Genetic Epidemiology 5

What makes a good genetic association study?

Andrew T Hattersley, Mark I McCarthy

Genetic association studies are central to efforts to identify and characterise genomic variants underlying susceptibility to multifactorial disease. However, obtaining robust replication of initial association findings has proved difficult. Much of this inconsistency can be attributed to inadequacies in study design, implementation, and interpretation—inadequately powered sample groups are a major concern. Several additional factors affect the quality of any given association study, with appropriate sample-recruitment strategy, logical variant selection, minimum genotyping error, relevant data analysis, and valid interpretation all essential to generation of robust findings. Replication has a vital role in showing that associations that are identified reflect interesting biological processes rather than methodological quirks. For an unbiased view of the evidence for and against any particular association, study quality, rather than significance value, needs to play the dominant part.

Lancet 2005; 366: 1315–23

This is the fifth in a Series of seven papers on genetic epidemiology.

Institute of Biomedical and Clinical Science, Peninsula Medical School, Exeter, UK (A T Hattersley DM); **Wellcome Trust Centre for Human Genetics, University of Oxford, UK** (Prof M I McCarthy MD); **and Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital Campus, Old Road, Headington, Oxford OX3 7LJ, UK** (Prof M I McCarthy)

Correspondence to:
Professor Mark McCarthy
mark.mccarthy@drl.ox.ac.uk

Introduction

The role of association studies in the detection and characterisation of genes contributing to common, multifactorial traits remains controversial. Theoretical analyses often emphasise the power of this methodology[1] but it has provided few novel insights so far.[2] A major source of frustration and confusion has been the frequency with which initial positive findings are not confirmed.[2–5] Part of the explanation could lie in biological factors. Differences between studies in the frequency of a particular susceptibility variant, genetic background, or environmental exposures could affect the capacity to replicate results. However, questions of study design, implementation, and interpretation can be important too.[3,4,6,7] Our aim is to allow the reader to assess the quality of association studies and better appreciate the inferences that should be drawn. We also hope to contribute to the development of standards for association studies, designed to raise their quality and to facilitate robust meta-analysis.[8]

What are we hoping for from an association study?

The objective is deceptively simple. Is there a statistical relation between genomic variation at one or more sites and phenotypic variation, usually represented by the presence or absence of a disease or by levels of a disease-related trait?[9] The archetypal case-control study compares two groups that are expected to differ in their prevalence of disease-susceptibility alleles. Provided the sampling has been appropriate (ensuring that cases and controls have similar ethnic background, for example),[10] the detection of a significant difference in variant frequency between groups (ie, a significant association) is consistent either with the hypothesis that the variant typed influences trait susceptibility or that a second variant in linkage disequilibrium with the first does so.

A key issue for understanding the limitations of many papers reporting association data for multifactorial traits is that small effect sizes are to be expected. The modest extent of familial clustering of many complex traits, and the failure, with notable exceptions, to detect large susceptibility effects through linkage studies[11] clearly indicate that large genetic effects are unlikely. Evidence from the few variants consistently shown to be associated with common disease endorses this view. The susceptibility variants identified so far are either common, but have only modest relative risk (eg, peroxisome proliferator activated receptor gamma [*PPARG*] and type 2 diabetes[12]), or are uncommon but carry substantial relative risks (eg, factor V Leiden and thrombosis[13]; table 1). In either case, detection by association is not straightforward.[1] Except for HLA effects on autoimmune disease, the only confirmed associations due to a common variant (>10% allele frequency) with a relative risk exceeding 2 are those between variation in the gene encoding apolipoprotein E (*APOE*) and Alzheimer's disease,[15] and between variation in the gene encoding complement factor H and age-related macular degeneration (table 1).

Many association studies have had only limited power to detect true susceptibility effects on this scale and even less power to exclude the involvement of a gene. Changing this dismal history is not only desirable but also possible, with collaborative efforts on larger sample sets, improved genotyping and analytical technologies, and advances in bioinformatics. Also required are attention to study design, implementation, and interpretation, and a better understanding of what makes a good association study.

