Chapter 11

Association Mapping and Disease: Evolutionary Perspectives

Søren Besenbacher, Thomas Mailund, and Mikkel H. Schierup

Abstract

In this chapter, we give a short introduction to the genetics of complex disease with special emphasis on evolutionary models for disease genes and the effect of different models on the genetic architecture, and finally give a survey of the state-of-the-art of genome-wide association studies.

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

1. Introduction

The phenotype of an individual is determined by a combination of its genotype and its environment. The degree to which the phenotype is determined by genotype rather than environment—the balance of nature versus nurture—varies from trait to trait, with some traits essentially independent of genotype and determined by the environment and others highly influenced by the genotype and independent of the environment.

A measure quantifying the importance of genotype as compared to the environment is the *heritability*. It is the fraction of the total variance in the population—referred to as the phenotypic variance—explained by variation in the genotype among the individuals in the population (1). An interesting trait, such as a common disease, that exhibits a nontrivial heritability, awakes an interest in finding the genetic explanation behind the trait, that is, identifying the genetic polymorphisms affecting the trait. The first step toward this is *association mapping*, searching for polymorphisms statistically associated with the trait. Polymorphisms associated

with the disease need not influence the trait directly, but it is among those that we will find the polymorphisms that do.

The variants at the various polymorphisms in the genome are correlated—they are in linkage disequilibrium (LD)—so we need not examine all polymorphisms. By analyzing a few hundred thousands to a million evident polymorphisms, we can capture most of the common variation in the entire genome (2–4). In finding such polymorphisms associated with disease risk, we locate a region of the genome that contains one or more polymorphisms that affect disease risk, and by examining such a region in more detail we may locate these.

In the following, we first discuss possible genetic architectures of complex diseases—mainly based on theoretical considerations since little is known about this—and then describe the state of the art in genome-wide association studies (GWASs).

2. The Allelic Architecture of Genetic Determinants for Disease

Many complex diseases show a rather large heritability. Each genetic variant that increases the risk of disease contributes to the measured heritability of the disease. A fraction of heritability can, thus, be attributed to each variant. Doing this for the disease-risk variants known at present, however, we only explain a small fraction of the total heritability (5). The allelic architecture of common diseases—in terms of the number of variants, their frequency, and the risk associated to each—is, thus, poorly understood.

To illustrate the difficulties of inferring the architecture, we consider two hypotheses: the common disease common variant (CDCV) hypothesis or the common disease rare variant (CDRV) hypothesis. 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.

2.1. Theoretical Models for the Allelic Architecture of Common Diseases

Arguments for the expected number of alleles, their frequency, and the risk associated with each allele are based on population genetics considerations. The frequency distribution of independent mutations under mutation–drift–selection balance in a stable population can be derived from diffusion approximations (see, e.g., Wright (6)). Central parameters are the mutation rate, μ , 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 $N\mu$ and subsequently their frequencies change in a stochastic manner. If a mutant allele functions like its origin, $s = 0$ and the allele is selectively neutral. It then rises and falls with equal

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 .

probability while if it is under selection it has a higher probability of increasing than decreasing in frequency for positive selection ($s > 0$) and conversely for negative selection ($s < 0$).

At very high or very low frequencies, selection has a very small effect on the change in frequency and the system evolves essentially completely 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 sufficiently strong selection that happens to reach moderate frequencies either halts its increase and drifts back to a low frequency or continues 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 very strong selection or in very 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 generally expected to be caused by a large number of different mutations in the genes, where damage conveys

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.

disease susceptibility. 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 the disease allele it can only

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.

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

Implicitly, this model assumes a population in mutation selection equilibrium, and this does not necessarily match the human population. The human population has recently expanded considerably in size, and changes in lifestyle, e.g., from hunter-gathers to farmers might have changed the adaptive landscape.

The number of variants at mutation–selection–drift balance is lower in a small population than in a large population. Therefore, in a large population (such as present-day humans), a deleterious mutation is not expected at high frequency unless the population has recently grown dramatically (9). This 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 an 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 hypothesis.

Similarly, a recent change in the selective landscape of a population might cause an allele previously held at low frequency to be under positive selection and rise in frequency while alleles previously at high frequencies can drop in frequency due to negative selection (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).

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.

