UCLA

UCLA Previously Published Works

Title

Discovering genes involved in disease and the mystery of missing heritability

Permalink

<https://escholarship.org/uc/item/9k26g4g7>

Journal

Communications of the ACM, 58(10)

ISSN

0001-0782

Author

Eskin, E

Publication Date

2015-10-01

DOI

10.1145/2817827

Peer reviewed

Discovering Genes Involved in Disease and the Mystery of Missing Heritability

Eleazar Eskin

Department of Computer Science, University of California, Los Angeles, Los Angeles, CA 90095

Department of Human Genetics, University of California, Los Angeles, Los Angeles, CA 90095

Abstract

It has been known for many decades that genetic differences among individuals account for a substantial portion of trait differences. Only recently has technology been developed to cost effectively measure genetic differences among individuals which creates the possibility of developing models that can predict disease related traits from an individual's genetics. These models are a key component of personalized medicine and developing such models raise many computational challenges which provide important opportunities for Computer Scientists to make contributions toward human health.

1 Introduction

We live in a remarkable time for the study of human genetics. Nearly 150 years ago, Gregor Mendel published his laws of inheritance which lay the foundation for understanding how the information that determines traits is passed from one generation to the next. Over 50 years ago, Watson and Crick discovered the structure of the DNA which is the molecule that encodes this genetic information. All humans share the same three billion length DNA sequence at more than 99% of the positions. Almost a hundred years ago, the first twin studies showed that this small fraction of genetic differences in the sequence accounts for a substantial fraction of the diversity of human traits. These studies estimate the contribution of the genetic sequence to a trait by comparing the relative correlation of traits between pairs of maternal twins (which inherit identical DNA sequences from their parents) and pairs of fraternal twins (which inherit a different mix of the genetic sequence from each parent)[11, 52]. This contribution is referred to as the “heritability” of a trait. For example, twin studies have shown that genetic variation accounts for 80% of the variability of height in the population[11, 48, 34]. The amount of information about a trait encoded in the genetic sequence suggests that it is possible to predict the trait directly from the genetic sequence and this is a central goal of human genetics.

Only in the past decade has technology developed to be able to cost effectively obtain DNA sequence information from individuals and a large number of the actual genetic differences have been identified and implicated in having an effect on traits. On the average, individuals who carry such a genetic difference, often referred to as a genetic variant, will have a different value for a trait

compared to individuals who do not carry the variant. For example, a recently published paper reporting on a large study to identify the genetic differences that affect height reported hundreds of variants in the DNA sequence which either increase or decrease an individual’s height if the individual carries the variant[5, 44]. Knowing these variants and their effects allows us to take the first steps in predicting traits only using genetic information. For example, if an individual carried many variants that increased height, we would predict that the individual’s height is higher than the population average. While predicting an easily measured trait such as height from genetic information seems like an academic exercise, the same ideas can be used to predict disease related traits such as risk of heart attack or response to a certain drug. These predictions can help guide selecting the best treatment options. Over a thousand genetic variants have been implicated in hundreds of traits including many human disease related traits of great medical importance[35, 53].

A majority of these discoveries were made using a type of genetic study called a genome wide association study (GWAS). In a GWAS, data from a large number of individuals is collected, including both a measurement of the disease related trait as well as information on genetic variants from the individual. GWAS estimate the correlation between the collected disease trait and the collected genetic variants to identify genetic variants that are “associated” with disease [49]. These associated variants are genetic variation that may have an effect on the disease risk of an individual.

While GWAS have been extremely successful in identifying variants involved in disease, the results of GWASs have also raised a host of questions. Even though hundreds of variants have been implicated to be involved in some traits, their total contribution only explains a small fraction of the total genetic contribution which is known from twin studies. For example, the combined contributions of the 50 genes discovered to have an effect on height using GWASs through 2009 with over tens of thousands individuals only account for $\sim 5\%$ of the phenotypic variation which is a far cry from the 80% heritability previously estimated from twin studies[55]. The gap between the known heritability and the total genetic contribution from all variants implicated in genome studies is referred to as “missing heritability”[36].

After the first wave of GWAS results reported in 2007 through 2009, it became very evident that the discovered variants are not going to explain a significant portion of the expected heritability and this observation was widely referred to as the “mystery of missing heritability.” A large number of possible explanations for the “missing heritability” were presented including interactions between variants, interactions between variants and the environments, or rare variants[36]. Missing heritability has very important implications for human health. A key challenge in personalized medicine is how to use an individual’s genomes to predict disease risk. Using only the genetic variants discovered from GWASs will only utilize a fraction of the predictive information which we know is present in the genome. In 2009 and 2010, a pair of papers shook the field by suggesting that the missing heritability wasn’t really “missing”, but actually accounted for in the common variants[55, 42]. More recent GWASes using tens of thousands of individuals have discovered even more variants involved in disease. The results of these studies provide a clearer picture of the genetic architecture of disease and motivate great opportunities for predicting disease risk for an individual using their genetic information. This review traces the history of the GWAS era from the first studies, through the mystery of missing heritability and to the current understanding of what GWAS has discovered.

What is exciting about the area of genetics is that many of these questions and challenges are “quantitative” in nature. Quantitative genetics is a field with a long and rich history dating back to the works of R.A. Fisher, Sewall Wright, and J. B. S. Haldane which are deeply intertwined with the development of modern statistics. With the availability of large scale genetic datasets[8, 1] including virtually all data from published GWASes, the critical challenges involve many computationally intensive data analysis problems. There are great opportunities for contributions to these important challenges from computer scientists.

2 The Relation between Genotypes and Phenotypes

The genomes of any two humans are approximately 99.9% identical and the small amount of differences in the remaining genomic sequence accounts for the full range of phenotypic diversity we observe in the human population. A genetic variant is a position in the human genome where individuals in the population have different genetic content. The most common type of genetic variation is referred to as a single nucleotide polymorphism (SNP). For example, the SNP rs9939609 refers to the position 53820527 on chromosome 16 which is in the FTO gene and was implicated in Type 2 Diabetes in one of the first genome wide studies performed[9]. For this SNP, 45% of the chromosomes in the European population have an “A” in that position while 55% have the “T” in that position[8]. The occurring genomic content (“A” or “T”) is referred to as the “allele” of the variant and the frequency of the rarer allele of the variant (0.45) is referred to as the minor allele frequency (MAF). The less common allele (in this case “A”) is referred to as the minor allele and the more common allele (in this case “T”) is referred to as the major allele. The specific allele present in an individual is referred to as the genotype. Because mutations occur rarely in human history, for the vast majority of SNPs, only two alleles are present in the human population. Since humans are diploid - each individual has two copies of each chromosome - the possible genotypes are “TT”, “AT” and “AA” typically encoded “0”, “1” and “2” corresponding to the number of minor alleles the individual carries.

There are many kinds of genetic variation which are present in addition to SNPs such as single position insertion and deletions, referred to as indels, or even larger variants, referred to as structural variants, encompassing such phenomenon as duplications or deletions of stretches of the genome or even inversions or other rearrangements of the genome. Virtually all GWASes collect SNP information because SNPs are by far the most common form of genetic variation in the genome and are present in virtually every region in the genome as well as amenable to experimental techniques that allow for large scale collection of SNP information[38, 13]. While other types of genetic variation types may be important in disease, since SNPs are so common in the genome, virtually every other type of genetic variant occurs near a SNP that is highly correlated with that variant. Thus genetic studies collecting SNPs can capture the genetic effects of both the SNPs that they collect as well as the other genetic variants which are correlated with these SNPs.

Genetic variation can be approximately viewed as falling into one of two categories: common and rare variation. The minor allele frequency threshold separating common and rare variation is obviously subjective and the threshold is usually defined in the range of 1% to 5% depending on the context. Variants that are more common tend to be more strongly “correlated” to other

variants in the region. The genetics community, for historical reasons, refers to this correlation by “linkage disequilibrium.” Two variants are “correlated” if whether or not an individual carries the minor allele at one variant provides information on carrying the minor allele at another variant. This correlation structure between neighboring variants is a result of human population history and the biological processes that pass variation from one generation to the next. The study of these processes and how they shape genetic variation is the rich field of population genetics[18].

