)}{\sqrt{Var(G_1)Var(G_2)}} = \frac{\rho_g}{\sqrt{h_1^2 h_2^2}}. \tag{3}$$

Here we denote the genetic covariance of two given traits by ρ_g. We refer to the terms $Var(G_i)$ as the phenotypic variation attributable to genetics, which

in turn, is commonly referred to as the heritability of a phenotype and denoted by h_i^2 for $i = 1, 2$. A more rigorous definition of the heritability along with a further discussion on modeling phenotypic variation is given in the Appendix, Section A.1.

2 Linkage disequilibrium score regression

Studying genetic overlap between phenotypes can lead to a better understanding of the etiology of complex traits. Non-zero genetic correlation can be an indicator of pleiotropy, i.e. of genetic variants affecting both traits but it can also be used to study heterogeneity within phenotypes. As an example, we might believe that a trait is genetically different between men and women or in young and old individuals. Here the null-hypothesis would instead be of a complete genetic overlap as opposed to a non existent one.

The goal of this paper is to assess what genetic overlap exists between individuals with RA and individuals with CVD based on summary statistic data obtained from GWASs. To our knowledge, neither rigorous definition nor established quantifying measure exists within the literature which has led to the development of several approaches to test for, or directly estimate, the genetic overlap between two different phenotypes. Available methods differ in underlying philosophy to assessing overlap, required form of genetic data as well as in their assumptions on the underlying genetic architecture of a phenotype. Some of these include testing for overlap through a genetic component [15], comparing direction of SNP associations between traits [16], or simply to estimate a correlation coefficient through a fitted model [17, 18].

In this section we cover how to use GWAS summary statistic data to estimate r_g (Equation 3) using a novel method called linkage disequilibrium score regression [18]. We give a background to the method in Section 2.1 discussing the development of the model as a way to control for bias in a GWAS and its subsequent extension to estimation of genetic variance components. In Section 2.2 we define the underlying model and prove the main results (Equations 5 and 6). We cover how to efficiently use these results to estimate trait heritabilities h^2 and genetic covariances ρ_g in Section 2.3. A brief review of other available methods for assessing genetic overlap is given in the Appendix, Section B.

2.1 Background

The basis of linkage disequilibrium score regression (hereafter LDSR) is of a linear relationship between GWAS test statistics of SNP association with trait and a quantity called the LD score. It estimates both trait heritability h^2 and genetic covariance ρ_g by regression of test statistics of SNP associations onto LD scores. Now the LD score l_j of a SNP j is given as an aggregated sum of the amount of genetic variation tagged by this SNP over the genome i.e. the amount of LD that SNP j is in. Formally it is defined as follows.

Definition 4. *Consider a set of M_l variants on the human genome. The LD score of a variant j is given by*

$$l_j = \sum_{k=1}^{M_l} r_{jk}^2, \tag{4}$$

where r_{jk}^2 is the r^2 measure of LD defined in Definition 2 for genetic variants j and k.

Note that we here write M_l to distinguish this quantity from the M SNPs that was genotyped for our GWAS. We often wish to compute LD scores using an external set such that $M_l \geq M$ to increase the precision in our estimates.

The method of LDSR was initially developed as a way to measure the amount of confounding in a GWAS due to population stratification, i.e. due to things such as cryptic relatedness and population substructures [19]. The goal was to improve upon the established genomic inflation factor [20] (see Section A.3 in the Appendix) but the method was subsequently extended to include estimation of trait heritability and genetic covariance for sets of GWAS summary statistic data [18,19]. The key result of the original paper was that, the expectation of a given test statistic from a GWAS, in the form of a χ^2-statistic, could be written as

$$E[\chi_j^2] = 1 + \frac{Nh^2}{M_l} l_j, \ j = 1, ..., M, \tag{5}$$

where N is the number of individuals in the GWAS sample, h^2 is the trait heritability, l_j is the LD score of SNP j and M_l is the number of SNPs considered in the computation of the LD scores [19]. There is thus a linear relationship between GWAS test statistics and LD scores in such a way that regressing χ^2 statistics onto LD scores would allow us to obtain an estimate of the trait heritability h^2 by rescaling the obtained estimate of the slope coefficient.

This model was later extended into cross-trait LDSR in which a similar relationship was derived but with respect to the term ρ_g of genetic covariance instead of the trait heritability. The authors of the original papers [18, 19] demonstrated that for two sets of GWAS summary statistic data containing z-scores of SNP association with trait (formally defined in Section 2.2), the expectation of the product of z-scores can be written as

$$E[z_{1j} z_{2j}] = \frac{\rho_s N_s}{\sqrt{N_1 N_2}} + \frac{\sqrt{N_1 N_2}\rho_g}{M_l} l_j, \ j = 1, ..., M, \tag{6}$$

where N_i is the number of individuals in the GWAS sample for trait i, N_s is the number of individuals contained in the sample set for both GWASs, ρ_s is the correlation, in phenotype, of the N_s overlapping individuals, l_j is the LD score of SNP j with M_l being the number of SNPs considered in the computation of the LD scores [18]. We prove this result in Section 2.2 and show that Equation 5 follows as a corollary.

A similar regression model to that discussed above is then appropriate for the estimation of genetic covariance. As quantities N_1, N_2 and M_l are generally

known we may obtain $\hat{\rho}_g$ through a rescaling of the estimated slope coefficient obtained by the regression of product of z-scores onto LD scores. These estimates can then be used in tandem to estimate the genetic correlation coefficient r_g as it was given in Definition 3.

2.2 Underlying model

In this section we define the underlying model for the results in Equations 5 and 6 and prove that these equations hold. These results, including the proof, are based on derivations done in the supplementary material of [18].

The model equation for phenotypic trait i is given by

$$Y_i = X^{(i)}\beta_i + \epsilon_i, \quad i = 1, 2. \tag{7}$$

Here, Y_i is a vector of phenotypic trait values with dimension $(N_i \times 1)$, β_i is a vector of standardized SNP effect sizes dimension $(M_l \times 1)$ and ϵ_i are residuals representing environmental effects and further noise. The matrix $X^{(i)}$ is a standardized genotype matrix, dimension $(N_i \times M_l)$ containing elements $X^{(i)}_{mj}$ of observed alleles at SNP j for individual m in the i'th sample, mean-centered and variance standardized to have expectation zero and variance one. Note that the number M_l is here allowed to be greater than the number M of SNPs genotypes in our GWAS.

Furthermore, we take all terms on the right hand side of Equation 7 as random and mutually independent both with each other as well as between traits, i.e. constituting a random effects model. The two exceptions are of β_1 and β_2 along with ϵ_1 and ϵ_2. For these we assume that (β_1, β_2) has mean zero and variance-covariance matrix given by

$$Var(\beta_1, \beta_2) = \frac{1}{M_l}\begin{bmatrix} h_1^2 I_1 & \rho_g I_1 \\ \rho_g I_1 & h_2^2 I_1 \end{bmatrix},$$

and that (ϵ_1, ϵ_2) has mean zero and variance-covariance matrix

$$Var(\epsilon_1, \epsilon_2) = \begin{bmatrix} (1-h_1^2) I_2 & \rho_e I_2 \\ \rho_e I_2 & (1-h_2^2) I_2 \end{bmatrix}.$$

Here, I_1 is an $(M_l \times M_l)$ identity matrix whereas I_2 has dimension $(N_1 \times N_2)$. As we make no assumptions on the relationship between study sample sizes N_1 and N_2 we allow for I_2 to not be square. As such we define I_2 as a matrix with 1's on the main diagonal and 0's in all other positions, i.e. the ij'th element is equal to 1 if $i = j$ and 0 otherwise.

Furthermore, we let N_s denote the number of individuals that belong to both studies (i.e. the number of identical rows between genotype matrices $X^{(i)}$ and let $\rho_s = \rho_e + \rho_g$. We define LD scores for variant j as

$$l_j = \sum_{k=1}^{M_l} r_{jk}^2 = \sum_{k=1}^{M_l} \left(E[X^{(1)}_{mj} X^{(1)}_{mk}]\right)^2 = \sum_{k=1}^{M_l} \left(E[X^{(2)}_{mj} X^{(2)}_{mk}]\right)^2,$$

where $X^{(i)}_{mj}$ denotes the mj'th element of the genotype matrix for sample i, and r^2_{jk} is the measure of LD defined in Definition 2 with indices denoting the loci in question.

This model is based on an assumption of an underlying polygenic architecture for the traits of interest, i.e. the assumption that many SNPs together make up the bulk of the genetic effect on trait as opposed to a few strongly associated SNPs accounting for the majority of genetic variation.

With the model in place we can prove the key result of LDSR.

Proposition 5. *Under the model detailed in the above text, it holds that for each of the z-scores of SNP association with trait*

$$E[z_{1j}z_{2j}] = \frac{\rho_s N_s}{\sqrt{N_1 N_2}} + \frac{\sqrt{N_1 N_2}\rho_g}{M_l} l_j, \quad j = 1, ..., M, \tag{6, revisited}$$

where z_{ij} is the j'th GWAS test statistic, as a z-score, from study i.

Proof. The z-score for a given SNP j from study i, is given by

$$z_{ij} = \left(X^{(i)}_j\right)^T Y_i / \sqrt{N_i},$$

where $X^{(i)}_j$ denotes the j'th column of $X^{(i)}$. Note that when we have $M < M_l$, only a subset of columns are used to create z-scores as j ranges from $1, ..., M$.

Conditioning again on the genotype matrices $X^{(1)}, X^{(2)}$ we get

$$E[z_{1j}z_{2j}] = E[E[z_{1j}z_{2j}|X^{(1)}, X^{(2)}]].$$

Working with the inner expectation we get the following

$$E[z_{1j}z_{2j}|X^{(1)}, X^{(2)}] = E\left[\frac{\left(X^{(1)}_j\right)^T Y_1}{\sqrt{N_1}} \left(\frac{\left(X^{(2)}_j\right)^T Y_2}{\sqrt{N_2}}\right)^T \middle| X^{(1)}, X^{(2)}\right]$$

$$= \frac{1}{\sqrt{N_1 N_2}} \left(X^{(1)}_j\right)^T E\left[Y_1 Y_2^T \middle| X^{(1)}, X^{(2)}\right] X^{(2)}_j$$

$$= \frac{1}{\sqrt{N_1 N_2}} \left(X^{(1)}_j\right)^T E\left[(X^{(1)}\beta_1 + \epsilon_1)(X^{(2)}\beta_2 + \epsilon_2)^T \middle| X^{(1)}, X^{(2)}\right] X^{(2)}_j$$

$$= \frac{1}{\sqrt{N_1 N_2}} \left(X^{(1)}_j\right)^T \left(X^{(1)} E[\beta_1\beta_2^T] \left(X^{(2)}\right)^T + X^{(1)} E[\beta_1\epsilon_2^T] + E[\epsilon_1\beta_2^T] \left(X^{(2)}\right)^T + E[\epsilon_1\epsilon_2^T]\right) X^{(2)}_j$$

$$= \frac{1}{\sqrt{N_1 N_2}} \left(X^{(1)}_j\right)^T \left(X^{(1)} \frac{\rho_g}{M_l} I_1 \left(X^{(2)}\right)^T + \rho_e I_2\right) X^{(2)}_j$$

$$= \frac{\rho_g}{M_l \sqrt{N_1 N_2}} \left(X^{(1)}_j\right)^T X^{(1)} \left(X^{(2)}\right)^T X^{(2)}_j + \frac{\rho_e}{\sqrt{N_1 N_2}} \left(X^{(1)}_j\right)^T I_2 X^{(2)}_j. \tag{8}$$

Consider first the random terms of the second part of the equation. Suppose,

without loss of generality, that $N_1 < N_2$. We then have that

$$\begin{aligned} E\left[\left(X_j^{(1)}\right)^T I_2 X_j^{(2)}\right] &= \sum_{m=1}^{N_1} E\left[X_{mj}^{(1)} X_{mj}^{(2)}\right] \\ &= \sum_{m\in Q} E[X_{mj}^2] + \sum_{m\notin Q} E[X_{mj}^{(1)}]E[X_{mj}^{(2)}] \\ &= \sum_{i\in Q} 1 = N_s, \end{aligned} \tag{9}$$

where we on the second line partition the sum over the individuals i that belong to the set Q where Q is the set of individuals which belong to both studies. Since individual genotypes are mean-centered and variance-standardized this becomes the number of elements in Q which by definition is N_s.

For the remaining term we again assume that $N_1 < N_2$. Then

$$\begin{aligned} E\left[\left(X_j^{(1)}\right)^T X^{(1)}\left(X^{(2)}\right)^T X_j^{(2)}\right] &= \sum_{k=1}^{M_l}\left(\sum_{m=1}^{N_1}\sum_{n=1}^{N_2} E\left[X_{mj}^{(1)} X_{mk}^{(1)} X_{nj}^{(2)} X_{nk}^{(2)}\right]\right) \\ &= \sum_{k=1}^{M_l}\left(\sum_{(m,n)\in Q} E\left[X_{mj}^2 X_{mk}^2\right] + \sum_{(m,n)\notin Q} E\left[X_{mj}^{(1)} X_{mk}^{(1)}\right] E\left[X_{nj}^{(2)} X_{nk}^{(2)}\right]\right) \\ &= \sum_{k=1}^{M_l}\left(\sum_{(m,n)\in Q}\left(Var(X_{mj}X_{mk}) + E\left[X_{mj}X_{mk}\right]^2\right) + \left(N_1N_2 - N_s\right) r_{jk}^2\right) \\ &= \sum_{k=1}^{M_l}\left(N_s\left(1 + r_{jk}^2\right) + N_1N_2 r_{jk}^2 - N_s r_{jk}^2\right) = M_l N_s + N_1 N_2 l_j, \end{aligned} \tag{10}$$

where we again partition the sum over Q, the set of individuals contained in both studies. Combining Equations 9 and 10 with the full results in Equation 8 gives

$$\begin{aligned} E[z_{1j}z_{2j}] &= \frac{\rho_g}{M_l\sqrt{N_1N_2}}\left(M_l N_s + N_1N_2 l_j\right) + \frac{\rho_e}{\sqrt{N_1N_2}} N_s \\ &= \frac{N_s}{\sqrt{N_1N_2}}\left(\rho_g + \rho_e\right) + \frac{\sqrt{N_1N_2}\rho_g}{M_l} l_j \\ &= \frac{\rho_s N_s}{\sqrt{N_1N_2}} + \frac{\sqrt{N_1N_2}\rho_g}{M_l} l_j, \end{aligned} \tag{11}$$

which completes the proof. □

Corollary 5.1. *If the two GWASs are the same, i.e. if $z_{1j} = z_{2j}$ for all j, then Proposition 5 becomes*

$$E[\chi_j^2] = 1 + \frac{Nh^2}{M_l} l_j, \tag{5, revisited}$$

where χ_j^2 is the GWAS test statistic of the j'th SNPs association with trait in the form of a chi-squared statistic.

