Chapter 4

Genome-Wide Association Studies

Abbas Dehghan

Abstract

Genetic association studies have made a major contribution to our understanding of the genetics of complex disorders over the last 10 years through genome-wide association studies (GWAS). In this chapter, we review the key concepts that underlie the GWAS approach. We will describe the "common disease, common variant" theory, and will review how we finally afforded to capture the common variance in genome to make GWAS possible. Finally, we will go over technical aspects of GWAS such as genotype imputation, epidemiologic designs, analysis methods, and considerations such as genomic inflation, multiple testing, and replication.

Key words Genome-wide association studies, Genetic association, Genotype imputation, Linkage disequilibrium

1 Introduction

It has long been known that the risk of complex disorders such as cardiovascular diseases, type 2 diabetes, or cancer is highly affected by the genetic background of the individual, however, the exact genetic structures that convey the risk were unknown. Researchers have applied different approaches in recent decades to pinpoint the genes that predispose individuals to complex disorders. In this chapter we focus on the genome-wide association study or GWAS, a novel approach that has revolutionized the study of genetics of complex disorders. This approach examines the whole genome in an agnostic system for regions where DNA sequence variations regulate a complex trait or affect the risk of the disease.

The findings of GWAS could have several implications. It could either be used to identify individuals who are at a higher risk of the disease or to shed light on pathways that underlie complex disease. The latter not only enhances our knowledge of the disease, but may also contribute to developing novel medications. Alternatively, this information could be used in the context of precision medicine to tailor the medication for better effects or less adverse effects. In this

Evangelos Evangelou (ed.), *Genetic Epidemiology: Methods and Protocols*, Methods in Molecular Biology, vol. 1793,
https://doi.org/10.1007/978-1-4939-7868-7_4,

chapter, we will briefly review the technology, study design, and analytical methods that are used in GWAS.

2 Genetic Association Versus Linkage Study

2.1 Genetic Variants

The genome or the totality of the genetic material of a cell varies from individual to individual. The variations could be existence of an excess piece of DNA (insertion), missing pieces (delete), or single nucleotide mutations [1]. When mutations are present in more than 1% of the population, they are called single nucleotide polymorphism or SNP. However, in recent years, mutations are referred to as rare or low-frequency SNPs in the literature. Given their simplicity, abundance, and dispersion across the whole genome, SNPs were the first and yet are the most common type of variation that is studied in GWAS. Insertion and deletions (Indel)s are also studied in recent GWAS next to SNPs.

2.2 Common or Rare Variants

Variants have different frequencies. Some are present in a small proportion of the population and some others are very common. There are also private variants that are only identified in one individual. So far millions of variants are discovered in humans and sequencing further individuals will discover more novel variants. The novel variants, of course, are likely to be rare variants in general population. However, any rare or low-frequency variant may be common in a specific ethnic group or an isolated population.

The frequency of the variants is commonly expressed by minor allele frequency (MAF). The fraction indicates the abundance of the less common variant in the pool of alleles in the reference population. For instance, a MAF of 0.3 means that 30% of the alleles carried by the populations are the one that is less common in the reference population. The frequencies could be different in study population than the reference population. As a result, MAF in a sample may sometimes exceed 0.5.

2.3 Common Disease Common Variant Hypothesis

Common disease, common variant hypothesis, is one of the foundations of GWAS. This hypothesis states that common disorders are likely to be influenced by common genetic variants. On one hand, given that common diseases occur in a large proportion of the population, the causal genes could not be rare. On the other hand, the causal variants should, in comparison with rare variants, have a small effect. Otherwise, nearly all who have inherited the deleterious variants should develop the disease which is in contrast to the multifactorial nature of the complex diseases. For instance, a single high penetrance variant with a MAF of 0.30 should lead to a disease that happens in nearly 30% of the population. Therefore, common variants by definition cannot have high penetrance. However, genetic studies have shown that complex disorders such as

cardiovascular diseases and cancer are highly heritable. The conclusion is that common diseases are caused by multiple genetic variants.

In recent decade GWAS has tested the common disease, common variant hypothesis for a wide range of traits and diseases [2]. Although the variants that are identified are continuously increasing, the small effect of genetic variants has led to small percentage of variance explained by these variants. This supports the common disease common variant hypothesis, although this does not exclude the role of rare variants in developing common diseases next to common variants.

