Genome-Wide Association Studies

Ammar Al-Chalabi

Cold Spring Harb Protoc 2009; doi: 10.1101/pdb.top66

Topic Introduction

Genome-Wide Association Studies

Ammar Al-Chalabi

INTRODUCTION

The goal of association studies is to discover genetic variation that differs in frequency between cases and controls or between individuals with different phenotypic values. Until a few years ago, the only method available for such studies was low-throughput analysis in which a single gene was selected and either genotyped for known genetic variants or sequenced to identify such variants. With the completion of the Human Genome Mapping Project, we have learned that single-nucleotide polymorphisms (SNPs) are frequent in the genome and that variants in physical proximity tend to correlate in genotype. Therefore, a major international effort was started to map this correlation in the form of the International HapMap Project. The concurrent advances in genetic laboratory techniques, statistical methods, and computing power, coupled with the information from the HapMap, have allowed large-scale microchip-based technologies to be used to assay large numbers of SNPs quickly and easily. Thus, truly genome-wide association studies (GWAS) can now be performed, analogous to linkage studies of Mendelian diseases in having no prior hypothesis of the chromosomal location responsible for disease. In this article, we will only discuss case-control studies in which family members are not analyzed, but the principles apply to large-scale family-based association studies as well.

RELATED INFORMATION

The Human Genome Mapping Project was completed in 2001 (International Human Genome Sequencing Consortium 2001; Venter et al. 2001). Assemblies and annotations of the human genome can be found at http://www.ncbi.nlm.nih.gov and http://genome.ucsc.edu (Kent et al. 2002). The International HapMap Project can be found at http://www.hapmap.org (International HapMap Consortium 2003). Other useful web resources include the database of genotypes and phenotypes (dbGaP) (http://www.ncbi.nlm.nih.gov/dbgap) and the database of SNPs (dbSNP) (http://www.ncbi.nlm.nih.gov/SNP/), which are both managed by the National Center for Biotechnology Information.

GWAS TEST THE COMMON DISEASE/COMMON VARIANT HYPOTHESIS

Although there is no hypothesis for the disease gene location, the use of microchip technologies to perform a GWAS is still testing a specific hypothesis: the common disease/common variant hypothesis. This is the idea that polymorphic variation in the population of more than ~5% frequency might increase susceptibility to common diseases (Lander 1996; Cargill et al. 1999; Chakravarti 1999; Reich and Lander 2001). Such variants would be in an "evolutionary shadow" either because the effect is only a little deleterious or because the diseases only affect individuals who have reached old age, something that was unusual until relatively recently in human history. Even though the relative risk imparted by such a variant might be very small, because it is common, such variation would have a significant effect on public health and disease frequency. The combination of many such variants and the effect of environmental factors would combine to cause disease. We now know that at least for several common diseases, this hypothesis appears to be true. The alternative hypothesis—the rare variant hypothesis—is not disproved even in these cases, however. This hypothesis states that low-

Adapted from *Genetics of Complex Human Diseases: A Laboratory Manual* (ed. Al-Chalabi and Almasy). CSHL Press, Cold Spring Harbor, NY, USA, 2009.
Cite as: Cold Spring Harb Protoc; 2009; doi:10.1101/pdb.top66

penetrance rare variants (<1% population frequency) impart a moderately large relative risk and are responsible for disease (Pritchard 2001).

TAG SNPs AND LINKAGE DISEQUILIBRIUM ARE THE BASIS OF GWAS

Tag SNPs

To use the common disease/common variant hypothesis to test for association still means that every common variant needs to be typed, which is no small undertaking given that the HapMap currently records more than 3.1 million common SNPs. One solution is to use the highly correlated structure of the genome to type a subset of SNPs that capture the variation in the untyped SNPs, thus greatly reducing the amount of work needed (Johnson et al. 2001). This is the reason for the HapMap.

A typed SNP with a genotype correlating with that of other, untyped SNPs is said to "tag" the information in the untyped SNPs. For example, take two loci, A and B, at each of which there is a SNP, and that the SNPs at these loci have strongly correlating genotypes. The genotype at SNP A predicts the genotype at SNP B, and therefore the genotype at SNP B can be estimated with a high degree of certainty by genotyping SNP A. Thus, SNP A is a tag SNP for SNP B. This relationship between SNPs is known as linkage disequilibrium (LD), and it can be estimated by comparison between the allele frequencies observed at the two loci and the haplotype frequencies observed. If there were no correlation between the genotypes at the two SNPs, the haplotype frequency would be the same as the product of the respective allele frequencies. The degree to which the haplotype frequency deviates from the product of the allele frequencies is a measure of LD.

Linkage Disequilibrium

For example, denoting alleles 1 and 2 at each locus as a subscript and signifying the haplotype frequencies with *H*, we can write the following table:

Haplotype	Observed frequency
A_1B_1	H_{11}
A_1B_2	H_{12}
A_2B_1	H_{21}
A_2B_2	H_{22}

The allele frequencies observed are denoted as follows:

Allele	Observed frequency
A_1	p_1
A_2	p_2
B_1	q_1
B_2	q_2

If we use the observed allele frequencies and assume that the alleles at each SNP are independent, we can calculate what the haplotype frequencies should be:

Haplotype	Calculated frequency
A_1B_1	$p_1 q_1$
A_1B_2	$p_1 q_2$
A_2B_1	$p_2 q_1$
A_2B_2	$p_2 q_2$

The difference between the calculated haplotype frequency, assuming independence, and that actually observed is a measure of the independence between the two loci and is usually denoted with a capital *D* as follows:

