Cold Spring Harbor Protocols

Introduction to Genetic Association Studies

Cathryn M. Lewis and Jo Knight

Cold Spring Harb Protoc 2012; doi: 10.1101/pdb.top068163

Topic Introduction

Introduction to Genetic Association Studies

Cathryn M. Lewis and Jo Knight

Genetic association studies are used to find candidate genes or genome regions that contribute to a specific disease by testing for a correlation between disease status and genetic variation. This article provides a broad outline of the design and analysis of such studies, focusing on case–control studies in candidate genes or regions.

INTRODUCTION

Genetic association studies test for a correlation between disease status and genetic variation to identify candidate genes or genome regions that contribute to a specific disease. A higher frequency of a single-nucleotide polymorphism (SNP) allele or genotype in a series of individuals affected with a disease can be interpreted as meaning that the tested variant increases the risk of a specific disease (although several other interpretations are also valid; see the following sections). SNPs are the most widely tested markers in association studies (and this term will be used throughout), but microsatellite markers, insertion/deletions, variable-number tandem repeats (VNTRs), and copy-number variants (CNVs) are also used.

Association studies are a major tool for identifying genes conferring susceptibility to complex disorders. These traits and diseases are termed “complex” because both genetic and environmental factors contribute to the susceptibility risk. Extensive experience in genetic studies for many complex disorders (such as diabetes, heart disease, autoimmune diseases, and psychiatric traits) confirms that many different genetic variants control disease risk, with each variant having only a subtle effect.

Associations with polymorphisms in candidate genes have been confirmed in many different diseases (Lohmueller et al. 2003), and genome-wide association studies (GWAS) are identifying many novel associations in genes that had not been strong a priori candidates for the disease under test (Wellcome Trust Case Control Consortium 2007). However, the modest increase in risk implies that large well-designed and analyzed studies are required to detect and confirm signals for association.

This article outlines the design and analysis of genetic association studies, but it focuses specifically on case–control studies in candidate genes or regions. Even in this era of genome-wide studies, case–control studies still form the majority of published reports. We illustrate the importance of quality control in performing these studies, describe basic analytical strategies for a SNP, and point the reader toward methods for analyzing haplotypes or multiple markers. We also highlight some of the pitfalls of performing powerful, accurate association studies and discuss how these challenges are reflected in the contradictory literature for many disease–gene investigations. In addition to GWAS, other approaches to genetic association studies include family-based association studies and quantitative trait locus studies; these approaches are not addressed in any detail here.

Adapted from *Genetics of Complex Human Diseases* (ed. Al-Chalabi and Almasy). CSHL Press, Cold Spring Harbor, NY, USA, 2009.

Cite this article as *Cold Spring Harbor Protoc*; 2012; doi:10.1101/pdb.top068163

INTERPRETING SIGNIFICANT GENETIC ASSOCIATION

Significant genetic association may be interpreted as either (1) direct association, in which the genotyped SNP is the true causal variant conferring disease susceptibility; (2) indirect association, in which a SNP in linkage disequilibrium (LD) with the true causal variant is genotyped; or (3) a false-positive result, in which there is either chance or systematic confounding, such as population stratification.

Distinguishing between direct and indirect association is challenging and may require resequencing of the candidate region, dense genotyping of all available SNPs, or functional studies to confirm the role of a putative mutation in disease.

FINDING DIRECT ASSOCIATION

Case–Control Study

The simplest study design used to test for association is the case–control study, in which a series of cases affected with the disease of interest are collected together with a series of control individuals. The specific choice of phenotype for the cases may define the exact hypothesis to be tested, and applying strict clinical criteria for ascertainment is necessary to ensure a homogeneous set of cases. Two standard methods are used for collecting controls: the use of either a series of individuals who have been screened as negative for presence of the disease or of controls randomly ascertained from the population, whose disease status is unknown. Both control sets form a valid test for association, and they will have similar power for a rare disease. For a more common disease, a study with screened unaffected controls (often termed "supernormal" controls) will have higher power to detect association compared with a study using population-based controls, and the increase in power is notable for diseases with high prevalence. For some diseases, screening controls for the presence or absence of the disease may be difficult, and using a larger sample of unscreened controls may be more efficient.

