Chapter 17

Association Mapping and Disease: Evolutionary Perspectives

Søren Besenbacher, Thomas Mailund, Bjarni J. Vilhjálmsson, and Mikkel H. Schierup

Abstract

In this chapter, we give a short introduction to the genetics of complex diseases emphasizing evolutionary models for disease genes and the effect of different models on the genetic architecture, and we give a survey of the state-of-the-art of genome-wide association studies (GWASs).

Key words Complex diseases, Association mapping, Genome-wide association studies, Common disease/common variant

1 Introduction

A combination of genes and environment determines our phenotype. The degree to which genotype or environment influences our phenotype—the balance of nature versus nurture—varies from trait to trait, with some traits independent of genotype and determined by the environment alone and others determined by the genotype alone and independent of the environment.

A measure quantifying the importance of genotype compared to the environment is the so-called heritability. It is the fraction of the total phenotypic variation in the population explained by variation in the genotype within the population [1]. A trait of interest, say a common disease, which exhibits a nontrivial heritability, tells us that genes are important for understanding this trait and that it is worthwhile to identify the specific genetic polymorphisms influencing the trait. The first step toward this is *association mapping*: searching for genetic polymorphisms that, statistically, associate with the trait. Polymorphisms associated with a given phenotype need not influence that phenotype directly, but it is among those associated genetic polymorphisms that we will find the causal ones.

Maria Anisimova (ed.), *Evolutionary Genomics: Statistical and Computational Methods*, Methods in Molecular Biology, vol. 1910, https://doi.org/10.1007/978-1-4939-9074-0_17,

Genetic variants are correlated, a phenomenon called *linkage disequilibrium* (LD), so by examining the trait association of a few variants, we learn about the association of many others. Examining the association between a phenotypic trait and a few hundred thousand to a million genetic variants suffices to capture how most of the common variation in the entire genome associates with the trait [2–4]. When we find a genetic variant associated with the trait, we have not necessarily located a variant that has any functional effect on the trait, but we have located a genomic region containing genetic variation that does. LD is predominantly a local phenomenon, so correlated genetic variants tend to be physically near each other on the genome. If we observe an association between the phenotype and a variant, and the variant is not causally affecting the trait but is merely in LD with a causal variant, the causal variant is likely nearby. Further examination of the region might reveal which variants affect the trait, and how, but that often involves functional characterization and is beyond association mapping. With association mapping, we merely seek to identify genetic variation that associates with a trait.

2 The Allelic Architecture of Genetic Determinants for Disease

Many complex diseases show a high heritability, typically ranging between 20% and 80%. Each genetic variant that increases the risk of disease contributes to the measured heritability of the disease and thus explains some fraction of the estimated total heritability of the trait. For most diseases investigated, many variants contribute, and the fraction of the heritability explained for each is therefore low. The number of contributing variants, their individual effects on the disease probability, their selection coefficient, and their dominance relations can be collectively termed the genetic architecture of a common disease. Insights into this architecture are slowly emerging and reveal differences between diseases [5].

Below we first consider two proposed genetic architectures based on theoretical arguments: the common disease common variant (CDCV) architecture and the common disease rare variant (CDRV) architecture. CDCV states that most of the heritability can be explained by a few high-frequency variants with moderate effects, while CDRV states that most of the heritability can be explained by moderate- or low-frequency variants with large effects. We present population genetic arguments for the two architectures and the consequences of the two architectures for association mapping. Later, in Subheading 5.1, we present empirical knowledge we have obtained about the genetic architectures of common diseases.

2.1 Theoretical Models for the Allelic Architecture of Common Diseases

Understanding the distribution of the number and frequency of genetic variants in a population is the purview of population genetics. Using diffusion approximations we can derive the expected frequency distribution of independent mutations under mutation-drift-selection balance in a stable population (*see*, e.g., Wright [6]). Central parameters are the mutation rate, u, and the selection for or against an allele, measured by s, scaled with the effective population size, N. Mutations enter a population with a rate determined by Nu, and subsequently, their frequencies change in a stochastic manner. If a mutant allele is not subject to natural selection, for example, if it does not lead to any change in function, it is selectively neutral. Its frequency then rises and falls with equal probability. If the allele is under selection, it has a higher likelihood of increasing in frequency than decreasing if it is under positive selection ($s > 0$) and conversely for negative selection ($s < 0$).

At very high or very low frequencies, selection has an insignificant effect on the change in frequency, and the system evolves essentially entirely stochastic (genetic drift). At moderate frequencies, however, the effect of selection is more pronounced, and given sufficiently strong selection (of an order $Ns \gg 1$), the direction of changes in the allele frequency is almost deterministically determined by the direction of selection. An allele subject to a sufficiently strong selection that happens to reach moderate frequencies either halts its increase if selection works against it, and drifts back to a low frequency, or if selection favors it, it rapidly rises to high frequencies, where eventually the stochastic effects again dominate (*see* Fig. 1).

The range of frequencies, where drift dominates, or selection dominates, is determined by the strength of selection (Ns) and the genotypic characteristics of selection, as, e.g., dominance relations between alleles. For strong selection or in large populations, the process is predominantly deterministic for most frequencies, while for weak selection or a small population, the process is highly stochastic for most frequencies. The time an allele can spend at moderate frequencies is also determined by Ns and selection characteristics.

Pritchard and Cox [7, 8] used diffusion arguments to show that common diseases are expected to be caused by a large number of distinct mutations. This implies that genes commonly involved in susceptibility exert their effect through multiple independent mutations rather than a single mutation identical by descent in all carriers (*see* Fig. 2). Each mutation, if under weak purifying selection, is unlikely to reach moderate frequencies, and since the population will only have few carriers of each disease allele, each can only explain little of the heritability. The accumulated frequency of several alleles, each kept to low frequency by selection, can, however, reach moderate frequencies. So the heritability can be