Proof. If the two studies are the same then $N_1 = N_2 = N$ and since ρ_g is the genetic covariance we get $\rho_g = Cov(G_1, G_1) = Var(G) = h^2$. Furthermore, since $N_1 = N_2 = N$ the number of overlapping individuals N_s must be N. Since the two study samples are the same their observed phenotypes must be the same in both studies, meaning that $\rho_s = 1$. Lastly, since z_{1j} is the z-score test statistic of SNP association for SNP j it holds that $z_{1j}^2 = \chi_j^2$ is the χ^2 test statistic of SNP association for SNP j. □

2.3 Estimating heritability and genetic covariance

As previously stated, obtaining estimates of heritability and genetic covariance is done by fitting a linear regression model. In matrix form, the model equation can be expressed as a simple linear regression model of

$$Y = Xb + \epsilon,$$

where Y is an $(M \times 1)$ vector of GWAS statistics for SNP associations with trait, X is a design matrix dimension $(M \times 2)$ with elements $(1, l_j)$ on row j, ϵ is a vector of independent and identically distributed residuals with $\epsilon_j \sim N(0, \sigma^2)$ for $j = 1, ..., M$ and b contains our coefficients here defined as $b = (\alpha, \beta)^T$ [21]. The contents of our response Y depends on whether we are attempting to estimate heritability ($y_j = \chi_j^2$) or genetic covariance ($y_j = z_{1j} z_{2j}$), i.e. whether we are using Equation 5 or Equation 6 but results are similar regardless.

Fitting the model and estimating the coefficients returns estimates of the regression slope coefficient β as

$$\hat{\beta} = \begin{cases} \dfrac{N\hat{h}^2}{M_l}, & \text{if } y_j = \chi_j^2 \\ \\ \dfrac{\sqrt{N_1 N_2}\hat{\rho}_g}{M_l}, & \text{if } y_j = z_{1j} z_{2j} \end{cases}$$

from which we then may solve for $\hat{h}^2$ and $\hat{\rho}_g$ by rescaling of $\hat{\beta}$ accordingly. The validity of the regression model equation depends on several factors and assumptions which all influence the performance and efficiency of $\hat{\beta}$.

In this subsection we comment on which assumptions of ordinary least squares regression is violated in LDSR and how we may correct and adjust for these to produce an efficient regression estimator. We discuss how weighted least squares can be utilized to account for heteroskedasticity in data and correlation among response variables. Furthermore, we discuss the concept of liability- and observed-scale estimates and how we may correct the scale of our estimates to make them comparable between studies while simultaneously adjusting for bias introduced by oversampling of cases.

2.3.1 Weighted least squares regression

One of the key assumptions in standard ordinary least squares regression is the assumption of constant variance (homoskedasticity), i.e. that the residuals ϵ_j

are all identically distributed with the same, constant variance σ^2. In LDSR, we regress GWAS test statistics of a SNPs association with phenotypic trait on LD score. Generally, SNPs with high LD scores will tend to have higher variance than SNPs with low LD scores, meaning that the assumption of homoskedasticity is likely to be violated [18, 19].

A second important assumption, also on the residuals ϵ_j, is that of independence, i.e. that $Cov(\epsilon_i, \epsilon_j) = 0$ for all $i \neq j$. However, due to the inherent correlation structure in SNPs caused by LD, SNPs are generally not independent of each other and thus the GWAS test statistics that we use as response variables tend to be correlated with each other. This correlation is further passed on to the residuals meaning that the assumption of independence among them are also violated [18, 19].

The standard way of dealing with these two problems simultaneously is through the use of generalized least squares [22]. Generalized least squares assumes the same linear model as given in Section 2.3 but relaxes the assumptions on the residuals. Instead of assuming that residuals are independent and identically normally distributed, we assume

$$E[\epsilon] = 0, \quad Var(\epsilon) = \Omega,$$

where Ω is, in this case, an $(M \times M)$ variance-covariance matrix [22]. In the setting of LDSR, we would generally like to take Ω to be the variance-covariance matrix of either the χ^2 statistics or multiplied z-scores, depending on our current target, but this matrix is intractable and we must resort to using a simpler approach [19].

The suggested approach is to use weighted least squares regression [21]. Weighted least squares regression is a specific case of generalized least squares where the variance-covariance matrix of ϵ is further assumed to be diagonal. This means that the assumptions on the residuals can be written as

$$E[\epsilon] = 0, \quad Var(\epsilon) = \Omega = \begin{bmatrix} 1/w_1 & 0 & \dots & 0 \\ 0 & 1/w_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & 1/w_M \end{bmatrix},$$

which corresponds to assuming a regression model where residuals are independent but allowing for non-constant variances [21].

We pick weights w_j to be $w_j = 1/\sigma_j^2$ where σ_j^2 is the variance of our response given the explanatory variables of LD scores [18, 19]. For $y_j = z_{1j}z_{2j}$ one can show that, under the assumption that the z-scores are jointly multivariate normal, the conditional variance σ_j^2 is given by

$$\begin{aligned} \sigma_j^2 &= Var(z_{1j}z_{2j}|l_j) \\ &= \left(1 + \frac{N_1 h_1^2}{M_l} l_j\right)\left(1 + \frac{N_2 h_2^2}{M_l} l_j\right) + \left(\frac{\rho_s N_s}{\sqrt{N_1 N_2}} + \frac{\sqrt{N_1 N_2}\rho_g}{M_l} l_j\right)^2. \end{aligned} \tag{12}$$

We do not prove this here but instead refer the reader to the supplementary material of [18] for further discussion and more details on the proof. Now if $z_{1j} = z_{2j}$ then we immediately get the conditional variance for the χ^2 statistics as a corollary of Equation 12:

$$\sigma_j^2 = Var(\chi_j^2|l_j) = 2\left(1 + \frac{Nh^2}{M_l} l_j\right)^2. \tag{13}$$

While weighted least squares generalizes ordinary least squares to allow for heteroskedasticity in data, it retains the assumption of independent residuals. The suggested approach to account for the non-independence of residuals in LDSR is a heuristic approach that uses a secondary weight to correct for their correlation [18, 19]. To this end we introduce the notation of $l_j(S)$ as the LD score over a set S, formally defined as

$$l_j(S) = \sum_{k \in S} r_{jk}^2. \tag{14}$$

To account for dependence in residuals we take as set S the set of all SNPs included in our regression i.e. $S = \{SNP_1, SNP_2, ..., SNP_M\}$. This means that the more a SNP is correlated with the other remaining SNPs in the set, the more its weight is reduced. However note that this is only an approximate solution to the issue of non-independence.

The full weights used in the weighted least squares regression of LDSR can then be written as

$$1/w_j = l_j(S)\sigma_j^2 = \begin{cases} l_j(S)Var(z_{1j}z_{2j}|l_j) & \text{if } y_j = z_{1j}z_{2j} \\ l_j(S)Var(\chi_j^2|l_j) & \text{if } y_j = \chi_j^2 \end{cases}$$

where S is the set of our GWAS SNPs, i.e. $S = \{SNP_1, SNP_2, ..., SNP_M\}$ and variances are as given in Equations 12 and 13.

2.3.2 Case-control adjustments

The LDSR methodology described above is valid for continuous phenotypes and does not immediately extend to binary phenotypes from case-control data. It does not naturally account for things such as underlying population prevalences of the phenotype or the oversampling of cases which occurs with rare phenotypes. Estimates produced through the original model on binary traits are said to be on the observed-scale and denoted by h^2_{obs} and $\rho_{g,obs}$ respectively. The scale of this estimate depends on three things: the binary scale of the phenotype, the population prevalence of the phenotype and the ascertainment of cases in the sampling of the phenotype. These factors make the estimates produced difficult to compare with those based on continuous phenotypes, binary phenotypes under different population prevalences and the same phenotype sampled in a different ratio [23].

Accounting for all three of these issues gives us estimates that are not plagued by scale and thus comparable with estimates for any other traits. Such scale-corrected estimates are said to be on the liability-scale. The usual workaround is to assume that binary coded phenotypes are generated through an underlying liability threshold model [23–25]. Such an assumption allows us to shift the binary phenotypes into a continuous setting through what is essentially a probit-transform.

The liability threshold model assumes that all individuals carry an unseen continuous liability to disease ψ which, upon exceeding a given threshold τ, gives an individual case-status. In general we say that the observed status (in the way of the binary case-control dichotomy) of individual y_j is given by

$$y_j = \mathbf{1}(\psi_j > \tau),$$

where $\mathbf{1}$ denotes the indicator function, ψ_j is the liability to trait of individual j and τ is the liability threshold [23–25]. Now if we assume that ψ_j is normally distributed, we can pick τ such that the liability threshold corresponds to the population prevalence of the phenotype. If we let K denote the true population prevalence of the phenotype, then specifying such a threshold τ is equivalent to taking $\tau = \Phi^{-1}(1-K)$ where Φ^{-1} is the inverse of the standard normal cdf [18].

Under the assumption of a liability threshold model as described above, one can show that transforming observed scale estimates h^2_{obs} and $\rho_{g,obs}$ to the desired liability scale is done through

$$\rho_{g,obs} = \rho_g \cdot \frac{\phi(\tau_1)\phi(\tau_2)\sqrt{P_1(1-P_1)P_2(1-P_2)}}{K_1(1-K_1)K_2(1-K_2)}, \tag{15}$$

where P_i is the sample prevalence for study i, i.e. the ratio of cases to individuals in the sample, K_i is the population prevalence for the trait in study i, ϕ denotes the probability density function of a standard normal and $\tau_i = \Phi^{-1}(1-K_i)$ with Φ^{-1} being the inverse of the cumulative density function for a standard normal. We do not cover the proof of this result but instead refer the reader to [23] or the supplementary material of [18] for further details.

A similar result holds for the liability scale heritability by taking study 1 equal to study 2 which gives

$$h^2_{obs} = h^2 \frac{\phi(\tau)^2 P(1-P)}{K^2(1-K)^2}, \tag{16}$$

where again P is sample prevalence, K is population prevalence and $\phi(\tau)$ is as given above [18].

As a result of Equations 15 and 16 we note that the genetic correlation coefficient r_g is independent of observed and liability scale, i.e. that

$$r_{g,obs} = \frac{\rho_{g,obs}}{\sqrt{h^2_{1,obs}h^2_{2,obs}}} = \frac{\rho_g}{\sqrt{h^2_1 h^2_2}} = r_g, \tag{17}$$

A formal proof of this is given in Section A.2 in the Appendix.

Table 1: Summary of the quality control steps taken for the genome-wide data of cases and controls with respect to RA-diagnosis. Further details on how the quality control steps were performed and which reasoning was used for the criteria set can be found in Section A of the Appendix.

Quality control step	**Criteria**	**No. Excluded**
Sex discrepancy	Discordance between reported and imputed	2 individuals
Individual genotyping rate	Missingness $> 5\%$	13 individuals
SNP genotyping rate	Missingness $> 5\%$	12 218 SNPs
Minor allele frequency	Frequency $< 1\%$	138 713 SNPs
Hardy-Weinberg equilibrium	$p > 10^{-10}$ for controls, $p > 10^{-6}$ for cases	1893 SNPs
Genomic relatedness	Relatedness > 0.1	310 individuals
Population outliers	6 standard deviations from origin on every principal component	316 individuals

3 Data and results from GWAS analysis

3.1 RA - data

Individual data was taken from the Swedish Epidemiological investigation of rheumatoid arthritis (EIRA) study, which contained a total of 7180 individuals out of which 4193 were cases and 2987 were controls. The study group in EIRA contains individuals in the age range of $18 - 70$ located in the middle and southern regions of Sweden. Cases were defined as individuals diagnosed with RA according to the 1987 American College of Rheumatology criteria [26]. Controls were randomly selected and matched on age, sex and residential area [27].

All 7180 individuals were genotyped for common genetic variants in the form of SNPs through the Illumina Infinium GSA chip for a total of 693 413 SNPs.

This makes up the full initial data set for the GWAS on RA-status. However, errors in data may arise for several reasons, including inadequate quality at genotyping, incorrectly handled samples, low quality of DNA or various other technical artifacts. Failing to account for things such as sample imperfections or mix-ups, confounding due to population structure and close relatedness may lead to spurious associations and biased results. As such, rigorous quality control of the genome-wide data must be carried out, at both individual- and SNP-level [28,29]. A detailed description of the quality control procedure employed for this data can be found in the Appendix, Section C. Here, we instead summarize the steps carried out and what observations were excluded in Table 1. All quality control steps were performed with the PLINK-software using the currently most recent version (v1.90, beta 6.10) [30].

The resulting data set used for the association analysis thus contains 6539 individuals divided into sets of 3766 cases and 2773 controls genotyped for a total of 540 319 SNPs.

3.2 RA - association analysis

Association analysis was performed in the way detailed in Section 1.2 on the set of individuals and SNPs that passed the quality control procedure covered in Table 1 and Section C of the Appendix. The full GWAS was performed with the PLINK software using the currently most recent version (v1.90, beta 6.10) [30].

We performed the analysis over three subsets of data. For the first set we filtered out cases that had tested negative for anti-citrullinated protein antibodies (ACPA), a type of antibody present in many, but not all, cases of RA [31]. This led to a subset of 4898 individuals, where 1608 ACPA-negative cases were removed and a further 33 individuals were removed due to missing phenotypes.