2.4 Genome-Wide Approaches for Monogenic and Complex Disorders

Genome-wide search for genetic risk factors has been done in two methods: genome-wide linkage study (GWLS) and GWAS. GWLS looks for physical segments of the genome that is linked to a given trait or disease. It compares the inheritance of traits or diseases with inheritance of DNA segments in a pedigree. GWLS was applied successfully to identify rare genetic variants that contribute to monogenic disorders or highly penetrant traits. It was also applied to multifactorial traits and diseases to map their regulating locus. Nevertheless, it had limited success when it was applied to common disorders like coronary artery disease, asthma, diabetes, or psychiatric disorders. Therefore, it was concluded that the genetic architecture of common disorders is different from rare disorders and will require different investigation approaches [3].

GWAS, however, is based on use of a large number of SNPs or other markers that are genotyped in known linkage regions and is studied in unrelated individuals. Compared to GWLS, GWAS have several advantages. First, it has a better genetic resolution. The resolution is in centimorgan range for GWLS and in kilobases for GWAS. Therefore, GWAS pinpoints the causal gene in a better way. In fact, the most significant SNP in GWAS is either the causal variant or is located in its vicinity. GWLS, however, highlights a large region that may include up to hundreds of genes. GLWS are also difficult to be used for late-onset diseases. A researcher should find family pedigrees including a couple of generations. However, GWAS could be applied to general populations with different age distributions. Finally, GWLS is the most efficient when one gene is inherited in a family but when it comes to multiple genes in general population, GWAS provide a better statistical power [4].

In conclusion, the most efficient approach to study genetics of a trait or disorder depends on the magnitude of effect and allele frequency of the variants that will be used. The variants with large effects are not likely to be common. Common variants with small effect are the ones that are targeted by GWAS and rare variants with large effect are best studied by GWLS. Rare variants with small effects are a real challenge to study and are not investigated much in recent years. Sequencing in large sample sizes may be an approach for this type of genetic effects.

3 Capturing the Common Variation in Genome

3.1 Linkage Disequilibrium

Genetic variants that are located on a chromosome are inherited together. However, this tie is broken apart through generations by genetic recombination. Genetic recombination involves the pairing of homologous chromosomes during meiosis. In a population with random mating, recombination events decrease the correlation between genetic variants and eventually all alleles in the population become independent. When two variants are inherited independent of each other, they are called "in linkage equilibrium." Likewise, the correlation that may remain between two variants is referred to as "linkage disequilibrium" or LD. LD describes the degree to which a genetic variant is inherited together with another genetic variant in a population over time. LD between two genetic variants could be different from one population to another depending on the distance from the founder population, and mating patterns. For instance, the genome of African and African-descent populations, due to being the oldest human population, have gone through more recombination events and therefore include smaller correlated regions compared to other ethnic groups such as Caucasians or Asians.

The level of linkage disequilibrium between two genes is measured by various indices [5]. The coefficient of linkage disequilibrium (D) is defined as

$$D = P_{AB} - (P_A \times P_B)$$

where P_A and P_B are the allele frequency at two loci and PAB is the frequency of A and B occurring together (AB haplotype). D is a difficult coefficient to interpret since its range of possible values depends on the frequencies of the two alleles. As an alternative, D' is defined as D divided by the maximum difference between the observed and expected allele frequencies ($D' = D/D_{min}$). D' varies between -1 and 1. A D' of 1 or -1 means that there is no evidence for recombination between the markers. If allele frequencies are the same, the two variants give the same information and could be used as surrogates for each other. A D' of 0 indicates that the two variants are inherited independent of each other (in perfect equilibrium).

An alternative to D' is the correlation coefficient (r^2) that is expressed as

$$r^2 = {}^{D}\Big/ \sqrt{P_A(1-P_A)P_B(1-P_B)}$$

Correlation coefficient or r^2 is between 0 and 1. Higher values indicate that the genetic variants are highly correlated and in essence include the same genetic variance. The implication of a high LD for genetic studies is that genotyping and study of only

one of the variants may be enough and the second variant includes redundant information.

Given that LD is usually high between close by variants in a region, the genome could be broken down into pieces with high LD. These pieces are called LD blocks. By use of this concept, one can study a limited number of variants and yet capture the whole genetic variation of the genome. The short listed genetic variants that are used in such an approach are called "tagging" variants.