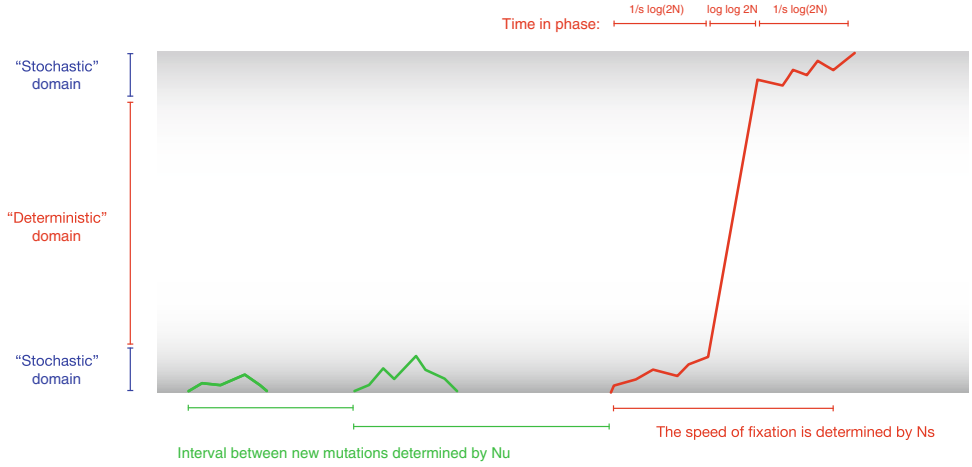

Fig. 1 Mutation, drift, and selection. New mutations enter a population at stochastic intervals, determined by the mutation rate, *u*, and the effective population size, *N*. For low or high frequencies, where the range of such frequencies is determined by the selection factor, *s*, and the effective population size, the frequency of a mutant allele changes stochastically. At medium frequencies, on the other hand, the frequency of the allele changes up or down, depending on *s*, in a practically deterministic fashion. If a positively selected allele reaches moderate frequency, it will quickly be brought to high frequency, at a speed also determined by *s* and *N*

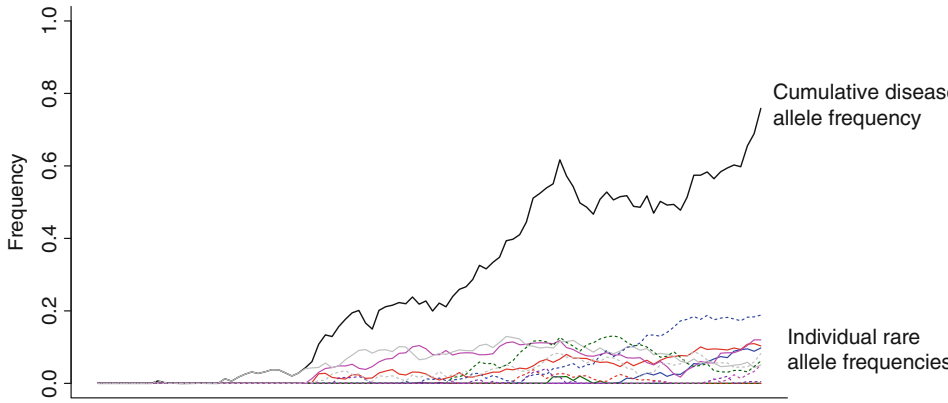

Fig. 2 Accumulation of several rare frequencies. If selection works against a set of alleles, each will be kept at a low frequency. Their accumulated frequency, however, can be high in the population

explained either by many recurrent mutations or many independent loci affecting the disease: the CDRV architecture.

Implicitly, this model assumes a population in mutation-selection equilibrium, and this does not necessarily match human populations. Humans have recently expanded considerably in numbers, and changes in our lifestyle, e.g., from hunter-gatherers to farmers might have changed the adaptive landscape driving selection of our genes.

The frequency range where drift, rather than deterministic selection, dominates is larger with a smaller population than with a larger population. We can think of the drift process as a birth–-death process operating on individual copies of genes, which is highly stochastic. Only when we consider a large number of these processes do we get an almost deterministic process. At low allele frequencies, the process is stochastic because we only have a few copies of the allele to consider. At higher frequencies, we have many copies, so we get the deterministic behavior. The same number of copies, however, constitutes a higher frequency of a small population than of a larger population. Consequently, selection is effective at much lower frequencies in a large population than it is in a small population; the absolute number of copies of a deleterious allele might be the same in a small and a large population, but they constitute a smaller fraction of the large population. In large populations, we expect to see deleterious mutations to be found at small frequencies unless, as is the case for most human populations, the large population size is a consequence of recent dramatic growth [9]. This effect is illustrated as the "transient period" in Fig. 3, where common genetic variants may contribute much more to disease than under stable demographic conditions. Following expansion, alleles that would otherwise be held at low frequency by selection may be at moderate frequencies and thus contribute a larger part of the heritability: the CDCV architecture.

Similarly, a recent change in the adaptive landscape of a population might cause an allele that was previously held at low frequency to be under positive selection and now rise in frequency [10]. In this transition period, an allele may be at a moderate frequency and therefore contributes significantly to the heritability of disease susceptibility (*see* Fig. 4).

Depending on which architecture underlies a given disease, different strategies are needed to discover the genetic variants

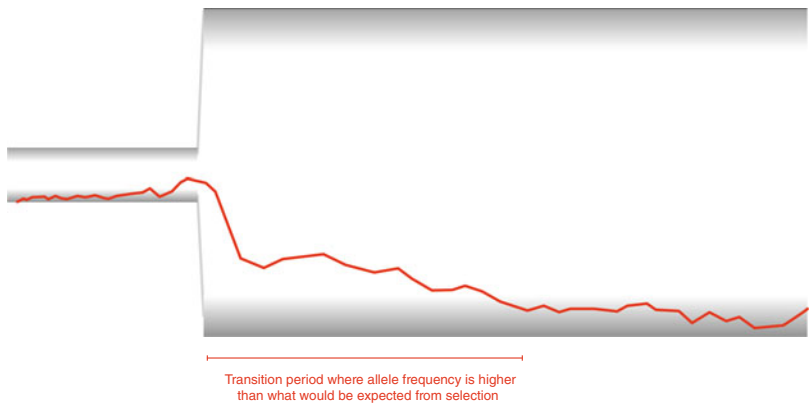

Fig. 3 A population out of equilibrium following an expansion. In a transition period following a population expansion, the allele frequency patterns are different from the patterns in a stable population

