Chapter 5

Bioinformatics Challenges in Genome-Wide Association Studies (GWAS)

Rishika De, William S. Bush, and Jason H. Moore

Abstract

Genome-wide association studies (GWAS) are a powerful tool for investigators to examine the human genome to detect genetic risk factors, reveal the genetic architecture of diseases and open up new opportunities for treatment and prevention. However, despite its successes, GWAS have not been able to identify genetic loci that are effective classifiers of disease, limiting their value for genetic testing. This chapter highlights the challenges that lie ahead for GWAS in better identifying disease risk predictors, and how we may address them. In this regard, we review basic concepts regarding GWAS, the technologies used for capturing genetic variation, the *missing heritability* problem, the need for efficient study design especially for replication efforts, reducing the bias introduced into a dataset, and how to utilize new resources available, such as electronic medical records. We also look to what lies ahead for the field, and the approaches that can be taken to realize the full potential of GWAS.

Key words Data imputation, Epistasis, Electronic medical records, Filtering, Gene–gene interactions, GWAS, Meta-analysis, Missing heritability, Replication

Abbreviations

EMR	Electronic medical record
GWAS	Genome-wide association study/studies
LD	Linkage disequilibrium
MAF	Minor allele frequency
SNP	Single nucleotide polymorphism

1 Introduction

In the field of genetics and epidemiology, genome-wide association studies (GWAS) have become a standard approach for querying the genetic basis of disease susceptibility. This study design measures and analyzes a million or more DNA sequence variations such as single nucleotide polymorphisms (SNPs) that capture

Ronald Trent (ed.), *Clinical Bioinformatics*, Methods in Molecular Biology, vol. 1168,
DOI 10.1007/978-1-4939-0847-9_5, © Springer Science+Business Media New York 2014

much of the common variation in the genome, in an effort to identify genetic risk factors for diseases [1]. Moreover, technological advances that have lowered the cost of genotyping have also fueled an increase in the number of GWAS over the years. In 2012 alone, the National Human Genome Research Institute (NHGRI) GWAS catalog recorded 1,350 published studies [2]. GWAS provide us with a unique opportunity to make disease risk predictions for the general population on the basis of the disease susceptibility loci that are identified. Knowledge of these loci may also provide clues to the biological basis for various diseases, and open up new avenues for prevention and treatment strategies. The key steps involved in conducting a GWAS are summarized in Fig. 1.

The GWAS approach is not *hypothesis-free*; it is based upon the Common Disease—Common Variant (CD–CV) hypothesis. This hypothesis ties together the basic principle of GWAS and the design of genotyping chips. It states that common diseases are caused in part by genetic variations that are also common in the population [3]. Testing the CD–CV hypothesis provides an insight into the underlying genetic architecture of common diseases, e.g., type 2 diabetes, rheumatoid arthritis, or essential hypertension, and some evidence that they are driven by multiple susceptibility alleles. If common variants have a small effect size but common diseases show a strong inheritance in families (high heritability), then almost by definition the disease must be influenced by multiple genetic factors. For example, if a disease shows a heritability of 30 %, this indicates that 30 % of the total variance in the disease risk comes from genetic factors. Hence, if a SNP has a modest effect on disease risk, it can only account for a small portion of the total variance due to genetic factors. Consequently, the total risk of disease due to common genetic variation then must be distributed over multiple susceptibility alleles.

Published concurrently with family-based linkage studies, one of the earliest GWAS success stories was the identification of Complement Factor H as a major risk factor for age-related macular degeneration [4–7]. This study not only showed that DNA sequence variations in the gene were associated with the disease but also provided a new insight into the biological basis for the disease. However, despite the moderate success of the risk variants identified for age-related macular degeneration, most loci identified by GWAS are known to be associated with small increases in disease risk, thereby limiting their value for genetic testing [2, 8, 9].

The example of breast cancer best highlights the failures and successes of GWAS during its tumultuous history. Familial breast cancer, a rare disease with high heritability, is believed to have a simple underlying genetic architecture. In 2007, Easton et al. identified five significant associations by GWAS that were also replicated in multiple independent samples [10]. In a follow-up study two additional susceptibility loci were identified.

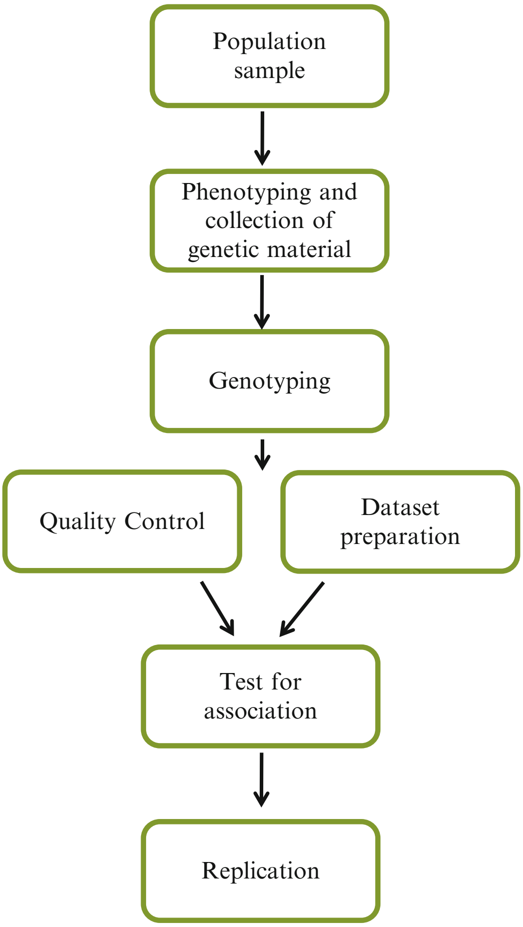

Fig. 1 Overview of the GWAS process. A sample of individuals, e.g., a group of families or cases/controls, is selected from the population to study a disease or phenotype of interest. After strict criteria have been established, phenotypic information and genetic material are collected from the study participants. This is followed by genotyping of the collected material using popular genotyping platforms such as those available from Illumina or Affymetrix. Genotypic data obtained for genetic variants such as SNPs (single nucleotide polymorphisms) are cleaned using quality control procedures such as MAF (minor allele frequency) or LD (linkage disequilibrium) filtering. Data are also adjusted for various covariates and population stratification if required. Next, single locus or multi-locus association tests can be performed to identify genetic variants associated with the phenotype of interest. Ultimately, identified genotype–phenotype associations must be replicated in an independent dataset to assert their credibility