3.2 Human HapMap Project

In order to achieve a shortlist of SNPs that could represent the whole genome, we needed a comprehensive set of information on the LD pattern of the genome. The HapMap international Project was an effort to draw the inheritance pattern of LD blocks in different ethnic groups and to interrogate the common variation in human genome [6]. The project conducted whole genome sequencing techniques to identify common SNPs and characterize their LD pattern. It was done primarily in a number of European descent populations, the Yoruba population of African origin, Han Chinese individuals from Beijing, and Japanese individuals from Tokyo. The data from the HapMap project indicated that more than 80% of the common variation in human genome could be captured by studying approximately 500,000–1,000,000 SNPs across the genome. The first wave of the GWAS were based on nearly 2,500,000 SNPs that were introduced by the HapMap project. Later, other sequencing projects such as the 1000 Genomes project or local sequencing efforts were used as a backbone for GWAS.

Although the HapMap project played a crucial role in making GWAS possible, its data browser is not available since June 2016. This is mainly due to the fact that more recent projects such as the 1000 Genomes project are becoming the standard for research in population genetics and genomics.

3.3 Aiming for Indirect Associations

GWAS were aiming to look up the whole genome for variants that modify the physiology of human body and regulate a trait or affect the risk of a disease. To this end, one should take a challenging and exhaustive effort of studying all genetic variants across the genome. However, the shortlist of SNPs provided by projects such as HapMap allowed us to study the association of such biologically functional variants even if the variant was not present in the shortlist. The LD between the HapMap chosen SNP and the functional variant allowed indirect examination of the association between the variant and the trait or disease of interest [7]. Although this approach increases the coverage of the genome, one should be careful when it comes to interpreting the results of a GWAS. The identified SNPs in GWAS are in most cases not the main functional variant that regulates the trait or causes the disease. It is in fact a tagging SNP that is in high LD with the functional variant in the region.

4 How Did We Afford to Cover the Whole Genome?

4.1 Genotyping Technologies

Although the HapMap project introduced a short list of few hundred thousand SNPs to cover the common variance of the genome, genotyping so many SNPs with low-throughput methods that was available in 1990s was a real challenge. In fact, the availability of microarray technology for high-throughput genotyping with a reasonable pricing gave birth to GWAS. Most of genotyping arrays are manufactured by two companies, Illumina (San Diego, CA) and Affymetrix (Santa Clara, CA). Illumina and Affymetrix use two different platforms. The first generations of these arrays were mainly designed for European descent populations. Therefore, their coverage of the common variation was better in Caucasians than in Asians or African descent populations [8].

4.2 Imputations

When genome-wide association studies became a possibility, it was soon clear that the sample sizes that are available at every center are not large enough to address the small effects of common variants for complex disorders and traits. Therefore, studies started to form consortia to combine their data in meta-analyses. One major challenge, however, was the differences between platforms. This meant that every study had a different set of SNPs and the overlapping SNPs were limited. It was known, however, that once the LD patterns are clear, the alleles for untyped variants could be estimated based on genotyped variants. This process was named genotype imputation since it estimates the missing variants that are not genotyped by the genotyping array. In early days, HapMap was the only reference panel that was available and the data imputed based on this reference panel gave birth to the first wave of GWAS. HapMap included nearly 2,500,000 SNPs and this set were the list of SNPs that all studies imputed their data. A few years later, the 1000 Genomes project provided an alternative imputation platform including a much larger set of SNPs as well as Indels [9]. Recently, the Haplotype Reference Consortium (HRC) has collected a large reference panel of human haplotypes by combining sequencing data from various populations. The HRC reference panels include a comprehensive bank of genetic variants and their haplotypes which not only increases the number of variants that could be imputed but also adds to the accuracy of the genotype imputation (especially for low-frequency variants) [10].

Genotype imputation is based on information provided by haplotypes. In the first step, the variants are linked together based on the most common haplotypes (phasing). Second, the haplotypes are compared to the reference panel. The haplotypes available at the reference panel are normally denser and include more variants compared to the genotyped data. The missing variants in the study population are filled out using the data from the reference

panel. In many instances, however, several haplotypes from the reference panel matches the data set. Several solutions could be applied in such instances. A simple method is to use the most likely allele. Such data is called "best guess" imputed data and is expressed as discrete numbers as 0, 1, or 2 (number of the coded alleles). An alternative is to form the data as a combination of the number of alleles and their probabilities, thus take the uncertainty into account. This data is expressed on a continuous scale from 0 to 2 and called "dosage data."

Every population should primarily be imputed using a reference panel with a similar ethnic background. However, a cosmopolitan reference panel that includes haplotypes from various ethnic groups may also improve the imputation quality since every individual may carry small haplotypes from a far ancestor from a different ethnic group.