Haplotype	Observed frequency	Calculated frequency	*D*
A_1B_1	H_{11}	p_1q_1	$H_{11} - p_1q_1$
A_1B_2	H_{12}	p_1q_2	$p_1q_2 - H_{12}$
A_2B_1	H_{21}	p_2q_1	$p_2q_1 - H_{21}$
A_2B_2	H_{22}	p_2q_2	$H_{22} - p_2q_2$

or alternatively:

	A_1	A_2	Total
B_1	$H_{11} = p_1q_1 + D$	$H_{21} = p_2q_1 - D$	q_1
B_2	$H_{12} = p_1q_2 - D$	$H_{22} = p_2q_2 + D$	q_2
Total	p_1	p_2	1

Although it is easy to calculate D, it is not a very intuitive measure of LD because it depends on the underlying allele frequencies and is maximal when the allele frequencies are each 0.5. Therefore, it is commonly standardized by dividing by the theoretical maximum when $D \geq 0$ to give $D' = D/D_{max}$, and the minimum when $D \leq 0$ to give $D' = D/D_{min}$. D_{max} is given by the smaller of the products p_1q_2 and p_2q_1, and D_{min} by the larger of the two products p_1q_1 and p_2q_2. The advantage of D' is that it is easily understood, with large absolute values implying strong LD between the two loci. D' is therefore a measure of the association of the two loci. This is not quite the same as the ability of one locus to predict the genotype at the other, however, as this is maximal when the allele frequencies at the two loci are the same. This is derived by taking the square of D' and standardizing by all allele frequencies as follows: $r^2 = D^2/p_1q_1p_2q_2$. Measuring LD using r^2 is also intuitive, with large values indicating high LD. The advantage of r^2 over D' is that it has useful statistical properties. For example, the change in sample size for a given power resulting from genotyping a tag SNP rather than the relevant causal SNP can be easily calculated by dividing the sample size by r^2. LD relationships between SNPs can be visualized using a matrix as shown in Figure 1.

Ancestral Mutation and Haplotypes

We can understand how LD is generated by considering how SNPs arise in the population. Imagine two loci, A and B, with no genetic variation. If a point mutation arises at locus A, then there are two types of haplotypes in the population: A_1-B and A_2-B. Now imagine that there is a second point mutation at site B. This must occur either in an A_1-B haplotype or an A_2-B haplotype. Let us imagine it has arisen on an A_2-B haplotype, generating a B_2 allele and designating the original allele as B_1. This means there are only three haplotypes in existence: A_1-B_1, A_2-B_1, and A_2-B_2. There is no A_1-B_2 at first because the B_2 allele arose on an A_2 background, and the identification of a B_2 reliably predicts an A_2 at the first locus, whereas an A_1 at the first locus reliably predicts a B_1 at the second. On the other hand, B_1 does not reliably predict the allele at A, and A_2 does not reliably predict the allele at B. This is a situation in which D' is 1 but r^2 is not. This complete LD gradually decays with time because of recombination between loci A and B, so that at some point it is likely that A_1-B_2 will be generated. It can also decay through mutation back to the original allele.

The HapMap records the genotypes at each locus in a selection of individuals of different ancestries representing large swaths of the worldwide population and allows the genotype correlations between SNPs to be analyzed. This information is used by chip manufacturers to generate DNA

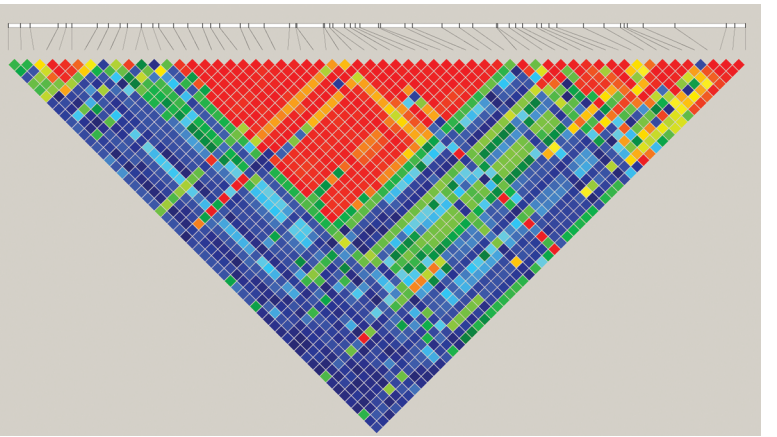

FIGURE 1. A triangle plot showing the LD relationships of SNPs as visualized using Haploview software. The strength of LD between any pair of SNPs is represented by coloring, with red being strong LD and blue weak. Plots can show any measure of LD, including r^2, D', LOD, and χ^2. (For color figure, see doi: 10.1101/pdb.top66 online at www.cshprotocols.org.)

microarrays that probe for tag SNPs preferentially, allowing, for example, a chip with just 300,000 SNPs to capture ~90% of the common variation in the genome. To put this in context, one needs to remember that, for this chip, there are 10,000 ungenotyped base pairs of genome for every SNP tested.

CHIPS AND PLATFORMS USED FOR GWAS

The two most common DNA microarray genotyping platforms currently in use are those manufactured by Affymetrix and Illumina, but several other companies use similar or identical technology, including NimbleGen and Perlegen. Each uses the concept of a chip containing probes that hybridize to the SNP of interest. The hybridization results in a fluorescent signal that indicates which allele is present and may include probes for specificity of binding as an error check. The Affymetrix chips use SNPs defined according to their proximity to various restriction enzyme cutting sites. The Illumina chips use SNPs preferentially selected in exons. The latest chips for both platforms also include SNPs in regions with variation in copy number and can assay more than 500,000 SNPs.