These two loci accounted for <1 % of the familial risk of breast cancer [11]. When these loci were combined with previously known genetic risk factors, together they were able to explain only 5.9 % of the familial risk of breast cancer. On the other hand, the *BRCA1* and *BRCA2* mutations together account for 20–40 % of familial breast cancer, and have been very successful as markers for genetic testing. Hence, although the susceptibility loci identified via GWAS analyses have been useful in providing a new insight regarding the biology of the disease, they have not resulted in new genetic tests.

Similar to the discouraging results with familial breast cancer, GWAS has had limited success in detecting genetic variants that account for a large portion of the heritability of any common disease trait. This chapter will highlight the challenges that lie ahead for GWAS in identifying genetic risk factors that are better classifiers of disease, and how these may be addressed.

First, we must address this *missing heritability* problem by specifically designing our studies to search for nonlinear interactions amongst SNPs. However, this can be a very computationally intensive problem due to the enormous number of pair-wise combinations possible from a GWAS dataset. Linkage disequilibrium (LD) patterns within a dataset may be used to devise strategies for prioritizing SNPs for inclusion in an analysis, reducing this computational burden. Second, we must improve the study design of our GWAS to ensure we increase our statistical power to detect true genetic effects and replicate them by using methods such as meta-analysis and data imputation. Additionally, we must make use of new resources such as electronic medical records (EMRs) to unravel a wealth of phenotypic detail that was previously unavailable. Third, to reduce the biases in GWAS design, we must establish strict criteria for defining phenotypes, adjust for confounding variables that may affect the phenotype of interest, and correct for multiple hypothesis testing (*see* also Chapter 4).

2 Materials

2.1 Genotyping Platforms

Much of the growth and success of GWAS reflects the technology behind the thumb-sized DNA microarray chips designed to probe one million or more SNPs dispersed throughout the genome.

The genotyping platforms used by most GWAS belong to one of two commercial companies: Illumina (San Diego, CA) or Affymetrix (Santa Clara, CA). The two companies' products differ slightly in their approaches to measure SNP variation, and provide researchers with options in terms of cost, coverage, amount of target DNA required, and protocol complexity.

Affymetrix chips use a printed array format, where each spot on the array, representing a locus or allele, contains a cluster of 25-mer oligonucleotides. This platform also offers a cost-effective approach for high-volume GWAS, as most costs are mainly up front. Illumina, however, produces chips that consist of an ordered array of beads, each representing 50-mer oligonucleotides. Even though this platform offers higher sensitivity, it comes at a cost—the arrays are more expensive and the protocols for decoding bead positions are time intensive. Ragoussis et al. provide an excellent review of these genotyping platforms and their unique strengths and weaknesses [12]. Ultimately, GWAS using either of these platforms have been equally successful in the search for genetic risk factors for common, complex diseases.

2.2 Electronic Medical Records

EMRs, which were primarily designed for hospital administrative processes, have recently given way to a new model of genetic discovery. These records are being used to extract relevant phenotypic information for a subject population. Medical centers can leverage this information for genetic studies by linking these data to biological samples to create large-scale biobanks.

EMRs are a rich resource for different types of information, such as—billing data, diagnosis codes, laboratory results, vital signs, provider documentation from reports and tests, and medication records. The billing data and certain laboratory results are made available as structured "name-value" pair data. The clinical documentation such as test results and medication records are provided as narrative or semi-narrative texts. Provider documentation and medical records form an important resource for correct phenotype characterization. Many hospitals are also installing barcodes to keep records of each drug administration for all patients, which may improve accuracy of pharmacogenomic traits [13].

3 Methods

3.1 Basic Concepts for GWAS

3.1.1 Single Nucleotide Polymorphisms and Minor Allele Frequency

A major goal for both the International HapMap Project as well as the 1000 Genomes Project was to capture and catalog sequence variation in the human genome [14, 15]. SNPs, which are single base pair changes in the DNA sequence, have now become the modern unit of genetic variation. Currently, the public catalog of variant sites (dbSNP Build 138) contains approximately 44 million SNPs [16].

SNPs have been shown to have important functional consequences such as affecting mRNA transcript stability or transcription factor binding affinity [17]. However, it is the ability of SNPs to explain much of the genetic diversity observed amongst humans that makes them ideal candidates for use as markers of a genomic region in GWAS.

For each SNP location, there are two or more allele possibilities. The frequency of the less common allele is referred to as minor allele frequency (MAF). The MAF, along with the minor allele, can be specific to a population. Variants can be classified as common or rare: SNPs with a MAF ≥ 5 % are usually referred to as common variants, and those with MAF <5 % are rare. For example, a SNP with a minor allele (A) frequency of 0.30 indicates that 30 % of the population carries the A allele at the SNP location, instead of the more common allele. Traditionally, most GWAS focus on common variants as witnessed by the long list of validated examples—*FTO* (type 2 diabetes and body mass index) [18–20], *GCKR* (triglycerides) [21], and *APOE4* (Alzheimer disease) [22]. Nevertheless, it has been suggested that rare variants play a role in disease and hereditary risk as well [23, 24].