Statistical Analysis of Case–Control Study

The genotypes of a single, biallelic SNP on a set of cases and controls can be summarized in a 2×3 contingency table of the genotype counts for each group, as shown in Figure 1. For a SNP with alleles G and T, we tabulate the number of cases and controls with each genotype GG, GT, and TT. Several different statistical analysis methods can be applied to this table. We will focus here on goodness-of-fit tests, rather than likelihood-based or regression methods. Pearson's chi-square test is used to assess departure from the null hypothesis that case and controls have the same the distribution of genotype counts. This test statistic has a chi-square distribution with two degrees of freedom on this 2×3 table.

This approach provides a valid statistical analysis of the data presented, but uses no genetic information. We have illustrated the data on a table with genotypes ordered as GG, GT, TT, with the inherent supposition that disease risk may increase (or decrease) as the number of T alleles increases. However, column order is not used in the test statistic, and reordering the table as GG, TT, GT gives the same value of the test statistic and p value. Other analysis methods that correspond to the underlying genetic models we expect to be acting in complex diseases may be preferred. First, the table may be decomposed from genotypes into alleles, with cell counts of the number of G alleles, and the number of T alleles carried by cases and controls (regardless of the genotype combination in which these alleles were carried) (Fig. 1A, upper left). This test is valid under the null hypothesis of no association, or when the true model of association is multiplicative (or log additive), so the genotype relative risks for GG, GT, and TT genotypes can be modelled as 1, r, and r^2, with relative risk increasing by a factor r for each T allele carried (Sasieni 1997). An alternative test for this model is the Cochran–Armitage test for trend (CATT) (Fig. 1A, upper right), which, as its name implies, tests for a trend in differences in cases and controls across the ordered genotypes in the table. This test is asymptotically equivalent to the allele test described previously, although it is more robust to departures from

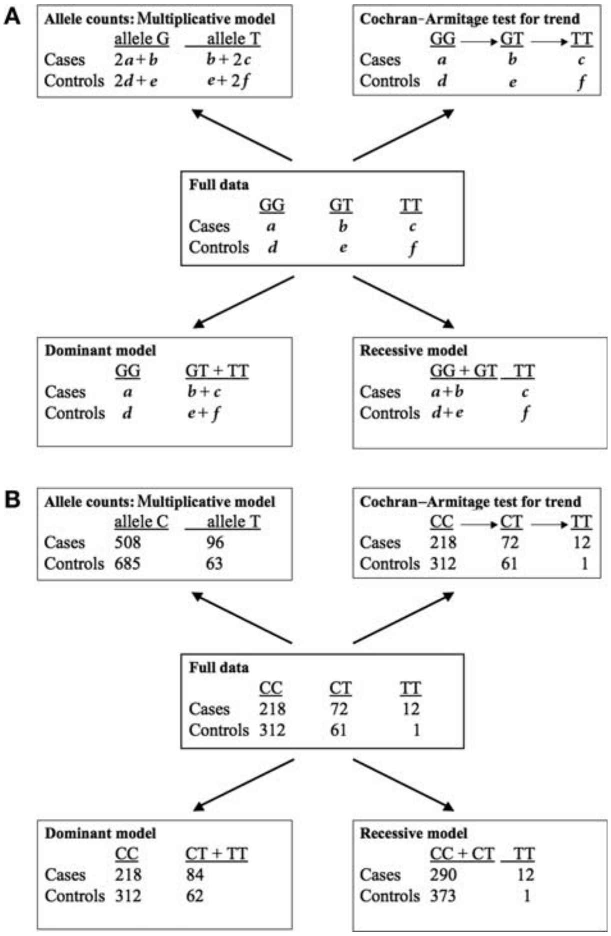

FIGURE 1. Analysis methods for single SNP association studies, testing under the assumption of specific genetic models for (*A*) arbitrary genotype counts and (*B*) the rheumatoid arthritis case–control study in Table 1.