In the second set we instead kept only seronegative cases, i.e. patients diagnosed with RA that tested negative for both ACPA and rheumatoid factor, a second antibody often present in cases of RA [31]. This subset contained 3478 individuals where 3044 cases were removed due to being seropositive and a further 17 individuals were removed due to missing phenotypes.

In the last set we performed no filtering. However the working set was still slimmed down to 6067 individuals where 428 were removed due to missing covariates and 44 due to missing phenotypes. Note that all three subsets contain the same set of controls and that filtering was done only on further classification of RA-status in cases.

In the performed GWAS, we adjust for covariates sex, age and population substructure. Here, population substructure is controlled for by adding as covariates, the principal components obtained in the final quality control step described in C.8. For each of the three RA-sets, we perform the GWAS a total of eleven times, using a different number of principal components in each round, the first containing no components and the last containing all ten.

To assess which GWAS performs the best, we measure the amount of bias introduced due to general population stratification, i.e. due to relatedness and population stratification, selecting the model with the lowest amount introduced into the association analysis. We do this by estimating the genomic inflation factor, $\lambda_{gc}^{(i)}$ for all combinations of number of principal components included as covariates ($i = 0, 1, ..., 10$), making our choice based on the obtained estimates. This is repeated for all three RA-sets.

The genomic inflation factor λ_{gc} measures amount of confounding and bias introduced by population substructure, cryptic relatedness and other sample population confounding in the way that more bias leads to a larger estimate of λ_{gc} [32]. As such, we expect it to vary with the number of principal components included as covariates as these should control for population structure. We thus make our choice of best performing GWAS to contain the number of

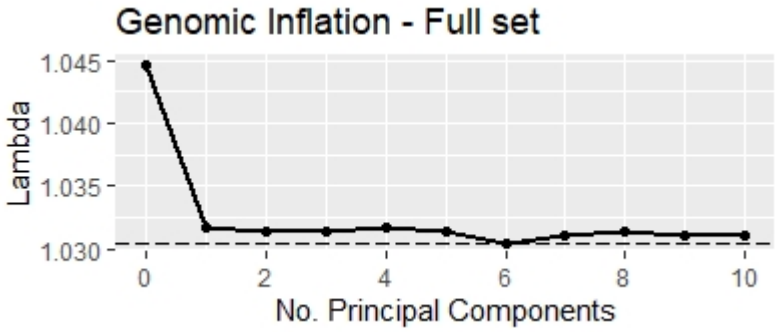

Figure 2: Genomic inflation factor $\hat{\lambda}_{gc}$ as a function of number of principal components on population stratification used as covariates in the association analysis on the full data set. The dashed line is the minimum of the curve at 6 components with an estimated genomic inflation $\hat{\lambda}_{gc}^{(6)} = 1.03$.

principal components which minimizes $\hat{\lambda}_{gc}^{(i)}$ for $i = 0, 1, ..., 10$. We do not give a mathematical detailing of the genomic inflation factor here but instead refer the reader to the Appendix, Section A.3 for a more thorough discussion.

For the full set we find that the estimated genomic inflation factor $\hat{\lambda}_{gc}^{(i)}$ reaches its minimum for 6 components at $\hat{\lambda}_{gc}^{(6)} = 1.03$. The full curve of the estimated genomic inflation factor as a function of number of included principal components can be found in Figure 2. For the ACPA-positive set we reach a minimum at the inclusion of the first component ($\hat{\lambda}_{gc}^{(1)} = 1.038$) and for the seronegative at the fifth component ($\hat{\lambda}_{gc}^{(5)} = 1.014$). Their genomic inflation curves can be found in the Appendix, Section D, Figure A1.

3.2.1 Results from the GWAS on RA

Traditionally, the results from a GWAS are presented through Manhattan plots, in which the significance level of SNPs are plotted against their positions on the genome. Figure 3 contains the Manhattan-plot for the full data set, in which SNP p-values are plotted at $-\log_{10}(p)$-level against their position over the chromosomes. Manhattan-plots for ACPA-positive individuals and seronegative individuals are found in the Appendix in Figures A2 and A3 (page 48). Results from the GWASs are in line with previous studies on SNP association with RA-status [33, 34].

We find a total of 559 SNPs that reach the genome-wide significance level of $p = 5 \cdot 10^{-8}$ in the full data set. Out of these, the majority (551 SNPs) reside in a tight region on chromosome 6 which we recognize as the human leukocyte antigen (HLA) locus. This region is the most well known genetic risk factor for RA [2, 34]. As the HLA complex is responsible for the regulation of the human immune system, it is not a surprise that the majority of our most strongly

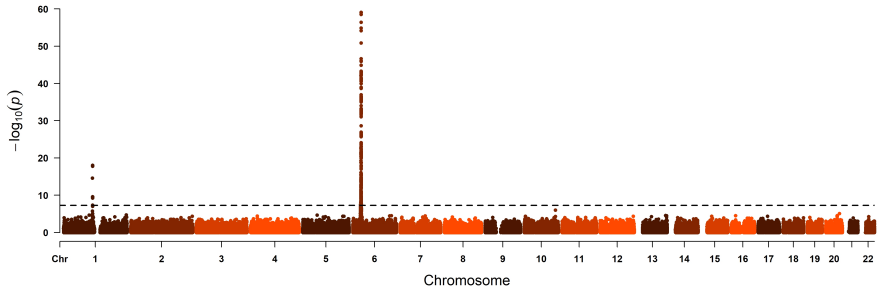

Figure 3: Manhattan-plot for the associations of the full set of RA data. The x-axis contains all autosomal SNPs studied, ordered after chromosome and chromosome position. The y-axis covers the p-values of SNP association with trait at $-\log$ base 10 scale. The dashed line represents the genome-wide significance level of $p = 5 \cdot 10^{-8}$ which is the common standard for recognizing a SNP association as significant in a GWAS.

associated SNPs should reside there as RA is an autoimmune disease, strongly linked to mechanisms of the immune system.

Furthermore, we note a small amount of significantly associated SNPs on chromosome 1, which corresponds to SNPs on the PTPN22 gene, another established genetic risk factor for RA [34, 35].

Similar results are obtained for the subset containing only the ACPA-positive cases (Figure A2, page 48). While the significantly associated regions are similar, the number of genome-wide significant SNPs are much larger at nearly 1200 SNPs. Despite this large increase, the number of significant SNPs at chromosome 1 are still the same.

For the seronegative set we found no SNPs that reached the genome-wide significance level (Figure A3, page 48). While it may seem surprising that no associations were found we note that this result replicates the findings of reference [33]. In their paper, based on a smaller set from the same EIRA-study, no genome-wide significant SNPs for ACPA-positive RA were replicated in the set for seronegative RA.

3.3 CVD - data

For RA, the genotyped data utilized for the GWAS was available from a previous study from our institution. However for cardiovascular disease (CVD), no in-house data was available. As such we opted to use publicly available GWAS summary statistic data. The vast amount of available public GWAS summary statistic data coupled with the fact that CVD is an umbrella term for a broad range of diseases, led to a long list of available data sets to consider. We made our choice from the available data by considering factors of study sample sizes,

the main ancestry of the genotyped individuals (see Section C.8 on confounding due to ancestrally different populations) and what type of underlying CVD-phenotype had been considered in the study.

Our phenotype of choice for this particular study was acute myocardial infarction (AMI, commonly known as heart attack). There are two main reasons behind this choice: firstly, as the basis for this study was an article in which RA patients demonstrated an increased risk for the event-based CVD of acute coronary syndrome [7], we wanted this event-based nature to be reflected in our CVD-phenotype. Secondly, we consider myocardial infarction to be a relatively "pure" phenotype in that there is relatively minor amounts of confounding due to things such as case misclassification or underlying sub-phenotypes. Thirdly, we know AMI to be a heritable disease, i.e. it has a trait heritability h^2 that ought to be non-zero. Looking at the genetic correlation between traits where one is not heritable may lead to spurious and nonsense correlations as we divide by an estimated heritability that would be near zero plus-minus noise.

Our GWAS summary statistic data was based on individual-level genotyped data from the UK Biobank [36]. The genotyped population consisted of approximately 500 000 individuals in the age range of $40 - 69$, sampled between $2006 - 2010$ in the United Kingdom. Cases were defined as individuals who had been diagnosed with AMIs by a physician and the remaining individuals in the sample were taken as controls. We chose to work with the UK Biobank set as it is a well established resource, containing a large group of genotyped individuals based on a northern European population.

All individuals were genotyped for common variants in the form of SNPs through the UK Biobank Axiom Array [37]. Further imputation was then performed to increase the number of available SNPs and subsequent quality control was applied to the data, controlling for outliers and badly performing DNA samples. We do not cover neither the imputation nor quality control procedures here and instead refer the interested reader to [36].

The resulting data set used for the association analysis, after quality control, imputation and removing individuals with missing pheontypes, consisted of 361 194 individuals divided into groups of 6412 cases and 354 782 controls genotyped for approximately 13 million SNPs.

3.4 CVD - association analysis

The obtained data consists of GWAS summary statistic data, i.e. the results from the corresponding association analysis of SNP allele frequency on the phenotypic trait of acute myocardial infarction. The association analysis was performed as detailed in Section 1.2 and the covariates used were age, $(\text{age})^2$, inferred sex, age $\cdot$ inferred sex, $(\text{age})^2 \cdot$ inferred sex and the first 20 principal components from a principal component analysis of the individuals ancestries [38].

We performed further SNP-level quality control on the GWAS output data to make sure that the SNPs analyzed for the AMI-data were held to the same standard as those for the RA-data. To this end, we filtered on SNP call rate, minor allele frequency and Hardy-Weinberg equilibrium, removing any SNPs

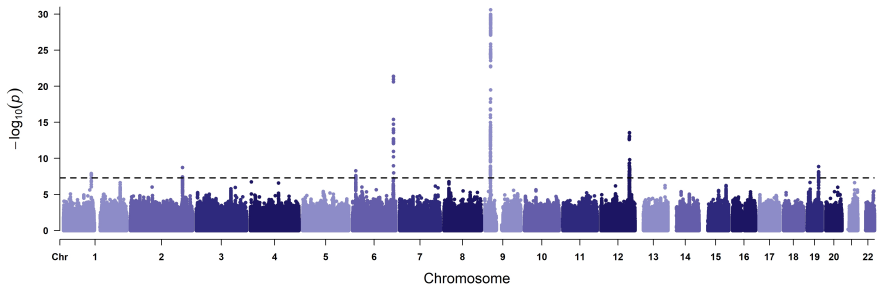

Figure 4: Manhattan-plot for the associations of the data on acute myocardial infarction. The x-axis contains all autosomal SNPs studied, ordered after chromosome and chromosome position. The y-axis covers the p-values of SNP association with trait at $-\log$ base 10 scale. The dashed line represents the genome-wide significance level of $p = 5 \cdot 10^{-8}$ which is the common standard for recognizing a SNP association as significant in a GWAS.

with either a SNP call rate below 0.95, observed minor allele frequency below 0.01 or deviating from Hardy-Weinberg equilibrium at a significance level of $p = 10^{-6}$. Further details on this type of filtering is covered in the Appendix, Sections C.3, C.4 and C.5.

This leaves us with a full GWAS summary statistic data set of approximately 9 million SNPs based on a set of 361 194 individuals. The estimated genomic inflation factor for the AMI-GWAS is $\hat{\lambda}_{gc} = 1.066$.

3.4.1 Results from the GWAS on CVD

The results from the GWAS on AMI is presented in Figure 4 as a Manhattan-plot. Out of the 9 million SNPs measured, 317 SNPs reached significance at the genome-wide level, i.e. had p-values below $p = 5 \cdot 10^{-8}$.

In contrast to the results from the GWAS on RA (Figure 3) we see genome-wide significant associations at a larger genomic spread than for the association analysis of RA in which the significant associations were concentrated on two parts of the genome. Here we instead see seven loci in which genome-wide significance is reached with an even greater amount exhibiting strong, albeit non-significant, association with the trait of AMI.

While there is a clear strong association for several regions at chromosome 6, there seems to be little effect in the HLA-region which demonstrated the strongest association with RA. Of the SNPs at genome-wide significance level for RA, none were found to be genome-wide significant for acute myocardial infarction. Among these, the strongest association was at around $p = 10^{-3}$ at which we found five SNPs. In Figure 5 we give a Manhattan-style plot of the SNP associations with AMI for the SNPs that reached genome-wide significance

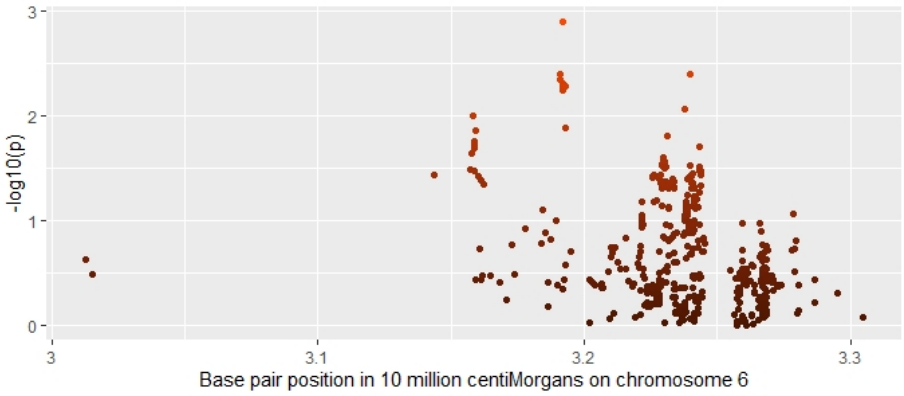

Figure 5: Manhattan-plot of SNP association with myocardial infarction for the UK Biobank sample on a small genomic region on chromosome 6. The y-axis is the base 10 $-log(p)$ transform of the p-values of SNP association. The x-axis is the SNPs position on the genome, given here as base-pair position in the scale of 10^7 centiMorgans. The points are the SNPs that reached genome-wide significance in the full RA set (Figure 3, page 22) here instead illustrating their association with AMI.

level at chromosome 6 for the full RA-set.

We further note a minor peak at chromosome 1 consisting of ten SNPs above the genome-wide significance level. Comparing these to the eight SNPs that reached a similar significance for RA we again find no overlapping SNPs, where for AMI, the most strongly associated SNPs have p-values around $p = 10^{-5}$.