The fluorescence at each probe is analyzed using a clustering algorithm, and the most likely genotype is reported by the software. Such an algorithm can be confused by some resulting patterns of fluorescence, and for this reason, it is wise to look at the actual raw output of the scanner for SNPs of interest, although this may not be possible if the work was outsourced or is public genotype data. In addition, in some cases, the unusual patterns of fluorescence arise because of copy-number variation at the locus, and direct analysis of the raw data allows an estimate of the copy number. Similarly, a rare minor allele (one with less than approximately 20-30 copies in the analysis) may not yield enough data points for the clustering algorithm to reliably call the genotype. SNPs this rare should probably be discarded from the analysis.

HOW TO PREPARE FOR GWAS ANALYSIS

Data Handling

The large scale of GWAS leads to potential problems with false-positive results because of multiple testing; false-negative results because of the much more stringent *P*-values, and, therefore, large sample sizes required; and the need for strict quality control to avoid multiplying up genotyping and other possible sources of error (NCI-NHGRI Working Group on Replication in Association Studies 2007). See Figure 2 for a summary of data handling and quality control.

The standard GWAS design is a case-control study in which controls are matched for ancestry and, if relevant, age and sex. Power calculations need to take into account the multiple testing inherent in analyzing millions of SNPs. Because the SNPs used are mainly tag SNPs, they theoretically do not correlate with each other, only with the untyped SNPs, and a strict multiple testing correction needs to be applied (e.g., Bonferroni correction) (Bonferroni 1936). As a result, unless the signal to be detected is large and there is little allelic or disease heterogeneity, thousands or tens of thousands of samples need to be analyzed to detect an effect. This means that there may be, for example, a million genotypes for 10,000 people, making 2 billion data points. This is a significant data-handling problem in terms of manipulation of the data for analysis and tracking of the results. In addition, tracking 10,000 DNA samples and their associated clinical and genotype data for the purposes of quality control is not trivial. Several software packages exist to handle these large data sets successfully, the most popular of which is PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml) (Purcell et al. 2007). Others include an R package, snpMatrix (http://www.bioconductor.org/packages/2.3/bioc/html/snpMatrix.html) (Clayton and Leung 2007), PBAT (http://www.biostat.harvard.edu/~clange/default.htm) (Lange et al. 2004; Van Steen and Lange 2005), SNPTEST (http://www.stats.ox.ac.uk/~marchini/software/gwas/snptest.html) (Marchini et al. 2007), and EIGENSTRAT/EIGENSOFT (http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm) (Patterson et al. 2006; Price et al. 2006). Haploview 4.0 (http://www.broadinstitute.org/mpg/haploview) (Barrett et al. 2005), which allows visualization of haplotype blocks in the genome and analysis of local regions for tagging and association, also allows the import of GWAS results from PLINK for reordering and plotting.

Quality Control

The first steps in analysis are quality control steps. Perhaps the most important quality control step is ensuring that the phenotype data are robust, because a problem here will make the subsequent

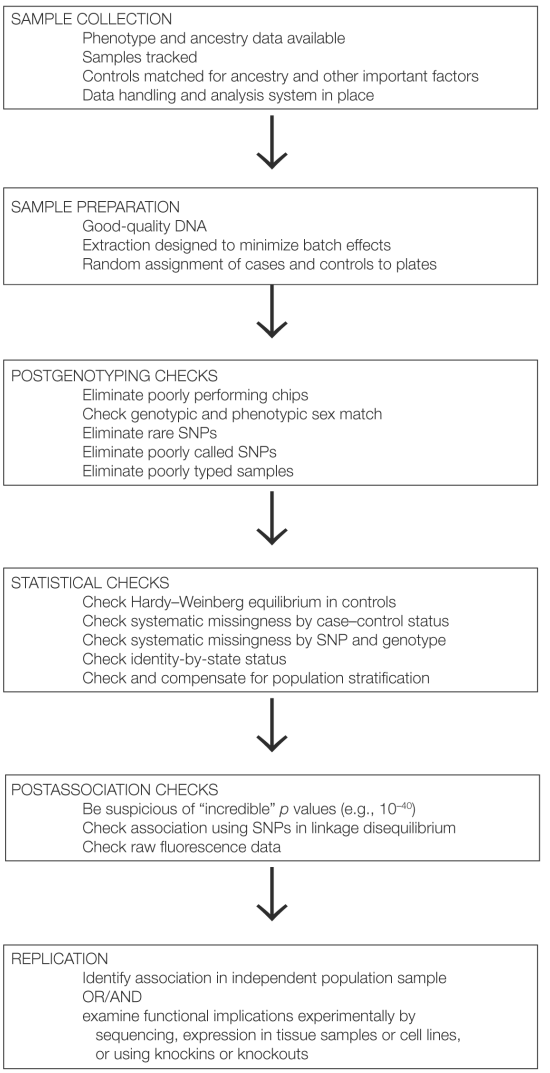

FIGURE 2. A summary of data handling and quality control considerations. Because of the large scale of GWAS, data considerations that do not apply to single candidate gene studies are important.

analyses meaningless. DNA samples should ideally be tracked using an automated tracking system from donation to analysis to reduce the risk of clerical errors. The use of bar codes is a common method. Good-quality DNA should be used where possible, and cases and controls should be drawn, tracked, and extracted in the same way, ideally at the same time and in the same center. These procedures reduce the risk of a systematic error affecting cases differently from controls because of handling. Plates used for genotyping should contain a random mixture of cases and controls in the same plate, again reducing the chance of a systematic error.

Once genotyping has occurred, chip call quality measures generated by the chip-scanning software should be used to filter out SNPs that are likely to be unreliable, and simple tests used to reduce the possibility of a genotype assigned to the wrong person. For example, the sex of each person should match with the genetic sex predicted by examining homozygosity on the X chromosome. Individuals for whom large numbers of SNPs could not be genotyped should be eliminated, because this suggests a problem with the DNA sample. Similarly, SNPs that failed to be genotyped in large numbers of individuals should be eliminated because this suggests that the SNP discovery probe was not reliable.