Hardy–Weinberg equilibrium (HWE) (see the following) (Sasieni 1997). Further tests can be used to test specific genetic hypotheses—for example, that the SNP alleles increase disease risk under a dominant or a recessive model (Fig. 1A, bottom). Assuming T is a high-risk allele, these tests compare GG genotypes to CT + TT genotypes (dominant model), or CC + CT to TT genotypes (recessive model).

Although the tests described previously are all valid methods for analysis of an association study, any such study should have a prespecified analysis plan because applying all tests will increase the probability of a false-positive result. Candidate gene association studies most commonly test for a difference in allele frequency. The allele frequencies in cases and controls provide useful, direct summary statistics for the data. The CATT has become popular in GWAS (e.g., O'Donovan et al. 2008). Other test statistics, such as analyzing under a dominant or recessive model, may also be applied to ensure that interesting findings are not missed because of the specific analysis method used. These tests are rarely a primary analysis tool for complex genetic disorders, but may be used as secondary analyses to explore the potential mode of inheritance of an associated SNP, or to test a prespecified hypothesis. When several analysis methods are used, a correction for multiple testing should be applied. This is not straightforward owing to the correlation between test statistics, and simulation studies may be required.

Example of a Statistical Analysis

PTPN22 is associated with several autoimmune phenotypes, with the strongest association seen at R602W. Table 1 shows the genotypes at this variant (SNP rs2476601, C1858T) in a study of

TABLE 1. *PTPN22* C1858T genotypes for rheumatoid arthritis (RA) case–control study

		Genotypes			
Cohort	No. of individuals	CC	CT	TT	Frequency of allele T
RA cases	302	218 (72.2%)	72 (23.8%)	12 (4.0%)	15.9%
Controls	374	312 (83.4%)	61 (16.3%)	1 (0.3%)	8.4%
OR (95% CI)		1	1.69 (1.15–2.48)	17.17 (2.22–133.06)	

Data from Steer et al. (2005).

London rheumatoid arthritis (RA) cases and randomly ascertained controls (Steer et al. 2005). Genotypes for both cases and controls were in HWE, with p values of 0.060 and 0.267, respectively. The genotype counts show that cases have a higher frequency of both CT and TT genotypes compared with controls. In RA cases, the frequency of allele T (15.9%) is higher than in controls (8.4%). Analyzing the allele counts contingency table shows strong evidence for association at this SNP ($p =$ 0.00003) (Fig. 1B). Significant evidence for association is also found using the CATT ($p = 0.00004$) (Fig. 1B).

A summary measure of the effect of this SNP on risk for RA can be obtained through calculating odds ratios (ORs). These can be calculated separately for CT and TT genotypes by comparing each to the baseline CC genotype, which is most common in the population. For the CT genotype, the OR is the odds of the CT genotype compared with the CC genotype in cases, divided by the same quantity in controls: (72/218) / (61/312) = 1.69. Confidence intervals on the OR can be calculated using the Woolf method: the standard error of ln(OR) is approximately

$$\sqrt{\frac{1}{a}+\frac{1}{b}+\frac{1}{c}+\frac{1}{d}},$$

in which a, b, c, and d are the entries in the relevant genotype subtable. The OR, or genotype relative risk, for CT and TT genotypes compared with CC genotypes confirm that both these genotypes have an elevated risk of RA (because neither confidence interval contains 1), although the confidence interval for TT genotypes is very wide because only a single TT control individual is observed (Table 1). Examination of ORs (1.69 and 17.2) suggests that a gene dosage model is acting, with much higher risk in mutation homozygotes (TT) than the heterozygotes (CT). The increase in risk from CT to TT is higher than would be expected under a multiplicative model, in which estimates from the allele count table give an OR of 2.06 for each T allele carried, implying an approximately fourfold increase risk for TT individuals. However, for an association of this strength, analyzing the genotype counts assuming a multiplicative model still results in highly significant evidence of association.