Comparing the results in Figure 4 with the results from the association analysis of the set of ACPA-positive cases (Figure A2, page 48) we found similar results regarding overlap. In the Appendix (Section D, Figure A4, page 49), we give the corresponding version of Figure 5, of SNP associations with AMI for the genome-wide significant SNPs from the GWAS on the ACPA subset. We note that these results closely resemble those in Figure 5.

Despite this immediate lack of overlap in genome-wide SNPs, it is still highly possible that an underlying genetic overlap exists. For two Mendelian traits (i.e. traits due to single genes) we would expect little to no genetic overlap when observing similar results. However, RA has been shown to be inherently polygenic in its genetic architecture, i.e. that a big part of the genetic effect is made up of a large amount of weak genetic effects [39]. As such, it is not unlikely that a greater overlap exists in a large group of individually non-significant SNPs that together account for a big part of the genetic effect on disease status. It is thus too early to draw conclusions about the existing genetic overlap of the two phenotypes and we will need further analysis to establish satisfying results.

4 Statistical analysis

All statistical implementation was done through the LDSC (LD Score) software using the currently most recent version (v.1.0.1, https://github.com/bulik/ldsc) [19]. LDSC is a command-line tool developed by the authors of the original papers on LDSR [18,19]. It uses code written in Python to allow for the estimation of LD scores l_j from individual-level genome-wide data and estimation of heritability h^2 and genetic covariance ρ_g for sets of GWAS summary statistic data by implementing LDSR. We used LDSC to produce estimates of trait heritability and genetic covariance for all combinations of our four sets of GWAS summary statistic data.

LD scores were not estimated from our own GWAS data but from an external reference panel. As we discussed in Section 1.3, the number of SNPs genotyped in a GWAS is normally only a fraction of the true amount of genetic variants on the human genome. Using only a subset of SNPs in estimation of a variants LD score can ultimately lead to downwards biased estimates [19].

For this study we used, as reference panel for LD score estimation, the European ancestry sample from the 1000 Genomes Project [40, 41]. We do not go into detail on the specifics of the sample as this has been done extensively elsewhere [41] but note that it is a well established and widely used data resource. Furthermore, its performance in LDSR has been covered in detail elsewhere and it has been shown to perform well as a reference panel for LD scores, with the authors of the original papers on LDSR noting that the European ancestry sample was an adequate match for genetic studies on northern European populations [19].

LD scores are computed by estimating the LD that a given SNP is in with all other SNPs in the reference panel where LD is measured through the r^2 measure given in Definition 2. In estimating r^2 we use an adjusted estimator. The standard estimator of r^2 is approximately unbiased but contributes, in LD score estimation, an error of order $O(M/N)$ where M and N are the number of SNPs and individuals respectively in the reference panel. As M is often several magnitudes larger than N this error may not be negligible. We thus hope to mitigate this error by using an adjusted estimator defined as

$$\hat{r}^2_{adj} = \tilde{r}^2 - \frac{1 - \tilde{r}^2}{N - 2}, \tag{18}$$

where $\tilde{r}^2$ is the standard estimator of r^2 for two arbitrary genomic loci. We discuss estimation of r^2 and its various properties in further detail in the Appendix, Section A.4, where we prove that $\hat{r}^2_{adj}$ performs better than $\tilde{r}^2$ for the purpose of LD score estimation.

Lastly, estimates of trait heritability h^2 and genetic covariance ρ_g were obtained on the liability scale by correcting for binary phenotypes and oversampling of cases through the use of Equations 15 and 16 per the discussion in Section 2.3.2. Population prevalences K of the four phenotypes were obtained through a combination of external information and within sample estimates. We used a previously estimated prevalence of RA in Sweden as $K = 0.0077$

Table 2: Estimates of heritability on the liability scale for four different phenotypes. The number in the parenthesis of the second column is the standard error of the estimate. The interval of the third column was computed as $\hat{h}^2 \pm 1.96 \cdot se$. All estimates were produced through the LDSC software (v.1.0.1) with LDSR intercepts constrained to 1.

Phenotype	$\hat{\boldsymbol{h}}^{\boldsymbol{2}}$	**Interval**
RA	0.1967 (0.0962)	(0.0082, 0.3853)
ACPA+ RA	0.4108 (0.2433)	(-0.0661, 0.8877)
SERO- RA	0.0958 (0.0786)	(-0.0583, 0.2499)
AMI	0.1598 (0.019)	(0.1226, 0.197)

corresponding to a population prevalence of 0.77% [1]. For the prevalences of ACPA-positive RA and seronegative RA we used crude estimates based on the population prevalence of RA as $K_{ACPA} = \frac{2K_{RA}}{3}$ and $K_{SERO} = 0.3\ K_{RA}$ where the scaling coefficients were based on their respective sample prevalences P. For AMI we used a population prevalence of $K_{AMI} = 0.0178$ based on an article on the epidemiology of CVDs in the United Kingdom [42]. We note that this value is close to the sample prevalence of $P_{AMI} = 0.018$ which would be reasonable for a population sample of that size.

As crude estimates were used for three of the four phenotypes we checked the heritability estimates for sensitivity to errors in population prevalence estimates. Figure A5 (page 50) in Section D of the Appendix contains the results for a grid of prevalences. We note that estimates are generally robust to minor errors in prevalences where only the estimates of ACPA-positive heritability show large variation. However, the prevalence of ACPA-positive RA is bounded upwards by the prevalence of general RA in the way of $K_{ACPA} \leq K_{RA}$, meaning that nonsense estimates of h^2 would be impossible even for absurd errors in estimated population prevalences.

5 Results

5.1 Heritability

Results in the form of heritability estimates for the four phenotypes are given in Table 2. The estimates were produced according to the discussion in Section 4.

All heritability estimates were obtained by constraining the LDSR regression intercept to 1 which has been shown to improve estimates by decreasing their standard errors [18]. This type of constraining corresponds to the assumption that little to no confounding due to population stratification, i.e. due to cryptic

relatedness or population substructures, is contained in the GWAS. We argue that all sets have gone through rigorous and adequate quality control with respect to both population substructures and genomic relatedness which would make such an assumption valid. As a further argument to the validity of this assumption we note that fitting the regression without constraining the intercept, all four regressions returned estimates near one with the greatest deviation obtained as $\hat{\alpha} = 1.0046$.

Heritability of the three types of RA seem to be in line with previous findings. We note that the magnitude of the estimates seem to concord with what was found through the Manhattan-plots, where ACPA-positive RA exhibits the strongest genetic effect and seronegative RA exhibits the weakest (3, 4, A2 and A3; pages 22, 24 and 48 respectively). However, the estimate of h^2_{ACPA} comes with a lack of precision as evidenced by the alarmingly high standard error and further care should be taken with the estimate for seronegative RA as the sample size N_{SERO-} was low and the estimate may as such be at low power. We note a modest heritability for AMI with a very small standard error which is what we had previously expected. Estimating genetic overlap using Definition 3 assumes that the phenotypes considered at the very least depend on a small genetic effect, as lack of genetic dependence would lead to low estimates of trait heritability which in turn will complicate the computation of r_g and may give spurious results. At $\hat{h}^2_{AMI} = 0.16$ we should be well equipped to detect a genetic overlap if it exists.

The true underlying heritability of RA varies between studies and populations but is often pinned somewhere around 50% [43, 44]. This puts our point-estimate of $h^2_{RA} = 0.2$ at a very long distance from the previously established estimates. However, this gap in estimates does not invalidate our results. In fact, it is well established that heritability estimates from genotyped data generally tend to underestimate the true underlying heritability of a phenotype [18,45,46]. Several explanations of this has been suggested which has spurred a search for this missing heritability [45]. In turn, this has led to the dubbing of heritability estimates produced by genotyped data of SNPs as SNP-based heritabilities h^2_{SNP} of which the relationship

$$\hat{h}^2_{SNP} \leq h^2,$$

has been established [46].

Furthermore, the heritability estimates presented in references [43, 44] are from family-based studies of heritability. Such studies have been shown to overestimate the heritability as their estimates may be inflated due to shared environmental effects, non-additive genetic variation and epigenetic factors which are not accounted for in our measures (see Section A.1) [46]. As such, we may extend the relationship on heritabilities given above as

$$\hat{h}^2_{SNP} \leq h^2 \leq \hat{h}^2_{Fam}$$

As a result, comparing estimates of SNP-based heritability as in our study, with estimates from family-based studies can be somewhat misleading. Contrasting instead with SNP-based estimates gives more promising comparisons.

Table 3: Estimates of heritability on the liability scale for four different phenotypes with the HLA-region excluded. The number in the parenthesis of the second column is the standard error of the estimate. The interval of the third column was computed as $\hat{h}^2_{-HLA} \pm 1.96 \cdot se$. All estimates were produced through the LDSC software (v.1.0.1) with LDSR intercepts constrained to 1.

Phenotype	$\boldsymbol{\hat{h}^2_{-HLA}}$	**Interval**
RA	0.0891 (0.0374)	(0.0158, 0.1624)
ACPA+ RA	0.1116 (0.0448)	(0.0238, 0.1994)
SERO- RA	0.0651 (0.0694)	(-0.071, 0.2011)
AMI	0.1568 (0.0195)	(0.1186, 0.195)

As an example, a previous study attempting to look at a broad range of phenotypes through LDSR estimated the heritability of RA to 0.161 with a standard error of 0.0215 in a sample of approximately 100 000 individuals [47] which is more in line with our findings [8].

5.1.1 Heritability without the HLA-region

We have previously mentioned the human leukocyte antigen (HLA) region on chromosome 6 as a major established risk factor for RA [2]. To investigate this dependence we re-ran all the regression models for the four phenotypes but with the HLA-region excluded to see its influence on estimates of h^2 for RA subtypes and AMI, comparing these results with the findings from the GWASs discussed in Sections 3.2.1 and 3.4.1. Results are presented in Table 3.

Estimates are again in line with what was expected based on previous Manhattan-plots (Figures 3, 4, A2 and A3; pages 22, 24 and 48 respectively). ACPA-positive RA shows the strongest dependence on the HLA-region, having its estimate reduced by nearly 75% with a modest reduction in standard RA and a minor decrease for seronegative RA. This is concordant with the results found in the Manhattan-plots mentioned above. The smallest change in heritability is observed for AMI which is near identical to its previous estimate, showcasing a very weak relationship between the alleles at the HLA loci and AMI. Note that this was previously hinted at in Figure 5 (page 25), where none of the top SNPs for RA, in the HLA-region, reached genome-wide significance for AMI.

Interesting to note is the much smaller standard error obtained for the ACPA+ phenotype, now on a magnitude comparable to the other types of RA. In LDSC, the ordinary regression estimate of the standard errors can be biased downward [19] and standard errors are instead estimated through a block jackknife procedure [48]. It is reasonable to assume that the large standard error exhibited in Table 2 occurs due to the strong dependence on the HLA-region noticed here and that estimates exhibit a greater fluctuation when blocks of

Table 4: Estimates of genetic covariance and genetic correlation between the phenotype of acute myocardial infarction (AMI) and the three types of RA given in the first column. Numbers in the parenthesis of the second and third columns are standard errors of the estimates. Intervals in the third column are computed as $\hat{r}_g \pm 1.96 \cdot se$. All estimates were produced through the LDSC software (v.1.0.1) with LDSR intercepts constrained to 0.

Phenotype	$\hat{\rho}_g$	$\hat{r}_g$	**Interval for** $\hat{r}_g$	p
RA	-0.0135 (0.0149)	-0.0762 (0.0798)	(-0.2327, 0.0802)	0.34
ACPA+ RA	-0.0111 (0.0167)	-0.0435 (0.0627)	(-0.1664, 0.0794)	0.488
SERO- RA	-0.0243 (0.0203)	-0.1967 (0.1772)	(-0.544, 0.1506)	0.267

the HLA-region are removed. As a further argument to this we note a similar, albeit minor, attenuation in the standard errors for ordinary RA but almost no change for AMI or seronegative RA. This explanation would be in line with our previous findings of a strong relationship between SNPs at the HLA loci and both ACPA+ RA as well as ordinary RA.

5.2 Genetic correlation

Results in the form of estimated genetic covariances $\hat{\rho}_g$ and genetic correlations $\hat{r}_g$ between AMI and the three types of RA are presented in Table 4. The estimates were produced according to the discussion in Section 4.

The genetic covariance estimates ρ_g were all obtained by constraining the regression intercept to 0. This type of constraining has been shown to reduce the standard errors of the subsequent estimates when valid [19]. Constraining the intercept to 0 is equivalent to the assumption that no individuals were genotyped for both GWASs. While we can not with full certainty establish that this is the case we find it unreasonable to assume otherwise and consider the error introduced into the model to be negligible at best.

We find no significant genetic correlation between any of the three RA subtypes and the phenotype of AMI. The estimated genetic correlation coefficient $\hat{r}_g$ is indicative of a potentially minor negative correlation existing between RA and AMI but if it exists we are not adequately powered to significantly detect it. As such, we fail to reject the null hypothesis of no genetic overlap between RA and AMI. This result replicates the findings of a previously published article in which the genetic correlation between RA and the CVD-phenotype of coronary artery disease was examined through LDSR. The study used a sample of nearly 100 000 cases and controls for RA [47] and an almost twice as large set of individuals for coronary artery disease [49] obtaining an estimate of $\hat{r}_g = -0.063$ at a p-value of $p = 0.4377$ which agrees with our estimate for AMI [8].

The goal of this study was to assess what kind of genetic overlap exists between RA and CVD. This question was spurred by a demonstrated shared

susceptibility for the CVD-phenotype of acute coronary syndrome in individuals diagnosed with RA and their direct siblings [7]. Such a shared susceptibility may be due to many factors, either genetic, environmental or through an interplay of both. Our results indicate that little to no pleiotropic risk factors exist for RA and AMI, i.e. that there are no genes which directly influence the risk of both RA and AMI. However, such a finding does not invalidate the role of genetics in the shared susceptibility and it may still be the case that siblings are genetically predisposed to a third phenotype that in turn influence the risk of both AMI and RA in a non-genetic fashion. An example of this would be through a genetic predisposition to smoking, which in turn would lead to an increased risk for both RA and AMI as tobacco intake is a large environmental risk factor for both diseases [2, 5].