Genotypes can be analyzed to flag likely genotyping problems. Hardy-Weinberg equilibrium (HWE) is expected in controls (Hardy 1908; Weinberg 1908). It is common practice to use deviation from HWE as a measure of genotyping error, because the assumption is that heterozygotes are difficult to identify and homozygotes are therefore overrepresented (Hosking et al. 2004). Although, theoretically, deviation from HWE is not a good measure of genotyping error (Zou and Donner 2006),

it has proven itself as a reasonable quality control metric in practice. However, because loss of HWE can occur for reasons other than genotyping error, such as population stratification or true association, SNPs that otherwise pass quality control should probably be reevaluated. Alternatively, a very stringent HWE threshold can be used, such as $p = 10^{-6}$. A lack of HWE in cases is not a problem and may even suggest association, unless the cases and controls were ascertained or genotyped on separate occasions, in which case it is possible that there is a systematic problem arising from the laboratory work on one day.

Such systematic "missingness," in which a genotype is missing in cases more frequently than controls or vice versa, can be looked for using, for example, PLINK. SNPs for which this is the case should be eliminated. Another more subtle form of systematic missingness is when one genotype is more difficult to call than others. For example, if heterozygotes are more likely to be missing, this may not show up on a Hardy-Weinberg test, but can be detected using flanking SNPs in LD to predict the missing genotype. If, for example, every time the flanking SNPs predict heterozygosity in the test SNP the genotype at the test SNP is missing, this suggests a systematic problem with calling heterozygotes, and those SNP results need to be treated with caution. Again, this is implemented in PLINK.

Finally, identity-by-state (IBS) measures should be used to flag duplicate samples (or identical twins) or related individuals using genome-wide SNP data. It is possible to prune out SNPs in LD without affecting the reliability of the results, thus reducing the data set and the time taken to perform the IBS analysis, which otherwise can take hours to days. (This data pruning is purely for generating an IBS file, and the SNPs should be restored for further analysis.) Genetic relatedness is ~0.5 for non-identical siblings and parent-child relationships; 0.25 for half-siblings, avuncular relationships, and grandparents; and 0.125 for first cousins. The statistical measure of relatedness is symbolized by $\hat{\pi}$. The analysis should reveal individuals with a high value for $\hat{\pi}$ easily, and it is often clear that there is a set of samples with $\hat{\pi}$ almost exactly 1, 0.5, 0.25, and 0.125, representing duplicate samples and the differing degrees of relatedness. The remaining samples may still show high degrees of relatedness, however, and a reasonable background relatedness can be taken as $\hat{\pi} = 0.05$ (although in founder populations this would obviously be higher).

Population Stratification

An extremely important element of quality control is ensuring the cases are well matched to the controls for ancestry. This is because hidden variables in an association analysis can cause false-positive results or even reverse the direction of the association, a problem known as Simpson's paradox (Simpson 1951). This can be illustrated with the following example: Imagine 1750 cases and 1800 controls. We genotype for a SNP, and the allele counts look like this:

	Allele 1	Allele 2
Cases	3000	500
Controls	2700	900

A quick calculation shows an odds ratio of 2.00 [i.e., (3000 × 900)/(500 × 2700)], with $p < 10^{-6}$. In other words, we find that allele 1 doubles the odds of being a case. Now we discover that there are really two populations here, one in which allele 1 is relatively common, and the other in which it is less common. If we now split the table to reflect these two populations, we find the following:

	Allele 1	Allele 2
Cases	2800	300
Controls	1100	50

	Allele 1	Allele 2
Cases	200	200
Controls	1600	850

The odds ratio is now 0.42, $p < 10^{-6}$ for the first population and 0.53, $p < 10^{-6}$ for the second population (you can check for yourself that the counts in both populations add up to the same as those in the original, combined group). In other words, the effect in both underlying populations is actually that allele 1 halves the odds of being a case (i.e., allele 2 is associated with being a case rather than allele 1). This example, in which association in one direction in the underlying populations is converted to association in the other direction when they are combined, is, of course, extreme, but less

extreme versions simply convert lack of association in the underlying populations into association in the combined population or destroy a true association. Even if we have collected our samples as diligently as possible, it is still possible to have an underlying population structure capable of distorting the results. The problem of population stratification is not confined to case-control designs and can also affect quantitative trait GWAS—for example, if the two populations discussed here, instead of having different frequencies of cases, had different mean trait levels. Fortunately, there are methods that can deal with this.

Correction for Population Stratification

A relatively simple correction method is to use "genomic control" (Devlin and Roeder 1999). With this method, an inflation factor is calculated from loci not relevant to the disease to correct the χ^2 statistics globally. When we test the hypothesis that there is no underlying difference between the cases and controls in a GWAS, for nearly all our tests this will be true. In other words, it will be as if we are testing association in two sample sets randomly drawn from the same population. In this situation, any test statistic value is as likely as any other. This means that if we were to rank our test statistics by size and plot the expected values against the observed values (a so-called quantile-quantile, or Q-Q, plot), we should have a straight line from the origin to (1, 1). If there is a systematic problem in our data, such as an underlying population structure, this has the effect of inflating (usually) the test statistic and the Q-Q plot deviates upward from the (0, 0)(1, 1) line. The correction factor λ_{GC} to revert the line back to the correct position is calculated using the ratio between the Cochran-Armitage trend test statistic and the χ^2 statistic for an allelic test of association as shown below. The counts are denoted as follows:

Alleles	aa	Aa	AA	Total
Case	r_0	r_1	r_2	R
Control	s_0	s_1	s_2	S
Total	n_0	n_1	n_2	N

The trend test is

$$Y^2 = \frac{N\left[N(r_1 + 2r_2) - R(n_1 + 2n_2)\right]^2}{N(n_1 + 4n_2) - (n_1 + 2n_2)^2},$$

and the χ^2 statistic for allelic association is

$$\chi^2 = \frac{2N\left[2N(r_1 + 2r_2) - R(n_1 + 2n_2)\right]^2}{4R(N - R)\left[2N(n_1 + 2n_2) - (n_1 + 2n_2)^2\right]},$$

with the two statistics being approximately equal in a population in HWE. The trend test statistic is inflated by stratification by a factor λ_{GC} so that $Y^2 \sim \lambda_{GC}\chi^2$. Another way to estimate λ_{GC} is to use the median or the mean of the trend statistics, which should be 1.0 if there is no inflation.