3.1.2 Linkage Disequilibrium

Linkage disequilibrium (LD) is a measure of correlation between SNP alleles at one site and the specific alleles carried at variant sites nearby. Likewise, a particular combination of alleles along a chromosome is termed a *haplotype*. The concept of LD is closely related to chromosomal linkage, where two markers on a chromosome are physically linked through multiple generations of a family. Both these properties can be eroded by recombination and mutation events across multiple generations, thereby breaking up any contiguous stretches of chromosome. The LD observed in a population is also dependent upon its ancestry. Consequently, populations of African descent show smaller regions of LD, as they are more ancestral compared to Asian and European populations and have undergone greater extents of recombination.

The most common measures for LD are—D′ and r^2. Both measures try to capture the difference in the observed frequency of two alleles that occur together, and how often they would be expected to occur together if they were independent of each other [14, 25].

Measures of LD are extremely useful in GWAS design. r^2 values are used to select *tag SNPs*, which are variants selected specifically because they are in strong LD with other variants surrounding them (*see* **Notes 1** and **2**). This advantageous property of tag SNPs allows them to be used for capturing the variation in that specific stretch of LD. Tag SNPs have been especially useful in reducing genotyping costs for GWAS. According to HapMap, more than 80 % of commonly occurring SNPs in populations of European descent can be captured by a set of 500,000 to a million SNPs spread across the genome [26, 27]. This is what forms the basis of selecting a panel of markers for genotyping chips.

An understanding of LD is also essential for correct comprehension of results from a GWAS analysis. Positive results from a GWAS may represent two types of associations—direct or indirect. A direct association involves a SNP that was directly genotyped in the study. Such a SNP is also referred to as a *functional SNP* or the causal variant. An indirect association is a positive association where the SNP of interest was not directly genotyped in the GWAS. This association represents a tag SNP that was genotyped in the study and is in strong LD with the variant altering the biology of the organism. Usually, follow-up tests, such as resequencing the specific genomic region or performing functional studies to examine the role of the variant in the disease, are required to distinguish between these two possibilities [1, 28].

3.2 GWAS Study Design

In this section we begin with an overview of GWAS design and the steps that we can take to reduce the bias introduced into a dataset—such as setting a rigorous criteria for defining phenotypes. Moreover, we talk about the exciting opportunity of extracting phenotypic information from EMRs and the unique challenges that presents.

3.2.1 Case–Control Design Versus Quantitative Design

Case–control GWAS utilize categorical phenotypes, which are often binary outcomes such as case/control or affected/unaffected. The case group includes individuals who have been diagnosed with the disease phenotype of interest. However, the control group can be chosen in one of two valid ways—individuals who are unaffected by the disease or randomly selected from the population. To avoid false positive results, cases and controls must be matched carefully (*see* **Note 3**). Overall, the case–control study design compares the frequency of SNPs or alleles between the two study groups. A higher frequency of a SNP within the cases instead of controls is indicative of that SNP being associated with increased disease risk [28, 29].

As the name suggests, *quantitative* study designs assess quantitative or *continuous* traits that can be measured, to obtain a quantitative value such as HDL (high-density lipoprotein) and LDL (low-density lipoprotein) cholesterol levels [30]. This study design is statistically more powerful for detecting genetic effects. Quantitative traits also make it easier for researchers to obtain a precise measurement, and provide an outcome that is clinically easier to interpret. Ultimately, such a study design measures if the frequency of a SNP or allele is associated with a certain amount of change in the quantitative trait being studied [31].

3.2.2 Standardizing Phenotype Criteria

For any GWAS, it is important to establish measures to standardize the criteria for defining the phenotype of interest. This is especially true for diseases that do not have well-established quantitative measures to describe the disease phenotype, such as multiple sclerosis. In such a situation, patients are usually classified as either being *affected* or *unaffected* by the disease in question. However, in these cases a simple misclassification error of categorizing someone as a *case* instead of a *control* can have more serious consequences than an error in recording a precise quantitative measure.

Despite a complex clinical phenotype that is difficult to diagnose, multiple sclerosis studies have been successful [32]. This is mainly because these studies use a rule list based on various clinical variables such as the McDonald criteria to establish case–control status [33]. This is especially important when studies are based on collaborations between multiple institutions and centers. In such cases, strict criteria ensure that phenotype definitions are applied uniformly across various clinicians, thereby avoiding any site-based effects. This brings to notice that the success of a GWAS does not always depend on the nature of the phenotypic outcome being analyzed, but rather on the awareness of the specific challenges each phenotypic category presents.

3.2.3 Extracting Phenotypes from EMRs

In recent years, the growth of EMR-linked DNA databanks has presented an exciting new avenue for genetic research. EMRs provide an alternative source for researchers to derive phenotype information about a large population of individuals. They are especially appealing as they contain a longitudinal record of robust clinical data due to routine clinical care of patients. Furthermore, EMR-linked DNA databanks provide researchers with the unique opportunity of reusing genetic information to investigate additional phenotypes. Nevertheless, identifying phenotypes from EMRs presents its own set of challenges, because these records were designed with the logistical problems and billing practices of hospitals in mind [13, 34].

The first step in phenotype extraction involves the use of an initial selection algorithm that chooses a subset of records from the bio-repository through text mining of unstructured text or by making use of structured data fields such as billing codes, in the EMR. The choice of billing codes available for use in the EMR is also important in ensuring the accuracy of the phenotype information extracted or the diagnosis established from the record. The CPT (Current Procedural Terminology) coding system is known to have a higher specificity and lower sensitivity, in comparison to the ICD (International Classification of Diseases) coding system. Though the availability of a single type of code is usually sufficient for identifying a phenotype, often a combination of the codes works better as ICD codes also provide the reason for a clinical encounter or procedure [13].