Depending on which hypothesis is valid, different mapping strategies are needed. Association mapping, however, has so far mainly assumed the CDCV hypothesis for two practical reasons. The first is caused by the fact that the LD patterns across the genome greatly restrict examination to only a small fraction of the total possible variation. This is also the effect that greatly reduces the cost of genome-wide studies by allowing a subset of polymorphisms to reflect the actual genetic variation in the human population due to polymorphisms segregating common alleles. Statistical analysis of association between polymorphism and disease is rather straightforward for moderate-frequency alleles but has far less power to detect association with low-frequency alleles. Thus, so far, only the CDCV hypothesis has been testable and the bulk of association studies have, therefore, used it as their working hypothesis.

2.2. The Allelic Frequency Spectrum in Humans

Empirically, the allelic frequency spectrum of SNPs in the human genome is known in great detail for relatively common alleles (minor allele frequency, MAF, $> 5\%$) from the HapMap project (11). The recently completed pilot project for the 1000 genome project (12) expands the knowledge on uncommon alleles ($1\% < \text{MAF} < 5\%$), which should all be identified during the next phase of the project. This will allow estimation, if not identification, of the number of rare variants and the number of variants carried by single individuals. Identification of very rare alleles awaits the sequencing of many thousand individuals for larger pieces of DNA. This may be achieved sooner for exons than for the rest of

the genome. There are already clear indications that the number of rare variants will be larger than a simple extrapolation of the common SNPs due to the complex demographic history of humans (12–15). Further, recent sequencing of 200 exomes in Europeans reported an enrichment of nonsynonymous variants over synonymous variants among rare polymorphisms (14), strongly suggesting that many nonsynonymous variants are kept in low frequency by natural selection. The proportion of these variants that are involved in complex diseases and perhaps selected against due to this effect is currently unknown.

The European population, where most GWASs so far have been carried out, reveals a site frequency distribution of synonymous variants that generally are shifted to more common alleles as compared to the African population. This is most likely due to a severe bottleneck connected to the out-of-Africa expansion, but also to the expected excess of rare variants in a demographically stable population of the same effective size under selective neutrality. Excess of low-frequency variants is a hallmark of recent population growth and/or weak selection against rare alleles. The latter is visible in the contrast between the frequency distribution for synonymous and nonsynonymous alleles as explained above.

3. The Basic GWAS

The first GWASs were published around 2006 (16, 17) 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 finding 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 is a wonderfully simple procedure that requires no complicated statistics or algorithms but only classical well-known statistical tests and a minimum of computing power. Despite the simplicity, a number of 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 to make sure that only good-quality genotype data is used and to take 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 marker separately under the assumption of an additive or multiplicative model. One way of doing that is by creating a 2×2 allelic

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	$(N_{\text{cases}} \times N_A)/N$	$(N_{\text{cases}} \times N_B)/N$	N_{cases}
Control	$(N_{\text{controls}} \times N_A)/N$	$(N_{\text{controls}} \times N_B)/N$	N_{controls}
	N_A	N_B	N

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 SNP unless we are looking at nonautosomal DNA. If there is no association between the SNP and the disease in question, we 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:

$$\chi^2 = \frac{\sum_{\text{Phenotype}} \sum_{\text{Allele}} (N_{\text{Phenotype, Allele}} - \text{EN}_{\text{Phenotype, Allele}})^2}{\text{EN}_{\text{Phenotype, Allele}}}.$$

This statistic approximates an 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 (18). The latter test is preferred by some since it does not require an assumption of Hardy–Weinberg equilibrium in cases and controls combined (19).

While a 1 degree of freedom test that assumes an additive or multiplicative model is usually the first analysis, many 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 column 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:

$$\text{OR} = \frac{(N_{\text{case}, A} / N_{\text{control}, A})}{(N_{\text{case}, B} / N_{\text{control}, B})} = \frac{N_{\text{case}, A} \times N_{\text{control}, B}}{N_{\text{case}, B} \times N_{\text{control}, A}}.$$

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:

$$\text{RR} = \frac{N_{\text{case}, A} / N_A}{N_{\text{case}, B} / N_B},$$

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 SNP specific or individual specific and inspection usually results in the removal of both problematic individuals and problematic SNPs 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 SNPs cannot be called and there is a higher risk of miscalling SNPs. It is, therefore, recommended that individuals lacking calls in more than 2–3% of the SNPs 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 (20).