Additionally, our results regarding the lack of pleiotropic risk factors is only valid for the phenotypes of RA and AMI studied here. It is still unclear whether there are CVD-phenotypes in which an overlap with RA exists or if these results are indicative of a general lack of genetic correlation. Studying the overlap of RA and CVDs for a wider range of phenotypes would be a topic for a future study.

5.3 Heterogeneity between RA-subtypes

For exploratory purposes, we attempted to investigate the genetic heterogeneity among the three types of RA by assessing the genetic correlation between them.

As expected, we found a strong genetic correlation between RA and ACPA-positive RA. Visual inspection of their Manhattan-plots (Figures 3 and A2, pages 22 and 48 respectively) hint about an overlap. Further inspection of the 559 SNPs that reached genome-wide significance in RA revealed that only 3 failed to be replicated in the ACPA-set with those 3 all being borderline significant. Genetic correlation was estimated at $\hat{r}_g = 0.8699$ and was found to be significant with a p-value in the range of 10^{-7}.

For the analysis of overlap of RA subtypes with seronegative RA, results were less meaningful. Correlation coefficient estimates were generally imprecise with large standard errors and point-estimates that reach out of bounds of the defined interval of $[-1, 1]$. As such, no meaningful conclusions could be drawn regarding the heterogeneity of seronegative RA and other RA subtypes in this study.

We note that there are several issues with the seronegative data in this study. Firstly, as stated previously, the seronegative set is at low power due to a relatively low amount of individuals. Secondly, the already low heritability of seronegative RA coupled with potential measurement errors means that computing the genetic correlation by dividing with the root of the heritability may lead to spurious results. Thirdly, there is a lack of polygenic effect in seronegative RA which breaks the model assumption in LDSR (see Section 2.2). Lastly, there may be an error in our constraining of the intercept for estimating the genetic covariance ρ_g as our pre-set value of the intercept differs largely from the model estimate when no constraints are imposed.

6 Discussion

6.1 Study summary

The basis for our study was an article which demonstrated that healthy, direct siblings of patients with RA was at an elevated risk for the CVD of acute coronary syndrome [7]. It has been previously established that patients with RA suffer an increased risk for a variety of CVDs [3], including acute coronary syndrome, but finding an increased risk in healthy siblings too indicates a common, shared susceptibility to both phenotypes. Whether the nature of this susceptibility is genetic, environmental, or due to both remains a question, as does if the susceptibility extends to other CVD-phenotypes or if the results of [7] are valid only for acute coronary syndrome. In this study our goal was to investigate the nature of this shared susceptibility by studying the genetic part of the two phenotypes and their interplay. We looked at the genetic overlap between RA and the CVD-phenotype of AMI, using AMI here as a proxy for CVDs in general, through parametric models that estimate the genetic correlation coefficient, r_g, defined as a correlation coefficient on the genetic part of the two phenotypes (see Definition 3). We employed GWAS summary statistic data for both RA and AMI and utilized a novel method called linkage disequilibrium score regression to obtain estimates of r_g for the phenotypes [18].

In studying the heritability of the four phenotypes considered, our findings mostly replicate previously published results regarding RA, AMI and the RA subtypes. However, what is worth noting are the modest discordances in heritability estimates between the RA subtypes, as well as the highly different GWAS results observable between Figures 3, A2 and A3 (pages 22 and 48 respectively). For instance, ACPA-positive RA exhibits a strong genetic dependence, especially towards SNPs contained in the HLA-region on chromosome 6, whereas seronegative RA is far less genetically dependent with no genome-wide significant SNP associations at all, and heritability only due to a polygenic effect of many variants offering a minor contribution. Our attempts to pinpoint this genetic discrepancy through assessment of within-disease heterogeneity led to inconclusive and nonsense estimates of the genetic correlation coefficient. Whether this genetic difference is due to incorrectly diagnosed RA, a faulty diagnosis criteria, an environmentally triggered phenotype or simply different diseases masquerading as RA is impossible to conclude here. However, this inconclusiveness is most likely due to small sample sizes, both N of individuals and M of SNPs, and a better sample could hopefully help elucidate the reason for this genetic difference.

In our study we found no significant genetic correlation for AMI with RA or with any of the investigated subtypes of RA (ACPA-positive RA and serologically negative RA). As previously stated, this implies that no pleiotropic risk factors simultaneously conferring risk for RA and AMI exist between the phenotypes. However, as mentioned in the previous section, genetics may still be a key factor in the elevated risk for siblings by working through a proxy third phenotype not here considered. For instance, a pair of siblings may be geneti-

cally predisposed towards such a phenotype that environmentally influences the risk of developing either both RA and AMI or simply AMI alone.

6.2 Study weaknesses

While we believe the study to be valid in general there are several areas in which the study could be improved. Firstly, individual sample size N could be improved. For RA, we have data on around 6000 individuals after quality control, with further diminishing counts for subtypes of ACPA-positive RA and seronegative RA. Low sample size N leads to a lack of power in detecting genome-wide significant SNP associations in our GWAS, which in turn lead to a decrease in power for estimation of r_g with LDSR. At around 6000 individuals, we are adequately powered to detect the strongest associations, but it usually takes samples of nearly ten times this size to detect the less obvious SNP associations. As such, the observed lack of genome-wide significant SNPs in seronegative RA (Figure A3, page 48) may be either due to a true lack of genetic signal or simply due to a lack of power. Meta-analysis with several sets of genotyped RA data may help mitigate these results but they add further complexities to the study as they often require using different populations which may lead to confounding of results due to differences in ancestry.

Secondly, the sample size M of SNPs is an immediate weakness in that low counts of SNPs can lead to a loss of precision in the LDSR estimates. At the beginning of our study we had data on approximately 600 000 SNPs which is already at the lower end compared with recent GWASs in which SNP sample sizes consistently reach over a million. Further reduction of the number of SNPs then occurs after quality control and merging with LD scores. As the SNPs genotyped for our RA data badly matched the SNPs genotyped in the reference panel [41] used for estimating LD scores, nearly two-thirds of our data material was lost. This can be adjusted, either by using better matching LD scores, imputing further SNPs into our data or a combination of both.

Thirdly, there may be confounding due to a mismatch between the populations considered. In comparing the Swedish population studied for RA with the UK population genotyped for AMI, we implicitly assume that these two populations are genetically similar. There is undoubtedly a genetic difference between these two but we assume this discrepancy to be negligible. Future studies should investigate this difference and make sure that the differences between the populations do not lead to significant confounding in our results. Furthermore, we implicitly assume that the reference population for our LD scores [41] is genetically similar to both our Swedish RA population and our UK AMI population. While this reference population has been considered an adequate match for general northern European populations [19], we did not investigate whether there are more appropriate reference panels available for the two populations considered in our study.

Fourth, the adequacy of the phenotype of AMI in the UK Biobank data could be questioned. Available information tells us that cases were classified as individuals who had experienced an AMI, based on a physicians diagnosis (as

opposed to self-reported). However, we were not able to find out whether the AMI cases were individuals who had experienced an AMI prior to the time of inclusion in the study or whether they had experienced AMIs during a followup period after inclusion. Our study assumes that it is the latter, as the former constitutes a different phenotype, not of experiencing an AMI but of surviving it which should lead to a different analysis.

Lastly, a better method for estimation of heritability and genetic correlation could have been utilized. The genomic restricted maximum likelihood method [17, 45, 46] has been shown to perform better than LDSR for samples of similar size by reducing standard errors [12, 18]. Unfortunately, such a method requires individual-level genome-wide data, which was not available to us for CVD-phenotypes.

Study improvements of this kind could hopefully increase the precision of estimates by reducing standard errors further and possibly through inflating heritability estimates so as to diminish the gap between our obtained estimates and some of the published family-based estimates. While we stand skeptical to whether the magnitude of estimates of r_g could increase, there is definitely a possibility that an improved study could detect a significant result, although we doubt that such a correlation would be anything more than minor.

A Supplementary theory

A.1 Modeling phenotypic variation

In genetics research of complex phenotypic traits, the trait heritability of h^2 is a key concept. It is formally recognized as a measure of the proportion of total phenotypic variance attributable to genetic variation [50]. As such it becomes a key concept in discussing phenotypic variation and genetic correlation between traits.

For an observed phenotype P we generally model it as

$$P = G + E,$$

where G denotes genotype and E environmental influence, both unobserved. A general form for the variance of a phenotype can then be given as

$$\sigma_P^2 = \sigma_G^2 + \sigma_E^2 + 2\sigma_{G,E} + \sigma_{G \cdot E}^2$$

The first two terms are of genotype and environmental variance with the two remaining terms being of covariance between genetics and environment and interaction between genetic and environmental factors. These two last terms of $\sigma_{G,E}$ and $\sigma_{G \cdot E}^2$ are usually ignored as they are highly difficult to measure and estimate [50]. For certain phenotypes, covariance and interaction may be relevant but we argue that this is not the case for the phenotypes studied in this text.

This leads to the definition of the broad-sense heritability as a ratio of genetic variance and phenotypic variance [50].

Definition A1. *We define the broad sense heritability H^2 as the phenotypic variation attributable to genetics. We write this as*

$$H^2 = \frac{\sigma_G^2}{\sigma_P^2},$$

where σ_P^2 is the phenotypic variation and σ_G^2 is the genetic variation.

The strength of the broad sense heritability H^2 is that it does not require any strong assumptions on the relationship between σ_G^2 and σ_P^2, neither does it require assumptions on how various different genetic effects contribute to σ_G^2. However, this flexibility of the broad sense heritability is also a limitation when it comes to estimation. Usually, two simplifying assumptions are made. Firstly, we assume that the genetic variation σ_G^2 can be partitioned as

$$\sigma_G^2 = \sigma_A^2 + \sigma_D^2 + \sigma_I^2,$$

into a sum of additive genetic effects (σ_A^2), dominant and recessive effects (σ_D^2) and of interaction effects from between variants (σ_I^2) [50].

The contribution of σ_A^2 is often assumed to be the largest in the sense of $\sigma_A^2 \gg \sigma_D^2$ and $\sigma_A^2 \gg \sigma_I^2$ respectively [50]. This assumption leads to a second definition of heritability.

Definition A2. *Under the assumptions given above, we define the narrow sense heritability h^2 as the phenotypic variation attributable to additive genetic effects. We write this as*

$$h^2 = \frac{\sigma_A^2}{\sigma_P^2}. \tag{19}$$

Note that if the quantities σ_D^2 and σ_I^2 are near zero we have $h^2 \approx H^2$, otherwise it holds that $h^2 \leq H^2$. As heritability can not be observed nor completely captured, only estimates are available. Most heritability estimates are of the narrow sense heritability h^2, as it is much easier to capture due to the simplifying assumptions. Different approaches exist for quantifying the heritability of a trait, where family-based sibling and twin studies have been the standard method for capturing h^2 of various phenotypic traits [50]. However for phenotypes such as rare diseases, it can be difficult to gather a large enough sample size to have the power to obtain reasonable estimates. As a result of this, many researchers have used genotyped data in various forms to quantify the heritability of rare, complex traits. Various methods to estimate the heritability of a trait based on genotyped data have been suggested and many of the approaches covered in Section B of methods for genetic overlap also have extensions for estimating the trait heritability h^2.

A.2 Scale independence of the genetic correlation coefficient

We commented previously in Section 2.3.2 on how the genetic correlation coefficient r_g is free from scale in the observed-liability sense. We prove this here.

Lemma A3. *If $r_{g,obs}$ denotes the observed scale genetic correlation coefficient of $r_{g,obs} = \rho_{g,obs}/\sqrt{h^2_{1,obs}h^2_{2,obs}}$ then*

$$r_{g,obs} = r_g, \tag{17, revisited}$$

i.e. the genetic correlation coefficient is independent of scale.

Proof. We have that

$$r_{g,obs} = \frac{\rho_{g,obs}}{\sqrt{h^2_{1,obs}h^2_{2,obs}}}$$

$$= \rho_g \frac{\phi(\tau_1)\phi(\tau_2)\sqrt{P_1(1-P_1)P_2(1-P_2)}}{K_1(1-K_1)K_2(1-K_2)} \sqrt{\frac{K_1^2(1-K_1)^2K_2^2(1-K_2)^2}{h_1^2\phi(\tau_1)^2P_1(1-P_1)h_2^2\phi(\tau_2)^2P_2(1-P_2)}}$$

$$= \frac{\rho_g}{\sqrt{h_1^2h_2^2}} \left(\frac{\phi(\tau_1)\phi(\tau_2)}{\sqrt{\phi(\tau_1)^2\phi(\tau_2)^2}}\right) \left(\frac{\sqrt{P_1(1-P_1)P_2(1-P_2)}}{\sqrt{P_1(1-P_1)P_2(1-P_2)}}\right) \left(\frac{\sqrt{K_1^2(1-K_1)^2K_2(1-K_2)^2}}{K_1(1-K_1)K_2(1-K_2)}\right)$$

$$= \frac{\rho_g}{h_1^2h_2^2} = r_g \tag{20}$$

□

A.3 Genomic inflation factor

Normally in a GWAS, we expect to see some inflation of test statistics corresponding to true genetic effect of SNP on phenotypic trait. However, further undesired inflation may occur due to population stratification or measurement errors inflicting bias. The, in the literature, standard way of quantifying the amount of inflation in GWAS test statistics is by $\hat{\lambda}_{gc}$, the estimated genomic inflation factor [32]. Mathematically, we define it as follows [20].

Definition A4. *Let $\chi^2_1, ..., \chi^2_M$ be χ^2-statistics of SNP effect on phenotypic trait for M SNPs. The estimated genomic inflation factor $\hat{\lambda}_{gc}$ is then given by*

$$\hat{\lambda}_{gc} = \frac{Median\{\chi^2_1, ..., \chi^2_M\}}{\chi^2_{0.5}}, \tag{21}$$

where $\chi^2_{0.5}$ denotes the 0.5 quantile of the χ^2 distribution with 1 degree of freedom.