Similarly, to complement the information from billing and procedure codes, they can be combined with free text in the EMR. Such free text can be parsed using Natural Language Processing (NLP) procedures, which apply syntactic and semantic rules to extract structured information. They do so by connecting the text with medical concepts from a controlled vocabulary such as the Unified Medical Language System (UMLS) or with medication information from vocabularies such as RxNorm [13, 35–37] (*see* also Chapter 16).

Ultimately, as a gold standard measure, clinicians and phenotype experts examine the accuracy of the results obtained from the subset of EMR records selected for the study. A measure of precision, the positive predictive value (PPV) of the initial selection algorithm is assessed. The algorithm is then continually refined based on the feedback from these experts. This process continues until the desired PPV is achieved [13, 34]. This approach has not only been applied to various pharmacogenomic and clinical conditions [38–41], but has also successfully replicated established genotype–phenotype relationships [42].

3.3 Testing for an Association

In addition to ensuring a strong study design, there are a few challenges that must be addressed at the level of association analysis in a GWAS. In this section we describe the steps involved in preparing a dataset prior to an association analysis and in adjusting for confounding variables that may also affect the phenotypic outcome of interest. Moreover, the *missing heritability* problem is addressed, as are the steps that can be taken during an association analysis to prioritize SNPs and search for nonlinear interactions.

3.3.1 Dataset Preparation Prior to an Association Analysis

Prior to testing for a genetic association with a disease outcome of interest, researchers must go through a few steps to prepare their dataset for this analysis. First, a method must be chosen for encoding the genotype information in the dataset, as this may have important implications on the statistical power of an association test. As such, association tests can test for either allelic or genotypic associations. Allelic associations look for an association between an allele and the phenotype of interest, whereas genotypic associations search for associations between genotypes or genotypic classes and the phenotype. There are several ways to form these genotypic classes—using a dominant, recessive, multiplicative, or additive model [29, 31].

Datasets must also be adjusted for a range of factors or covariates—age, sex, clinical covariates like Body Mass Index (BMI) or the study site used for data collection—that are known to affect the phenotype outcome, to prevent spurious associations from being detected. Regression methods are a popular choice for covariate adjustment; logistic regression is used for binary traits and linear regression for examining quantitative traits. These methods calculate the "residuals" for the trait of interest, after covariate adjustment. This is the portion of the trait that is not accounted for by the covariates [43].

Population substructure is one of the more important covariates to address in a dataset, especially when the population comprises various ethnicities. The prevalence of a disease phenotype, as well as allele frequencies, can vary between different human subpopulations. Due to this, within a dataset of multiple ethnicities, ethnic-specific SNPs may show up to be associated with a trait due to population stratification [44]. To prevent any false associations, the ancestry of each subsample needs to be measured using one of various methods such as STRUCTURE [45] or EIGENSTRAT [46]. These methods compare genome-wide allele frequencies with ethnic-specific frequencies on HapMap. This allows for samples to be excluded if they are found to be similar to a nontarget population. As an alternative, EIGENSTRAT can also use a statistical method such as principle component analysis (PCA) to generate principle component values or *ethnicity scores*, which can then be used as covariates for adjustments.

3.3.2 Testing for an Association: Single Locus Versus Multi-locus

The popular approach for analyzing GWAS data includes a series of single-locus statistical tests, which compare the genotype distributions for cases and controls, one SNP at a time. On the whole, these methods aim to identify an association between a SNP and the disease/phenotype of interest. However, the type of association test chosen is dependent upon the phenotypic class (case–control or quantitative) being studied.

Binary traits and case–control study designs are analyzed using a contingency table method or logistic regression. For a set of cases and controls, a contingency table summarizes the number of individuals within each genotypic group for a single biallelic SNP [28]. It searches for a deviation from the *null hypothesis* that there is no association between the phenotype and genotype. Popular statistical tests using this method are the chi-square test or the Fisher's exact test. In addition, contingency tables can be analyzed using standard statistical software packages such as SAS, SPSS, Stata, or Microsoft Excel [29]. As for logistic regression, it is an extension of linear regression where the phenotypic outcome studied is transformed using a logistic function. This method predicts the probability of an individual having a *case* status, given their genotype class. Moreover, logistic regression is often the method of choice as it allows for covariate adjustment.

A popular method for analyzing quantitative traits is the Analysis of Variance (ANOVA), which is similar to linear regression with a categorical predictor variable. For single-SNP analysis, ANOVA functions under the null hypothesis, which states that there is no difference between the trait means for any genotype group. However, ANOVA does function on the basis of certain assumptions: it assumes that the trait is normally distributed, the variance of the trait is the same within each group, and that the groups are independent.

For such GWAS analysis, PLINK is a popular and useful software. It has robust features to handle large amounts of data. It can perform association tests per SNP using either the allelic or inheritance model, or by using the Cochran-Armitage test (a contingency table method). Most importantly, PLINK provides a very detailed user manual that is easy to follow [47].

As mentioned earlier in this chapter, the field of GWAS has had limited success in detecting genetic variants that explain a large portion of the heritability for any given trait. This has led researchers to propose potential sources of *missing heritability*. One such possibility is that *missing heritability* may be found within epistatic interactions between various genes [48]. *Epistasis* is usually defined in one of two ways—biological or statistical. Biological epistasis refers to the physical interactions between biomolecules that are influenced by multiple genetic variants. Statistical epistasis is the term for the nonadditive interactions between multiple genes, each of which affects disease susceptibility, and the environment [49, 50].