For a given SNP, 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 SNPs, therefore, reveal a high fraction of missing data and SNPs' 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 (21).

3.4. Confounding Factors

Confounding factors are differences between cases and controls unrelated to the disease. For instance, if cases are gathered primarily from one part of a country and controls from another part, false association signals could be created because of genetic differences between the two parts of the country. This confounding error is particularly likely to occur when samples mix different ethnicities. If the source of the data is mainly samples from one population, then samples not originating from that population should be excluded whenever possible. Methods for inferring population substructure, such as principal components analysis, are useful for detecting outliers that ought to be removed from the data (22). When a data set includes individuals from distinct subpopulations, the association analysis should be performed separately in each subpopulation and subsequently combined using a procedure that does not assume that the frequencies in the two subpopulations are the same.

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.

Although analyzing distinct populations separately and excluding outliers go a long way, it does not remove all problems caused by population structure. A general inflation of test statistics due to population substructure and cryptic relatedness should be kept in mind, especially when analyzing diseases with high familial aggregation, since that causes cases to be more closely related than controls. 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 case of no true positives and

Fig. 5. QQ plots from an 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.

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 as an increase in values above the line in the right tail of the distribution but do not affect the rest of the points since only a small fraction of the SNPs are 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). A standard adjustment for the inflation of the test statistic is to shrink the range of their distribution to make the median coincide with the expected value. This procedure is called genomic control (23).

Population substructure is not the only difference between cases and controls that can cause trouble. The pooling of data genotyped using different kinds of equipment may produce similar problems. A recent study on longevity by Sebastiani et al. (24) 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 (25). 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, but unfortunately the chip with the bad SNPs was used in twice as many cases as controls.

3.5. Replication

The best way to make sure that a finding is real is to replicate it. If you find the same signal in another set of cases and controls, it means that the association was not caused by a confounding factor specific to your data set. Likewise, if you still see the association 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 is

tagging a true causal variant in one population might not be tagging the same variant in a population of different ethnicity, where the LD structure can be different. This is especially a problem when trying to replicate an association found in a non-African population in an African population (26). A marker might easily have 20 completely correlated markers in a European population, but no good correlates in an African population. This means that if you see a significant association with an SNP that has 20 equivalent SNPs in the European population it is not enough to try to replicate only that SNP, but in an African population you have to test all 20. This, however, also offers a way to fine map the signal and possibly find the causative variant (27).

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. Decisions in cases like this may be further validated by investigating markers that according to HapMap are correlated to the marker in question.

4. Imputation: Squeezing More Information Out of Your Data

The current generation of SNP chip types include only 0.3–2 million of the 9–10 million common SNPs in the human (that is, SNPs with an 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 anyway). Although the Illumina HumanHap300 chip only directly tests about 3% of the 10 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 (11). 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). Secondly, 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

SNPs. 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 (28).

4.1. Selection of Reference Data Set

The two important choices when performing imputation is the reference data set to use and the software to use. Usually, a publicly available reference data set, such as the HapMap (11) or 1000 genomes project (12), is used. Alternatively, researchers type a part of their study cohort on a larger SNP chip with a denser coverage and thus creating 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 is 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 few 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 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).

Whole-genome sequencing projects, such as the 1000 genomes project, coupled with imputation will soon make it possible to use the SNP chips to test many structural variants that are not being (routinely) tested today (29).

4.2. Imputation Software

The commonly applied genotype imputation methods, such as IMPUTE (30), MACH (31), and BAMBAM (32, 33), 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 (28, 34). The basic HMMs used in these programs are similar to earlier HMMs developed to model LD patterns and estimate recombination rates.

Since imputation is based on probabilistic models, its output is merely a probability for each genotype of an SNP unknown in a given individual. That is, instead of reporting the genotype of individual as AG, 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 imputation is to use the “best guess” genotype, i.e., assume the genotype with the highest probability and ignore the others. In the example above, 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 $1.03 (2 \times p_{AA} + p_{AG})$. The method developed on this basis has proved to be surprisingly robust at least when the effect of the risk allele is small (35), which is the case for most of the variants found using GWA. An even better solution is to use methods that fully account for the uncertainty of the imputed genotypes (30, 35, 36).