Using Quantitative Measures

Some complex phenotypes, such as high blood pressure, height, and obesity, are better characterized by quantitative rather than qualitative measures. Several options are available for the analysis of such data. The quantitative measure can be tested for association in a linear regression framework, assessing whether the genotypes (as an explanatory variable) predict trait value. Similar to the analysis options described in case–control studies previously discussed, genotypes may be coded as a three-level factor, or as a count of T alleles carried (0, 1, 2), or as a dominant or recessive model. Quantitative measures may be analyzed in a case–control framework by dichotomizing the sample. However, this method may result in a loss of power because all information on the distance of an individual's observed phenotype from the dichotomizing threshold is lost. The power of a quantitative trait association study may be increased by ascertaining individuals only from the extremes of the distribution (Slatkin 1999).

FINDING INDIRECT ASSOCIATION

Linkage Disequilibrium

LD measures the correlation between SNP alleles at sites in the same region of the genome. Dependence between SNPs arises because any novel SNP (e.g., a change from base pair C A at a genomic site) occurs on a background of fixed alleles at other SNPs in the region. For example, for five flanking SNPs in each direction, the existing chromosomal haplotype may have been ACTCG-**C**-GGATC, which becomes ACTCG-**A**-GGATC. The A allele at this SNP is fully correlated with these flanking haplotypes, so that initially all copies of allele A have allele G at the neighboring SNP. As this chromosome is transmitted through the generations, the length of this haplotype is diminished by recombination, and different copies of the original A allele will have different recombination patterns, and be flanked by different lengths of DNA from the original chromosome. Figure 2 illustrates this process, in which a disease mutation D occurs in meiosis from generation 1 to generation 2, on a specific background chromosome. This mutation (if it is not lost to the population through nontransmission) is transmitted through the generations, with recombinations reducing the length of the original ancestral chromosome. However, all copies of this mutation D arising from the same mutation event will harbor some portion of the ancestral chromosome, with the length of retained chromosome depending on the pattern of recombination events through the generations.

LD is an important phenomenon in association testing because it induces correlation in short regions of the genome. In Figure 2, mutation D occurs close to a polymorphic marker bearing the M allele. In the current generation, most chromosomes carrying mutation D also carry allele M. Thus, we have two opportunities to detect association with the disease, by genotyping either M or D. Genotyping the true disease mutation D (direct association) should have higher power to detect association, but where M and D are in strong LD, and sample sizes are adequate, significant association should be detectable by genotyping M (indirect association).

Intuitively, LD measures the correlation between SNP alleles. Given a chromosome with a specific SNP allele, how does this influence the probability distribution of alleles carried at other SNPs within the same genetic region? Many different statistical measures to quantify LD between two SNPs have been proposed (Devlin and Risch 1995), with D' and r^2 being most widely used. The International HapMap Project is a valuable resource for study design, allowing researchers to investigate LD in a region and to select an informative subset of available SNPs to be genotyped in an association study (http://www.hapmap.org).

Analysis of Multiple Markers and Haplotypes

Although high-throughput genotyping has increased the number of SNPs it is feasible to genotype, studies still consider only a subset of available SNPs and test for indirect association using such a subset. In such circumstances, statistical analysis of individual SNPs (as described previously) may not be the most effective strategy and may lack the ability to detect association at an ungenotyped

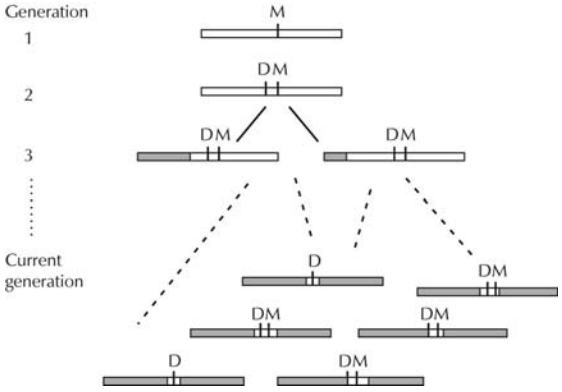

FIGURE 2. Association with disease through direct association (D) and indirect association (M). Disease susceptibility mutation D arises on an ancestral chromosome (white) close to a SNP marker, M. The ancestral chromosome flanking D is lost through recombination through the generations. Observing chromosomes in the current generations shows that all copies of D carry some region of the ancestral (white) chromosome and many of these will also carry the marker allele M.