The *missing heritability* problem may be exacerbated by GWAS approaches that use a linear modeling framework to analyze SNPs one at a time, thereby failing to recognize the genetic and environmental context of each SNP [51, 52]. Hence, this has led to the adoption of more holistic approaches that recognize the complex landscape of the genotype–phenotype relationship and examine nonlinear interactions between genetic variants throughout the genome. This is referred to as a *multi-locus analysis*, which brings with it a new of set of challenges [53, 54]. Amongst these, the biggest challenge is that the exhaustive examination of all pair-wise interactions involving 500,000 SNPs can be very computationally intensive. This often makes it necessary to use specific criteria to filter the 500,000 markers to make the problem computationally tractable.

Traditionally, most GWAS approaches using a chip of this size perform an initial filtering based on MAF, LD, and other initial quality control checks [47]. Even though these steps reduce the number of markers greatly, a researcher may still be left with about 300,000 SNPs in the dataset. In such cases, a single SNP analysis can be performed to select markers with main effects (these are single SNPs that show a strong association with the disease outcome), based on an arbitrary threshold set as the *significance criteria*. This creates a manageable data subset for an unbiased search of all pair-wise interactions.

Conversely, the dataset can also be filtered so that only those multi-marker interactions will be examined that fit within a certain biological context such as a biological pathway, protein family, and group of genes or proteins involved in a certain molecular function. For example, the Biofilter algorithm combines biomedical knowledge from multiple public repositories with statistical methods such as logistic regression or multifactor dimensionality reduction (MDR) method to analyze SNP–SNP combinations [55]. MDR is a novel method that detects and characterizes higher order combinations of genetic and environmental factors that may be predictive of a phenotype or clinical outcome of interest [56]. Another similar method is INTERSNP, which uses logistic regression, log-linear, and contingency table methods to assess SNP–SNP models [57]. However, it is important to keep in mind that any dataset filtering based on particular criteria will introduce its own biological bias into the dataset (*see* **Notes 4** and **5**).

3.3.3 Post Analysis: Correcting for Multiple Hypothesis Testing

A *p*-value is defined as the probability of observing a test statistic that is equal to or greater than the observed test statistic, if the null hypothesis is true. It is generated for each statistical test that is carried out. A common *p*-value cut off (α) that is used in scientific literature is 0.05. When a *p*-value is equal to or falls below this α cut off, the null hypothesis is rejected. This means that 5 % of the time, when the null hypothesis is rejected, it will actually be true, representing a false positive. This probability value is with regard to a single hypothesis or statistical test. However, for a GWAS study that tests numerous hypotheses and applies many statistical tests, each of these tests has their own false positive probability. Hence, the combined likelihood of a GWAS result being a false positive is a lot higher than from one test. This brings to light the importance of correcting for multiple hypothesis testing and adjusting the *p*-value threshold accordingly.

There are a few popular ways to approach correction for multiple testing:

- *The Bonferroni correction*. This is the most stringent of the three; it assumes that each association test in a GWAS is independent of all the others. It corrects an $\alpha = 0.05$ to $\alpha = (0.05/k)$, where k is the number of statistical tests performed. However, this assumption of independence between all the association tests is not necessarily true, due to the presence of LD between markers. For a GWAS with 500,000 markers, the statistical significance threshold for an association would be corrected to 1e–7.
- *Adjusting the False Discovery Rate (FDR)*. Developed by Benjamini and Hochberg this provides an estimate of the proportion of the statistically significant results that are false positives, at an $\alpha = 0.05$ [58]. The approach essentially corrects for this expected number of *false discoveries*, giving the user an idea of the proportion of true associations within their results. The FDR approach is less stringent than the Bonferroni correction as it allows for a proportion of false positive results rather than calculating the probability of observing one or more false positive results over the entire analysis. These procedures have been used extensively in GWAS and also extended in a variety of ways [59].
- *Using permutation testing to adjust the significance threshold*. Although it is computationally intensive, it is the best approach for generating an empirical distribution of test statistics for a given dataset when the null hypothesis is true. The dataset is permuted by rearranging the phenotype labels for all the individuals, but leaving the genotypic information intact. This breaks up any genotype–phenotype relationship within the dataset. However, this technique ensures that the inherent genotype architecture of the dataset is kept intact. This rearrangement of the phenotype labels is done *N* times (a prespecified number). Each time the labels are rearranged, it represents a new permuted dataset, i.e., a possible sampling of individuals under the null hypothesis. There are a number of software packages that can perform permutation testing for GWAS such as—PLINK [60], PRESTO [61], and PERMORY [62].

3.4 Replication of Results

The biggest concern regarding GWAS results has been the lack of replication of genotype–phenotype associations in an independent study. But an equally formidable challenge is to ensure that a replication study has sufficient statistical power to detect the initial finding. Accordingly, meta-analysis and data imputation procedures can help to tackle this type of challenge.

3.4.1 Statistical Replication

The sole purpose of a replication study is to evaluate an initial positive finding from a GWAS and replicate it to assert its validity and give the association higher credibility. But, despite the general consensus regarding its importance, what actually constitutes a replication is still up for debate. This was the topic of a National Human Genome Research Institute (NHGRI) working group—to outline various criteria involved in defining a replication of a GWAS result [63].

One of the first criteria for establishing a positive replication is that the sample size of the replication study be large enough to detect the effect of the susceptibility allele. This is especially crucial because the effects detected in the original GWAS are often overestimated in the study population it was identified in, as compared to the general population, due to a phenomenon called *winner's curse* [64]. Hence, in reality the sample size required to detect this effect, would have to be much larger than the original study population. This is especially true when trying to distinguish the proposed effect from no effect.