5. Current Status

Association mapping has for the last 5 years had a strong focus on GWAS using SNP chips with 500,000 – 1,000,000 SNPs, based on the HapMap identification of common human variation in Europe, Asia, and Africa. Hundreds of SNPs have been found to be associated with common diseases in a discovery cohort of affected individuals with matched controls and at least one further population for replication of the initial finding (37). These studies have typically found increased risks of 5–20% for each variant. While initially considered a great proof of concept for the CDCV model, there is a growing awareness that these are unlikely to explain most of the genetic effect unless there is a very large number of common alleles with very small ORs that have escaped detection using the current cohort sizes (typically, 2,000–20,000 individuals). Hence, there is a renewed focus on the site frequency spectrum of disease alleles as discussed in Subheading 2. A clear pattern observed in findings so far is that most variants identified are common and that the inferred ORs increase with rarity of the variant. This cannot be taken as evidence that rare variants have higher ORs, though, since, as demonstrated by Iles (38), we can only detect the rarer variants if they have higher ORs. However, if analysis is restricted to nonsynonymous disease SNPs, then rare variants do seem to have a generally larger OR (39). An analysis of the site frequency spectrum as a function of functional classification of SNPs using PolyPhen also found that rare variants should be more damaging (which through selection would explain their rarity) (13). Thus, it cannot be ruled out that the bulk of heritability not explained by GWASs so far could be explained by many rare variants, each with ORs larger than common variants identified (perhaps, ORs in the 2–10 range which are still sufficiently low to be easily missed by linkage studies).

6. Perspectives

Future data will provide identification of most, if not all, SNPs and CNVs in a large set of individuals. To further our quest toward understanding the genetic etiology of common diseases, methods are needed that expand information on the role of rare variants and variants of small effect. This requires statistically powerful ways of handling information from rare SNPs, rare LD blocks, and amassing distributed local effects. Several promising methods have recently emerged as ways to add signals together locally (39–43). It will be of great interest to use these approaches even in cases, where association with common variants has been shown, since it is possible that some of these associations are due to synthetic association, i.e., that several rare variants accidentally are associated with the same more common variants (44). Searching for variants adding up to a risk in a certain gene does not identify any specific causal variants, but it points to causal genes or regions that can then be further scrutinized either statistically in replication cohorts, by bioinformatics pathway and functional prediction, or experimentally. With full sequencing, we know that the causal variants have been included, but many other variants will be associated due to LD. LD will, thus, be more of a burden than of an asset in future studies and populations with least LD should be most easily amenable for association mapping. However, other approaches to distinguish real from associated variants that are based on biological information on their putative function will be very useful. A wealth of annotation to each position in the genome will soon be available, including the epigenetic context (e.g., nucleosome positioning and modifications, transcription factor and enhancer binding, DNase structure) and the structure of the protein, including its position in biological pathways and interaction with other proteins. Thus, each putative variant can be assigned a posterior probability of true association, which can be used as hypothesis generator as well as a prior probability in replication studies.