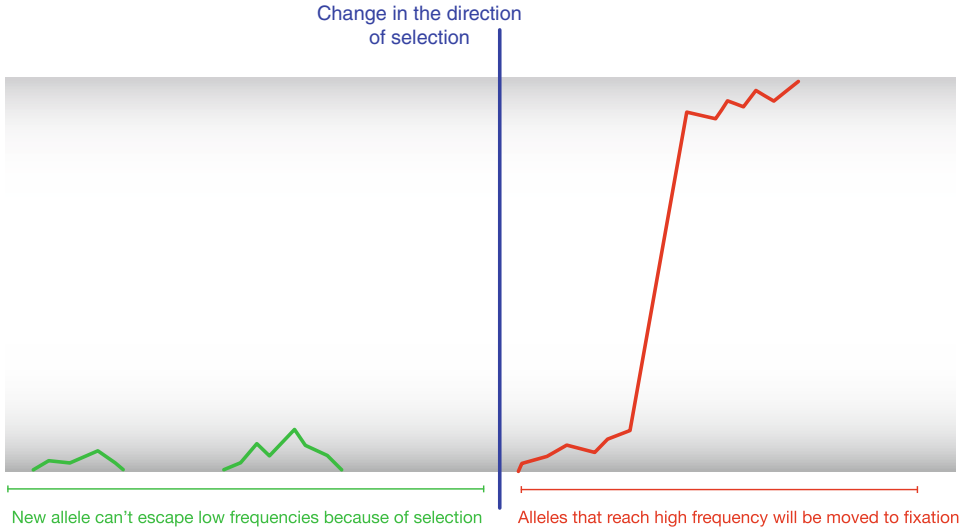

Fig. 4 A population out of equilibrium following changes in the selective landscape. If the selection of an allele changes direction, so the positively selected allele becomes negatively selected and vice versa, it will eventually move through moderate frequencies. Following a change in the selective landscape, it is thus possible to find alleles at moderate frequencies that would not otherwise be found

involved. When genome-wide association mapping was proposed as a strategy for discovering disease variants, the proposal was based on the hypothesis that, at least for some common diseases, the CDCV architecture underlies them. GWAS relies on the CDCV hypothesis for two practical reasons. The first is that the LD patterns across the genome greatly restrict examination to only a small fraction of the total possible variation. It is feasible to probe the common variants of a genome from a small selection of representative variants, but the association with rare variants is far less detectable. Second, statistical analysis of the association between polymorphism and disease is rather straightforward for moderate-frequency alleles but has far less power to detect association with low-frequency alleles.

While the GWAS approach is only practical as an approach for variant discovery for common alleles, it was necessary to hypothesize that the CDCV architecture would be underlying diseases of interest. The actual genetic architecture behind common diseases was unknown, but there were no alternative methods aimed at CDRV, so GWAS was the only show in town.

2.2 The Allelic Frequency Spectrum in Humans

The vast majority of human nucleotide variation is very rare because of our history of population bottlenecks followed by rapid growth. For instance, in the 2500 individuals of the 1000 genomes study, 64 million SNVs have frequency $<0.5\%$, and 20 million SNVs have frequency $>0.5\%$ [11]. Nevertheless the majority of heterozygous variants observed within a single individual are not rare [11]. The

rare variants are most often very recent and therefore specific to populations, and they are also more often deleterious because selection has not yet acted on them [12]. This is particularly clear for loss-of-function variants and other protein-coding variants. A study of 2636 Icelanders found that the fraction of variants with a minor allele frequency (MAF) below 0.1% was 62% for protein-truncating variants, 46% for missense variants, and 38% for synonymous variants [13].

The strong recent population expansions have also allowed variants to increase in frequency by surfing on the population expansion wave front even if they would be selected against in a population with stable size. Thus, rare variants with large effects on disease may exist. The GWAS studies so far have been successful in identifying a large set of common variants associated with disease, so common variants contributing to disease do exist. It is likely that rare variants with large phenotypic effects also contribute to the heritability of many common diseases, but the extend is likely to be disease specific.

3 The Basic GWAS

The first GWASs were published around 2006 [14, 15] when Illumina and Affymetrix first introduced genotyping chips that made it possible to test hundreds of thousands of SNPs quickly and inexpensively. The GWASs' approach to find susceptibility variants for diseases boils down to testing approximately 0.3–-2 million SNPs (depending on chip type) for differences in allele frequencies between cases and controls, adjusting for the high number of multiple tests. This approach is a wonderfully simple procedure that requires no complicated statistics or algorithms but only well-known statistical tests and a minimum of computing power. Despite the simplicity, some issues remain, such as faulty genotype data and confounding factors that can result in erroneous findings if not handled properly. The most important aspects of any GWAS are, therefore, thorough quality control of the data used and measures to avoid and reduce the effect of confounding factors.

3.1 Statistical Tests

The primary analysis in an association study is usually testing each variant separately under the assumption of an additive or multiplicative model. One way of doing that is by creating a 2×2 allelic contingency table as shown in Table 1 by summing the number of A and B alleles seen in all case individuals and all control individuals. Be aware that we are counting alleles and not individuals in this contingency table, so N_{cases} will be equal to two times the number of case individuals because each individual carries two copies of each variant unless we are looking at non-autosomal DNA. If there is no association between the variant and the disease in question, we

Table 1
Contingency table for allele counts in case/control data

	Allele A	Allele B	
Case	$N_{\text{case,A}}$	$N_{\text{case,B}}$	N_{cases}
Control	$N_{\text{control,A}}$	$N_{\text{control,B}}$	N_{controls}
	N_{A}	N_{B}	N

Table 2
Expected allele counts in case/control data

	Allele A	Allele B	
Case	$(N_{\text{cases}} \cdot N_{\text{A}})/N$	$(N_{\text{cases}} \cdot N_{\text{B}})/N$	N_{cases}
Control	$(N_{\text{controls}} \cdot N_{\text{A}})/N$	$(N_{\text{controls}} \cdot N_{\text{B}})/N$	N_{controls}
	N_{A}	N_{B}	N

would expect the fraction of cases that have a particular allele to match the fraction of controls that have that allele. In that case, the expected allele count (EN) would be as shown in Table 2. To test whether the difference between the observed allele counts (in Table 1) and the expected allele counts (in Table 2) is significant, a Pearson χ^2 statistic can be calculated:

$$X^2 = \Sigma_{\text{Phenotype}}\Sigma_{\text{Allele}}\left(N_{\text{Phenotype,Allele}} - \text{EN}_{\text{Phenotype,Allele}}\right)^2/\text{EN}_{\text{Phenotype,Allele}}$$

This statistic approximates a χ^2 distribution with 1 degree of freedom, but if the expected allele counts are very low (<10), the approximation breaks down. This means that if the MAF is very low or if the total sample size, N, is small, an exact test, such as the Fisher's exact test, should be applied. An alternative to the tests that use the 2×2 allelic contingency table and thereby assumes a multiplicative model is the Cochran–Armitage trend test that assumes an additive risk model [16]. This test is preferred by some since it does not require an assumption of Hardy–Weinberg equilibrium in cases and controls combined [17].

While a 1 degree of freedom test that assumes an additive or multiplicative model is usually the first analysis, some studies also perform a test that would be better at picking up associations following a dominant or recessive pattern, for instance, by performing a 2 degrees of freedom test of the null hypothesis of no association between rows and columns in the 2×3 contingency table that counts genotypes instead of alleles.