The replication study must be carried out in an independent dataset derived from the same population to avoid any introduction of bias due to differences in ethnicity. Additionally, identical criteria should be used in the replication set to define the phenotype in question. Since the ultimate goal is to replicate a statistical model—a given SNP with a given phenotypic effect—using even slightly different phenotypic definitions can adversely affect the interpretation of the replication results.

Since GWAS markers are chosen based on LD patterns, researchers should aim to replicate a *genomic region*, and not necessarily the original SNP from the initial study. All SNPs in high LD with the original SNP would be considered as candidates for replication. However, a strong rationale should be provided regarding the SNPs being selected for replication, based on linkage disequilibrium, published literature, or putative functional significance. To be considered a successful replication, the magnitude and direction of the genetic effect should be similar across both discovery and replication studies.

3.4.2 Meta-analysis

Meta-analysis is a statistical method for combining several different studies to provide one summary result. It is a widely applied technique in the GWAS field; it allows researchers to increase the power to detect association signals by increasing sample size and examining a larger number of variants across the genome. Ultimately this helps reduce the chances of false positive findings. An essential component to combining multiple GWAS for a meta-analysis is that all the studies should be *examining the same hypothesis*. A key advantage to the meta-analysis method is the inherent protection of patient and clinical data. It only requires the transfer of statistical results and not the original data that other parties may not have permission for.

In the initial stages of a meta-analysis, researchers should set up strong collaborative agreements ahead of time. Accordingly, a detailed analysis plan should be formulated to avoid any heterogeneity being introduced into the study (*see* **Note 6**). There are various statistical measures to quantify heterogeneity and to measure how much the various combined studies differ from each other. Some typical measures of heterogeneity are Cochran's Q or the I^2 statistic [65, 66]. The Cochran's Q statistic aims at revealing whether there is statistically significant heterogeneity or not. It is the weighted sum of squared differences between individual study effects and the summary effect across studies. However, the statistic is often underpowered when too few studies are involved in the meta-analysis.

The I^2 statistic, which is favored more in recent studies, measures the proportion of heterogeneity between studies that is true and not due to chance. A major advantage is that the power of the statistic is not dependent upon the number of studies combined in the meta-analysis. I^2 values may fall within low (<25), medium (>25 and <75) and high (>75) heterogeneity values. These ranges are helpful in identifying which studies may need to be removed from the meta-analysis (*see* **Note 7**).

3.4.3 Data Imputation

A meta-analysis aims to examine the effect of the same allele across all studies. However, this proves difficult when the combined studies have been carried out using different genotyping platforms, each using a different set of markers. To ease this challenge, GWAS can use data or genotype imputation to generate results for a common set of SNP across all the combined studies. The imputation procedure makes use of the known LD and haplotype patterns in reference panels such as HapMap and the 1000 Genomes project, to estimate genotypes for SNPs that were not directly genotyped within a study (*see* **Note 8**) [67, 68].

Some popular algorithms for genotype imputation are BimBam [69], IMPUTE [70], MaCH [71], and Beagle [72]. The underlying principle for these algorithms is similar to that of haplotype phasing algorithms, which estimate the contiguous set of alleles that lie on a specific chromosome. Genotype imputation algorithms identify the shared underlying haplotypes between the study population and the reference panel. This set of shared haplotypes is then used to calculate haplotype frequencies within the genotyped SNPs. The phased haplotypes are next compared with a reference set of haplotypes such as those from the HapMap or 1000 Genomes projects. The matched reference haplotypes are also able to provide genotypic information for surrounding markers that were not directly genotyped. Additionally, haplotypes from the study sample may match more than one reference haplotype. In such cases, the surrounding genotypes are given a score or probability of a match, based on the amount of overlap. These scores are also useful for getting an idea about the amount of uncertainty in the genotype imputation process.

4 Future Directions

Irrespective of its victories and failures, GWAS have ushered in an exciting era in the field of genetics and has added new knowledge to our understanding of various diseases and their underlying mechanisms. Although, as the content of genotyping chips, cohort sizes, and biobanks grow even larger, the challenges of data manipulation, quality control, strong study design, and strict phenotypic definitions grow more complex. Hence, moving forward human geneticists will have to develop bioinformatics infrastructure and expertise to overcome such challenges. Most importantly, scientists will have to combine their bioinformatics efforts with genetics, biochemistry and cell biology to confirm the functional consequence and biological relevance of the genotype–phenotype associations that are identified. Ultimately, the translation of GWAS findings into clinical practice will rely upon correct assumptions regarding the genetic architecture of complex traits especially in the context of gene–gene and gene–environment interactions.

5 Notes

1. An r^2 value of 1 is a sign of complete LD and that the alleles at these two associated markers have identical frequencies. To select a tag SNP, an r^2 value of 0.8 or greater is considered to be high and appropriate for using one SNP to tag another in a GWAS [73, 74].
2. LD structures vary between populations, hence, tag SNPs picked for one population may not work for another. Accordingly, populations with high LD will require fewer tag SNPs to capture their variation.
3. Appropriate matching of cases and controls in a GWAS is crucial for preventing any genetic difference between the two groups from being detected due to biased sampling. Researchers must ensure that cases and controls share the same ethnicity, and, if possible, come from the same geographical area.