In a GWAS with little to no inflation of test statistics, we expect $\hat{\lambda}_{gc}$ to be close to 1. However, as stated above, we expect some amount of test statistics inflation, either due to true SNP effect or bias. Still, as the number of genotyped SNPs M tends to be large, the observed deviation from 1 in any GWAS ought to be small. As such, we may use the estimated $\hat{\lambda}_{gc}$ to assess a kind of goodness-of-fit for our GWAS.

We note that the authors in [32] are somewhat critical of using $\hat{\lambda}_{gc}$ as a measure of the confounding bias introduced into a GWAS by population structures, claiming that λ_{gc} depends upon far more factors than traditionally accounted for. They point to a dependence upon variables such as sample size, disease prevalence and heritability to name a few. We agree that this would make $\hat{\lambda}_{gc}$ non-optimal in general, and especially difficult to compare for association studies on differing populations, phenotypic traits or study groups. However, in this text we use it to compare the goodness of fit within study, trait and population, keeping most of the factors detailed in [32] fixed. As such, a change in genomic inflation factor would be solely due to population structure as adjusted for by principal components, making it a reasonable measure for selecting the best performing set of GWAS summary statistic data in our particular situation.

A.4 Bias in estimates of r^2

For individual-level genome-wide data, the LD of two genomic regions measured in r^2, where r^2 is as given in Definition 2 is generally estimated as

$$\tilde{r}^2_{jk} = \left(\frac{1}{N} \sum_{i=1}^{N} x_{ij} x_{ik} \right)^2, \tag{22}$$

where $x_{ij} \in \{0, 1\}$ denotes the genotype of individual i at SNP j and analogously for x_{ik}. One can show (as is done in the supplementary material of [19]) that

the expectation of this estimator is given by

$$E[\tilde{r}^2_{jk}] \approx r^2_{jk} + \frac{1 - r^2_{jk}}{N},$$

which means that the LD score of SNP j using the standard estimator of r^2 confers a bias of

$$E[\tilde{l}_j] = E\left[\sum_{j=1}^{M_l} \tilde{r}^2_{jk}\right] \approx \sum_{j=1}^{M_l} r^2_{jk} + \frac{M_l - \sum_{j=1}^{M_l} r^2_{jk}}{N} = l_j + \frac{M_l - l_j}{N},$$

As an estimator of the amount of LD between two genomic regions, $\tilde{r}^2$ is approximately unbiased with the error tending to zero as sample sizes grow larger. However, as an estimator in the LD scores, the error grows significantly larger, ultimately conferring an upwards error of order $O(M_l/N)$ which might not be negligible depending on the study sample sizes.

The suggested correction is then an adjusted estimator of

$$\hat{r}^2_{adj} = \tilde{r}^2 - \frac{1 - \tilde{r}^2}{N - 2}. \qquad \text{(18, revisited)}$$

Using this estimator, the LD scores instead contribute bias on the order of $O\left(\frac{M_l}{N(N-2)}\right)$. We prove this below.

Lemma A5. *Estimating LD scores with the $\hat{r}^2_{adj}$ estimator given in Equation 18 confers a smaller amount of bias than estimates produced using the standard r^2 estimator of $\tilde{r}^2$ given in Equation 22.*

Proof. For two genomic regions j and k we have that

$$\begin{aligned} E\left[\hat{r}^2_{jk}\right] &= E\left[\tilde{r}^2_{jk}\right] - \frac{1 - E[\tilde{r}^2_{jk}]}{N - 2} \\ &= r^2_{jk} + \frac{1 - r^2_{jk}}{N} - \frac{1 - \left(r^2_{jk} + \frac{1 - r^2_{jk}}{N}\right)}{N - 2} \\ &= r^2_{jk} + \frac{r^2_{jk} - 1}{N(N-2)}. \end{aligned}$$

Now taking the sum over k we get that the error in the LD scores are given by

$$E\left[\hat{l}_j\right] = E\left[\sum_{k=1}^{M_l} \hat{r}^2_{jk}\right] = l_j + \frac{l_j - M_l}{N(N-2)} = l_j + O\left(\frac{M_l}{N(N-2)}\right),$$

which completes the proof. □

B Methods for genetic overlap

In this section we cover a few of the available approaches to assessing genetic overlap between two phenotypic traits. We do not aim to give an extensive review of what has been previously published but hope to give a short introduction to some of the alternatives to the method of linkage disequilibrium score regression that is covered in Section 2.

All of the methods presented in this section require two sets of genetic data, either in the form of individual-level genome-wide data or as GWAS summary statistic data. By individual-level genome-wide data, we mean data consisting of observed genotypes for each individual with phenotypic measurements on each person. We can think of this data as a matrix of dimension $N \times (M + 1)$ where each row represents an individual with the first column corresponding to their observed phenotype and the remaining M columns containing their genotypes at each of the M genotyped SNPs.

Furthermore, for GWAS summary statistic data we assume that individual-level genome-wide data has been processed according to what is described in Section 1.2. We define the output of the GWAS, i.e. what is here referred to as the GWAS summary statistic data, as containing a test of association (and direction of association), standard error and p-value of association significance for each of the M SNPs.

Further methods exist that dissect the genetic part of any phenotype using other formats of genome-wide data but we do not consider these here.

B.1 Polygenic risk scores

Polygenic risk scores are essentially risk scores based on genetic data that give an individuals propensity towards developing a phenotype of interest. The name polygenic comes from the idea that an individuals propensity towards a phenotype is due to the collective effect of a large set of both minor and major genetic effects, as opposed to only a small set of major genetic effects. Computing the risk scores requires two sets of genetic data: one set of GWAS summary statistic data on the phenotype of interest, often called the training data, and a set of individual-level genome-wide data, often called the testing data, for the individuals of which we wish to obtain the risk scores. Based on the GWAS summary statistic data, a set Q is then created containing all the SNPs which can be seen as risk-influencing, often taken as all the SNPs that were found to be associated with the phenotype of interest at a given significance level. From this set Q we may then compute the polygenic risk score of individual i, denoted here as S_i, by

$$S_i = \sum_{j \in Q} x_{ij} w_j, \tag{23}$$

where w_j is a weight based on the effect of the j'th SNP in Q and $x_{ij} \in \{0, 1, 2\}$ is the number of risk-influencing alleles observed at the j'th SNP in Q for individual i in the individual-level genome-wide data.

This quantity allows researchers to estimate a healthy individuals genetic propensity towards developing a given disease or other phenotypic trait. However, depending on the population sampled for the individual-level genome-wide data, we may utilize the polygenic risk scores S_i, $i = 1, ..., N$ to test for genetic overlap between the phenotype in the training GWAS summary statistic data set and the testing individual-level genome-wide data. In general, if y_i denotes the observed phenotype of individual i with corresponding polygenic risk score S_i, then we may test S_i for association with phenotype y_i by regressing S_i onto y_i. Significant association would then be an indication that the SNPs contained in Q based on the training GWAS summary statistic data, are associated with the population phenotype of the testing individual-level genome-wide data.

This is best illustrated by an example. Say we are, as in this study, interested in the genetic overlap between RA and AMI. Suppose we have GWAS summary statistic data on the phenotype of AMI, i.e. we have a set of M SNPs, their individual associations with the phenotype of AMI as effect sizes w_j and their subsequent p-values of significance of association. We may then create the set Q as the SNPs that surpass a threshold of significance in the GWAS on AMI, here denoted by p_T, as

$$Q = \{SNP_j \ ; \ p_j \leq p_T, j = 1, ..., M\},$$

where p_j denotes the p-value of the j'th SNP. Suppose next that the population sampled for the testing, individual-level genome-wide data set, consists of individuals diagnosed with RA (cases) and healthy individuals (controls) as described in Section B above. Polygenic risk scores for each of the N individuals in the testing sample of RA cases and controls may then be computed through the use of Equation 23. Now testing for genetic overlap simply amounts to testing observed phenotype in the RA-data, for association with the polygenic risk scores S_i. The standard way to do this in practice is by fitting the model of

$$y_i = \alpha + S_i\beta + \epsilon_i, \ i = 1, ..., N$$

where y_i denotes observed phenotype of individual i, and testing the hypothesis of $\beta = 0$.

Polygenic risk scores were first presented in a seminal paper in which they were used to study the genetics of schizophrenia. In the paper, the authors demonstrated an association between the polygenic risk score, based on training GWAS summary statistic data for schizophrenia, with the phenotype of bipolar disorder [15]. Today, a wider applicability of the method has seen use, including in estimation of heritability h^2, studying the genetic architecture of a trait and in prediction of disease development. Further extensions have been made to the original approach to computing risk scores detailed above, allowing researchers to further increase the numbers of risk-influencing SNPs contained in the set Q to allow for even more of the M genotyped SNPs to be included in the analysis. Such methods include a Bayesian approach to estimating the SNP weights w_j by incorporating linkage disequilibrium [51] and an approach employing penalized regression aiming to adequately shrink effect size estimates w_j to allow for further SNPs to be incorporated [52].

B.2 Genomic restricted maximum likelihood

The method here referred to as genomic restricted maximum likelihood is known in the literature under various different names such as genome-wide complex trait analysis or simply linear mixed models. It aims to model SNP effect on phenotype through a linear mixed model where SNPs are taken as the random effects. Estimates of genetic correlation r_g are then produced as in Definition 3 by fitting a bivariate linear mixed model. Estimates of heritability h^2 and genetic covariance ρ_g are then obtained as estimates of the variance components of the model [17, 45].

The model does not utilize any GWAS summary statistic data and instead requires two sets of individual-level genome-wide data. The bivariate linear mixed model equations can be given as:

$$Y_k = X_k\beta_k + Z_k g_k + \epsilon_k,\ k = 1, 2,$$

where Y_k is an $(N_k \times 1)$ vector of phenotypic trait values, β_k is an $(p_k \times 1)$ vector of fixed effects, g_k is a vector of total genetic SNP effects dimension $(M \times 1)$ where $g_k \sim N(0, Ah_k^2)$ and ϵ_k is the residual vector $\epsilon \sim N(0, 1 - h_k^2)$ for phenotypic trait k. Lastly, the variables X_k and Z_k are both incidence matrices for fixed and random effects respectively [17].

Here, A is a matrix commonly referred to as the genomic relationship matrix, each of its elements corresponding to correlation coefficients between the individuals in their respective sets. The elements of A are given as

$$a_{jk} = \frac{1}{M}\sum_{i=1}^{M}\frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}, \tag{24}$$

where $x_{ij} \in \{0, 1, 2\}$ denotes the number of copies of the reference allele at SNP i for individual j and p_i is the observed allele frequency of the reference allele at the i'th SNP [53].

The variance-covariance matrix is given by

$$V = \begin{bmatrix} Z_1 A Z_1^T h_1^2 + I(1 - h_1^2) & Z_1 A Z_2^T \rho_g \\ Z_2 A Z_1^T \rho_g & Z_2 A Z_2^T h_2^2 + I(1 - h_2^2) \end{bmatrix}$$

The method of genomic restricted maximum likelihood then fits this type of model and estimates the variance components of h_1^2, h_2^2 and ρ_g through restricted maximum likelihood estimation out of which one may then compute the genetic correlation coefficient r_g as it is defined in Definition 3.

Currently, genomic restricted maximum likelihood is implemented in a popular software tool known as GCTA, or genome-wide complex trait analysis, which contains tools for estimation of heritability, estimation of genetic correlation, simulating GWAS data and various other tools for genetic analysis of phenotypic traits [53]. It has since reached wide-spread popularity and been employed in several seminal papers studying topics such as the interplay of psychiatric diseases [54], the change in cognitive ability from young age to old [55] and in further improving upon the previously biased downwards SNP-based heritability estimates of several complex traits [45] to simply name a few.

B.3 SNP effect concordance analysis

Commonly abbreviated SECA, SNP effect concordance analysis was initially developed to allow researchers to test for genetic overlap using only GWAS summary statistic data. The method aims to test for overlap by checking for concordance in direction of association for each SNP between the two phenotypes. The null hypothesis of no genetic overlap is then rejected if the amount of concordance among the two sets is greater than what is expected by chance [16].

In practice, let $\hat{\theta}_{ij}$ denote the estimated effect of SNP i on phenotype j from the GWAS, $i = 1, ..., M$ and $j = 1, 2$. We may then sort the data into a 2×2 contingency table as

	$\hat{\theta}_{i1} < 1$	$\hat{\theta}_{i1} \geq 1$
$\hat{\theta}_{i2} < 1$	n_{11}	n_{12}
$\hat{\theta}_{i2} \geq 1$	n_{21}	n_{22}

where n_{ij} denotes the observed counts of the various cells. Testing the null hypothesis of no genetic overlap is then equivalent to testing for independence of rows and columns. As such, we may test for an excess of concordance in direction of SNP associations through either a chi-squared test or a Fisher's exact test depending on the magnitudes of the x_{ij} counts.

The method is implemented in a web-based application [16] and has since its development been used to investigate the heterogeneity between subtypes of migraine [56] and to test for genetic overlap in collagenous colitis and different inflammatory bowel diseases [57].

C Quality control of RA data

As mentioned in Section 3.1, rigorous quality control must be carried out on the genome-wide data to avoid potentially spurious findings in the association analysis. In this section of the Appendix we give a detailed description of all the steps taken to ascertain quality of the RA-data for the GWAS. We do this to offer transparency of the process as well as to present our arguments for the decisions made in our quality control [58].

C.1 Sex discrepancy

We checked individuals for discordance between their reported, and from data imputed, sex. Such a discordance may occur for several reasons but it may be an indicator of sample mix-ups or mishandling of data [28]. We do this by computing observed homozygozity rates for each of the individuals in the set and comparing these to the expected homozygozity rates based on reported sex. Typically, we expect males to have a rate of 1 while females are expected to have a rate below 0.2. Two individuals were found to strongly discord with the expected rates (females with observed rates above 0.6) in a way that could not be explained by errors in genotyping rate. As we could not confirm that sex

had been reported incorrectly for these two individuals we decided to exclude them.

C.2 Individual genotyping rate

We expect the genotyping rate of SNPs in individuals to be uniform and homogeneous with respect to the amount of SNPs captured. As such, a low genotyping rate would be an indicator of low DNA quality in samples or genotyping complications [28]. This rate is decided within-sample, where the genotyping rate of the individual with the highest amount of genotyped SNPs is 1. To establish a high standard we excluded any individuals who exhibited a missingness greater than 5%, which is a common threshold for missingness in quality control for association analysis [28]. This lead to the exclusion of 13 individuals who failed to reach a genotyping rate of 95%.