The field of genetics assumes standard a mathematical model for the relationship between genetic variation and traits or phenotypes. This model is called the polygenic model. Despite its simplicity, the model is a reasonable approximation of how genetic variation affects traits and provides a rich starting point for understanding genetic studies. Here we describe a variant of the classic polygenic model.

We assume that our genetic study collects N individuals and the phenotype of individual j is denoted y_j . We assume that a genetic study collects M variants and for simplicity, we assume that all of the variants are independent of each other (not correlated). We denote the frequency of variant i in the population as p_i . We denote the genotype of the i th variant in the j th individual as $g_{ij} \in \{0, 1, 2\}$ which encodes the number of minor alleles for that variant present in the individual. In order to simplify the formulas later in this article, without loss of generality, we normalize the genotype values such that $x_{ij} \in \{\frac{-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{1-2p_i}{\sqrt{2p_i(1-p_i)}}, \frac{2-2p_i}{\sqrt{2p_i(1-p_i)}}\}$ since the mean and variance of the column vector of genotypes (g_i) is $2p_i$ and $2p_i(1-p_i)$ respectively. Because of the normalization, the mean and variance of the vector of genotypes at a specific variant i denoted X_i is 0 and 1 respectively and $X_i^T X_i = N$.

The phenotype can then be modeled using

$$y_j = \mu + \sum_{i=1}^M \beta_i x_{ij} + \epsilon_j \quad (1)$$

where the effect of each variant on the phenotype is β_i , the model mean is μ and ϵ_j is the contribution of the environment on the phenotype is assumed to be normally distributed with variance σ_e^2 , denoted $\epsilon_j \sim \mathcal{N}(0, \sigma_e^2)$. We note that inherent to this model is the “additive” assumption in that the variants all contribute linearly to the phenotype value. More sophisticated models which include non additive effects or gene-by-gene interactions are an active area of research.

If we denote the vector of phenotypes y and vector of effect sizes β , the matrix of normalized genotypes X and the vector of environmental contributions $\mathbf{e}$, then the model for the study population can be denoted

$$y = \mu \mathbf{1} + \sum_i \beta_i X_i + \mathbf{e} \quad (2)$$

where $\mathbf{1}$ is a column vector of 1s, and $\mathbf{e}$ is a random vector drawn from the multivariate normal distribution with mean 0 and covariance matrix $\sigma_e^2 \mathbf{I}$, denoted as $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$.

3 Genome Wide Association Studies

Genome-wide association studies (GWAS) collect disease trait information, referred to as phenotypes, and genetic information, referred to as genotypes, from a set of individuals. The phenotype is either a binary indicator of disease status or a quantitative measure of a disease related trait such as an individual’s cholesterol level. Studies that collect binary trait information are referred to as case/control studies and typically collect an equal number of individual with and without the disease. Studies that collect quantitative measures are often performed on a representative sample of a population, referred to as a population cohort, and collect individuals using a criteria designed to be representative of a larger population (e.g. all individuals who were born in a specific location in a specific year[46]).

GWASes focus on discovering the common variation that is involved in disease traits. Because of the correlation structure of the genome, GWASes only collect a subset of the common variation typically in the range of 500,000 variants. Studies have shown that collecting only this fraction of the common variants “captures” the full set of common variants in the genome. For the vast majority of common variants in the genome, at least one of the 500,000 variants that is collected is correlated with the variant. GWASes typically collect genotype information on these variants in thousands of individuals along with phenotypic information.

The general analysis strategy of GWAS is motivated by the assumptions of the polygenic model (equation (1)). In a GWAS, genotypes and phenotypes are collected from a set of individuals with the goal of discovering the associated variants. Intuitively, a GWAS identifies a variant involved in disease by splitting the set of individuals based on their genotype (“0”, “1” or “2”) and computing the mean of the disease related trait in each group. If the means are significantly different, then this variant is declared associated and maybe involved in the disease. More formally, the analysis of GWAS data in the context of the model in equation (1) corresponds to estimating the vector β from the data and we refer to the estimated vector as $\hat{\beta}$ following the convention that estimates of unknown parameters from data are denoted with the “hat” over the parameter. Since the number of individuals is at least an order of magnitude smaller than the number of variants, it is impossible to simultaneously estimate all of the components of β . Instead, in a typical GWAS, the effect size for each variant is estimated one at a time and a statistical test is performed to determine whether or not the variant has a significant effect on the phenotype. This is done by estimating the maximum likelihood parameters of the following equation

$$y = \mu^* + \beta_k X_k + e^* \quad (3)$$

which results in estimates of $\hat{\mu}$ and $\hat{\beta}_k$ and performs a statistical test to see if the estimated value of $\hat{\beta}_k$ is non-zero. See Box 1 for more details on association statistics. The results of an association study is then the set of significantly associated variants, which we denote using the set A , and their corresponding effect size estimates $\hat{\beta}_i$.

The results of GWASes can be directly utilized for personalized medicine. In personalized medicine, one of the challenges is to identify individuals which have high genetic risk for a particular disease. In our model from equation (1), each individual’s phenotype can be decomposed into a genetic mean ($\mu + \sum_{i=1}^M \beta_i x_{ij}$) and an environmental component (ϵ_j). The genetic mean, which

is a function of the effect sizes and the individual’s genotypes, can be thought of as a measure of the individual’s genetic risk. Thus, inferring this genetic mean is closely related to identifying individuals at risk for a disease and since the environmental contribution has mean 0, predicting the genetic mean and the phenotype are closely related problems.

Knowing nothing about an individual’s genotypes or the effect sizes, the best prediction for an individual’s phenotype would be the prediction of the phenotypic mean of $\hat{\mu}$. The more information we have on an individual’s genotypes and the effects sizes, the more closely our phenotype prediction is to the true phenotype. Using the results of a GWAS and the genotypes of a new individual x^* , we can use the discovered associated loci to make a phenotype prediction, y^* , for the individual using $y^* = \hat{\mu} + \sum_{i \in A} \hat{\beta}_i x_i^*$. As we discuss below, while the prediction of a trait from GWAS is more informative than just using the mean, unfortunately, the predictions are not accurate enough to be clinically useful.

4 What GWAS has discovered and the Mystery of Missing Heritability

In the genetics community, how much genetics influences a trait is quantified using “heritability” which is the proportion of disease phenotypic variance explained by the genetics. The heritability of a trait can be measured using other approaches taking advantage of related individuals. One approach for measuring heritability is taking advantage of twin studies. Twin studies measure the same trait in many pairs of twins. Some of these pairs of twins are monozygotic (MZ) twins, often referred to as maternal twins and some of them pairs as dizygotic (DZ) twins, often referred to as fraternal twins. The difference between MZ twins and DZ twins is that MZ twins have virtually identical genomes, while DZ twins only share about 50% of their genomes. By computing the relative correlation between trait values of MZ twins vs DZ twins, heritability of the trait can be estimated[52]. Intuitively, if the MZ twins within a pair have very similar trait values while DZ twins within a pair have different trait values, then the trait is very heritable. If the difference in trait values with pairs of MZ twins is approximately the same as the difference between values within pairs of DZ twins, then the trait is not very heritable.

In our model of the relationship between genotype and phenotype (equation (2)), the heritability refers to the proportion the term $\sum_i \beta_i X_i$ contributes to the overall phenotypic variance. In our model, the total phenotypic variance $\text{Var}(y)$ can be decomposed into a genetic component and environmental component. The variance corresponding to the environment is σ_e^2 . Since the genotypes are normalized, the phenotypic variance accounted for by each variant is β_i^2 , thus the total genetic variance is $\sum_{i=1}^M \beta_i^2$. The heritability, which is denoted h^2 for historical reasons, is then

$$h^2 = \frac{\sum_{i=1}^M \beta_i^2}{\sum_{i=1}^M \beta_i^2 + \sigma_e^2} \quad (4)$$

Unfortunately, we do not know the true values of β_i or σ_e^2 . The studies using twins have been shown to closely approximate the heritability as defined in equation (4).