3.2 Effect Estimates

A commonly used way of measuring the effect size of an association is the allelic odds ratio (OR), which is the ratio of the odds of being a case given that you carry n copies of alleles A to the odds of being a case if you carry $n-1$ copies of allele A. Assuming a multiplicative model, this can be calculated as:

$$\begin{aligned}\mathrm{OR} &= \left(N_{\mathrm{case,A}}/N_{\mathrm{control,A}}\right)/\left(N_{\mathrm{case,B}}/N_{\mathrm{control,B}}\right)\\ &= N_{\mathrm{case,A}}N_{\mathrm{control,B}}/N_{\mathrm{case,B}}N_{\mathrm{control,A}}\end{aligned}$$

Another measure of effect size that is perhaps more intuitive is the relative risk (RR), which is the disease risk in carriers divided by the disease risk in noncarriers. This measure, however, suffers from the weakness that it is harder to estimate. If our cases and controls were sampled from the population in an unbiased way, the allelic RR could be calculated as:

$$\mathrm{RR} = \left(N_{\mathrm{case,A}}/N_A\right)/\left(N_{\mathrm{case,B}}/N_B\right)$$

but it is very rare to have an unbiased population sample in association studies because the studies are generally designed to deliberately oversample the cases to increase the power. This oversampling affects the RR as calculated by the formula above but not the OR which is one of the reasons why the OR is usually reported in association studies instead of the RR.

3.3 Quality Control

Data quality problems can be either variant specific or individual specific, and inspection usually results in the removal of both problematic individuals and problematic variants from the data set.

Individual-specific problems can be caused by low DNA quality or contamination by foreign DNA. A sample of low DNA quality results in a high rate of missing data, where particular variants cannot be called, and there is a higher risk of miscalling variants. It is, therefore, recommended that individuals lacking calls in more than 2–3% of the variants are removed from the analysis. Excess heterozygosity is an indicator of sample contamination, and individuals displaying that should also be disregarded. Sex checks and other kinds of phenotype tests might also be applied to remove individuals, where the genotype information does not match the phenotype information due to a sample mix-up [18].

For a given variant, the data from an individual can be suspicious in two ways: it can fail to be called by the genotype-calling program or it can be miscalled. Typically, a conservative cutoff value is used in the calling process securing that most problems show up as missing data rather than miscalls. Most problematic variants, therefore, reveal a high fraction of missing data, and variants missing calls above a given threshold (typically, 1–5%) are removed. Miscalls typically occur when the homozygotes are hard to distinguish from the heterozygotes, and some of the heterozygotes are being misclassified as homozygotes or vice versa. Both biases

manifest as deviation from Hardy–Weinberg equilibrium, and SNPs that show large deviations from Hardy–Weinberg equilibrium within the controls should be removed [19].

3.4 Confounding Factors

Confounding in GWAS can arise if there are genotyping batch effects or if there is population or family structure in the sample. For example, if cases and controls in GWAS are predominantly collected from geographically distinct areas, association signals could arise due to genetic differences caused by geographic variation, and most of such genetic signals are unlikely to be causal. Such confounding due to population structure typically occurs when samples have different genetic ancestry, e.g., if the sample contains individuals of both European and Asian ancestry. Population structure confounding can also happen when the population structure is more subtle, especially for large sample sizes. Methods for inferring population substructure, such as principal components analysis, are useful for detecting outliers we can remove from the data [20]. However, this approach is not suitable when dealing with subtle structure, as a small bias can become significant in a large enough sample of individuals of similar genetic ancestry.

Confounding in GWAS can be detected as inflation of the test statistics, beyond what is expected due to truly causal variants. A useful way of visualizing such inflation of test statistics is the so-called quantile–quantile (QQ) plot. In this plot, ranked values of the test statistic are plotted against their expected distribution under the null hypothesis. In the case of no true positives and no inflation of the test statistic due to population structure or cryptic relatedness, the points of the plot lie on the $x = y$ line (*see* Fig. 5a). True positives show an increase in values above the line in the right tail of the distribution but do not affect the rest of the points since

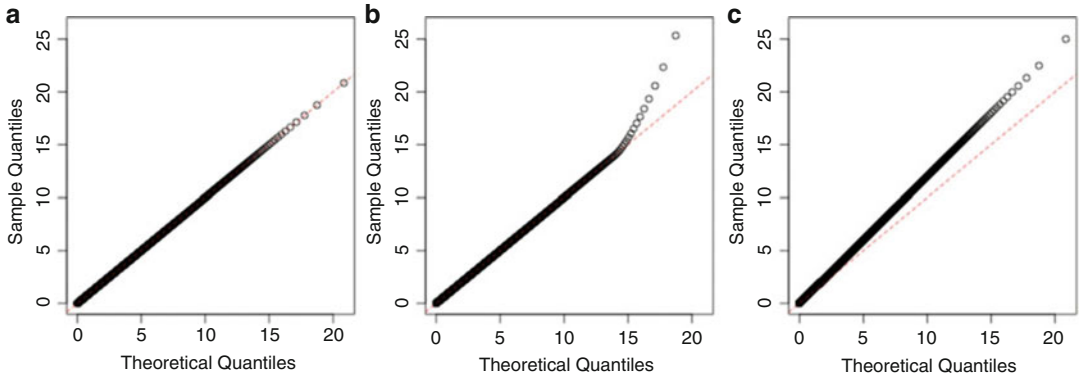

Fig. 5 QQ plots from a χ^2 distribution. (**a**) A QQ plot, where the observation follows the expected distribution. (**b**) A QQ plot, where the majority of observations follow the expected distribution, but where some have unexpectedly high values, i.e., are statistically significant. (**c**) A QQ plot, where the observations all seem to be higher than expected, which is an indication that the observations are not following the expected distribution

only a small fraction of the SNPs is expected to be true positives (Fig. 5b). Cryptic relatedness and population stratification lead to a deviation from the null distribution across the whole distribution and can, thus, be seen in the QQ plot as a line with a slope larger than 1 (Fig. 5c).

Several approaches accounting for population structure in GWAS have been proposed. Devlin and Roeder [21, 22] proposed *genomic control*, i.e., to shrink the observed χ^2 test statistic to make the median coincide with the expected value under the null model. However, studies by Yang et al. [23] and Bulik-Sullivan et al. [24] pointed out that the median and mean χ^2 statistic is expected to be inflated for polygenic traits, even when there is no population structure confounding. With that in mind, we recommend adjusting for the confounders in the statistical model instead of performing genomic control. One such approach is to include covariates that capture the relevant structure in the model. Price et al. [25] proposed including the largest principal components as covariates in the model to adjust for population structure. This approach has proved to be effective in most cases. However, if the sample includes related individuals or if it is very large, controlling for the top PCs may not be able to capture subtle structure. An alternative approach is to use mixed models [26, 27], where the expected genetic relatedness between the individuals is included in the model. Advances in computational efficiency of mixed models [28] now enable analysis of very large and complex data sets, such as the UK biobank data set [29].