C.3 SNP genotyping rate

A similar check as the one above for individuals is then performed for SNPs. Again, we expect the genotyping to perform similarly and low SNP call rates may be another indicator of complications or errors in the genotyping process. Failure to account for these may lead to false positives, reducing the ability to detect true associations in the GWAS [28].

The rate is established in-sample by measuring the number of individuals in which the allele at the position of the SNP was genotyped. We allow for at most 5% missingness and as such exclude all SNPs with a call rate below 95% which is again a common threshold [28]. This type of procedure leads to the exclusion of 12 218 SNPs.

C.4 Minor allele frequency

The minor allele frequency is the observed allele frequency for the allele in minority at each of the SNPs. We filtered on these, excluding all SNPs which had a minor allele frequency $< 1\%$ leading to the exclusion of 138 713 SNPs.

Note that this is a big chunk of the data material, accounting for about 20% of the total SNPs. While the amount is large, it is not a surprising find and the stringent criteria for exclusion is in fact a commonly used threshold for association studies [28].

A SNP with low minor allele frequency may occur due to measurement error, i.e. incorrect detection of a SNP. This would be a false positive and removal of the SNP would be correct. If the SNP is not due to measurement error, we would have a true rare variant. However, these would be less robust than non-rare variants and would lack power to detect association in the resulting GWAS meaning their removal should not strongly affect the outcome of the study. Furthermore, for this particular data, the material is still quite large for GWAS purposes meaning that despite the exclusion of SNPs in this step we should still retain enough power to detect meaningful associations.

C.5 Hardy-Weinberg equilibrium

Hardy-Weinberg equilibrium (HWE) is an assumption on the distributional relationship between allele and genotype frequencies in a population. Mathematically we define it as follows. Consider a specific locus on the genome which has alleles **A** and **a**. Suppose the underlying allele frequencies of these are p and $(1-p)$ for **A** and **a** respectively. If a population is under Hardy-Weinberg equilibrium then it holds that the genotypic frequencies π_1, π_2 and π_3 for **AA**, **Aa** and **aa** respectively are given by

$$\pi_1 = p^2, \quad \pi_2 = 2p(1-p), \quad \pi_3 = (1-p)^2.$$

For genome-wide data, failure to meet HWE can be an indication of genotyping, or genotyping-call, error [28].

Testing for HWE is done through a Fisher's exact test. Generally for large samples, a chi-squared test is preferred over the exact test but in this setting, such tests can have inflated type I error rates and we thus opt for the Fisher's exact test [59]. Filtering on HWE is done by removing SNPs that reach significance under testing with H_0 : HWE holds. In controls we set a threshold of 10^{-6} while we allow for a more lenient threshold in cases of 10^{-10} as is done in [29]. There's differing opinions on what significance threshold should be set when testing for HWE but most agree that we should be less stringent for cases. This is due to the fact that deviation from HWE may occur due to selection meaning that removing SNPs due to deviations in cases may remove true signal and as such, some authors argue that we should allow all deviation present in cases [28]. However, as the method we utilize for estimating genetic overlap assumes that we can standardize observed genotypes with respect to HWE, we choose to remove SNPs that exhibit extreme deviation in cases.

This type of filtering leads to the removal of 1893 SNPs.

C.6 Pruning by linkage disequilibrium

The last steps needed for our quality control is to account for genomic relatedness of individuals and population substructures. The methods employed here function best if performed on a set of independent SNPs, i.e. SNPs in linkage equilibrium with each other (see Section 1.3) [28, 29]. To account for LD we perform two types of pruning on the SNPs in the data set. Note that this pruning is only for the sake of estimating relatedness and population substructures: the SNPs pruned here will be returned into the data set for the subsequent association analysis.

The first pruning is done by completely removing known problematic and complicated regions of the genome. One such region on each of four chromosomes are pruned, including the human leukocyte antigen (HLA) region on chromosome 6. The areas targeted for the genomic regions are given in Table A1 and are based on regions pruned in a similar quality control in [60]. In total, 9792 SNPs are removed.

Table A1: The genomic positions on the chromosomes of the regions targeted for pruning. Distance measure Mb refers to megabases where a distance of one megabase corresponds to a sequence length of one million nucleotides. Regions are based on a similar exclusion performed in reference [60].

Chromosome	**Region**
5	44 Mb - 51.5 Mb
6	25 Mb - 33.5 Mb
8	8 Mb - 12 Mb
11	45 Mb - 57 Mb

To make sure that the remaining SNPs are in approximate linkage equilibrium, further pruning must be employed. This pruning procedure looks at a window of markers of a given size, pruning so that no SNPs in the window are above a certain cutoff level, measured in LD, and then moving the window a given step length after a window has been successfully pruned. In our quality control we used a window containing 1000 markers, moving it a length of 10 markers at each step and pruning SNPs that have $r^2 > 0.2$ where the quantity r^2 is the one defined in Definition 2 in Section 1.3. Over the whole genome, this procedure removed 327 207 SNPs in total, leaving a set of 213 112 SNPs in approximate linkage equilibrium for population-based quality control.

C.7 Genomic relatedness

For GWAS purposes, we require sampled individuals to be essentially unrelated. Failing to account for relatedness may lead to the introduction of bias as within family genotypes can be over represented. Filtering on genomic relatedness amounts to estimating the relationship of individuals based on the individual-level genome-wide data and excluding one individual from each pair reaching a certain threshold. Here, genomic relatedness is measured as a correlation coefficient sorted into an $N \times N$ symmetric matrix A known as the genomic relationship matrix, where N is the number of sampled individuals. The matrix A has elements a_{jk} of the level of genomic relatedness for individuals j and k given by

$$a_{jk} = \frac{1}{M} \sum_{i=1}^{M} \frac{(x_{ij} - 2p_i)(x_{ik} - 2p_i)}{2p_i(1 - p_i)}, \tag{25}$$

where $x_{ij} \in \{0, 1, 2\}$ is the number of copies of the reference allele at SNP i for individual j and p_i is the observed allele frequency for the same allele at the i'th SNP, $i = 1, ..., M$ [53]. Note that this is the same matrix as was modeled in the method of genomic restricted maximum likelihood (Section B.2, Equation 24).

An estimated relatedness of 1 approximately corresponds to a duplicated individual or monozygotic twin, estimated relatedness of 0.5 approximately corresponds to first-degree relatives and so forth [28]. We set our threshold to 0.1

which approximately corresponds to pairs of third-degree relatives, leading to the exclusion of 310 individuals.

C.8 Population substructures

In any research on the genetics of a sampled population, we implicitly assume that there is an underlying genetic homogeneity within the population. Genetic heterogeneity between the individuals in a sampled population often occurs when individuals have evolved from different ancestries. As an example, sampling individuals from both Asian and European populations for a genetic study may lead to inconclusive or erroneous findings due to the large differences in general ancestry between individuals of Asian descent and individuals of European descent [61]. To account for potential heterogeneity within our study population, our RA data has been sampled from a geographically tight population that we assume to be genetically similar. However, an unknown substructure may still exist in our population so to ensure that as little confounding as possible enters our analysis, we check our data for population substructures, filtering on individuals that differ greatly from the observed average [28].

We use principal components analysis to correct for outliers with respect to population substructures. Here we consider an individual to be an outlier if they are more than 6 standard deviations from the origin over all the components. Filtering is done iteratively by fitting the principal component analysis, removing all outliers with respect to the above definition, re-fitting the model and filtering again and so forth. We stop after 5 iterations unless no outliers were detected earlier. This procedure is based on the default approach of the EIGENSTRAT software, developed as a tool to account for population heterogeneity in genetic studies [62].

The resulting procedure runs for 5 iterations removing a total of 316 individuals. The resulting principal components from the final performed analysis are then saved and kept as variables. These components will be used as covariates to control for population substructures in the performed GWAS for our RA data.

D Supplementary figures

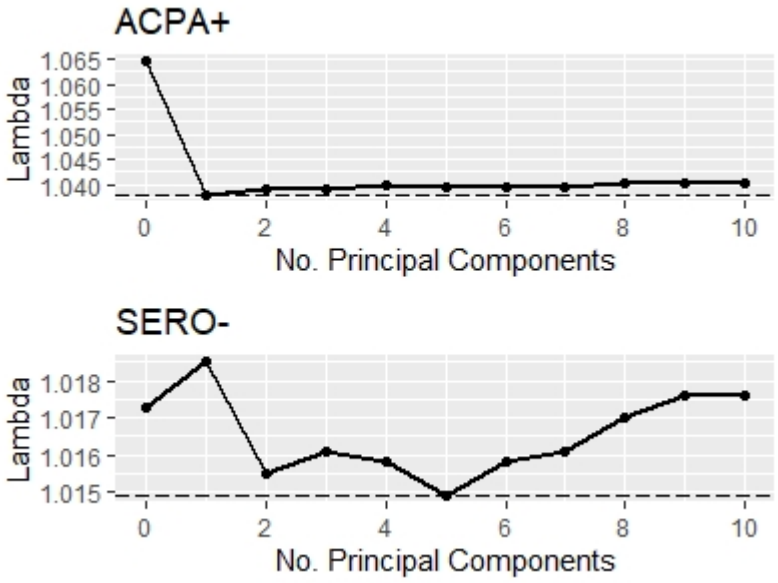

Figure A1: Genomic inflation factor $\hat{\lambda}_{gc}$ as a function of number of principal components on population stratification used as covariates in the association analysis on the data sets containing ACPA-positive cases and seronegative cases respectively. The dashed line is the minimum of the curve. Minimum for ACPA-set is reached at $\hat{\lambda}_{gc}^{(1)} = 1.038$ and for seronegative-set at $\hat{\lambda}_{gc}^{(5)} = 1.014$, corresponding to 1 and 5 components respectively.

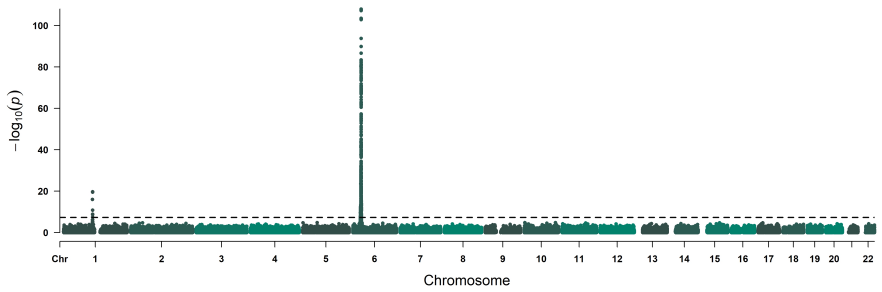

Figure A2: Manhattan-plot for the associations of the ACPA-set. The x-axis contains all autosomal SNPs studied, ordered after chromosome and chromosome position. The y-axis covers the p-values of SNP association with trait at $-\log$ base 10 scale. The dashed line represents the genome-wide significance level of $p = 5 \cdot 10^{-8}$ which is the common standard for recognizing a SNP association as significant in a GWAS.

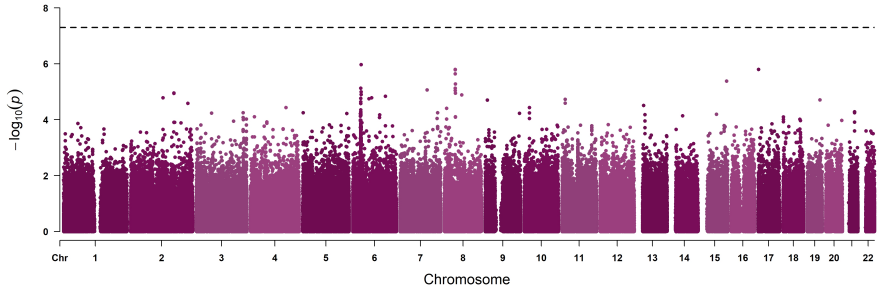

Figure A3: Manhattan-plot for the associations of the seronegative set. The x-axis contains all autosomal SNPs studied, ordered after chromosome and chromosome position. The y-axis covers the p-values of SNP association with trait at $-\log$ base 10 scale. The dashed line represents the genome-wide significance level of $p = 5 \cdot 10^{-8}$ which is the common standard for recognizing a SNP association as significant in a GWAS.

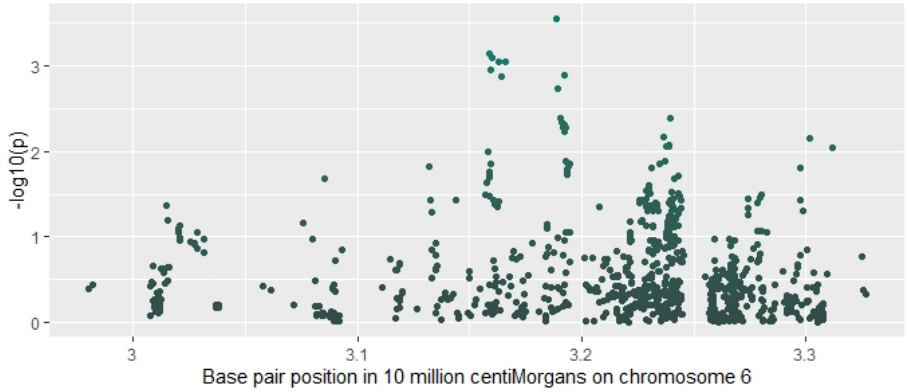

Figure A4: Manhattan-plot of SNP association with myocardial infarction for the UK Biobank sample on a small genomic region on chromosome 6. The y-axis is the base 10 $-log(p)$ transform of the p-values of SNP association. The x-axis is the SNPs position on the genome, given here as base-pair position in the scale of 10^7 centiMorgans. The points are the SNPs that reached genome-wide significance in the ACPA set (Figure A2) here instead illustrating their association with AMI.