GWAS have been tremendously successful in discovering variation involved in traits. The initial studies found a few variants in disease. For example, one of the first GWASes was the Wellcome

Trust Case Control Consortium study which used 3000 healthy individuals and 2000 individuals from each of 7 diseases[9]. They found 24 associations. As samples sizes increased, more discoveries were found particularly because many smaller GWASes were combined to enable a meta-analysis of a larger population. The results of all GWASes are catalogues at the National Human Genome Research Institute (<http://www.genome.gov/gwastudies>) and as of 11/2013, GWASes have identified 11,996 variants associated with 17 disease categories[23].

While the large number of associations discovered can lead to new insights about the genetic basis of common diseases, the vast majority of discovered loci have very small effect sizes. Yet it is well known that genetics plays a large role in disease risk. For example, for many diseases, it is known that parental disease history is a strong predictor of disease risk.

Now let us use the results of GWAS to estimate the heritability. We can also estimate the total phenotypic variance by estimating the variance of our phenotypes directly, $\text{Var}(y)$, which is a reasonable approximation for $\text{Var}(y)$. Let A be the set of associated variants and for these variants, the estimate $\hat{\beta}_i$ is a reasonable estimate for β_i . We can use them to estimate the heritability explained by GWAS which we denote $\hat{h}_G^2$

$$\hat{h}_G^2 = \frac{\sum_{i \in A} \hat{\beta}_i^2}{\text{Var}(y)} \quad (5)$$

We note that the main difference between $\hat{h}_G^2$ and h^2 is that there are only $|A|$ terms in the numerator of $\hat{h}_G^2$ while there are M terms in h^2 . For this reason, $\hat{h}_G^2 < h^2$. This difference is intuitively the difference between what has been discovered in GWAS compared to what is known to compose of the genetic effect.

A landmark survey in 2009 compared the heritability estimates from twin studies to the heritability explained by GWAS results[36]. In this study, they showed that the 18 variants implicated by GWAS in Type 2 Diabetes only explained 6% of the known heritability. Similarly, the 40 variants implicated to be involved in height at that time only explained 5% of the heritability. The large gap between the heritability is referred to as the “missing heritability” and a large amount of effort has gone into finding this missing heritability.

Part of the picture of missing heritability can be explained by analyzing the statistical power of GWASes. Figure 1(a) and 1(b) shows the effect of minor allele frequency and study size on the power of discovering associations (See Box 1 for a discussion and definition of statistical power). As can be shown in the figure, for small minor allele frequencies, even very large studies have very low power to detect associations. A possible explanation for missing heritability is that a very large number of variants with very small effects are present throughout the genome accounting for the remaining heritability and simply could not be discovered by GWAS due to power considerations. If this is the case, as study samples increase, more and more of these variants will be discovered and the amount of heritability explained by the GWAS results will slowly approach the total heritability of the trait. Unfortunately, there is a practical limit to how large GWASes can become due to cost considerations. Even putting cost aside, for some diseases, there are simply not enough individuals with the disease on the planet to perform large enough GWASes to discover all of the variants involved with the disease.

Figure 1: The effect of minor allele frequency and study size on the statistical power of a GWAS. Power is shown for studies of size 1000, 5000, 10,000, 50,000 and 100,000 as a function of minor allele frequency for an effect equal to 10% of a standard deviation of the phenotype (a) and a larger effect equal to 20% of a standard deviation of the phenotype (b).

Without the ability to perform even larger GWASes, it was not clear if we could identify whether there are enough small effect size variants in the genome corresponding to the missing heritability or the missing heritability was due to some other reasons such as interactions between variants, structural variation, rare variants or interactions between genetics and environment.

5 Mixed Models for Population Structure and Missing Heritability

Another insight into missing heritability emerged from what initially seemed like an unrelated development addressing an orthogonal problem in association studies. GWAS statistics (Box 1) make the same assumptions as linear regression which assumes that the phenotype of each individual is independently distributed. Unfortunately, this is not always the case. The reason is due to the discrepancy the statistical model that actually generated the data (equation (3)) and the statistical model that is assumed when performing a GWAS (equation (3)). The term which is missing from the testing model, $\sum_{i \neq k} \beta_i x_{ij}$, is referred to as an unmodeled factor. This unmodeled factor corresponds to the effect of variants in the genome other than the variant which is being tested in the statistical test.

If the values for the unmodeled factor is independently distributed among individuals, then the factor will increase the amount of variance, but not violate the independently distributed assumption of the statistics. The effect of the unmodeled factor is that it will increase the variance estimate of $\sigma_{e^*}^2$ in equation (3) compared to the true environmental variance σ_e^2 in equation (1). However, if the unmodeled factor is not independently distributed, then this will violate the assumptions of

the statistical test in equation (3).

Unfortunately, in association studies, the effect of the rest of the genome on a trait is not independent when individuals who are related are present in the association studies. Consider a pair of siblings who are present in an association study as well as a pair of unrelated individuals. Since siblings share about half of their genome, for half of the genome, they will have identical genotypes. Many of these variants will have an affect on the phenotype. The values of $\sum_{i \neq k} \beta_i x_{ij}$ will be much closer to each other for siblings compared to a pair of unrelated individuals. This applies for more distant relationships as well. This problem is referred to as “population structure” where differing degrees of relatedness between individuals in the GWAS cause an inflation of the values of the association statistics leading to false positives. Many methods for addressing population structure have been presented over the years including genomic control[10] which scales the statistics to avoid inflation, principal component based methods[41] and most recently mixed model methods[27, 25, 33, 61].

The basis of the mixed model approach to population structure is the insight that the proportion of the genome shared corresponds to the expected similarity in the values of the unmodeled factors. In fact, the amount of similarity between the unmodeled factors in association studies will be proportional to the amount of the genome shared between individuals, particularly under some standard assumptions made about the effect sizes of the variants and the assumption that each variant has equal likelihood of being causal. More precisely, the covariance of the unmodeled factors is proportional to the amount of the genome shared. The amount of genome shared is referred to as the “kinship matrix” and since the genotypes are normalized, the kinship is simply $\mathbf{K} = \mathbf{X}\mathbf{X}^T/M$ where $\mathbf{X}$ is the $N \times M$ matrix of the normalized genotypes. We then add a term to the statistical model to capture these unmodeled factors resulting in the statistical model

$$y = \mu \mathbf{1} + \beta_k x_k + \mathbf{u} + \mathbf{e} \quad (6)$$

where x_k is a column vector of normalized genotypes for variant k , $\mathbf{e} \sim \mathcal{N}(0, \sigma_e^2 \mathbf{I})$ and $\mathbf{u} \sim \mathcal{N}(0, \sigma_g^2 \mathbf{K})$ represents the contributions of the unmodeled factors. When performing an association, mixed model methods estimate the maximum likelihood for parameters μ , β_k , σ_g^2 and σ_e^2 using the likelihood

$$L(N, y, x_k, \mu, \beta_k, \sigma_e^2, \sigma_g^2, \mathbf{K}) = (2\pi)^{-\frac{N}{2}} |\sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I}|^{-\frac{1}{2}} e^{-\frac{1}{2}(y - \mu - \beta_k x_k)^T (\sigma_g^2 \mathbf{K} + \sigma_e^2 \mathbf{I})^{-1} (y - \mu - \beta_k x_k)} \quad (7)$$

and compare this maximum likelihood to the maximum likelihood when β_k is restricted to 0. By comparing these likelihoods, mixed model methods can obtain a significance for the association at variant k correcting for population structure. Mixed models were shown to perform well for a wide variety of population structure scenarios and correct for the structure in studies involving closely related individuals[28] to studies with more distant relationships[25].

A major development related to the mystery of missing heritability was when the connection was made between the mixed model estimates of σ_g^2 and σ_e^2 . In a seminal paper, it was pointed out that these estimates from GWAS data for a population cohort can be used to estimate the heritability[55]. We refer to this estimate as h_M^2 where

$$h_M^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} \quad (8)$$

This method was applied to estimate the heritability of height from the full set of GWAS data and obtained an estimate of 0.45 which is dramatically higher than the estimate from the results of the association studies (h_G^2) which was 0.05. This study suggests that the common variants capture a much larger portion of the heritability than just the associated variants, which provides strong support that the main reason cause of missing heritability is simply many variants with very small effects spread throughout the genome.