Besides population structure, family structure or cryptic relatedness can also confound the analyses. Here one can identify closely related individuals by calculating a genetic relatedness matrix and prune the data so that it does not contain any close relatives. Lastly, sequencing batch effects due to incomplete randomizations can lead to structure, unrelated to genetics, which confounds the analysis. A study on polygenic prediction of longevity by Sebastiani et al. [30] serves as a warning. The researchers applied two different kinds of chips and failed to remove several SNPs that exhibited bad quality on only one of the chips [31]. If the fraction of the two different kinds of chips had been the same in both cases and controls that would probably not have resulted in false signals, unfortunately, the chip with the bad SNPs was used in twice as many cases as controls. When this genotyping batch effect was discovered, the authors had to retract their publication from Science. Type and frequency of errors that may happen during sample preparation and SNP calling are likely to vary through time and space, so case and control samples should be completely randomized as early as possible in the procedure of genotypic typing. Failure to carefully plan this aspect of an investigation introduces errors in the data that are hard, if not impossible, to disclose, and they may reduce interesting findings to mere artifacts.

3.5 Meta-analysis of GWAS

The statistical power to detect association depends directly on the sample size used, all other things being equal. This fact has driven researchers to collaborate across institutions and countries in GWAS consortia, where they combine multiple cohorts in one large analysis. However, for logistic and legal reasons, it may not be possible to share individual-level genotypes, which are required for all of the GWAS approaches covered so far. Meta-analyses of GWASs performed in each cohort are a solution to this problem. These require coordination between the researchers, where they share GWAS summary statistics instead of individual-level genotypes. These summary statistics are then meta-analyzed using statistical approaches that either assume a constant effect across cohorts or not. In recent years many large-scale GWAS meta-analyses have been published, and the resulting summary statistics of these are often made public, providing a treasure trove for understanding genetics of common diseases and traits [32].

3.6 Replication

The best way to make sure that a finding is real is to replicate it. If the same signal is found in an independent set of cases and controls, it means that the association is unlikely to be the result of a confounding factor specific to the original data. Likewise, if the association persists after typing the markers using another genotyping method, it means that it is not a false positive due to some artifact of the genotyping method used.

When trying to replicate a finding, the best strategy is to try to replicate it in a population of similar ancestry. A marker that correlates with a true causal variant in one population might not be correlated with the same variant in a population of different ethnicity, where the LD structure can be different. This is especially problematic when trying to replicate an association found in a non-African population in an African population [33]. A marker might easily have 20 completely correlated markers in a European population, but no good correlates in an African population. To replicate a finding in the European population of one of these variants, it does not suffice to test one of the variants in an African population; all 20 variants must be tested. This, however, also offers a way to fine map the signal and possibly find the causative variant [34].

Before spending time and effort to replicate an association signal in a foreign cohort, it is a good idea to search for the existing partial replication of the marker within the data. Usually, a marker is surrounded by several correlated markers on the genotyping chip, and if one marker shows a significant association, then the correlated markers should show an association too. If a marker is significantly associated with a disease, but no other marker in the region is, then it should be viewed as suspicious.

4 Imputation: Squeezing More Information Out of Your Data

The current generation of SNP chip types includes only 0.3–-2 million of the nine to ten million common SNPs in the human (i.e., SNPs with a MAF of more than 5%). Because of the correlation between SNPs in LD, however, the SNP chips can still claim to assay most of the common variants in the genome (in European populations at least). Although the Illumina HumanHap300 chip only directly tests about 3% of the ten million common SNPs, it still covers 77% of the SNPs in HapMap with a squared correlation coefficient (r^2) of at least 0.8 in a population of European ancestry [35]. The corresponding fraction in a population of African ancestry is only 33%, however.

These numbers expose two limitations of the basic GWAS strategy. First, there is a substantial fraction of the common SNPs that are not well covered by the SNP chips even in European populations (23% in the case of the HumanHap300 chip). Second, we rely on tagging to test a large fraction of the common SNPs, and this diluted signal from correlated SNPs inevitably causes us to overlook true associations in many instances. An efficient way of alleviating these limitations is genotype imputation, where genotypes that are not directly assayed are predicted using information from a reference data set that contains data from a large number of variants. Such imputation improves the GWAS in multiple ways: It boosts the power to detect associations, gives a more precise location of an association, and makes it possible to do meta-analyses between studies that used different SNP chips [36].

4.1 Selection of Reference Data Set

The two important choices when performing imputation are the reference data set to use and the software to use. Usually, a publicly available reference data set, such as the 1000 Genomes Project [11] or the large Haplotype Reference Consortium [37], is used. Alternatively, researchers sequence a part of their study cohort and thus create their own reference data set. The latter strategy has the advantage that one can be certain that the ancestry of the reference data matches the ancestry of the study cohort. It is important that the reference data be from a population that is similar to the study population. If the reference population is too distantly related to the study population, the reliability of the imputed data will be reduced. The quality and nature of the reference data also limit the quality of the imputed data in other ways. A reference data set consisting of only a small number of individuals is not able to reliably estimate the frequency of rare variants and that in turn means that the imputation of rare variants lacks in accuracy. This means that there is a natural limit to how low a frequency a variant can have and still be reliably imputed.

The largest publicly available reference data set is the Haplotype Reference Consortium (HRC) that combines whole-genome sequence data from 20 studies of predominantly European ancestry [37]. The first release of this reference panel has data from 32,611 samples at 39,235,157 SNPs. The large sample size means that variants with minor allele frequencies as low as 0.1% can correctly be imputed using this data set.

The use of imputation methods does not only offer the possibility of increased SNP coverage, but, given the right reference data, also eases the analysis of common non-SNP variation, such as indels and copy number variations (CNVs). So far some reference panels have, however, only include SNVs and disregarded indels and structural variants. The increasing quality of whole-genome sequencing and software for calling structural variants means that better data sets that include structural variants should soon become available. Imputation will then make it possible to use the SNP chips to test many indels and structural variants that are not being (routinely) tested today [38].

4.2 Imputation Software

The commonly applied genotype imputation methods, such as IMPUTE2 [39], BIMBAM [40], MaCH-Admix [41], and minimac3 [42], are all based on hidden Markov models (HMMs). Comparisons of these software packages have shown that they produce data of broadly similar quality but that they are superior to imputation software based on other methodological approaches [36, 43]. The basic HMMs used in these programs are similar to earlier HMMs developed to model LD patterns and estimate recombination rates.