5 Epidemiologic Design of GWAS

GWAS could be done in different epidemiologic designs depending on the characteristics of the phenotype and data. Phenotypes could either be quantitative (e.g., height) or categorical (often dichotomous, e.g., diseased/healthy). Quantitative traits could also be broken down into categorical variables (e.g., recoding BMI into normal weight, overweight, and obese), however, this is not recommended from a statistical perspective since information is lost due to the categorization and statistical power is reduced. Quantitative traits could be studied in a cross-sectional design. Given that genetic data is constant over time. It is yet acceptable if DNA samples were collected in a different round of the study than phenotype measurement. Nevertheless, the potential effect of survival between the two rounds on the results, if relevant, should not be overlooked. Binary outcomes are commonly studied in a case-control design. Such designs are popular since they allow the investigator to collect a large number of diseased cases from disease registries, hospital admissions, or large epidemiologic studies. A relevant set of individuals are used as controls. Such designs, however, mostly rely on cross-sectional identification of the diseased cases which are called "prevalent cases." The downside of using prevalent cases is that they do not represent all those who have developed the disease in a population. For instance, prevalent cases of coronary artery disease do not include cases of sudden cardiac death or under represent those who have passed away shortly after MI due to arrhythmias. If the survival after the disease is affected by genetic factors, a GWAS on prevalent cases could be misleading. In such an instance, the alleles that are associated with a better survival after disease could be mistakenly picked up as risk allele for the disease since they are enriched in prevalent cases. This is known as

Neyman's bias or incidence-prevalence bias [11]. To avoid this bias, a prospective setting suits the study best to ensure that a representative set of cases are included in the study.

6 Statistical Analysis of GWAS

6.1 Genetic Model

One of the first assumptions that should be made for a GWAS is the genetic inheritance model. Single variants could affect the phenotype or disease in an additive, recessive/dominant, or multiplicative model. The additive model assumes that there is a linear uniform increase in the risk by adding further copies of the risk allele. In GWAS the additive model is most commonly used model since the exact inheritance model is not known the variants and additive model has reasonable power to detect variants that have additive or dominant effect [12]. The power of this approach, however, is limited if the inheritance model is recessive. Moreover, applying an additive model does not allow identifying the underlying genetic model. Some GWAS examine the best inheritance model fit of their findings in a secondary analysis. Alternatively, some studies repeat their analysis based on several inheritance models but adjust their significance threshold for the number of tests.

6.2 Univariate Analysis

The main analysis in GWAS is normally a regression model. Depending on the nature of the phenotype, a linear, logistic, or Cox regression model is applied. Quantitative phenotypes are commonly analyzed using linear regression models. The genetic variants are the independent factors and the quantitative trait is the dependent variable in the model. Normal distribution is not a strict prerequisite for a linear regression model. However, transformations are used when the phenotype is severely skewed. Although transformation will make the beta estimates difficult to interpret, it helps in avoiding the results to be driven by outliers. Dichotomous phenotypes such as diseases are analyzed either using logistic regression models or if time to event data is provided, a Cox regression model.

GWAS are mainly done primarily in an age and sex adjusted model. Further adjustment, if applicable, could be done for study site or population substructure. Given that genetic variants are inherited randomly, confounding by environmental risk factors is not a major issue. However, confounding by population substructure should be evaluated and adjusted. Every population may be composed of people with different ancestral backgrounds and therefore allele frequencies could vary across subpopulations. When the phenotype or the risk of disease is different among these subpopulations, the test statistics will be inflated across the genome. To illustrate this inflation QQ-plots are used to plot the distribution of the observed test statistics against the distribution of

the test statistics under a null hypothesis. The deviation of the observed test statistics could be measured and expressed as λ. This index is equal to 1, when there is no genomic inflation. Measures above 1.05 are commonly unacceptable in HapMap imputed data and are dealt with either by adjusting for principle components representing population stratification in the regression model or correcting the test statistics for the genomic inflation.

6.3 Multivariate Adjustments

Although the findings in an age and sex adjusted model are not likely to be driven by confounding bias, researchers are sometimes interested in examining the effect of adjustment for certain factors mainly, aiming to examine their potential mediatory role. It should be noted that adjustment comes at the cost of higher degrees of freedom and may negatively affect the statistical power.