Around the same time, another study showed that if the criterion for including variants in a prediction model is not as stringent as genome-wide associated, but reducing the significance threshold, the predictions of the model are more accurate[42]. In this study, not only significant associated variants, but variants that had weaker effects were included in the model and the resulting predictive model showed better performance when it was evaluated using cross-validation. This further suggests that many weakly associated variants are contributing the missing heritability. This concept of including more variants in the predictive model is analogous to the tradeoff related to prediction accuracy and overfitting when including more features in a machine learning classifier.

While mixed model approaches are a step toward understanding the mystery of missing heritability, there are still many unanswered questions. There is still a significant discrepancy between estimates from related individuals and mixed model estimates. For example, height is estimated to have a heritability of 0.8 while mixed models only explain 0.45. One theory for the remaining heritability argues that the remaining portion of the missing heritability can be explained by rare variants having small effects which are poorly correlated with the common variants used to compute kinship matrices[55]. Other theories postulate that interactions between variants account for the remaining missing heritability[62, 6]. Additional questions are related to the fact that the interpretation of the mixed model estimate of heritability is only correct under the assumption that only the causal variants are used for estimating the kinship matrices[62]. Unfortunately, which variants are causal is unknown and various approaches have been proposed to address this issue[12].

The developments in mixed models provide interesting opportunities for phenotype prediction which is a problem with a rich history in genetics, particularly in the literature on the best linear unbiased predictor (BLUP)[22, 45, 39]. Consider the scenario where we have a population of individuals with known phenotypes y and genotypes X . Given a new individuals genome x^* , we can predict the individual’s phenotype y^* using mixed models. In order to make predictions, we first estimate the parameters of the mixed model σ_g^2 and σ_e^2 . We then compute the kinship values between the new individual and the set of individuals with known genotypes and phenotypes. We can then treat the new individual’s phenotype as missing and compute the most likely value for this phenotype value given the mixed model likelihood value.

6 Conclusion: The Future of Phenotype Prediction

Phenotype prediction from genetic information is currently an active area of research. Clearly phenotype prediction using only the results of GWAS ignores the information from the polygenic score obtained from mixed models and only leverages the information from the portion of the heritability that is accounted for in GWASes. However, using only the polygenic score from mixed models ignores variants that are clearly involved in the trait. Several strategies are utilizing both

types of information by first utilizing the associated SNPs and then using a polygenic score from the rest of the genome[43, 60]. However, even these combined strategies seem to be missing out on information because variants that are just below the significance threshold have a higher chance of having an effect on the phenotype than other variants, yet all variants are grouped together when estimating the kinship matrix and the polygenic score from variants that are not associated. This problem is closely related to the standard classification problem widely investigated in the machine learning community.

Phenotype and genotype data for massive amounts of individuals is widely available. The actual disease study datasets are available through a National Center for Biotechnology Information database called the database of Genotypes and Phenotypes (dbGaP) available at <http://www.ncbi.nlm.nih.gov/gap>. Virtually all US government funded GWASes are required to submit their data into the dbGaP database. A similar project, the European Genome-Phenome Archive (EGA) hosted by European Bioinformatics Institute (EBI) is another repository of genome wide association study data available at <https://www.ebi.ac.uk/ega/>. For both of these databases, investigators must apply for the data in order to guarantee that they comply with certain restrictions on the use of the data due to the inherent privacy and ethical concerns. Literally hundreds of large datasets are available through both of these resources.

In addition, many computational problems in genetics can be investigated using carefully crafted simulations which closely approximate collected data. These simulations typically use real data as a starting point from publicly available reference datasets such as the HapMap data[8] or the 1000 Genomes project[1]. These reference datasets can be combined with a simulator such as HAPGEN[50] to generate realistic association study simulations.

Missing heritability and phenotype prediction is only one of the interesting computational questions in human genetics which have opportunities for computer scientists and where computer scientists have made contributions. Other problems in genetics where computer scientists have made contributions are listed in Box 2.

The combination of the impact of addressing the challenges in human genetics will have on human health, combined with the quantitative nature of these challenges as well as the readily available datasets provide tremendous opportunities for important contributions from Computer Scientists.

7 Box 1: GWAS Statistics and Power

In GWAS, the effect of each variant on the trait is examined independently of the other variants. When analyzing variant k , the following model for the effect of variant k on the phenotype is utilized

$$y_j = \mu^* + \beta_k x_{kj} + \epsilon_j^* \quad (9)$$

and in vector notation

$$y = \mu^* \mathbf{1} + \beta_k x_k + \mathbf{e}^* \quad (10)$$

where x_k is a column vector of normalized genotypes for variant k and $\mathbf{e}^* \sim \mathcal{N}(0, \sigma_{e^*}^2 \mathbf{I})$. We note that the equations (9) and (1) differ by the omission of the terms $\sum_{i \neq k} \beta_i x_{ij}$ which results in the

values of these terms being absorbed in the mean and residual which is why we use the notation μ^* and ϵ_j^* instead of μ and ϵ_j . We describe the implications of this omission in more detail below.

Using equation (9), we can use the observed data to obtain an estimate of β_k . This reduces to a simple regression problem where the resulting estimates are $\hat{\mu} = \frac{1}{N} \mathbf{1}^T y$, $\hat{\beta}_k = (x_k^T x_k)^{-1} x_k^T y = \frac{x_k^T y}{N}$ since x_k is normalized so $x_k^T x_k = N$. The estimated residuals $\hat{\mathbf{e}} = y - \hat{\mu} \mathbf{1} - \hat{\beta}_k x_k$ can be used to estimate the standard error $\hat{\sigma} = \sqrt{\frac{\hat{\mathbf{e}}^T \hat{\mathbf{e}}}{n-2}}$. Since the studies are large, the association statistic

$$S_k = \frac{\hat{\beta}_k}{\hat{\sigma}} \sqrt{N} \sim \mathcal{N}\left(\frac{\beta_k}{\sigma_{e^*}} \sqrt{N}, 1\right) \quad (11)$$

will approximately follow the standard normal distribution under the null hypothesis of no association and can be used to determine whether or not the statistic is significant.

Since GWASes collect many markers, the significance threshold is adjusted for multiple hypothesis testing. The community has settled on using the significance threshold of $\alpha_s = 5 \times 10^{-8}$ as the genome-wide significance threshold which takes into account the large number of SNPs in the human genome. Since our statistics, S_k , are normally distributed, a GWAS simply computes S_k for each variant and then checks to see if $\Phi(S_k) < \alpha_s/2$ or $\Phi(S_k) > 1 - \alpha_s/2$ in which case the variant is associated. $\Phi(x)$ computes the cumulative standard normal distribution. Figure 2 shows an example of checking for significance of a variant using the normal distribution.

The statistical power measures the probability of detecting an association under the assumption that an association is present with a certain effect size. Intuitively, the power measures the probability that the truly associated variants will be discovered. Since statistical power depends on both the effect size and the number of individuals in the study, statistical power can be used to guide the choice of study size as well as provide expectations on what effect sizes can and can not be discovered in association studies.

Using the distributions above, we can also estimate the statistical power of an association study. We assume that the effect size at variant i is β_i and the residual variance from equation (9) is $\sigma_{e^*}^2$. Since we know that the statistic S_k follows the distribution in equation (11), the question is how often this statistic is significant (i.e. $\Phi(S_k) < \alpha_s/2$ or $\Phi(S_k) > 1 - \alpha_s/2$). This probability can be estimated using the following

$$P(\alpha_s, \beta, \sigma, N) = \Phi(-\Phi^{-1}(\alpha_s/2) + \frac{\beta}{\sigma} \sqrt{N}) + 1 - \Phi(\Phi^{-1}(\alpha_s/2) + \frac{\beta}{\sigma} \sqrt{N}) \quad (12)$$

Note that the power depends on what is referred to as the non-centrality parameter (in our case $\frac{\beta}{\sigma} \sqrt{N}$) which is the mean of the distribution of the statistic under the assumption of the genetic effect. A visualization of estimating the power is shown in Figure 3.