An alternative method is to use genome-wide data, or a subset of ancestry informative SNPs, to analyze the population structure. There are various methods to do this implemented in programs such as *structure* (http://pritch.bsd.uchicago.edu/software.html) (Pritchard et al. 2000), EIGENSTRAT (http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm) (Price et al. 2006), and PLINK (http://pngu.mgh.harvard.edu/~purcell/plink/download.shtml) ((Purcell et al. 2007). The first step is usually to dissect out the ancestry of each individual using some sort of data reduction method such as principal components analysis. It is then relatively straightforward to identify groups of individuals with the same ancestry. If there are only a few individuals with a different ancestry, these can be eliminated from the analysis. If there are many, the analysis can be stratified by two or more groups. Alternatively, the first two or more components of the ancestry information can be used as covariates in an analysis, thus effectively accounting for the population structure.

DATA ANALYSIS METHODS USED IN GWAS

The most basic test of association for a discrete trait, such as disease affection, is a χ^2 test. There are six genetic models that could be tested at each SNP. These are allelic (or multiplicative) in which allele counts are used to make a 2 × 2 table; additive in which genotypes are counted in a 3 × 2 table and

the null hypothesis is that no genotype is associated with disease making a two degrees of freedom (df) test; a trend to test AA > Aa > aa (usually the Cochran-Armitage trend test is used) in which the genotypes are counted in a 3 × 2 table but the test is a 1-df test for a dose effect; a dominant model (AA and Aa vs. aa); a recessive model (AA vs. Aa and aa); and an overdominant model (also called heterozygous advantage) in which Aa is compared with AA and aa. If the samples are from different centers or there is some other reason for the data to be analyzed in strata, a Cochran/Mantel-Haenszel χ^2 test can be used in which a 2 × 2 × *k* table is tested, where *k* is the number of strata. This will correctly handle confounding by a categorical variable such as ancestral group.

Regression methods can also be used. For example, discrete traits can be analyzed by logistic regression, the advantage being that other variables may be used as covariates. Situations in which there is no control category because quantitative traits such as blood pressure, cholesterol level, or IQ are being investigated can be tested by linear regression. Again, covariates may be included in the model. Similarly, association testing can be used in case-only studies to examine disease modifier genes. For example, age of onset for diabetes could be studied by regressing the age of onset against the genotype.

Complex disease genes under the common disease common variant hypothesis generally have small effect sizes, with odds ratios of the order of 1.3 or smaller (Bodmer and Bonilla 2008). Because the number of tests is large, *P*-values of the order of 10^{-7} are needed to guard against large numbers of false-positive associations from the statistical noise generated by multiple testing. This, in turn, requires large numbers of people to be studied to have sufficient power. There are various statistical strategies to help overcome this multiple testing problem, but there are also some genetic methods that can help to increase confidence in a finding as real. The simplest is to use a gene-based test of association. This reduces the number of independent tests from hundreds of thousands to tens of thousands. One such test is Hotelling's T^2 test in which all the SNPs in a gene are considered jointly for analysis (Hotelling 1931). A second strategy is to investigate a putative association by examining association using surrounding SNPs. This is implemented in the program PLINK as the proxy-haplotype test. In this procedure, three SNPs on either side of the associated variant are used to form a seven-SNP haplotype. Association is then tested using the haplotype without the associated variant. If the association is not caused by some genotyping error at the associated SNP, it should hold. In addition, nearby SNPs in LD with the associated SNP should also show association, thus further increasing confidence that the association is technically real. Finally, if the association with the surrounding haplotype is stronger than that with the associated SNP, there may be a causal SNP that is untyped on the associated haplotype.

A third strategy is to use permutation testing. Permutation of case-control status can be used to show that the association at a SNP is not something that is likely to have occurred by chance. This is the most robust method of dealing with multiple testing as the probability distribution is derived empirically from the data, but it has the disadvantage of being computationally intensive. More recently, methods in which pathways or networks of genes are analyzed as a unit have been developed as a means of pulling out a likely signal (Wang et al. 2007). Even though the multiple testing burden remains high, the finding that more genes in a particular pathway or network tend to be associated than would be expected by chance increases the confidence with which it can be said that the pathway is relevant. Finally, replication of the association in a second independent sample is the ideal confirmation. This is most powerfully analyzed jointly with the original data (Skol et al. 2006).

HAPLOTYPE ANALYSES CAN LOCATE A FUNCTIONAL VARIANT

Identifying Haplotypes

When an association is identified, it can be useful to examine haplotype associations that include the SNP of interest as a means to locate a functional variant. As described in the previous section on "Ancestral Mutation and Haplotypes," the association between a mutant allele and the haplotype upon which it arose can only be destroyed through recombination or (less commonly) mutation. Thus, the implication of a strong haplotypic association is that the true causal variant lies on the associated haplotype background and may not have been typed. This is similar to the outcome from imputation, which also uses local LD patterns to attempt to locate a causal variant, the main difference being that imputation uses a reference set of LD data such as the HapMap to impute missing data and therefore the genotype at the untyped SNP. In a family-based study, a haplotype is a segment of chromosome between two crossover points. In a population-based study, the crossover points

are unknown because previous generations are not genotyped, so a likely haplotype has to be estimated. Haplotypes can be assigned to an individual using a systematic method such as the expectation-maximization (E-M) algorithm, which uses existing data to estimate missing data (Fig. 3; Dempster et al. 1977). Some haplotypes can be identified with certainty. For example, if we take two loci, A and B, with alleles A, a, B, and b, and denote separate chromosomes with a slash, a doubly homozygous individual A/A and B/B must have two AB haplotypes. Similarly, someone who is A/A and B/b must have one AB and one Ab haplotype. In the case of a double heterozygote A/a and B/b, there are two equally likely possibilities, AB/ab and Ab/aB. Given a series of genotypes, we can estimate the likely phase of the uncertain haplotypes using information from the known haplotypes.