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 – concepts and misconceptions. *Nat Rev Genet* 9: 255–266.
2. Cardon LR, Abecasis GR (2003) Using haplotype blocks to map human complex trait loci. *Trends Genet* 19: 135–140.
3. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, et al. (2005) Efficiency and power in genetic association studies. *Nat Genet* 37: 1217–1223.
4. Daly MJ, Rioux JD, Schaffner SF, Hudson TJ, Lander ES (2001) High-resolution haplotype structure in the human genome. *Nat Genet* 29: 229–232.
5. Maher B (2008) Personal genomes: The case of the missing heritability. *Nature* 456: 18–21.
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.
8. Pritchard JK, Cox NJ (2002) The allelic architecture of human disease genes: common disease-common variant...or not? *Hum Mol Genet* 11: 2417–2423.
9. Reich DE, Lander ES (2001) On the allelic spectrum of human disease. *Trends Genet* 17: 502–510.
10. Di Rienzo A, Hudson RR (2005) An evolutionary framework for common diseases: the ancestral-susceptibility model. *Trends Genet* 21: 596–601.
11. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature* 449: 851–861.
12. Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, et al. (2010) A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.
13. Gorlov IP, Gorlova OY, Sunyaev SR, Spitz MR, Amos CI (2008) Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am J Hum Genet* 82: 100–112.
14. Li Y, Vinckenbosch N, Tian G, Huerta-Sanchez E, Jiang T, et al. (2010) Resequencing of 200 human exomes identifies an excess of low-frequency non-synonymous coding variants. *Nat Genet* 42: 969–972.
15. Pelak K, Shianna KV, Ge D, Maia JM, Zhu M, et al. (2010) The characterization of twenty sequenced human genomes. *PLoS Genet* 6.
16. Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, et al. (2005) Complement factor H polymorphism in age-related macular degeneration. *Science* 308: 385–389.
17. Duerr RH, Taylor KD, Brant SR, Rioux JD, Silverberg MS, et al. (2006) A genome-wide association study identifies IL23R as an inflammatory bowel disease gene. *Science* 314: 1461–1463.
18. Lewis CM (2002) Genetic association studies: design, analysis and interpretation. *Brief Bioinform* 3: 146–153.
19. Balding DJ (2006) A tutorial on statistical methods for population association studies. *Nat Rev Genet* 7: 781–791.
20. WTCCC (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447: 661–678.
21. McCarthy MI, Abecasis GR, Cardon LR, Goldstein DB, Little J, et al. (2008) Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 9: 356–369.

22. Patterson N, Price AL, Reich D (2006) Population structure and eigenanalysis. *PLoS Genet* 2: e190.
23. Devlin B, Roeder K (1999) Genomic control for association studies. *Biometrics* 55: 997–1004.
24. Sebastiani P, Solovieff N, Puca A, Hartley SW, Melista E, et al. (2010) Genetic Signatures of Exceptional Longevity in Humans. *Science*.
25. Alberts B (2010) Editorial expression of concern. *Science* 330(6006): 912. DOI: 10.1126/science.330.6006.912-b.
26. Teo YY, Small KS, Kwiatkowski DP (2010) Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet* 11: 149–160.
27. Zaitlen N, Pasaniuc B, Gur T, Ziv E, Halperin E (2010) Leveraging genetic variability across populations for the identification of causal variants. *Am J Hum Genet* 86: 23–33.
28. Marchini J, Howie B (2010) Genotype imputation for genome-wide association studies. *Nat Rev Genet* 11: 499–511.
29. Sudmant PH, Kitzman JO, Antonacci F, Alkan C, Malig M, et al. (2010) Diversity of human copy number variation and multicopy genes. *Science* 330: 641–646.
30. Marchini J, Howie B, Myers S, McVean G, Donnelly P (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 39: 906–913.
31. Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR (2010) MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 34: 816–834.
32. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. *Am J Hum Genet* 78: 629–644.
33. Servin B, Stephens M (2007) Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet* 3: e114.
34. Nothnagel M, Ellinghaus D, Schreiber S, Krawczak M, Franke A (2009) A comprehensive evaluation of SNP genotype imputation. *Hum Genet* 125: 163–171.
35. Guan Y, Stephens M (2008) Practical issues in imputation-based association mapping. *PLoS Genet* 4: e1000279.
36. Stephens M, Balding DJ (2009) Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 10: 681–690.
37. Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, et al. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 106: 9362–9367.
38. Iles MM (2008) What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet* 4: e33.
39. Kryukov GV, Pennacchio LA, Sunyaev SR (2007) Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am J Hum Genet* 80: 727–739.
40. Li B, Leal SM (2008) Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data. *Am J Hum Genet* 83: 311–321.
41. Madsen BE, Browning SR (2009) A groupwise association test for rare mutations using a weighted sum statistic. *PLoS Genet* 5: e1000384.
42. Morris AP, Zeggini E (2010) An evaluation of statistical approaches to rare variant analysis in genetic association studies. *Genet Epidemiol* 34: 188–193.
43. Price AL, Kryukov GV, de Bakker PI, Purcell SM, Staples J, et al. (2010) Pooled association tests for rare variants in exon-resequencing studies. *Am J Hum Genet* 86: 832–838.
44. Dickson SP, Wang K, Krantz I, Hakonarson H, Goldstein DB (2010) Rare variants create synthetic genome-wide associations. *PLoS Biol* 8: e1000294.