8 Box 2: Computational Problems in Genetics

- how to efficiently estimate mixed model parameters, which has been an active area of research for several years[27, 25, 33, 61]

Figure 2: (Figure should be located in Box 1) Significance testing in association studies. The null distribution is shown which is the standard normal distribution and the expected distribution of the association statistics under the assumption that the effect size is 0. For each variant, that association statistic in equation (11) is computed and its significance is evaluated using the null distribution. If the statistic falls in the significance region of the distribution, the variant is declared associated. In this example, S_1 is not significant and S_2 and S_3 are significant. The exact location of the threshold is defined as the location on the x axis where the tail probability area equals the significance threshold (α_s). This is denoted using the quantile of the standard normal $\pm\Phi^{-1}(\alpha_s/2)$.

Figure 3: (Figure should be located in Box 2) Power of association studies. The power is defined by considering the expected distribution of the association statistics assuming a specific effect size which is referred to as the alternative distribution. This effect size as well as the number of individuals in the study define the non-centrality parameter (NCP) of the alternative distribution. The area of the alternative (noncentral) distribution outside the significance threshold defined by the null distribution is the probability that an observation under the assumption of the specific effect size will be declared significant. This probability is referred to as the power.

- how to efficiently identify pairs of segments in individuals which were inherited from a recent common ancestry[7, 14, 19]
- predicting haplotypes or the sequence of alleles on a chromosome from the genotype information which mixes the two chromosomes[15, 3, 16, 2, 20, 56]
- identifying the population origin of each region of an individual’s genome for individuals who are admixed or a mixture of multiple ancestral populations[51, 47].
- identifying the geographical origin of an individual[40, 57, 4]
- identifying pairs of genetic variants which have a larger effect on the trait than each individually[59, 58]
- inferring the genetic relationships between individuals[30, 32, 21]
- inferring the genetic history of a region of the genome in the population which is referred to as an ancestral recombination graph[54]
- predicting missing genotype data in an association study[37, 26, 24]
- correcting for multiple testing in genome-wide association studies[29, 17]
- efficiently computing association statistics[31]

References

- [1] Goncalo R. Abecasis, Adam Auton, Lisa D. Brooks, Mark A. DePristo, Richard M. Durbin, Robert E. Handsaker, Hyun Min Kang, Gabor T. Marth, and Gil A. McVean. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56–65, 11 2012.
- [2] Derek Aguiar and Sorin Istrail. Haplotype assembly in polyploid genomes and identical by descent shared tracts. *Bioinformatics*, 29(13):i352–i360, 7 2013.
- [3] Vineet Bafna, Dan Gusfield, Sridhar Hannenhalli, and Shibu Yooseph. A note on efficient computation of haplotypes via perfect phylogeny. *J Comput Biol*, 11(5):858–66, 2004.
- [4] Yael Baran, Inés Quintela, Angel Carracedo, Bogdan Pasaniuc, and Eran Halperin. Enhanced localization of genetic samples through linkage-disequilibrium correction. *Am J Hum Genet*, 5 2013.
- [5] Sonja I. Berndt, Stefan Gustafsson, Reedik Mägi, Andrea Ganna, Eleanor Wheeler, Mary F. Feitosa, Anne E. Justice, Keri L. Monda, Damien C. Croteau-Chonka, Felix R. Day, Tõnu Esko, Tove Fall, Teresa Ferreira, Davide Gentilini, Anne U. Jackson, Jian’an Luan, Joshua C. Randall, Sailaja Vedantam, Cristen J. Willer, Thomas W. Winkler, Andrew R. Wood, Tsegaselassie Workalemahu, Yi-Juan J. Hu, Sang Hong Lee, Liming Liang, Dan-Yu Y. Lin, Josine L.

Min, Benjamin M. Neale, Gudmar Thorleifsson, Jian Yang, Eva Albrecht, Najaf Amin, Jennifer L. Bragg-Gresham, Gemma Cadby, Martin den Heijer, Niina Eklund, Krista Fischer, Anuj Goel, Jouke-Jan J. Hottenga, Jennifer E. Huffman, Ivonne Jarick, Åsa Johansson, Toby Johnson, Stavroula Kanoni, Marcus E. Kleber, Inke R. König, Kati Kristiansson, Zoltán Kutalik, Claudia Lamina, Cecile Lecoeur, Guo Li, Massimo Mangino, Wendy L. McArdle, Carolina Medina-Gomez, Martina Müller-Nurasyid, Julius S. Ngwa, Ilja M. Nolte, Lavinia Paternoster, Sonali Pechlivanis, Markus Perola, Marjolein J. Peters, Michael Preuss, Lynda M. Rose, Jianxin Shi, Dmitry Shungin, Albert Vernon Smith, Rona J. Strawbridge, Ida Surakka, Alexander Teumer, Mieke D. Trip, Jonathan Tyrer, Jana V. Van Vliet-Ostaptchouk, Liesbeth Vandenput, Lindsay L. Waite, Jing Hua Zhao, Devin Absher, Folkert W. Asselbergs, Mustafa Atalay, Antony P. Attwood, Anthony J. Balmforth, Hanneke Basart, John Beilby, Lori L. Bonnycastle, Paolo Brambilla, Marcel Bruinenberg, Harry Campbell, Daniel I. Chasman, Peter S. Chines, Francis S. Collins, John M. Connell, William O. Cookson, Ulf de Faire, Femmie de Vegt, Mariano Dei, Maria Dimitriou, Sarah Edkins, Karol Estrada, David M. Evans, Martin Farrall, Marco M. Ferrario, Jean Ferrières, Lude Franke, Francesca Frau, Pablo V. Gejman, Harald Grallert, Henrik Grönberg, Vilmundur Gudnason, Alistair S. Hall, Per Hall, Anna-Liisa L. Hartikainen, Caroline Hayward, Nancy L. Heard-Costa, Andrew C. Heath, Johannes Hebebrand, Georg Homuth, Frank B. Hu, Sarah E. Hunt, Elina Hyppönen, Carlos Iribarren, Kevin B. Jacobs, John-Olov O. Jansson, Antti Jula, Mika Kähönen, Sekar Kathiresan, Frank Kee, Kay-Tee T. Khaw, Mika Kivimäki, Wolfgang Koenig, Aldi T. Kraja, Meena Kumari, Kari Kuulasmaa, Johanna Kuusisto, Jaana H. Laitinen, Timo A. Lakka, Claudia Langenberg, Lenore J. Launer, Lars Lind, Jaana Lindström, Jianjun Liu, Antonio Liuzzi, Marja-Liisa L. Lokki, Mattias Lorentzon, Pamela A. Madden, Patrik K. Magnusson, Paolo Manunta, Diana Marek, Winfried März, Irene Mateo Leach, Barbara McKnight, Sarah E. Medland, Evelin Mihailov, Lili Milani, Grant W. Montgomery, Vincent Mooser, Thomas W. Mühleisen, Patricia B. Munroe, Arthur W. Musk, Narisu Narisu, Gerjan Navis, George Nicholson, Ellen A. Nohr, Ken K. Ong, Ben A. Oostra, Colin N. A. Palmer, Aarno Palotie, John F. Peden, Nancy Pedersen, Annette Peters, Ozren Polasek, Anneli Pouta, Peter P. Pramstaller, Inga Prokopenko, Carolin Pütter, Aparna Radhakrishnan, Olli Raitakari, Augusto Rendon, Fernando Rivadeneira, Igor Rudan, Timo E. Saaristo, Jennifer G. Sambrook, Alan R. Sanders, Serena Sanna, Jouko Saramies, Sabine Schipf, Stefan Schreiber, Heribert Schunkert, So-Youn Y. Shin, Stefano Signorini, Juha Sinisalo, Boris Skrobek, Nicole Soranzo, Alena Stančáková, Klaus Stark, Jonathan C. Stephens, Kathleen Stirrups, Ronald P. Stolk, Michael Stumvoll, Amy J. Swift, Eirini V. Theodoraki, Barbara Thorand, David-Alexandre A. Tregouet, Elena Tremoli, Melanie M. Van der Klauw, Joyce B. J. van Meurs, Sita H. Vermeulen, Jorma Viikari, Jarmo Virtamo, Veronique Vitart, Gérard Waeber, Zhaoming Wang, Elisabeth Widén, Sarah H. Wild, Gonneke Willemsen, Bernhard R. Winkelmann, Jacqueline C. M. Witteman, Bruce H. R. Wolffenbuttel, Andrew Wong, Alan F. Wright, M. Carola Zillikens, Philippe Amouyel, Bernhard O. Boehm, Eric Boerwinkle, Dorret I. Boomsma, Mark J. Caulfield, Stephen J. Chanock, L. Adrienne Cupples, Daniele Cusi, George V. Dedoussis, Jeanette Erdmann, Johan G. Eriksson, Paul W. Franks, Philippe Froguel, Christian Gieger, Ulf Gyllensten, Anders Hamsten, Tamara B. Harris, Christian Hengstenberg, Andrew A. Hicks, Aroon Hingorani,