The E-M Algorithm

We start by assigning weighted counts to the haplotypes (Fig. 3). A haplotype that can be identified with certainty is given a weight of 1. A haplotype that is of uncertain phase is assigned an estimated weight based on, for example, the prior probability that the haplotype could exist. This is the "expectation" step. The number of people with each haplotype can now be counted (including fractions for partial weights), generating the haplotype frequencies. This is the "maximization" step. The new haplotype frequencies can now be used to weight the counts again, and the counts can again be used to modify the haplotype frequencies. This process continues iteratively until there is no change in the result. Thus, a likely haplotype phase can be assigned to an individual using the information from individuals with known phase.

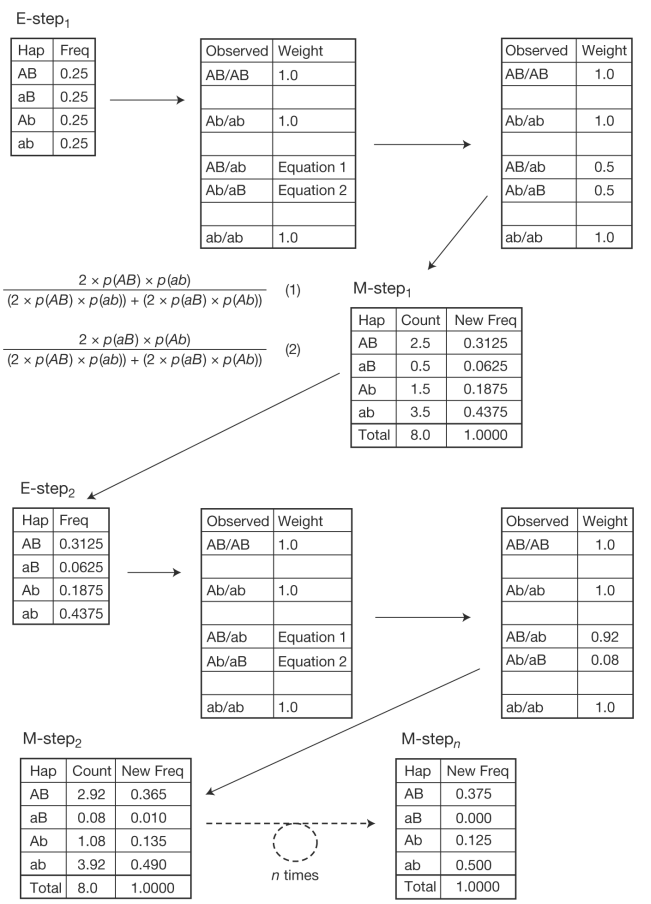

FIGURE 3. The E-M algorithm. Information from the haplotypes that are known with certainty is used to inform the likely haplotypes with ambiguous phase. The expectation step is used to calculate likely weightings (using Equations 1 and 2, which generate the probabilities over all possibilities), and these are used to count the likely haplotypes in the maximization step. This generates new frequencies that are used to calculate the new weights and so on. In this case, the AB/ab haplotype is chosen as the correct phase and the aB haplotype does not exist. (Example courtesy of Shaun Purcell.)

Haplotype Blocks

Increasing numbers of markers can be included in the haplotype. The point at which it is reasonable to say one haplotype ends and another begins is largely a matter of opinion, but there are various techniques to try to define haplotype blocks objectively. For example, one method is to use measures of D' between SNP alleles to define where ancestral recombinations have occurred and therefore where a haplotype block ends (Gabriel et al. 2002). Another method is to search for haplotypes in which the fourth haplotype (the A_1-B_2 haplotype in the example from the earlier section on "Ancestral Mutation and Haplotypes") is missing, because this implies that no recombination has occurred since the mutation arose (Wang et al. 2002). For analysis purposes, another strategy is to use a sliding window of, say, three SNPs. Because of the very large number of ways that haplotypes can be made and tested, permutation testing is particularly important in deciding if a haplotypic association is significant.

GWAS HAVE PRODUCED SOME KEY RESULTS

Age-Related Macular Degeneration

The first association reported from a GWAS was for age-related macular degeneration and complement factor H. Three independent studies found an association with what is, in retrospect, a surprisingly large effect size (Edwards et al. 2005; Haines et al. 2005; Klein et al. 2005). In one of the studies, 96 cases and 50 controls were sufficient to find the associated SNP with a P-value $< 10^{-7}$. The 95% confidence limits for the odds ratio for the causal variant in one study was between ~2.45 and 5.57, which is extremely high for a complex disease and is the reason such a small study could detect the signal.