4. A created dataset based on SNPs that show main effects, enriches for markers that first show a strong association on their own, before searching for pair-wise interactions. This will prevent the detection of certain *purely epistatic* multi-marker interactions—i.e., interactions between markers which by themselves may not have a detectable main effect, and a large part of the heritability is concentrated in their interaction, not individual effects [53].
5. An obvious drawback of filtering datasets based on biological criteria is the reliance upon existing biomedical knowledge, and the quality of this knowledge in public databases. However, SNP combinations identified from the examination of such a data subset are easier to interpret within a biological context.
6. There are several measures that can be taken to avoid introducing heterogeneity in a meta-analysis. The general design of each included study, the quality control procedures, covariate adjustment, and phenotypic definition applied should be the same across all studies. Similarly, the SNP analysis strategies at the level of each individual study should also follow near-identical procedures. Most importantly, the samples added from each study should be independent of each other. Lastly, all results from the individual studies should be reported relative to a common genomic build and reference allele [66, 75].
7. As is true with using any statistical values, these measures should only be used as guides to identify studies introducing an obvious bias. For example, a study may examine a different hypothesis or it may be unduly influential as an outlier. Furthermore, removing a study based solely on a statistical score increases the chances for false discoveries, as it does not make correct use of an agnostic statistical procedure designed to reduce such bias.
8. The reference panel chosen for genotype imputation should be derived from a population with the same ethnicity as the study population to avoid poor quality of the haplotype matches. Additionally, the reference allele for each SNP must be identical between the study population and the reference panel used.

References

1. Hirschhorn JN, Daly MJ (2005) Genome-wide association studies for common diseases and complex traits. Nat Rev Genet 6: 95–108
2. Hindorff L, MacArthur J, Morales J et al. A catalog of published genome-wide association studies. www.genome.gov/gwastudies/
3. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502–510
4. Edwards AO, Ritter R, Abel KJ et al (2005) Complement factor H polymorphism and age-related macular degeneration. Science 308: 421–424

5. Haines JL, Hauser MA, Schmidt S et al (2005) Complement factor H variant increases the risk of age-related macular degeneration. Science 308:419–421
6. Klein RJ, Zeiss C, Chew EY et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308: 385–389
7. Maller J, George S, Purcell S et al (2006) Common variation in three genes, including a noncoding variant in CFH, strongly influences risk of age-related macular degeneration. Nat Genet 38:1055–1059
8. Williams SM, Canter JA, Crawford DC et al (2007) Problems with genome-wide association studies. Science 316:1841–1842
9. Jakobsdottir J, Gorin MB, Conley YP et al (2009) Interpretation of genetic association studies: markers with replicated highly significant odds ratios may be poor classifiers. PLoS Genet 5:e1000337
10. Easton DF, Pooley KA, Dunning AM et al (2009) Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087–1093
11. Ahmed S, Thomas G, Ghoussaini M et al (2009) Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. Nat Genet 41:585–590
12. Ragoussis J (2009) Genotyping technologies for genetic research. Annu Rev Genomics Hum Genet 10:117–133
13. Denny JC (2012) Mining electronic health records in the genomics era. PLoS Comput Biol 8:e1002823
14. The International HapMap Consortium (2005) A haplotype map of the human genome. Nature 437:1299–1320
15. The 1000 Genomes Project Consortium, Abecasis GR, Altshuler D, Auton A et al (2010) A map of human genome variation from population-scale sequencing. Nature 467:1061–1073
16. Sherry ST, Ward MH, Kholodov M et al (2001) dbSNP: the NCBI database of genetic variation. Nucleic Acids Res 29:308–311
17. Griffith OL, Montgomery SB, Bernier B et al (2008) ORegAnno: an open-access community-driven resource for regulatory annotation. Nucleic Acids Res 36:D107–D113
18. The Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature 447:661–678
19. Scuteri A, Sanna S, Chen W-M et al (2007) Genome-wide association scan shows genetic variants in the FTO gene are associated with obesity-related traits. PLoS Genet 3:e115
20. Frayling TM, Timpson NJ, Weedon MN et al (2007) A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. Science 316:889–894
21. Saxena R, Voight BF, Lyssenko V et al (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. Science 316:1331–1336
22. Corder EH, Saunders AM, Strittmatter WJ et al (1993) Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. Science 261:921–923
23. Bansal V, Libiger O, Torkamani A et al (2010) Statistical analysis strategies for association studies involving rare variants. Nat Rev Genet 11:773–785
24. Gibson G (2012) Rare and common variants: twenty arguments. Nat Rev Genet 13:135–145
25. Devlin B, Risch N (1995) A comparison of linkage disequilibrium measures for fine-scale mapping. Genomics 29:311–322
26. Li M, Li C, Guan W (2008) Evaluation of coverage variation of SNP chips for genome-wide association studies. Eur J Hum Genet 16: 635–643
27. Distefano JK, Taverna DM (2011) Technological issues and experimental design of gene association studies. Methods Mol Biol 700:3–16
28. Lewis CM, Knight J (2012) Introduction to genetic association studies. Cold Spring Harb Protoc 3:297–306
29. Lewis CM (2002) Genetic association studies: design, analysis and interpretation. Brief Bioinform 3:146–153
30. Teslovich TM, Musunuru K, Smith AV et al (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466:707–713
31. Bush WS, Moore JH (2012) Genome-wide association studies. PLoS Comput Biol 8: e1002822
32. Habek M, Brinar VV, Borovečki F (2010) Genes associated with multiple sclerosis: 15 and counting. Expert Rev Mol Diagn 10: 857–861
33. Polman CH, Reingold SC, Edan G et al (2005) Diagnostic criteria for multiple sclerosis: 2005 revisions to the "McDonald Criteria". Ann Neurol 58:840–846
34. Kohane IS (2011) Using electronic health records to drive discovery in disease genomics. Nat Rev Genet 12:417–428

35. Sager N, Lyman M, Bucknall C et al (1994) Natural language processing and the representation of clinical data. J Am Med Inform Assoc 1:142–160