6.4 GWIS

Next to the single variant analysis, researchers are sometimes interested in studying the interaction effect between genetic variants or between the variants and environmental risk factors. Such an analysis for the whole genome is called genome-wide interaction analysis or GWIS. Although valid interaction could be valuable and may have clinical and public health implications, the very small interaction effects have so far hampered the efforts to identify robust interactions. Significant, validated, and robust interactions are very scarce. Applying GWIS to study gene-gene interaction has an extra challenge. Given that every GWAS includes hundreds of thousands of genetic variants, the interaction between all variants will include billions of tests which is computationally exhaustive and statistically underpowered. To prune the list of SNPs some investigators use single variant analysis results and pick up the most significant variants, presumably with an arbitrary significance threshold. However, this approach has the downside of overlooking variants that are purely epistatic, i.e., the effect is only shown in the presence of a certain allele of the other interaction genetic variant. Such associations are likely to be overlooked in single variant analysis. Another approach is to limit the analysis to a specific pathway or make a short list of the variants based on their biological relevance.

6.5 Conditional Analysis

In GWAS, commonly, every identified locus is represented by the most significant genetic variant in a genomic region. It is assumed that either the other genetic variants are showing a signal due to their correlation with the sentinel variant or the sentinel SNP is capturing the largest amount of variance from the functional variant in the region. In practice, however, there could be multiple causal variants and the variants in the array could capture different fractions of the variance of the causal variant. Therefore, multiple variants could represent different associations that are independent of each other. Identifying independent variants in a region could

help to increase the proportion of variance that could be explained by the genetic variants.

Conditional analysis is the conventional analytical method to identify independent associations in one locus. To this end, the analysis is repeated for all variants in that locus, adjusted for the sentinel SNP. If the statistical power is large enough, further genetic variants could be identified. This procedure should be conducted over and over to identify further independent associations. Although this procedure is straightforward when it is done for a single study, it would be administratively cumbersome and time consuming when a large meta-analysis of summary statistics is done. The researcher needs to contact the participating studies to conduct the analysis, collect the data, run the meta-analysis, and perform the cycle over and over to make sure that no further signals are left. An alternative approach is introduced where summary-level statistical data and a LD reference panel is used to identify multivariant loci. The method is implemented in GCTA statistical software that is nowadays used for this purpose [13, 14].

6.6 Multiple Testing

Statistical tests are considered significant in classic epidemiology when the p value is smaller than 0.05. This threshold, however, should be adjusted when the hypothesis is examined using multiple tests since the chances of false positive or spurious findings increase by the number of tests. Therefore, adjustment for multiple testing is very crucial to the validity of the findings. Although conservative approaches toward multiple testing could ensure the validity of the findings, an ultimate approach should not hamper the statistical power of the study to identify genetic variants with small effects.

The most commonly applied method to deal with multiple testing is the Bonferroni correction where the significance threshold is divided by the number of tests. In GWAS, millions of variants are tested to identify the one that is associated with the phenotype of interest. In a GWAS where 500,000 variants were genotyped, the significance threshold will be $0.05/500{,}000 = 1 \times 10^{-7}$. The HapMap imputed GWAS, however, are commonly using 5×10^{-8} as the genome-wide significant threshold. This threshold is justified based on an assumption that the contemporary arrays include correlated variants and effectively include one million tests [15]. Although GWAS based on extended reference panels such as 1000 Genomes should consider more stringent significance threshold, many of them are yet using 5×10^{-8}.

An alternative approach to take care of multiple testing is false discovery rate (FDR). The FDR estimates the rate of type I error and enables the investigator to set a threshold where the proportion of false positive results are under a certain limit. In practice it is very common to choose an FDR of 5%. This means that 5% of the associations above this threshold are likely to be false positive (null hypothesis wrongly rejected) [16].

A third option is to perform permutation. To this end, the phenotype of interest is shuffled hundreds or thousands of times across the population to produce databases where the genotype and phenotype are distributed similar to the original dataset but they are not associated with each other. The analysis is repeated each time and the test statistics represent an empirical distribution of the test statistics under null hypothesis. The test statistic is compared to the null distribution and significance is deduced. Permutation could be done by several statistical packages including PLINK which is popular in running GWAS [17].

6.7 Replication

GWAS are hypothesis free studies that examine the whole genome in an agnostic approach. The function of GWAS could therefore be considered hypothesis generating. To test this hypothesis, the association should be validated in an independent sample. This step is known as replication. Although the value of the replication for GWAS findings is widely appreciated, there are inconsistencies in identifying the associations that deserve replication, defining a proper replication study and criterion for refuting the finding based on the replication results.