Wellcome Trust Case Control Consortium

The Wellcome Trust Case Control Consortium (WTCCC) paper is a landmark because the large numbers involved, the validation of the GWAS approach, and the public health interest of the several diseases studied have provided many lessons (Wellcome Trust Case Control Consortium 2007). Despite the very large numbers of individuals studied, no convincing signal was seen for hypertension, probably because of disease heterogeneity. The diseases in which an association was found had SNPs with odds ratios of the order of 1.3 or smaller, confirming the findings of other studies that complex diseases need large sample sizes to have the power to detect the effect sizes seen under the common disease/common variant hypothesis. Imputation was used to estimate the likely genotypes at the untyped SNPs present in the HapMap, resulting in 2.5 million genotypes per person (Marchini et al. 2007). The study design also showed that a "universal control" strategy, in which one well-characterized set of controls can be used for several studies, could be successful. The generation of so many genotypes from the United Kingdom population also helped confirm that if people with non-European ancestry are excluded from the sample, there is only modest population stratification in the British population.

Type 1 Diabetes

In the original WTCCC study of 5000 people and a partly overlapping study of nonsynonymous SNPs in 7000 people, six chromosomal regions were associated with type 1 diabetes (T1D). Those six and a further six top hits were followed up in 18,000 more people in a second study (Todd et al. 2007). In total, four of these loci were convincingly associated with T1D, possible only because of the huge numbers of people studied, enabling small effect sizes to be seen above the noise of multiple testing. Eight were regarded as a small effect or false positive. These gigantic sample sets were also used to explore the proportion of T1D that is genetic and environmental and to shed light on the T1D metabolic pathway. They were also used to study geographical changes in allele frequency and to look for gene-gene interaction by searching for deviation from a multiplicative model.

MY STUDY IS NEGATIVE—HELP!

The difficulties non-geneticists have with leaving behind the notion of one gene, one disease and Mendelian genetics are compounded when GWAS are involved, as the misunderstanding of many

people is that "gene chips" effectively assay the genome for mutations. The interpretation of a GWAS in which there is no significant association is that the disease does not have a significant genetic component, rather than the far more likely interpretations that the study was underpowered, rare variants were involved, copy-number variants were involved, other difficult-to-tag variations such as microsatellites were responsible, or gene-gene interactions were important. Thus, it is vital to reflect on the hypothesis that is being tested with the typical GWAS, which is the common disease/common variant hypothesis, and to reinforce that even a well-powered negative study simply means that this hypothesis is likely to be false, but says nothing about rare variants or the other variation described because these cannot be tested with tag SNPs.

OTHER STRATEGIES ARE IMPORTANT

GWAS currently consist of a search for common variants predisposing to disease. At least two other strategies are becoming more important—a search for structural variants (copy-number variants) in the genome that might be more common in cases than controls and a search for rare variants that might be less common but have a larger effect size.

Copy-Number Variants

Structural variation in the genome is now regarded as an important cause of phenotypic variation (Iafrate et al. 2004; Sebat et al. 2004). This is defined as a region of duplication or deletion greater than 1 kb, and such loci are regarded as variant in copy number. There are several methods for analysis of copy-number variants. Recently, an integrated approach has been developed using a software package, Birdsuite, which can test for association of copy-number variants using the PLINK software package (http://www.broadinstitute.org/science/programs/medical-and-population-genetics/birdsuite/birdsuite-0) (Korn et al. 2008). The best methods for interpreting the contribution of copy-number variants to disease are still being debated but are likely to be resolved soon.

Rare Variants

Genetic sequence variations that are of <1% frequency are likely to be neutral or deleterious, but they may take some time to be eliminated from the population if only mildly deleterious. In fact, statistical theory suggests that rare variants are likely to be a common component of the genetic contribution to disease (Bodmer and Bonilla 2008). Whereas the odds ratio for common variants is typically ~1.3 or less, the odds ratio for rare variants based on an examination of the current literature is likely to be 2 or more, with a current average of 3.84. Rare variants do not need to cause familial clustering. Because the variants we are interested in do not cause Mendelian disease by definition (we are interested in complex diseases), any one variant must have reduced penetrance. We can use the binomial distribution to estimate the penetrance that would produce a sufficiently low familial rate. For example, given a dominant model (because the variant is rare, a homozygote is unlikely) and a sibship size of 4, a penetrance of 0.2 means there is only a 0.05 probability of more than one affected in the sibship. Even a penetrance of 0.5 means that only one-quarter of such sibships would have more than one affected. If the sibship is size 3, then the probability drops to ~1 in 6, and in today's smaller families, only 1 in 16 sibships of size 2 would have both affected.

There are several striking consequences of this observation. The first is that rare variants with even quite high penetrance could be responsible for complex diseases. Second, as family sizes reduce, more diseases will appear sporadic, even if the variant is highly penetrant. For example, assuming a dominant mode of inheritance and 90% penetrance, 80% of sibships of size 2 would have only one affected (although, of course, in this example, one parent would be very likely affected). Third, many diseases have a familial and sporadic form. For example, in many neurological diseases, ~10% of cases are familial. This could be accounted for by a gene with ~30% penetrance because many of the gene carriers would not be manifesting.

Genetic technology is now moving toward whole-genome sequencing at a rapid rate. The current generation of sequencers is capable of sequencing the entire coding sequence (55 Mb) in ~1 wk, and it is likely that in the next decade we will have access to huge quantities of sequence data that can be mined for rare variants. A χ^2chi-square test for association can be performed by collaps-

ing all variants in a single gene into one class, but the interpretation of which variants are causal and which are neutral will require bioinformatics and functional analyses on a large scale.

CONCLUSION

GWAS are relatively new and have already taught us many lessons in the design and interpretation of such studies and in the causes of complex traits. Larger data sets, cheaper technologies, and greater international cooperation will mean that smaller and smaller effect sizes will be dissected out over the next few years, greatly increasing our understanding of the causes of complex diseases.

REFERENCES

Barrett JC, Fry B, Maller J, Daly MJ. 2005. Haploview: Analysis and visualization of LD and haplotype maps. *Bioinformatics* **21:** 263–265.