SNP. The pattern of alleles at multiple markers is usually better able to predict the allele at an untyped locus; hence, simultaneous analysis of multiple markers can improve the power of association studies (de Bakker et al. 2005).

Perhaps the most obvious method for analyzing multiple markers simultaneously is multiple logistic regression. Logistic regression is an adaptation of linear regression in which a logit transformation is used to allow for analysis of a binary outcome (i.e., case–control status). In the equation below, p is the probability of having disease, β_0 represents the intercept, β_1 and β_2 represent the main effect of each marker on the trait, and β_3 represents the interaction term. The variables x_1 and x_2 contain information about the genotype at the two markers and can be coded in a number of different ways—for example, –1, 0, and 1. The interaction term $(x_1{}^*x_2)$ can also be coded in a number of different ways:

$$\text{logit}(p) = \ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3(x_1{}^*x_2).$$

Coefficients β_i can be estimated for each SNP as well as for interactions between them. Stepwise regression can be used systematically to compare different genetic models and to investigate whether multiple markers have independent effects on the trait or are simply in LD with each other, with either marker capturing evidence for association and no improvement in model fit when both markers are included (Cordell and Clayton 2002).

An alternate analysis approach is to phase genotypes into haplotypes and use these as the unit of analysis. This method is attractive because the haplotype is the functional unit of the gene. It is often impossible to be certain about the combination of haplotypes carried by any one individual. However, it is straightforward to determine all possible combinations, and techniques like the E–M algorithm can be used to assign a probability to each haplotype pair (Excoffier and Slatkin 1995). Haplotype effects can be estimated using regression techniques adapted to handle phase uncertainty—for example, a weighted regression technique in which the likelihood function of a finite mixture regression is a weighted sum over all possible haplotypes for each individual (Sham et al. 2004).

Interaction between Genes in Disease Risk

The previous discussion of analysis of multiple markers is focused on markers within a short genetic region (and potentially in LD), but analysis of multiple markers across the genome is also important to identify interaction between genes in disease risk. Interaction is most simply defined as the interdependence of effects at two loci. If the disease risk conferred by the presence of risk alleles at two markers can be inferred from the marginal effects of the presence of each risk allele individually, then no interaction is present. When the joint effect of risk alleles at both markers is much larger (or smaller) than implied by the marker-specific effects, then interaction exists. Statistical interaction (as defined previously) may differ from biological interaction between two genes (Cordell 2002).

The presence of interaction between loci may make each locus more difficult to identify in single SNP tests. Despite the increase in numbers of tests, regression techniques with interaction terms are both computationally feasible and powerful for GWAS (Marchini et al. 2005). Another method for analysis of interaction is multifactor dimensional reduction (MDR). This is a nonparametric approach with a focus on overcoming low numbers of observations in high-order data sets (Ritchie et al. 2003). This technique has been applied to several different traits but, as of this writing, none of the results has yet been replicated independently (Milne et al. 2008).

ADDRESSING PROBLEMS IN ANALYSIS

Quality Control

One disadvantage of a case–control study design compared with family-based association studies is the lack of an internal check on genotyping quality. Standard laboratory practice of assigning both cases and controls to each plate, checking for differences in genotype frequency across plates, and genotyping duplicate samples can help eliminate systematic errors. Testing for HWE in controls can also identify problems with genotyping quality.

Hardy–Weinberg Equilibrium

Under HWE, alleles segregate randomly in the population, allowing expected genotype frequencies to be calculated from allele frequencies. A comparison of the expected and observed genotype frequencies provides a test of HWE (e.g., using a chi-square statistic). For alleles G and T, in which the frequency of allele G is p and the frequency of allele T is $q = (1 - p)$, the expected frequencies of genotypes GG, GT, and TT are p^2, $2pq$, and q^2. Allele frequencies (p, q) are usually estimated from the genotype sample under test, rather than obtained from external genotyping data.