When the sample size is large, imputation using these HMM-based methods imposes a high computational burden. One possible way of decreasing this burden is to pre-phase the samples so that resolved haplotypes are used as input for the imputation software instead of genotypes [44]. But even with pre-phasing, the computational task is far from trivial, and whole-genome imputation is not a task that can be performed on a single computer. This computational problem can be solved by using one of the two free imputation services that have recently been launched (https://imputationserver.sph.umich.edu, https://imputation.sanger.ac.uk). These services allow users to upload their data through a web interface and choose between a set of reference panels. The data set will then be imputed on a High Performance Computing Cluster, and the user will receive an email when the imputed data is ready for download.

4.3 Testing Imputed Variants

Since imputation is based on probabilistic models, the output is merely a probability for each genotype for the unknown variants in a given individual. That is, instead of reporting the genotype of an individual as AG, say, the program reports that the probability of

the genotype being AA is 5%, that of being AG is 93%, and that of being GG is 2%. This nature of the output data challenges the GWAS. The simplest way of analyzing the imputed data is to use the "best guess" genotype, i.e., assume the genotype with the highest probability and ignore the others. In the example above, the individual would be given the genotype AG at the SNP in question, and usually, an individual's genotype would be considered as missing if none of the genotypes have a probability larger than a certain threshold (e.g., 90%). The use of "best guess" genotype is problematic since it does not take the uncertainty of the imputed genotypes into account, may introduce a systematic bias, and lead to false positives and false negatives. A better way is to report a logistic regression on the expected allele count—in the example above, the expected allele count for allele A would be 1.03 ($2p_{AA} + p_{AG}$). This method has proved to be surprisingly robust at least when the effect of the risk allele is small [45], which is the case for most of the variants found through GWAS. An even better solution is to use methods that fully account for the uncertainty of the imputed genotypes [45–47].

5 Current Status

After the first GWAS saw publication in 2005, it was followed by many more studies, and today almost 4000 such studies of human diseases or traits have been published (Fig. 6a). The first GWASs had moderate sample sizes with hundreds of samples, but over the years the sample sizes and thereby the power of the studies have gradually been increasing (Fig. 6b). Imputation and later also next-generation sequencing have resulted in a rapid increase in the

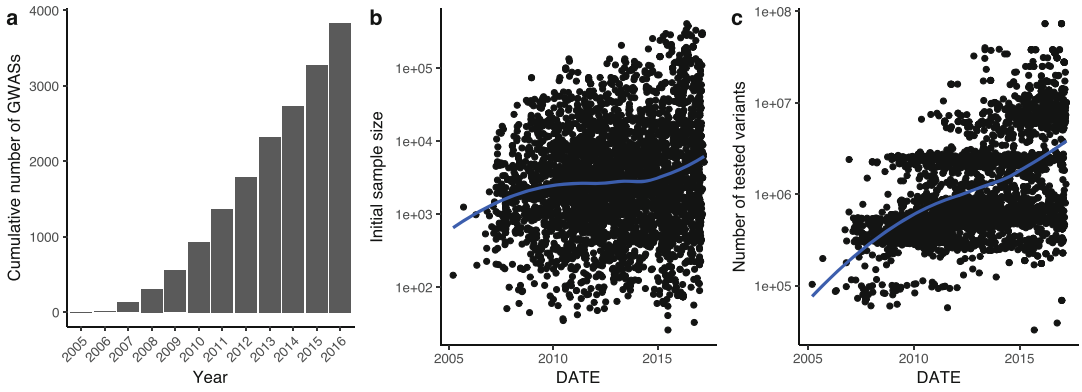

Fig. 6 GWAS statistics from the NHGRI-EBI GWAS Catalog [63] (accessed June 2017). (**a**) The cumulative number of GWASs published since 2005. (**b**) The initial sample sizes of the GWASs. For dichotomous traits the combined number of cases and controls is shown. Replication samples are not counted. (**c**) The number of tested variants in each study

number of variants that are tested in a GWAS (Fig. 6c). All these GWASs published in the last decade have increased our knowledge about the genetic architecture of common diseases a lot. In this section, we will go through some of the insights that have been revealed by these studies.

5.1 Polygenic Architecture of Common Diseases

GWASs have consistently shown that most complex traits and diseases have very polygenic architectures with a large number of causal variants with small effects. The small effect sizes mean that enormous sample sizes are needed to detect the associated variants and that each variant only explains a small fraction of the heritability. Even though large sample sizes have led to the discovery of many loci affecting common diseases, the aggregated effect of all these loci still only explains a small fraction of the heritability.

A good example is type 2 diabetes where researchers by 2012 had identified 63 associated loci that collectively only explained 5.7% of the liability-scale variance [48]. Such results led to much discussion about the possible source of the remaining "missing heritability" [49, 50]. A significant contribution to this debate was when researchers in 2010 started using mixed linear models to estimate the heritability explained by all common variants not only those that surpass a conservative significance threshold. These studies showed that a significant fraction of the so-called missing heritability was not truly missing from the GWAS data sets but only hidden due to small effect sizes. This was first illustrated in height where 180 statistically significant SNPs could only explain 10% of the heritability, but this fraction increased to 45% when all genotyped variants were considered [51].

For common diseases, such analyses have typically shown that around half of the heritability can be explained by considering all common variants. Given the small individual contribution of each of the discovered variants and that the individual contribution of the yet to be found variants will be even smaller, it is likely that the actual number of causal variants will be much more than a thousand for many common diseases. Recent data shows that in many diseases these causal variants are relatively uniformly distributed along the genome. It has, for instance, been estimated that 71–100% of 1 MB windows in the genome contribute to the heritability of schizophrenia [52]. Another article recently estimated that most 100 kB windows contribute to the variation of height and that more than 100,000 markers have an independent effect on height. This strikingly large number leads the authors to propose a new "omnigenic" model in which most genes expressed in a cell type that is relevant for a given disease have a nonzero contribution to the heritability of that disease [53].

5.2 Pleiotropy

The variants that have been discovered by GWASs so far reveal numerous examples where one genetic locus affects multiple often seemingly unrelated traits [54, 55]. One explanation for such a

shared association between a pair of traits is mediation where the shared locus affects the risk of one of the traits, and that trait is causal for the other. Another possible explanation is pleiotropy where the shared locus is independently causal for both traits. It is possible to distinguish between mediation and true pleiotropy by adjusting or stratifying for one trait while testing the other. In the case of mediation, it is also possible to determine the direction of the causation. In general, it is difficult to make such causal inference from observational data, but Mendelian randomization, which uses significantly associated variants as instrumental variables, can in some circumstances be used to assess a causal relationship between a potential risk factor and a disease. For instance, Voight and colleagues used SNPs associated with lipoprotein levels to assess whether the correlation between different forms of lipoprotein and myocardial infarction risk was causal [56]. They found that while low-density lipoprotein (LDL) had a causal effect on disease risk, high-density lipoprotein (HDL) did not.