Bodmer W, Bonilla C. 2008. Common and rare variants in multifactorial susceptibility to common diseases. *Nat Genet* **40:** 695–701.

Bonferroni CE. 1936. Teoria statistica delle classi e calcolo delle probabilità. *Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze* **8:** 3–62.

Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane CR, Lim EP, Kalyanaraman N, et al. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* **22:** 231–238.

Chakravarti A. 1999. Population genetics—making sense out of sequence. *Nat Genet* **21 (Suppl 1):** 56–60.

Clayton D, Leung HT. 2007. An R package for analysis of whole-genome association studies. *Hum Hered* **64:** 45–51.

Dempster AP, Laird NM, Rubin DB. 1977. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Statist Soc Ser B* **39:** 1–38.

Devlin B, Roeder K. 1999. Genomic control for association studies. *Biometrics* **55:** 997–1004.

Edwards AO, Ritter R III, Abel KJ, Manning A, Panhuysen C, Farrer LA. 2005. Complement factor H polymorphism and age-related macular degeneration. *Science* **308:** 421–424.

Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, et al. 2002. The structure of haplotype blocks in the human genome. *Science* **296:** 2225–2229.

Haines JL, Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, et al. 2005. Complement factor H variant increases the risk of age-related macular degeneration. *Science* **308:** 419–421.

Hardy GH. 1908. Mendelian proportions in a mixed population. *Science* **28:** 49–50.

Hosking L, Lumsden S, Lewis K, Yeo A, McCarthy L, Bansal A, Riley J, Purvis I, Xu CF. 2004. Detection of genotyping errors by Hardy-Weinberg equilibrium testing. *Eur J Hum Genet* **12:** 395–399.

Hotelling H. 1931. The generalization of Student's ratio. *Ann Math Statist* **2:** 360–378.

Iafrate AJ, Feuk L, Rivera MN, Listewnik ML, Donahoe PK, Qi Y, Scherer SW, Lee C. 2004. Detection of large-scale variation in the human genome. *Nat Genet* **36:** 949–951.

The International HapMap Consortium. 2003. The International HapMap Project. *Nature* **426:** 789–796.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

Johnson GC, Esposito L, Barratt BJ, Smith AN, Heward J, Di Genova G, Ueda H, Cordell HJ, Eaves IA, Dudbridge F, et al. 2001. Haplotype tagging for the identification of common disease genes. *Nat Genet* **29:** 233–237.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D. 2002. The Human Genome Browser at UCSC. *Genome Res* **12:** 996–1006.

Klein RJ, Zeiss C, Chew EY, Tsai JY, Sackler RS, Haynes C, Henning AK, SanGiovanni JP, Mane SM, Mayne ST, et al. 2005. Complement factor H polymorphism in age-related macular degeneration. *Science* **308:** 385–389.

Korn JM, Kuruvilla FG, McCarroll SA, Wysoker A, Nemesh J, Cawley S, Hubbell E, Veitch J, Collins PJ, Darvishi K, et al. 2008. Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nat Genet* **40:** 1253–1260.

Lander ES. 1996. The new genomics: Global views of biology. *Science* **274:** 536–539.

Lange C, DeMeo D, Silverman EK, Weiss ST, Laird NM. 2004. PBAT: Tools for family-based association studies. *Am J Hum Genet* **74:** 367–369.

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.

NCI-NHGRI Working Group on Replication in Association Studies. 2007. Replicating genotype-phenotype associations. What constitutes replication of a genotype-phenotype association, and how best can it be achieved? *Nature* **447:** 655–660.

Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet* **2:** e190 doi: 10.1371/journal.pgen.0020190.

Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet* **38:** 904–909.

Pritchard JK. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* **69:** 124–137.

Pritchard JK, Stephens M, Donnelly P. 2000. Inference of population structure using multilocus genotype data. *Genetics* **155:** 945–959.

Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, Maller J, Sklar P, de Bakker PI, Daly MJ, et al. 2007. PLINK: A tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet* **81:** 559–575.

Reich DE, Lander ES. 2001. On the allelic spectrum of human disease. *Trends Genet* **17:** 502–510.

Sebat J, Lakshmi B, Troge J, Alexander J, Young J, Lundin P, Maner S, Massa H, Walker M, Chi M, et al. 2004. Large-scale copy number polymorphism in the human genome. *Science* **305:** 525–528.

Simpson EH. 1951. The interpretation of interaction in contingency tables. *J Roy Statist Soc Ser B* **13:** 238–241.

Skol AD, Scott LJ, Abecasis GR, Boehnke M. 2006. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat Genet* **38:** 209–213.

Todd JA, Walker NM, Cooper JD, Smyth DJ, Downes K, Plagnol V, Bailey R, Nejentsev S, Field SF, Payne F, et al. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet* **39:** 857–864.

Van Steen K, Lange C. 2005. PBAT: A comprehensive software package for genome-wide association analysis of complex family-based studies. *Hum Genomics* **2:** 67–69.

Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, et al. 2001. The sequence of the human genome. *Science* **291:** 1304–1351.

Wang N, Akey JM, Zhang K, Chakraborty R, Jin L. 2002. Distribution of recombination crossovers and the origin of haplotype blocks: The interplay of population history, recombination, and mutation. *Am J Hum Genet* **71:** 1227–1234.

Wang K, Li M, Bucan M. 2007. Pathway-based approaches for analysis of genomewide association studies. *Am J Hum Genet* **81:** 1287–1283.

Weinberg W. 1908. Ber den Nachweis der Vererbung Beim Menchen. *Jahreshefte Verein* **64:** 368–382.

The Wellcome Trust Case Control Consortium. 2007. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447:** 661–678.

Zou GY, Donner A. 2006. The merits of testing Hardy-Weinberg equilibrium in the analysis of unmatched case-control data: A cautionary note. *Ann Hum Genet* **70:** 923–933.