- Anke Hinney, Albert Hofman, Kees G. Hovingh, Kristian Hveem, Thomas Illig, Marjo-Riitta R. Jarvelin, Karl-Heinz H. Jöckel, Sirkka M. Keinänen-Kiukaanniemi, Lambertus A. Kiemeny, Diana Kuh, Markku Laakso, Terho Lehtimäki, Douglas F. Levinson, Nicholas G. Martin, Andres Metspalu, Andrew D. Morris, Markku S. Nieminen, Inger Njolstad, Claes Ohlsson, Albertine J. Oldehinkel, Willem H. Ouwehand, Lyle J. Palmer, Brenda Penninx, Chris Power, Michael A. Province, Bruce M. Psaty, Lu Qi, Rainer Rauramaa, Paul M. Ridker, Samuli Ripatti, Veikko Salomaa, Nilesh J. Samani, Harold Snieder, Thorkild I. A. Sorensen, Timothy D. Spector, Kari Stefansson, Anke Tönjes, Jaakko Tuomilehto, André G. Uitterlinden, Matti Uusitupa, Pim van der Harst, Peter Vollenweider, Henri Wallaschofski, Nicholas J. Wareham, Hugh Watkins, H-Erich E. Wichmann, James F. Wilson, Goncalo R. Abecasis, Themistocles L. Assimes, Inês Barroso, Michael Boehnke, Ingrid B. Borecki, Panos Deloukas, Caroline S. Fox, Timothy Frayling, Leif C. Groop, Talin Haritunian, Iris M. Heid, David Hunter, Robert C. Kaplan, Fredrik Karpe, Miriam F. Moffatt, Karen L. Mohlke, Jeffrey R. O’Connell, Yudi Pawitan, Eric E. Schadt, David Schlessinger, Valgerdur Steinthorsdottir, David P. Strachan, Unnur Thorsteinsdottir, Cornelia M. van Duijn, Peter M. Visscher, Anna Maria Di Blasio, Joel N. Hirschhorn, Cecilia M. Lindgren, Andrew P. Morris, David Meyre, André Scherag, Mark I. McCarthy, Elizabeth K. Speliotes, Kari E. North, Ruth J. F. Loos, and Erik Ingelsson. Genome-wide meta-analysis identifies 11 new loci for anthropometric traits and provides insights into genetic architecture. *Nat Genet*, 45(5):501–12, 5 2013.
- [6] Joshua S. Bloom, Ian M. Ehrenreich, Wesley T. Loo, Thúy-Lan Võ L. Lite, and Leonid Kruglyak. Finding the sources of missing heritability in a yeast cross. *Nature*, 494(7436):234–7, 2 2013.
- [7] Brian L. Browning and Sharon R. Browning. Improving the accuracy and efficiency of identity by descent detection in population data. *Genetics*, 194(2):459–71, 3 2013.
- [8] The International HapMap Consortium and The International HapMap Consortium. A haplotype map of the human genome. *Nature*, 437(7063):1299, 10 2005.
- [9] Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661–78, 6 2007.
- [10] Bernie Devlin and Kathryn Roeder. Genomic control for association studies. *Biometrics*, 55(4):997–1004, 12 1999.
- [11] Ronald Aylmer Fisher et al. The correlation between relatives on the supposition of mendelian inheritance. *Transactions of the Royal Society of Edinburgh*, 52:399–433, 1918.
- [12] David Golan and Saharon Rosset. Accurate estimation of heritability in genome wide studies using random effects models. *Bioinformatics*, 27(13):i317–23, 7 2011.
- [13] Kevin L. Gunderson, Frank J. Steemers, Grace Lee, Leo G. Mendoza, and Mark S. Chee. A genome-wide scalable SNP genotyping assay using microarray technology. *Nat Genet*, 37(5):549–54, 5 2005.

- [14] Alexander Gusev, Jennifer K. Lowe, Markus Stoffel, Mark J. Daly, David Altshuler, Jan L. Breslow, Jeffrey M. Friedman, and Itsik Pe'er. Whole population, genome-wide mapping of hidden relatedness. *Genome Res*, 19(2):318–26, 2 2009.
- [15] Dan Gusfield. Haplotyping as perfect phylogeny: Conceptual framework and efficient solutions. In *Proceedings of the Sixth Annual International Conference on Computational Biology, RECOMB '02*, pages 166–175, New York, NY, USA, 2002. ACM.
- [16] Eran Halperin and Eleazar Eskin. Haplotype reconstruction from genotype data using imperfect phylogeny. *Bioinformatics*, 20(12):1842–9, 8 2004.
- [17] Buhm Han, Hyun Min Kang, and Eleazar Eskin. Rapid and accurate multiple testing correction and power estimation for millions of correlated markers. *PLoS Genet*, 5(4):e1000456, 4 2009.
- [18] D. L. Hartl and A. G. Clark. Sunderland, MA: Sinauer Associates, 2007.
- [19] Dan He. Ibd-groupon: an efficient method for detecting group-wise identity-by-descent regions simultaneously in multiple individuals based on pairwise ibd relationships. *Bioinformatics*, 29(13):i162–i170, 7 2013.
- [20] Dan He, Buhm Han, and Eleazar Eskin. Hap-seq: An optimal algorithm for haplotype phasing with imputation using sequencing data. *J Comput Biol*, 20(2):80–92, 2 2013.
- [21] Dan He, Zhanyong Wang, Buhm Han, Laxmi Parida, and Eleazar Eskin. Iped: Inheritance path-based pedigree reconstruction algorithm using genotype data. *J Comput Biol*, 20(10):780–91, 10 2013.
- [22] C. R. Henderson. Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31(2):423–47, 6 1975.
- [23] Lucia A. Hindorff, Praveen Sethupathy, Heather A. Junkins, Erin M. Ramos, Jayashri P. Mehta, Francis S. Collins, and Teri A. Manolio. Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci U S A*, 106(23):9362–7, 6 2009.
- [24] Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini, and Gonçalo R. Abecasis. Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nat Genet*, 44(8):955–9, 2012.
- [25] Hyun Min Kang, Jae Hoon Sul, Susan K. Service, Noah A. Zaitlen, Sit-Yee Y. Kong, Nelson B. Freimer, Chiara Sabatti, and Eleazar Eskin. Variance component model to account for sample structure in genome-wide association studies. *Nat Genet*, 42(4):348–54, 4 2010.
- [26] Hyun Min Kang, Noah A. Zaitlen, and Eleazar Eskin. Eminim: an adaptive and memory-efficient algorithm for genotype imputation. *J Comput Biol*, 17(3):547–60, 3 2010.