Departure from HWE is generally tested for by using the Pearson chi-square test to assess goodness of fit (of the observed genotype counts to their expectation under HWE). Table 2 shows the step-by-step calculation with observed counts for genotypes GG, GT, and TT of a, b, c, and an application to a data set of 100 control genotypes (GG: 60, GT: 30, TT: 10). The estimated frequency of allele G is 0.75 ($= [2 \times 60 + 30]/200$), noting the division by the number of alleles ($2N$) here, not genotypes (N). The chi-square goodness-of-fit test statistic is then calculated from summing $(\mathrm{O} - E)/E^2$ across genotypes, giving chi-square = 4.0. Under the null hypothesis of no departure from HWE, the test statistic has one degree of freedom (not two degrees of freedom, as implied by the table dimensions), because the allele frequency p has been estimated from the observed data. In this test data set, a p value of 0.046 is obtained, giving slight evidence of departure from HWE, with a deficit in the number of observed heterozygotes.

Departures from HWE in control samples may be caused by the following:

1. Genotyping error. In many genotyping platforms, calling heterozygotic individuals is more challenging than homozygotic individuals, and a higher rate of missing individuals for this genotype can distort HWE.
2. Assortative mating. HWE requires random mating for the SNP under test, which is reasonable for a random SNP across the genome, but may be violated for SNPs that affect mate choice, such as height.
3. Selection. Any genotype increasing the risk of fetal loss or early death is likely to be underrepresented.
4. Population stratification. Control samples that arise from a combination of genetically distinct subpopulations may not be in HWE.
5. Chance. HWE p values for studies of more than one SNP should be corrected appropriately for multiple testing.

Departures from HWE may be caused by any of these factors, but also by the genotyped SNP playing a role in disease susceptibility. Case genotypes for a disease mutation will only be in HWE if the genetic model is multiplicative, with genotype relative risks of 1, r, r^2. However, for modest effect sizes, the power to detect departures from HWE may be low in cases.

TABLE 2. Testing for departure from Hardy–Weinberg equilibrium

	Genotype counts				
	GG	GT	TT	Total	Estimated frequency of G allele
General					
Observed (O)	a	b	c	$N = a + b + c$	$p = (2a + b)/(2N)$
Expected (E)	Np^2	$2Np(1 - p)$	$N(1 - p)^2$		
Test data set					
Observed (O)	60	30	10	100	$p = (2 \times 60 + 30)/200 = 0.75$
Expected (E)	56.25	37.5	6.25		

No standard guidelines for rejecting SNPs that depart from HWE have been developed. In practice, all SNPs for which HWE *p* values decrease below a predetermined threshold should be checked manually for genotyping quality. Investigators should also be aware of SNPs showing significant association in which HWE *p* values are close to this threshold and unsupported by neighboring SNPs in LD.

Missing Genotypes

Another indication of poor genotyping quality is low call rates, with many missing genotypes for each SNP or each individual. This is a major issue in GWAS, but it is also applicable to candidate gene association studies. Genotypes that are missing at random will not bias a test, but poor genotype call rates may indicate nonrandom missingness, with one specific genotype (often heterozygotes) having a lower call rate. This may bias tests of association. Differential rates of missingness between cases and controls (for example, because of differences in DNA extraction and storage) may also be a problem (Clayton et al. 2005).

Population Stratification

Population stratification arises in case–control studies when the two study groups are poorly matched for genetic ancestry. Confounding then occurs between disease state (case, control) and genetic ancestry, with a subsequent increase in false-positive associations. For population stratification to occur, the underlying populations must differ in SNP allele frequency and be represented at different frequencies in the case and control groups. Detecting and controlling for population stratification is important, particularly in GWAS, in which even subtle differences between cases and controls can have major effects on the analysis. Several methods are available to detect and correct for population stratification, including genomic control, the Cochran/Mantel–Haenszel test, and the transmission disequilibrium test.

Genomic control (GC) assumes that population stratification inflates the association test statistics by a constant factor λ, which can be estimated from the median or mean test statistic from a series of unlinked SNPs genotyped in both cases and controls (Devlin and Roeder 1999). Test statistics are then divided by λ and compared with a chi-square distribution or an F distribution) to test for association (Devlin et al. 2004). Genotypes at SNPs uncorrelated with disease status can also be used to infer population ancestry, assigning the samples to distinct population groups, which can then be controlled for in the analysis (Pritchard et al. 2000). In GWAS, population substructure can be identified through a principal components analysis, which models ancestral genetic differences between cases and controls and then corrects for this in the analysis (Price et al. 2006).