The fact that pleiotropy is widespread has several implications. One is that variants that have already been found to affect one trait can be prioritized in other studies since they are more likely also to affect another trait than a random variant is. Another implication is that we cannot always examine the effect of selection by studying one trait in isolation. There are multiple examples of antagonistic pleiotropy where a variant increases the risk of one disease while decreasing the risk of another.

5.3 Differences Between Diseases

Because of differences in age of onset and severity, we do not expect identical allelic architectures in all common diseases. Using the currently available GWAS data sets, we can now start to identify these differences in the allelic architectures, but because of the significant differences in samples sizes and the number of tested variants, this is not an easy task.

The data available to date show that the degree of polygenicity differs between diseases with schizophrenia, for example, having more predicted loci than immune disorders [57] and hypertension [52]. Results also show that rare variants play a larger role in some diseases compared to others. Rare variants, for example, have a greater role in amyotrophic lateral sclerosis than in schizophrenia [58] and are even less important in lifestyle-dependent diseases such as type 2 diabetes [59].

6 Perspectives

The price of whole-genome sequencing is still declining, and it is not unreasonable to expect that at some point in the future, a majority of people will get their genomes sequenced. At that point the availability of genetic data will no longer be a limiting

factor in studies of common human diseases. In order to make the most of such huge data sets, the genetic information needs to be combined with high-quality phenotypic and environmental information. If that is achieved, we will be able to explain most—if not all—of the additive genetic variance for the common human diseases. Having large population data sets where genetic data is combined with extensive phenotypic data including information about lifestyle, diet and other environmental risk factors will also enable much better studies of pleiotropy and gene–environment interactions. A few large population data sets are already available now with the UK Biobank [29]—a prospective study of 500,000 individuals—being the best example.

While GWASs have found a lot of loci that are associated with common diseases, the actual causal variant and the functional mechanism driving the causation are still unknown for a large fraction of the loci. In order to understand the functional mechanism of a specific locus, it is necessary to combine sequence data with other types of data. This includes gene expression data (from the correct tissue) and epigenetic data such as methylation. Such data sets are fortunately also becoming cheaper to produce and thus more abundant as a result of falling sequencing costs. Furthermore large consortium data sets such as GTEx [60], ENCODE [61], and Roadmap Epigenomics [62] mean that each lab studying these mechanisms will not have to produce all the data themselves but can in part rely on these public data sets. It is thus likely that we in the future not only will find many more GWAS loci for each common disease but we will also have a much better understanding of how each of these loci affects the disease.

7 Questions

1. How can you distinguish causal variants from other variants when all variants have been typed? Is there any statistical way of distinguishing between correlation and causality just from genotype data? Could you use functional annotations?
2. Consider a GWAS data set, where in the top ten ranked statistics you have five markers that are close together and the remaining five scattered across the genome. Would you consider the five close markers more or less likely to be a true positive? Why? If one of them is a false positive, what would you think about the others?
3. Why is the RR but not the OR estimate affected by a biased case/control sample?
4. How would you test for, e.g., dominant or recessive effects in a contingency table?

References

1. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era |[mdash]| concepts and misconceptions. Nat Rev Genet 9:255. https://doi.org/10.1038/nrg2322
2. Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. Trends Genet 19:135
3. de Bakker PIW, Yelensky R, Pe'er I et al (2005) Efficiency and power in genetic association studies. Nat Genet 37:1217–1223. https://doi.org/10.1038/ng1669
4. Daly MJ, Rioux JD, Schaffner SF et al (2001) High-resolution haplotype structure in the human genome. Nat Genet 29:229–232. https://doi.org/10.1038/ng1001-229
5. Shi H, Kichaev G, Paşaniuc B (2016) Contrasting the genetic architecture of 30 complex traits from summary association data. Am J Hum Genet 99:139–153. https://doi.org/10.1016/j.ajhg.2016.05.013
6. Wright S (1931) Evolution in mendelian populations. Genetics 16:97–159
7. Pritchard JK (2001) Are rare variants responsible for susceptibility to complex diseases? Am J Hum Genet 69:124–137. https://doi.org/10.1086/321272
8. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease–common variant… or not? Hum Mol Genet 11:2417
9. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. Trends Genet 17:502
10. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. Trends Genet 21:596–601. https://doi.org/10.1016/j.tig.2005.08.007
11. 1000 Genomes Project Consortium, Auton A, Brooks LD et al (2015) A global reference for human genetic variation. Nature 526:68–74. https://doi.org/10.1038/nature15393
12. Quintana-Murci L (2016) Understanding rare and common diseases in the context of human evolution. Genome Biol 17:225. https://doi.org/10.1186/s13059-016-1093-y
13. Gudbjartsson DF, Helgason H, Gudjonsson SA et al (2015) Large-scale whole-genome sequencing of the Icelandic population. Nat Genet 47:435. https://doi.org/10.1038/ng.3247
14. Klein RJ, Zeiss C, Chew EY et al (2005) Complement factor H polymorphism in age-related macular degeneration. Science 308:385–389. https://doi.org/10.1126/science.1109557
15. Duerr RH, Taylor KD, Brant SR et al (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. Science 314:1461–1463. https://doi.org/10.1126/science.1135245
16. Lewis CM (2002) Genetic association studies: design, analysis and interpretation. Brief Bioinform 3:146–153
17. Balding DJ (2006) A tutorial on statistical methods for population association studies. Nat Rev Genet 7:781–791. https://doi.org/10.1038/nrg1916
18. 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. https://doi.org/10.1038/nature05911
19. Mccarthy MI, Abecasis GCAR, Cardon LR et al (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. Nat Rev Genet 9:356. https://doi.org/10.1038/nrg2344
20. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. PLoS Genet 2:e190. https://doi.org/10.1371/journal.pgen.0020190
21. Devlin B, Roeder K (1999) Genomic control for association studies. Biometrics 55:997–1004
22. Devlin B, Roeder K, Wasserman L (2001) Genomic control, a new approach to genetic-based association studies. Theor Popul Biol 60:155–166. https://doi.org/10.1006/tpbi.2001.1542
23. Yang J, Weedon MN, Purcell S et al (2011) Genomic inflation factors under polygenic inheritance. Eur J Hum Genet 19:807–812. https://doi.org/10.1038/ejhg.2011.39
24. Bulik-Sullivan BK, Loh P-R, Finucane HK et al (2015) LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. Nat Genet 47:291–295. https://doi.org/10.1038/ng.3211
25. 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. https://doi.org/10.1038/ng1847
26. Yu J, Pressoir G, Briggs WH et al (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38:203–208. https://doi.org/10.1038/ng1702
27. Price AL, Zaitlen NA, Reich D, Patterson N (2010) New approaches to population

stratification in genome-wide association studies. Nat Rev Genet 11:459–463. https://doi.org/10.1038/nrg2813