- [27] Hyun Min Kang, Noah A. Zaitlen, Claire M. Wade, Andrew Kirby, David Heckerman, Mark J. Daly, and Eleazar Eskin. Efficient control of population structure in model organism association mapping. *Genetics*, 178(3):1709–23, 3 2008.
- [28] Eimear E. Kenny, Minseung Kim, Alexander Gusev, Jennifer K. Lowe, Jacqueline Salit, J. Gustav Smith, Sirisha Kovvali, Hyun Min Kang, Christopher Newton-Cheh, Mark J. Daly, Markus Stoffel, David M. Altshuler, Jeffrey M. Friedman, Eleazar Eskin, Jan L. Breslow, and Itsik Pe’er. Increased power of mixed models facilitates association mapping of 10 loci for metabolic traits in an isolated population. *Hum Mol Genet*, 20(4):827–39, 12 2010.
- [29] Gad Kimmel, Michael I. Jordan, Eran Halperin, Ron Shamir, and Richard M. Karp. A randomization test for controlling population stratification in whole-genome association studies. *Am J Hum Genet*, 81(5):895–905, 11 2007.
- [30] Bonnie Kirkpatrick, Shuai Li, Richard Karp, and Eran Halperin. Pedigree reconstruction using identity by descent. In Vineet Bafna and S. Sahinalp, editors, *Research in Computational Molecular Biology*, volume 6577 of *Lecture Notes in Computer Science*, pages 136–152. Electrical Engineering and Computer Sciences, University of California, Berkeley and International Computer Science Institute, Berkeley, USA, Springer Berlin / Heidelberg, 2011.
- [31] Emrah Kostem and Eleazar Eskin. Efficiently identifying significant associations in genome-wide association studies. In *Research in Computational Molecular Biology*, pages 118–131. University of California, Springer Berlin Heidelberg, 2013.
- [32] Sofia Kyriazopoulou-Panagiotopoulou, Dorna Kashef Haghighi, Sarah J. Aerni, Andreas Sundquist, Sivan Bercovici, and Serafim Batzoglou. Reconstruction of genealogical relationships with applications to phase III of hapmap. *Bioinformatics*, 27(13):i333–41, 7 2011.
- [33] Christoph Lippert, Jennifer Listgarten, Ying Liu, Carl M. Kadie, Robert I. Davidson, and David Heckerman. Fast linear mixed models for genome-wide association studies. *Nat Methods*, 8(10):833–5, 2011.
- [34] Stuart Macgregor, Belinda K. Cornes, Nicholas G. Martin, and Peter M. Visscher. Bias, precision and heritability of self-reported and clinically measured height in australian twins. *Hum Genet*, 120(4):571–80, 11 2006.
- [35] Teri A. Manolio, Lisa D. Brooks, and Francis S. Collins. A hapmap harvest of insights into the genetics of common disease. *J Clin Invest*, 118(5):1590–605, 5 2008.
- [36] Teri A. Manolio, Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, Erin M. Ramos, Lon R. Cardon, Aravinda Chakravarti, Judy H. Cho, Alan E. Guttmacher, Augustine Kong, Leonid Kruglyak, Elaine Mardis, Charles N. Rotimi, Montgomery Slatkin, David Valle, Alice S. Whittemore, Michael Boehnke, Andrew G. Clark, Evan E. Eichler, Greg Gibson, Jonathan L. Haines, Trudy F. C. Mackay, Steven A. McCarroll, and Peter M. Visscher. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–53, 10 2009.

- [37] Jonathan Marchini, Bryan Howie, Simon Myers, Gil McVean, and Peter Donnelly. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet*, 39(7):906–13, 7 2007.
- [38] Hajime Matsuzaki, Shoulian Dong, Halina Loi, Xiaojun Di, Guoying Liu, Earl Hubbell, Jane Law, Tam Berntsen, Monica Chadha, Henry Hui, Geoffrey Yang, Giulia C. Kennedy, Teresa A. Webster, Simon Cawley, P. Sean Walsh, Keith W. Jones, Stephen P. A. Fodor, and Rui Mei. Genotyping over 100,000 SNPs on a pair of oligonucleotide arrays. *Nat Methods*, 1(2):109–11, 11 2004.
- [39] T. H. Meuwissen, B. J. Hayes, and M. E. Goddard. Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157(4):1819–29, 4 2001.
- [40] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R. Boyko, Adam Auton, Amit Indap, Karen S. King, Sven Bergmann, Matthew R. Nelson, Matthew Stephens, and Carlos D. Bustamante. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 11 2008.
- [41] Alkes L. Price, Nick J. Patterson, Robert M. Plenge, Michael E. Weinblatt, Nancy A. Shadick, and David Reich. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet*, 38(8):904–9, 8 2006.
- [42] Shaun M. Purcell, Naomi R. Wray, Jennifer L. Stone, Peter M. Visscher, Michael C. O’Donovan, Patrick F. Sullivan, and Pamela Sklar. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460(7256):748–52, 8 2009.
- [43] Barbara Rakitsch, Christoph Lippert, Oliver Stegle, and Karsten Borgwardt. A lasso multi-marker mixed model for association mapping with population structure correction. *Bioinformatics*, 29(2):206–14, 11 2012.
- [44] Joshua C. Randall, Thomas W. Winkler, Zoltán Kutalik, Sonja I. Berndt, Anne U. Jackson, Keri L. Monda, Tuomas O. Kilpeläinen, Tõnu Esko, Reedik Mägi, Shengxu Li, Tsegaselassie Workalemahu, Mary F. Feitosa, Damien C. Croteau-Chonka, Felix R. Day, Tove Fall, Teresa Ferreira, Stefan Gustafsson, Adam E. Locke, Iain Mathieson, Andre Scherag, Sailaja Vedantam, Andrew R. Wood, Liming Liang, Valgerdur Steinthorsdottir, Gudmar Thorleifsson, Emmanouil T. Dermitzakis, Antigone S. Dimas, Fredrik Karpe, Josine L. Min, George Nicholson, Deborah J. Clegg, Thomas Person, Jon P. Krohn, Sabrina Bauer, Christa Buechler, Kristina Eisinger, Amélie Bonnefond, Philippe Froguel, Jouke-Jan J. Hottenga, Inga Prokopenko, Lindsay L. Waite, Tamara B. Harris, Albert Vernon Smith, Alan R. Shuldiner, Wendy L. McArdle, Mark J. Caulfield, Patricia B. Munroe, Henrik Grönberg, Yii-Der Ida D. Chen, Guo Li, Jacques S. Beckmann, Toby Johnson, Unnur Thorsteinsdottir, Maris Teder-Laving, Kay-Tee T. Khaw, Nicholas J. Wareham, Jing Hua Zhao, Najaf Amin, Ben A. Oostra, Aldi T. Kraja, Michael A. Province, L. Adrienne Cupples, Nancy L. Heard-Costa, Jaakko Kaprio, Samuli Ripatti, Ida Surakka, Francis S. Collins, Jouko Saramies, Jaakko Tuomilehto, Antti Jula, Veikko Salomaa, Jeanette Erdmann, Christian Hengstenberg, Christina Loley, Heribert Schunkert,