Where individuals can be classified into known subgroups (e.g., by birthplace), analysis can be performed within each subgroup and combined using a Cochran/Mantel–Haenszel test (Clayton et al. 2005). The issue of population stratification can be avoided by using family-based studies. The most widely used method is the transmission disequilibrium test (TDT) (Spielman et al. 1993), which tests for non-Mendelian transmission of SNP alleles from heterozygous parents to affected offspring; overtransmission suggests that the SNP allele increases risk of disease.

PITFALLS AND PROBLEMS OF ASSOCIATION STUDIES

A major challenge in association studies of candidate genes has been nonreplication of significant findings. For many diseases and genes, the literature contains papers with little consistent pattern in the results obtained. Typically, this comprises an initial report showing significant association, with follow-up studies showing little or no evidence of association. We discuss here reasons for these discrepancies between studies.

False-Positive Finding

The initial report of association may have been a false-positive finding that arose by chance or systematic bias in the study. The “Quality Control” section discussed several problems that can lead to such results, and each of these should be checked (population stratification and genotyping errors). False-positive results may arise through a failure to correct for multiple testing across the number of genes, SNPs, statistical analysis methods, or phenotypic subgroups tested, although this can be difficult to determine from a publication. For independent tests (e.g., multiple genes that are not in LD), a Bonferroni correction may be applied to the p values. Where tests are correlated, an appropriate correction may be difficult to determine, but permutation tests can be used to determine empirical levels of significance. A noted phenomenon is that the first published study tends to overestimate the effect size, with subsequent studies detecting more moderate contribution of the genotyped variant to disease risk (Ioannidis et al. 2001).

Replication Study Lacks Power

Alternatively, replication studies may lack power to detect the true association. Most genes contributing to complex disorders confer only a very modest increase in disease risk, and to detect these with high power requires large sample sizes. For example, for a SNP of 10% frequency, under a multiplicative model with heterozygote relative risk of 1.3, at least 1146 cases and controls are required to obtain 80% power at a significance level of 5% with no correction for multiple testing (Purcell et al. 2003). Including the multiple testing correction greatly increases the numbers needed. Many association studies have used samples of hundreds, not thousands, of cases and controls, and therefore lack the ability to detect such associations. Meta-analysis of published data provides a possible solution, and such studies have confirmed many associations that were unclear from individual study reports (Altshuler et al. 2000; Ioannidis et al. 2001; Lohmueller et al. 2003).

Heterogeneity between Studies

Another problem is that heterogeneity between studies may validly lead to different conclusions about the role of a SNP in disease risk. Sources of heterogeneity include the precise clinical criteria used in case definition for each study, differences in disease severity, disease subtype, age of diagnosis, or duration of disease. If a genetic variant contributes predominantly to a specific subphenotype of disease, then the mix of cases ascertained in different studies will substantially affect the power of each study to detect association. Information from family or twin studies on heritability of different components of disease definition can help refine the hypothesis to be tested, with some studies choosing to ascertain cases likely to be more heavily genetically predisposed, for example, those with a family history of disease, or early onset (Antoniou and Easton 2003).

Heterogeneity across Studies

Population heterogeneity across studies may also lead to differences in study outcomes. Variations in SNP frequencies are seen across the major population groups because of random drift, novel mutations, and (less commonly) selection. However, meta-analyses of replicated genetic association studies suggest that even when the SNP frequency differs across populations, the effect size of mutations remains approximately constant (Ioannidis et al. 2004). Some mutations may be absent in specific population groups; for example, NOD2 mutations, which are present in >30% of Crohn's disease patients in European populations are absent in Asian populations (Mathew and Lewis 2004).

CONCLUSION

This article has given a broad outline of the design and analysis of genetic association studies, as well as the pitfalls of performing powerful, accurate association studies. These challenges are reflected in the contradictory literature for many disease–gene investigations. However, consistent findings of disease–gene associations have been detected, and the realization that most mutations confer only modest increases in risk has led to an improvement in study design. Larger studies are now being performed and internal replication of significant findings is becoming standard practice.