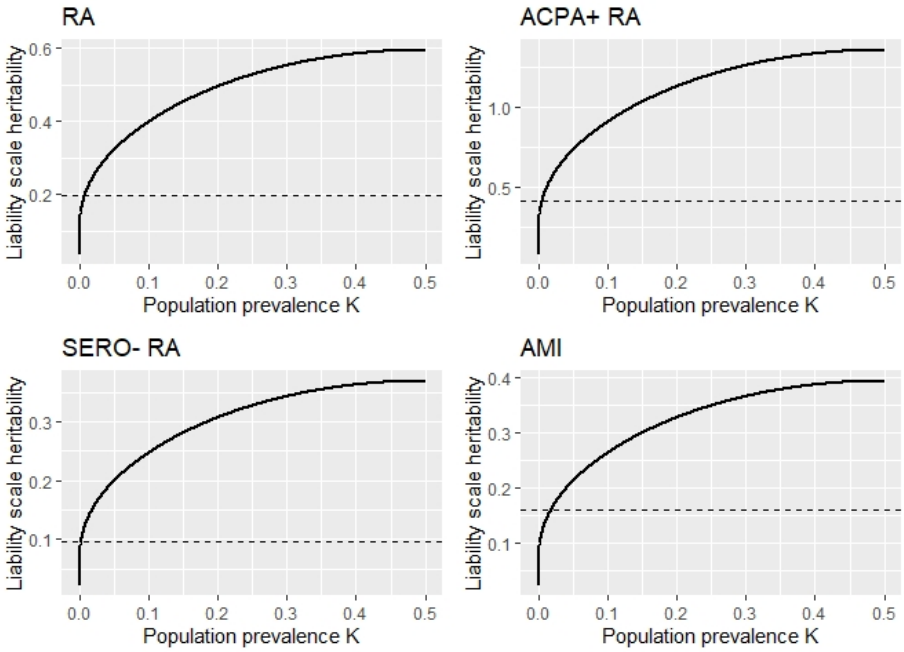

Figure A5: Plots of the resulting liability scale estimates of heritability for the four phenotypes when the population prevalences K are allowed to vary. Illustrates the sensitivity in liability scale conversion due to errors in population prevalence estimates. Formula used for correction to liability scale is given in Equation 16.

Glossary

Acute myocardial infarction (AMI): A type of cardiovascular disease. Commonly referred to as heart attack.

Anti-citrullinated protein antibodies (ACPA): A type of autoantibody present in many, but not all, patients with rheumatoid arthritis.

Cardiovascular disease (CVD): Umbrella term for a set of diseases characterized by involving blood vessels and the heart. Acute myocardial infarction (AMI) is a type of cardiovascular disease.

Epidemiological investigation of rheumatoid arthritis (EIRA): Case-control study of rheumatoid arthritis in a Swedish population consisting of individuals in middle and southern Sweden in the age range of $18 - 70$.

Genome-wide association study (GWAS): Observational study of genetic variants over the entire human genome, i.e. genome-wide genetic variants. Aims to locate genetic variants that are associated with a given phenotypic trait of interest. See Section 1.2 for further details.

Human leukocyte antigen (HLA): A gene complex on chromosome 6 responsible for the regulation of the immune system in humans. Individuals with certain polymorphisms in the human leukocyte antigen-region are known to be more likely to develop certain autoimmune diseases such as rheumatoid arthritis [2].

LD Score (LDSC): Not to be confused with Definition 4, LDSC here refers to the software implementing linkage disequilibrium score regression (LDSR) developed by the authors of the method (https://github.com/bulik/ldsc).

Linkage disequilibrium (LD): Non-random association of alleles at different loci on the genome. Can be seen as a correlation structure between the alleles at these different regions. See Section 1.3 for further details.

Linkage disequilibrium score regression (LDSR): Recent multi purpose method that uses genome-wide association study (GWAS) data to estimate heritability h^2 for single phenotypes and genetic correlation r_g for pairs of phenotypes by regressing GWAS test statistics on a quantity known as the linkage disequilibrium score. Can also be used to estimate a correction factor to account for bias induced by population substructures in a GWAS. See Section 2 for further details.

PLINK: Open source software commonly used for analysis and handling of large scale, genome-wide genetic data [30].

Rheumatoid arthritis (RA): Chronic autoimmune disorder primarily manifesting in the joints of the body. In untreated patients, the disease will lead to joint deterioration and physical disability.

Serologically negative (SERO-): Here referring to an individual diagnosed with RA that tested negative for both rheumatoid factor and anti-citrullinated protein antibodies (ACPA), two antibodies present in many, but not all, cases of RA [31].

Single nucleotide polymorphism (SNP): Pronounced "snips". Common genetic variant, defined as a base-pair change in a single nucleotide in the DNA sequence, present in at least 1% of the human population.

References

[1] M. Neovius, J. Simard, and J. Askling. Nationwide prevalence of rheumatoid arthritis and penetration of disease-modifying drugs in sweden. *Ann. Rheum. Dis*, 70:624–629, 2010.

[2] A. Silman and J. Pearson. Epidemiology and genetics of rheumatoid arthritis. *Arthritis Res*, 4:265–272, 2002.

[3] K. Michaud and F. Wolfe. Comorbidities in rheumatoid arthritis. *Best Pract Res Clin Rheumatol*, 21:885–906, 2007.

[4] S. Kathirsean and D. Srivastava. Genetics of human cardiovascular disease. *Cell.*, 148:1242–1257, 2012.

[5] M. Law, J. Morris, and N. Wald. Environmental tobacco smoke exposure and ischaemic heart disease: an evaluation of the evidence. *BMJ*, 315:973–980, 1997.

[6] A. Timmis et al. European society of cardiology: Cardiovascular disease statistics 2017. *Eur. Heart. J.*, 39:508–577, 2018.

[7] H. Westerlind, M. Holmqvist, and L. Ljung et al. Siblings of patients with rheumatoid arthritis are at increased risk of acute coronary syndrome. *Ann. Rheum. Dis.*, 78:683–687, 2019.

[8] J. Zheng et al. LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. *Bioinformatics*, 33:272–279, 2017.

[9] W. Bush and J. Moore. Genome-wide association studies. *PLoS Comput. Biol.*, 8, 2012.

[10] P. Visscher, M. Brown, M. McCarthy, and J. Yang. Five years of GWAS discovery. *Am. J. Hum. Genet.*, 90:7–24, 2012.

[11] B. Pasaniuc and A. Price. Dissecting the genetics of complex traits using summary association statistics. *Nat. Rev. Genetics*, 18:117–127, 2017.

[12] R. Maier, P. Visscher, M. Robinson, and N. Wray. Embracing polygenicity: a review of methods and tools for psychiatric genetic research. *Psychological Medicine*, 48:1055–1067, 2017.

[13] N. Risch and K. Merikangas. The future of genetic studies of complex human diseases. *Science*, 273:1516–1517, 1996.

[14] R. Deonier, S. Tavaré, and M. Waterman. *Computational Genome Analysis*. Springer, 2005.

[15] S. Purcell et al. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–752, 2009.

[16] D. Nyholt. SECA: SNP effect concordance analysis using genome-wide association summary results. *Bioinformatics*, 30:2086–2088, 2014.

[17] S.H. Lee, J. Yang, M.E. Goddard, P.M. Visscher, and N.R. Wray. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics*, 28:2540–2542, 2012.

[18] B.K. Bulik-Sullivan et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet.*, 47:1236–1241, 2015.

[19] B.K. Bulik-Sullivan et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.*, 47:291–295, 2015.

[20] B. Devlin and K. Roeder. Genomic control for association studies. *Biometrics*, 55:997–1004, 1999.

[21] M. Kutner, C. Nachtscheim, J. Neter, and W. Li. *Applied Linear Statistical Models*. McGraw-Hill, 2004.

[22] A. Atiken. On least-squares and linear combinations of observations. *Proceedings of the Royal Society of Edinburgh*, 55:42–48, 1934.

[23] S. Lee, N. Wray, M. Goddard, and P. Visscher. Estimating missing heritability for disease from genome-wide association studies. *Am J. Hum. Genet*, 88:294–305, 2011.

[24] D. Falconer. The inheritance of liability to certain diseases, estimated from the incidence among relatives. *Ann. Hum. Genet.*, 29:51–76, 1965.

[25] K. Pearson and A. Lee. On the inheritance of characters not capable of quantitative measurements. *Philosophical transactions of the royal society of London*, 195:79–150, 1901.

[26] F.C. Arnett, S.M. Edworthy, D.A. Bloch, D.J. McShane, J.F. Fries, and N.S. Cooper et al. The american rheumatism association 1987 revised criteria for the classification of rheumatoid arthritis. *Arthritis Rheum.*, 31:315–324, 1988.

[27] P. Stolt, C. Bengtsson, and B. Nordmark et al. Quantification of the influence of cigarette smoking on rheumatoid arthritis results from a population based case-control study using incident cases. *Ann. Rheum. Dis*, 62:835–841, 2003.

[28] C.A. Anderson, F. H. Pettersson, G. M. Clarke, L. R. Cardo, A. P. Morris, and K. T. Zondervan. Data quality control in genetic case-control association studies. *Nat. Protoc.*, 5:1564–1573, 2010.

[29] A. T. Marees et al. A tutorial on conducting genome-wide association studies: quality control and statistical analysis. *Int. J. Methods. Psychiatr. Res.*, 27, 2018.

[30] S. Purcell et al. PLINK: a tool set for whole-genome association and population-based linkage analysis. *Am. J. Hum. Genet.*, 81:559–575, 2007.

[31] G. Steiner and J. Smolen. Autoantibodies in rheumatoid arthritis and their clinical significance. *Arthritis Res.*, 4:1–5, 2002.

[32] J. Yang et al. Genomic inflation factors under polygenic inheritance. *Eur. J. Hum. Genet.*, 19:807–812, 2011.

[33] L. Padyukov et al. A genome-wide association study suggests contrasting associations in ACPA-positive versus ACPA-negative rheumatoid arthritis. *Ann. Rheum. Dis*, 70:259–265, 2011.

[34] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447:661–678, 2007.

[35] A. Hinks, J. Worthington, and W. Thomson. The association of PTPN22 with rheumatoid arthritis and juvenile idiopathic arthritis. *Rheumatology (Oxford)*, 45:365–368, 2006.

[36] C. Bycroft, C. Freeman, and D. Petkova et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562:203–209, 2018.

[37] UK Biobank Axiom array - content summary. http://www.ukbiobank.ac.uk/wp-content/uploads/2014/04/UK-Biobank-Axiom-Array-Content-Summary-2014-1.pdf, 2014. [Accessed: 2019-11-25].

[38] Neale Lab - UK Biobank GWAS. http://www.nealelab.is/uk-biobank/. [Accessed: 2019-11-18].

[39] E. Stahl et al. Bayesian inference analyses of the polygenic architecture of rheumatoid arthritis. *Nat. Genet.*, 44:483–489, 2012.

[40] The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature*, 467:1061–1073, 2010.

[41] The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature*, 526:68–74, 2015.

[42] P. Bhatnagar, K. Wickramasinghe, and J. Williams et al. The epidemiology of cardiovascular disease in the UK 2014. *Heart*, 101:1182–1189, 2015.

[43] T. Frisell et al. Familial risks and heritability of rheumatoid arthritis. *Arthritis Rheum.*, pages 2773–2782, 2013.

[44] A.J. MacGregor, H. Sneider, A.S. Rigby, M. Koskenvuo, J. Kaprio, and K. Aho. Characterizing the quantitative genetic contribution to rheumatoid arthritis using data from twins. *Arthritis Rheum.*, pages 30–37, 2000.

[45] J. Yang et al. Common SNPs explain a large proportion of the heritability for human height. *Nat. Genet.*, 42:565–569, 2010.

[46] J. Yang et al. Concepts, estimation and interpretation of SNP-based heritability. *Nat. Genet.*, 49:1304–1311, 2017.

[47] Y. Okada. Genetics of rheumatoid arthritis contributes to biology and drug discovery. *Nature*, 506:376–381, 2013.

[48] B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, 9:586–596, 1981.

[49] M. Nikpay et al. A comprehensive 1,000 genomes-based genome-wide association meta-analysis of coronary artery disease. *Nat. Genet.*, 47:1121–1130, 2015.

[50] P. Visscher, W. Hill, and N. Wray. Heritability in the genomics era - concepts and misconceptions. *Nat. Rev. Genet.*, 9:255–266, 2008.

[51] B. Vilhjálmsson et al. Modeling linkage disequilibrium increases accuracy of polygenic risk scores. *Am. J. Hum. Genet.*, 97:576–592, 2015.

[52] T. S. H. Mak et al. Polygenic scores via penalized regression on summary statistics. *Genet. Epidemiol.*, 41:469–480, 2017.

[53] J. Yang, S.H. Lee, M. E. Goddard, and P. M. Visscher. GCTA: A tool for genome-wide complex trait analysis. *Am. J. Hum. Genet.*, 88:76–82, 2011.

[54] Cross-Disorder Group of the Psychiatric Genomics Consortium. Genetic relationship between five psychiatric disorders estimated from genome-wide SNPs. *Nat. Genet*, 45:984–994, 2013.

[55] I. Deary and J. Yang et al. Genetic contributions to stability and change in intelligence from childhood to old age. *Nature*, 482:212–215, 2012.

[56] D. Nyholt et al. Concordance of genetic risk across migraine subgroups: Impact on current and future genetic association studies. *Cephalalgia*, 35:489–499, 2015.

[57] H. Westerlind, M-R. Melander, and F. Bresso et al. Dense genotyping of immune-related loci identifies HLA-variants associated with increased risk of collagenous colitis. *Gut*, 66:421–428, 2017.

[58] J. Little, JP. Higgins, JP. Ioannidis, D. Moher, F. Gagnon, and E. von Elm et al. STrengthening the REporting of Genetic Association studies (STREGA) - an extension of the STROBE statement. *PLoS Medicine*, 6:151–162, 2009.

[59] J. Wiggington, D. Cutler, and G. Abecasis. A note on exact tests of Hardy-Weinberg Equilibrium. *Am. J. Hum. Genet.*, 76:887–893, 2005.

[60] J. Fellay et al. Common genetic variation and the control of HIV-1 in humans. *PLoS Genet.*, 5, 2009.

[61] J. Marchini et al. The effects of human population structure on large genetic association studies. *Nat. Genet.*, 36:512–517, 2004.

[62] A. L. Price et al. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, 38:904–909, 2006.