36. Friedman C, Hripcsak G, Shablinsky I (1998) An evaluation of natural language processing methodologies. Proc AMIA Symp 855–859

37. Haug PJ, Ranum DL, Frederick PR (1990) Computerized extraction of coded findings from free-text radiologic reports. Work in progress. Radiology 174:543–548

38. Kullo IJ, Fan J, Pathak J et al (2010) Leveraging informatics for genetic studies: use of the electronic medical record to enable a genome-wide association study of peripheral arterial disease. J Am Med Inform Assoc 17:568–574

39. Ding K, de Andrade M, Manolio TA et al (2013) Genetic variants that confer resistance to malaria are associated with red blood cell traits in African-Americans: an electronic medical record-based genome-wide association study. G3 (Bethesda) 3:1061–1068

40. Wilke RA, Berg RL, Linneman JG et al (2010) Quantification of the clinical modifiers impacting high-density lipoprotein cholesterol in the community: Personalized Medicine Research Project. Prev Cardiol 13:63–68

41. McCarty CA, Wilke RA (2010) Biobanking and pharmacogenomics. Pharmacogenomics 11:637–641

42. Ritchie MD, Denny JC, Crawford DC et al (2010) Robust replication of genotype–phenotype associations across multiple diseases in an electronic medical record. Am J Hum Genet 86:560–572

43. Dubé JB, Hegele RA (2013) Genetics 100 for cardiologists: basics of genome-wide association studies. Can J Cardiol 29:10–17

44. Price AL, Zaitlen NA, Reich D et al (2010) New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11:459–463

45. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567–1587

46. Price AL, Patterson NJ, Plenge RM et al (2006) Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38:904–909

47. Sale M, Mychaleckyj JC, Chen W (2009) Planning and executing a genome wide association study (GWAS). Methods Mol Biol 590:403–418

48. Eichler EE, Flint J, Gibson G et al (2010) Missing heritability and strategies for finding the underlying causes of complex disease. Nat Rev Genet 11:446–450

49. Cordell HJ (2002) Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. Hum Mol Genet 11:2463–2468

50. Moore JH, Williams SM (2005) Traversing the conceptual divide between biological and statistical epistasis: systems biology and a more modern synthesis. Bioessays 27:637–646

51. Manolio TA, Collins FS, Cox NJ et al (2009) Finding the missing heritability of complex diseases. Nature 461:747–753

52. Moore JH, Asselbergs FW, Williams SM (2010) Bioinformatics challenges for genome-wide association studies. Bioinformatics 26: 445–455

53. Moore J, Ritchie M (2004) The challenges of whole-genome approaches to common disease. J Am Med Assoc 291:1642–1643

54. Moore JH (2004) Computational analysis of gene–gene interactions using multifactor dimensionality reduction. Expert Rev Mol Diagn 4:795–803

55. Bush WS, Dudek SM, Ritchie MD (2009) Biofilter: a knowledge-integration system for the multi-locus analysis of genome-wide association studies. Pac Symp Biocomput 368–379

56. Ritchie MD, Hahn LW, Roodi N et al (2001) Multifactor-dimensionality reduction reveals high-order interactions among estrogen-metabolism genes in sporadic breast cancer. Am J Hum Genet 69:138–147

57. Herold C, Steffens M, Brockschmidt FF et al (2009) INTERSNP: genome-wide interaction analysis guided by a priori information. Bioinformatics 25:3275–3281

58. Hochberg Y, Benjamini Y (1990) More powerful procedures for multiple significance testing. Stat Med 9:811–818

59. van den Oord EJ (2008) Controlling false discoveries in genetic studies. Am J Med Genet Part B Neuropsychiatr Genet 147B:637–644

60. 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 81:559–575

61. Browning BL (2008) PRESTO: rapid calculation of order statistic distributions and multiple-testing adjusted P-values via permutation for one and two-stage genetic association studies. BMC Bioinformatics 9:309

62. Pahl R, Schäfer H (2010) PERMORY: an LD-exploiting permutation test algorithm for powerful genome-wide association testing. Bioinformatics 26:2093–2100

63. Chanock SJ, Manolio T, Boehnke M et al (2007) Replicating genotype–phenotype associations. Nature 447:655–660

64. Zollner S, Pritchard JK (2007) Overcoming the winner's curse: estimating penetrance parameters from case-control data. Am J Hum Genet 80:605–615

65. Huedo-Medina TB, Sánchez-Meca J, Marín-Martínez F et al (2006) Assessing heterogeneity in meta-analysis: Q statistic or I^2 index? Psychol Methods 11:193–206

66. Evangelou E, Ioannidis JP (2013) Meta-analysis methods for genome-wide association studies and beyond. Nat Rev Genet 14:379–389

67. Li Y, Willer C, Sanna S et al (2009) Genotype imputation. Annu Rev Genomics Hum Genet 10:387–406

68. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511

69. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. PLoS Genet 4:e1000279

70. Howie BN, Donnelly P, Marchini J (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS Genet 5:e1000529

71. Biernacka J, Tang R, Li J et al (2009) Assessment of genotype imputation methods. BMC Proc 3(Suppl 7):S5

72. Browning BL, Browning SR (2009) A unified approach to genotype imputation and haplotype-phase inference for large data sets of trios and unrelated individuals. Am J Hum Genet 84:210–223

73. Barrett JC, Cardon LR (2006) Evaluating coverage of genome-wide association studies. Nat Genet 38:659–662

74. Pe'er I, de Bakker PI, Maller J et al (2006) Evaluating and improving power in whole-genome association studies using fixed marker sets. Nat Genet 38:663–667

75. Zeggini E, Ioannidis JP (2009) Meta-analysis in genome-wide association studies. Pharmacogenomics 10:191–201