Any replication effort should be done under the same circumstances as in the discovery. The inheritance model, definition of the phenotype, and covariate adjustment should be identical. One major challenge, however, is to provide sufficient sample size. Associations are commonly stronger in GWAS than replication studies, a phenomenon known as the winner's curse that complicates the sample size estimation for replication studies [18]. Lack of replication in a small population set is always difficult to interpret. It is not possible to find out whether the association is absent due to the false positive association in discovery panel or lack of power in the replication set.

The replication study should also be done in an identical sample that is independent of the discovery set. Once the finding is replicated in a similar population, the association could be extended to other ethnic groups by replicating it in those populations. Some studies use the latter both as a mean for replication and generalization. Although replicated associations could be considered replicated and generalized, lack of association, for instance in a different ethnic group, is difficult to interpret. It may be due to a difference in LD pattern across populations or false positive finding in the discovery panel.

7 Concluding Note

It is no exaggeration to say that GWAS have revolutionized the field of human genetics. Thousands of genetic loci are introduced in association with various complex traits and disorders in recent

decade using GWAS. Many of the findings refer to pathways and mechanisms that were not in the radar due to our limited biological knowledge. The discoveries are expected to continue as larger sample sizes and better imputation platforms are becoming available. At the same time, next generation sequencing seems to move GWAS one step forward by providing a comprehensive DNA sequence readout of the genome. Despite this advancement, genotyping technologies are likely to keep their role as a valid technique for GWAS due to their cheaper prices, larger available sample sizes, and simpler analytical methods. In fact, sequencing further individuals may improve current imputation reference panels and help the microarray genotyping technology as a rival for sequencing technologies by advancing the imputation quality of low-frequency variants.

References

1. 1000 Genomes Project Consortium, Abecasis GR, Altshuler D et al (2010) A map of human genome variation from population-scale sequencing. Nature 467(7319):1061–1073. https://doi.org/10.1038/nature09534
2. Hindorff LA, Sethupathy P, Junkins HA et al (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. Proc Natl Acad Sci U S A 106(23):9362–9367. https://doi.org/10.1073/pnas.0903103106
3. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6(2):95–108. https://doi.org/10.1038/nrg1521
4. Risch N, Merikangas K (1996) The future of genetic studies of complex human diseases. Science 273(5281):1516–1517
5. Guo SW (1997) Linkage disequilibrium measures for fine-scale mapping: a comparison. Hum Hered 47(6):301–314
6. International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437(7063):1299–1320. https://doi.org/10.1038/nature04226
7. Wang DG, Fan JB, Siao CJ et al (1998) Large-scale identification, mapping, and genotyping of single-nucleotide polymorphisms in the human genome. Science 280 (5366):1077–1082
8. Li M, Li C, Guan W (2008) Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur J Hum Genet 16 (5):635–643. https://doi.org/10.1038/sj.ejhg.5202007
9. 1000 Genomes Project Consortium, Abecasis GR, Auton A et al (2012) An integrated map of genetic variation from 1,092 human genomes. Nature 491(7422):56–65. https://doi.org/10.1038/nature11632
10. McCarthy S, Das S, Kretzschmar W et al (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48 (10):1279–1283. https://doi.org/10.1038/ng.3643
11. Hill G, Connelly J, Hebert R et al (2003) Neyman's bias re-visited. J Clin Epidemiol 56 (4):293–296
12. Lettre G, Lange C, Hirschhorn JN (2007) Genetic model testing and statistical power in population-based association studies of quantitative traits. Genet Epidemiol 31(4):358–362. https://doi.org/10.1002/gepi.20217
13. Yang J, Lee SH, Goddard ME et al (2011) GCTA: a tool for genome-wide complex trait analysis. Am J Hum Genet 88(1):76–82. https://doi.org/10.1016/j.ajhg.2010.11.011
14. Yang J, Ferreira T, Morris AP et al (2012) Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. Nat Genet 44(4):369–375., S361-363. https://doi.org/10.1038/ng.2213
15. Pe'er I, Yelensky R, Altshuler D et al (2008) Estimation of the multiple testing burden for genomewide association studies of nearly all common variants. Genet Epidemiol 32 (4):381–385. https://doi.org/10.1002/gepi.20303

16. van den Oord EJ (2008) Controlling false discoveries in genetic studies. American journal of medical genetics part B, neuropsychiatric genetics: the official publication of the international society of. Psychiatr Genet 147B (5):637–644. https://doi.org/10.1002/ajmg.b.30650
17. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet 81(3):559–575. https://doi.org/10.1086/519795
18. Zollner S, Pritchard JK (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80(4):605–615. https://doi.org/10.1086/512821