Claudia Lamina, H. Erich Wichmann, Eva Albrecht, Christian Gieger, Andrew A. Hicks, Asa Johansson, Peter P. Pramstaller, Sekar Kathiresan, Elizabeth K. Speliotes, Brenda Peninx, Anna-Liisa L. Hartikainen, Marjo-Riitta R. Jarvelin, Ulf Gyllensten, Dorret I. Boomsma, Harry Campbell, James F. Wilson, Stephen J. Chanock, Martin Farrall, Anuj Goel, Carolina Medina-Gomez, Fernando Rivadeneira, Karol Estrada, André G. Uitterlinden, Albert Hofman, M. Carola Zillikens, Martin den Heijer, Lambertus A. Kiemeney, Andrea Maschio, Per Hall, Jonathan Tyrer, Alexander Teumer, Henry Völzke, Peter Kovacs, Anke Tönjes, Massimo Mangino, Tim D. Spector, Caroline Hayward, Igor Rudan, Alistair S. Hall, Nilesh J. Samani, Antony Paul Attwood, Jennifer G. Sambrook, Joseph Hung, Lyle J. Palmer, Marja-Liisa L. Lokki, Juha Sinisalo, Gabrielle Boucher, Heikki Huikuri, Mattias Lorentzon, Claes Ohlsson, Niina Eklund, Johan G. Eriksson, Cristina Barlassina, Carlo Rivolta, Ilja M. Nolte, Harold Snieder, Melanie M. Van der Klauw, Jana V. Van Vliet-Ostaptchouk, Pablo V. Gejman, Jianxin Shi, Kevin B. Jacobs, Zhaoming Wang, Stephan J. L. Bakker, Irene Mateo Leach, Gerjan Navis, Pim van der Harst, Nicholas G. Martin, Sarah E. Medland, Grant W. Montgomery, Jian Yang, Daniel I. Chasman, Paul M. Ridker, Lynda M. Rose, Terho Lehtimäki, Olli Raitakari, Devin Absher, Carlos Iribarren, Hanneke Basart, Kees G. Hovingh, Elina Hyppönen, Chris Power, Denise Anderson, John P. Beilby, Jennie Hui, Jennifer Jolley, Hendrik Sager, Stefan R. Bornstein, Peter E. H. Schwarz, Kati Kristiansson, Markus Perola, Jaana Lindström, Amy J. Swift, Matti Uusitupa, Mustafa Atalay, Timo A. Lakka, Rainer Rauramaa, Jennifer L. Bolton, Gerry Fowkes, Ross M. Fraser, Jackie F. Price, Krista Fischer, Kaarel Krjutskov, Andres Metspalu, Evelin Mihailov, Claudia Langenberg, Jian'an Luan, Ken K. Ong, Peter S. Chines, Sirkka M. Keinänen-Kiukaanniemi, Timo E. Saaristo, Sarah Edkins, Paul W. Franks, Göran Hallmans, Dmitry Shungin, Andrew David Morris, Colin N. A. Palmer, Raimund Erbel, Susanne Moebus, Markus M. Nöthen, Sonali Pechlivanis, Kristian Hveem, Narisu Narisu, Anders Hamsten, Steve E. Humphries, Rona J. Strawbridge, Elena Tremoli, Harald Grallert, Barbara Thorand, Thomas Illig, Wolfgang Koenig, Martina Müller-Nurasyid, Annette Peters, Bernhard O. Boehm, Marcus E. Kleber, Winfried März, Bernhard R. Winkelmann, Johanna Kuusisto, Markku Laakso, Dominique Arveiler, Giancarlo Cesana, Kari Kuulasmaa, Jarmo Virtamo, John W. G. Yarnell, Diana Kuh, Andrew Wong, Lars Lind, Ulf de Faire, Bruna Gigante, Patrik K. E. Magnusson, Nancy L. Pedersen, George Dedoussis, Maria Dimitriou, Genovefa Kolovou, Stavroula Kanoni, Kathleen Stirrups, Lori L. Bonnycastle, Inger Njolstad, Tom Wilsgaard, Andrea Ganna, Emil Rehnberg, Aroon Hingorani, Mika Kivimäki, Meena Kumari, Themistocles L. Assimes, Inês Barroso, Michael Boehnke, Ingrid B. Borecki, Panos Deloukas, Caroline S. Fox, Timothy Frayling, Leif C. Groop, Talin Haritunians, David Hunter, Erik Ingelsson, Robert Kaplan, Karen L. Mohlke, Jeffrey R. O'Connell, David Schlessinger, David P. Strachan, Kari Stefansson, Cornelia M. van Duijn, Gonçalo R. Abecasis, Mark I. McCarthy, Joel N. Hirschhorn, Lu Qi, Ruth J. F. Loos, Cecilia M. Lindgren, Kari E. North, and Iris M. Heid. Sex-stratified genome-wide association studies including 270,000 individuals show sexual dimorphism in genetic loci for anthropometric traits. *PLoS Genet*, 9(6):e1003500, 6 2013.

- [45] G. K. Robinson. That blup is a good thing: The estimation of random effects. *Statistical Science*, 6(1):15–32, 1991.

- [46] Chiara Sabatti, Susan K. Service, Anna-Liisa L. Hartikainen, Anneli Pouta, Samuli Ripatti, Jae Brodsky, Chris G. Jones, Noah A. Zaitlen, Teppo Varilo, Marika Kaakinen, Ulla Sovio, Aimo Ruokonen, Jaana Laitinen, Eveliina Jakkula, Lachlan Coin, Clive Hoggart, Andrew Collins, Hannu Turunen, Stacey Gabriel, Paul Elliot, Mark I. McCarthy, Mark J. Daly, Marjo-Riitta R. Järvelin, Nelson B. Freimer, and Leena Peltonen. Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nat Genet*, 41(1):35–46, 1 2009.
- [47] Sriram Sankararaman, Srinath Sridhar, Gad Kimmel, and Eran Halperin. Estimating local ancestry in admixed populations. *Am J Hum Genet*, 82(2):290–303, 2 2008.
- [48] Karri Silventoinen, Sampo Sammalisto, Markus Perola, Dorret I. Boomsma, Belinda K. Cornes, Chayna Davis, Leo Dunkel, Marlies De Lange, Jennifer R. Harris, Jacob V. B. Hjelmborg, Michelle Luciano, Nicholas G. Martin, Jakob Mortensen, Lorenza Nisticò, Nancy L. Pedersen, Axel Skytthe, Tim D. Spector, Maria Antonietta Stazi, Gonneke Willemsen, and Jaakko Kaprio. Heritability of adult body height: a comparative study of twin cohorts in eight countries. *Twin Res*, 6(5):399–408, 10 2003.
- [49] Daniel O. Stram. *Design, Analysis, and Interpretation of Genome-Wide Association Scans - Springer*. 2013.
- [50] Zhan Su, Jonathan Marchini, and Peter Donnelly. Hapgen2: simulation of multiple disease SNPs. *Bioinformatics*, 27(16):2304–5, 8 2011.
- [51] Andreas Sundquist, Eugene Fratkin, Chuong B. Do, and Serafim Batzoglou. Effect of genetic divergence in identifying ancestral origin using hapaa. *Genome Res*, 18(4):676–82, 4 2008.
- [52] Jenny van Dongen, P. Eline Slagboom, Harmen H. M. Draisma, Nicholas G. Martin, and Dorret I. Boomsma. The continuing value of twin studies in the omics era. *Nat Rev Genet*, 7 2012.
- [53] Danielle Welter, Jacqueline MacArthur, Joannella Morales, Tony Burdett, Peggy Hall, Heather Junkins, Alan Klemm, Paul Flicek, Teri Manolio, Lucia Hindorff, and Helen Parkinson. The nhgri gwas catalog, a curated resource of snp-trait associations. *Nucleic Acids Res*, 42(Database issue):D1001–6, 1 2014.
- [54] Yufeng Wu. *Association Mapping of Complex Diseases with Ancestral Recombination Graphs: Models and Efficient Algorithms*, pages 488–502. Springer Berlin Heidelberg, 2007.
- [55] Jian Yang, Beben Benyamin, Brian P. McEvoy, Scott Gordon, Anjali K. Henders, Dale R. Nyholt, Pamela A. Madden, Andrew C. Heath, Nicholas G. Martin, Grant W. Montgomery, Michael E. Goddard, and Peter M. Visscher. Common SNPs explain a large proportion of the heritability for human height. *Nat Genet*, 42(7):565–9, 7 2010.
- [56] Wen-Yun Y. Yang, Farhad Hormozdiari, Zhanyong Wang, Dan He, Bogdan Pasaniuc, and Eleazar Eskin. Leveraging multi-SNP reads from sequencing data for haplotype inference. *Bioinformatics*, 7 2013.

- [57] Wen-Yun Y. Yang, John Novembre, Eleazar Eskin, and Eran Halperin. A model-based approach for analysis of spatial structure in genetic data. *Nat Genet*, 44(6):725–31, 2012.
- [58] Xiang Zhang, Shunping Huang, Fei Zou, and Wei Wang. Team: efficient two-locus epistasis tests in human genome-wide association study. *Bioinformatics*, 26(12):i217–27, 6 2010.
- [59] Xiang Zhang, Feng Pan, Yuying Xie, Fei Zou, and Wei Wang. *COE: A General Approach for Efficient Genome-Wide Two-Locus Epistasis Test in Disease Association Study*, pages 253–269. Springer Berlin Heidelberg, 2009.
- [60] Xiang Zhou, Peter Carbonetto, and Matthew Stephens. Polygenic modeling with bayesian sparse linear mixed models. *PLoS Genet*, 9(2):e1003264, 2013.
- [61] Xiang Zhou and Matthew Stephens. Genome-wide efficient mixed-model analysis for association studies. *Nat Genet*, 44(7):821–4, 2012.
- [62] Or Zuk, Eliana Hechter, Shamil R. Sunyaev, and Eric S. Lander. The mystery of missing heritability: Genetic interactions create phantom heritability. *Proc Natl Acad Sci U S A*, 109(4):1193–8, 1 2012.