28. Loh P-R, Tucker G, Bulik-Sullivan BK et al (2015) Efficient Bayesian mixed-model analysis increases association power in large cohorts. Nat Genet 47:284–290. https://doi.org/10.1038/ng.3190
29. Sudlow C, Gallacher J, Allen N et al (2015) UK biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. PLoS Med 12: e1001779. https://doi.org/10.1371/journal.pmed.1001779
30. Sebastiani P, Solovieff N, Puca A et al (2010) Genetic signatures of exceptional longevity in humans. Science. https://doi.org/10.1126/science.1190532
31. Alberts B (2010) Editorial expression of concern. Science 330:912. https://doi.org/10.1126/science.330.6006.912-b
32. Zheng J, Erzurumluoglu AM, Elsworth BL et al (2017) LD Hub: a centralized database and web interface to perform LD score regression that maximizes the potential of summary level GWAS data for SNP heritability and genetic correlation analysis. Bioinformatics 33:272–279. https://doi.org/10.1093/bioinformatics/btw613
33. Teo Y-Y, Small KS, Kwiatkowski DP (2010) Methodological challenges of genome-wide association analysis in Africa. Nat Rev Genet 11:149–160. https://doi.org/10.1038/nrg2731
34. Zaitlen N, Paşaniuc B, Gur T et al (2010) Leveraging genetic variability across populations for the identification of causal variants. Am J Hum Genet 86:23–33. https://doi.org/10.1016/j.ajhg.2009.11.016
35. International HapMap Consortium, Frazer KA, Ballinger DG et al (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449:851–861. https://doi.org/10.1038/nature06258
36. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. Nat Rev Genet 11:499–511. https://doi.org/10.1038/nrg2796
37. McCarthy S, Das S, Kretzschmar W et al (2016) A reference panel of 64,976 haplotypes for genotype imputation. Nat Genet 48:1279–1283. https://doi.org/10.1038/ng.3643
38. Sudmant PH, Kitzman JO, Antonacci F et al (2010) Diversity of human copy number variation and multicopy genes. Science 330:641–646. https://doi.org/10.1126/science.1197005
39. 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. https://doi.org/10.1371/journal.pgen.1000529
40. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. PLoS Genet 3: e114. https://doi.org/10.1371/journal.pgen.0030114
41. Liu EY, Li M, Wang W, Li Y (2013) MaCH-admix: genotype imputation for admixed populations. Genet Epidemiol 37:25–37. https://doi.org/10.1002/gepi.21690
42. Das S, Forer L, Schönherr S et al (2016) Next-generation genotype imputation service and methods. Nat Genet 48:1284–1287. https://doi.org/10.1038/ng.3656
43. Nothnagel M, Ellinghaus D, Schreiber S et al (2009) A comprehensive evaluation of SNP genotype imputation. Hum Genet 125:163–171. https://doi.org/10.1007/s00439-008-0606-5
44. Howie B, Fuchsberger C, Stephens M et al (2012) Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. Nat Genet 44:955–959. https://doi.org/10.1038/ng.2354
45. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. PLoS Genet 4:e1000279. https://doi.org/10.1371/journal.pgen.1000279
46. Marchini J, Howie B, Myers S et al (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. Nat Genet 39:906–913. https://doi.org/10.1038/ng2088
47. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. Nat Rev Genet 10:681–690. https://doi.org/10.1038/nrg2615
48. Morris AP, Voight BF, Teslovich TM, Ferreira T et al (2012) Large-scale association analysis provides insights into the genetic architecture and pathophysiology of type 2 diabetes. Nat Genet 44:981
49. Maher B (2008) Personal genomes: the case of the missing heritability. Nature 456:18–21. https://doi.org/10.1038/456018a
50. Manolio TA, Collins FS, Cox NJ, Goldstein DB (2009) Finding the missing heritability of complex diseases. Nature 461:747

51. Yang J, Benyamin B, McEvoy BP, Gordon S (2010) Common SNPs explain a large proportion of the heritability for human height. Nat Genet 42:565
52. Loh P-R, Bhatia G, Gusev A et al (2015) Contrasting genetic architectures of schizophrenia and other complex diseases using fast variance-components analysis. Nat Genet 47:1385–1392. https://doi.org/10.1038/ng.3431
53. Boyle EA, Li YI, Pritchard JK (2017) An expanded view of complex traits: from polygenic to omnigenic. Cell 169:1177
54. Price AL, Spencer CCA, Donnelly P (2015) Progress and promise in understanding the genetic basis of common diseases. Proc R Soc B 282:20151684. https://doi.org/10.1098/rspb.2015.1684
55. Pickrell JK, Berisa T, Liu JZ et al (2016) Detection and interpretation of shared genetic influences on 42 human traits. Nat Genet 48:709–717. https://doi.org/10.1038/ng.3570
56. Voight BF, Peloso GM, Orho-Melander M (2012) Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380:572
57. Ripke S, O'Dushlaine C, Chambert K et al (2013) Genome-wide association analysis identifies 13 new risk loci for schizophrenia. Nat Genet 45:1150–1159. https://doi.org/10.1038/ng.2742
58. Van Rheenen W, Shatunov A, Dekker AM (2016) Genome-wide association analyses identify new risk variants and the genetic architecture of amyotrophic lateral sclerosis. Nat Genet 48:1043
59. Gorlov IP, Gorlova OY, Amos CI (2015) Allelic spectra of risk SNPs are different for environment/lifestyle dependent versus independent diseases. PLoS Genet 11: e1005371
60. GTEx Consortium (2013) The Genotype-Tissue Expression (GTEx) project. Nat Genet 45:580–585. https://doi.org/10.1038/ng.2653
61. ENCODE Project Consortium, Bernstein BE, Birney E et al (2012) An integrated encyclopedia of DNA elements in the human genome. Nature 489:57–74. https://doi.org/10.1038/nature11247
62. Roadmap Epigenomics Consortium, Kundaje A, Meuleman W et al (2015) Integrative analysis of 111 reference human epigenomes. Nature 518:317–330. https://doi.org/10.1038/nature14248
63. MacArthur J, Bowler E, Cerezo M et al (2017) The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res 45:D896–D901. https://doi.org/10.1093/nar/gkw1133