REFERENCES

Altshuler D, Hirschhorn JN, Klannemark M, Lindgren CM, Vohl MC, Nemesh J, Lane CR, Schaffner SF, Bolk S, Brewer C, et al. 2000. The common PPARg Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* **26:** 76–80.

Antoniou AC, Easton DF. 2003. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet Epidemiol* **25:** 190–202.

Clayton DG, Walker NM, Smyth DJ, Pask R, Cooper JD, Maier LM, Smink LJ, Lam AC, Ovington NR, Stevens HE, et al. 2005. Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat Genet* **37:** 1243–1246.

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

Cordell HJ, Clayton DG. 2002. A unified stepwise regression procedure for evaluating the relative effects of polymorphisms within a gene using case/control or family data: Application to HLA in type 1 diabetes. *Am J Hum Genet* **70:** 124–141.

de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D. 2005. Efficiency and power in genetic association studies. *Nat Genet* **37:** 1217–1223.

Devlin B, Risch N. 1995. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29:** 311–322.

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

Devlin B, Bacanu SA, Roeder K. 2004. Genomic control to the extreme. *Nat Genet* **36:** 1129–1130; author reply 1131.

Excoffier L, Slatkin M. 1995. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12:** 921–927.

Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. 2001. Replication validity of genetic association studies. *Nat Genet* **29:** 306–309.

Ioannidis JP, Ntzani EE, Trikalinos TA. 2004. "Racial" differences in genetic effects for complex diseases. *Nat Genet* **36:** 1312–1318.

Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. 2003. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* **33:** 177–182.

Marchini J, Donnelly P, Cardon LR. 2005. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat Genet* **37:** 413–417.

Mathew CG, Lewis CM. 2004. Genetics of inflammatory bowel disease: Progress and prospects. *Hum Mol Genet* **13:** R161–R168.

Milne RL, Fagerholm R, Nevanlinna H, Benitez J. 2008. The importance of replication in gene-gene interaction studies: Multifactor dimensionality reduction applied to a two-stage breast cancer case-control study. *Carcinogenesis* **29:** 1215–1218.

O'Donovan MC, Craddock N, Norton N, Williams H, Peirce T, Moskvina V, Nikolov I, Hamshere M, Carroll L, Georgieva L, et al. 2008. Identification of loci associated with schizophrenia by genome-wide association and follow-up. *Nat Genet* **40:** 1050–1055.

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

Purcell S, Cherny SS, Sham PC. 2003. Genetic Power Calculator: Design of linkage and association genetic mapping studies of complex traits. *Bioinformatics* **19:** 149–150.

Ritchie MD, Hahn LW, Moore JH. 2003. Power of multifactor dimensionality reduction for detecting gene–gene interactions in the presence of genotyping error, missing data, phenocopy, and genetic heterogeneity. *Genet Epidemiol* **24:** 150–157.

Sasieni PD. 1997. From genotypes to genes: Doubling the sample size. *Biometrics* **53:** 1253–1261.

Sham PC, Rijsdijk FV, Knight J, Makoff A, North B, Curtis D. 2004. Haplotype association analysis of discrete and continuous traits using mixture of regression models. *Behav Genet* **34:** 207–214.

Slatkin M. 1999. Disequilibrium mapping of a quantitative-trait locus in an expanding population. *Am J Hum Genet* **64:** 1764–1772.

Spielman RS, McGinnis RE, Ewens WJ. 1993. Transmission test for linkage disequilibrium: The insulin gene region and insulin-dependent diabetes mellitus (IDDM). *Am J Hum Genet* **52:** 506–516.

Steer S, Lad B, Grumley JA, Kingsley GH, Fisher SA. 2005. Association of R602W in a protein tyrosine phosphatase gene with a high risk of rheumatoid arthritis in a British population: Evidence for an early onset/disease severity effect. *Arthritis Rheum* **52:** 358–360.

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

Cite this article as *Cold Spring Harbor Protoc*; 2012; doi:10.1101/pdb.top068163