How good a candidate is the gene?

Since there are up to 30 000 genes in the human genome, and it is unlikely that more than a few hundred make a meaningful contribution to variation in any single phenotype, the a priori probability that any gene selected at random will influence a given trait is very low. Evidence from a range of sources (table 2) can be used to identify genes (so-called candidates) with higher prior odds for phenotypic involvement, and almost all

	Gene	Polymorphism	Approximate frequency of disease-associated allele	Approximate odds ratio for disease-associated allele	Reference
Thrombophilia	*F5*	Leiden Arg506Gln	0·03	4	13
Crohn's disease	*CARD15*	3 SNPs	0·06 (composite)	4·6	14
Alzheimer's	*APOE*	e2/3/4	0·15	3·3	15,16
Osteoporotic fractures	*COL1A1*	Sp1 restriction site	0·19	1·3	17,18
Age-related macular degeneration	*HF1/CFH*	Tyr402His	0·30	2·5	19–23
Type 2 diabetes	*KCNJ11*	Glu23Lys	0·36	1·23	24
Type 1 diabetes	*CTLA4*	Thr17Ala	0·36	1·27	25,26
Graves' disease	*CTLA4*	Thr17Ala	0·36	1·6	27
Type 1 diabetes	*INS*	5′ variable number of tandem repeats	0·67	1·2	28
Bladder cancer	*GSTM1*	Null (gene deletion)	0·70	1·28	29
Type 2 diabetes	*PPARG*	Pro12Ala	0·85	1·23	12

Examples in order of frequency of disease-associated allele (in controls).

Table 1: **Examples of some polymorphisms or haplotypes that have shown consistent association with complex disease**

association studies have featured genes selected on these criteria. Prior information of this kind should also inform the interpretation of association findings. It seems reasonable to demand a higher level of statistical evidence for genes with very little supporting biological information before interest is provoked or significance attributed.[1,7]

However, assessing candidacy is a notoriously imprecise art. Poor understanding of the molecular mechanisms underlying most complex traits (itself one of the main justifications for gene discovery efforts) means that the prior odds associated with any given gene can rarely be calculated precisely. Also, many of the methods listed in table 2, especially if used in isolation, are not very good at picking out true susceptibility genes.[33] Furthermore, detection of a true association demands not only that the gene product be involved in pathways relevant to the development of the trait of interest but also that the gene contains variants capable of influencing its regulation or function.

How strong is the case for the genetic variants that have been typed?

To guarantee detection of all possible disease-associated variants at a given gene it would be necessary to examine, in large samples, every base at which variation might conceivably alter gene function or expression. Only then could we be confident that an association had not been missed just because the wrong markers had been typed.[6] This pursuit of perfection is, for the time being at least, unrealistic: genotyping remains too expensive to examine many hundreds of variants in several thousand people for every gene of interest. So choices have to be made, and the strategy used to define the subset of variants to be typed has a substantial effect on the power and quality of the study. Greater understanding of genomic variation has allowed more logical choices. Nevertheless, variant selection is always a pragmatic compromise.

There are two complementary approaches to the selection of variants. Tagging exploits the extensive

	Explanation	Advantages	Disadvantages	Examples
Biology	Function of protein encoded is implicated in biology of disease or trait	Good when pathophysiology known (eg, autoimmune disease or thrombophilia)	Less good when primary pathophysiology unknown (eg, hypertension or type 2 diabetes)	Factor V Leiden in thrombophilia[13]
Pharmacology	Gene encodes protein implicated in mechanism of action of disease-modifying or trait-modifying drug	Evidence that modification of pathway by small molecules can influence trait suggests that genetic variation could do likewise	Treatment might not act on aetiological pathway	*PPARG* in type 2 diabetes[12]
Animal models	Identification of genes influencing related traits in animal models offers candidates for testing in man	Provides clear functional links between gene dysfunction and whole body phenotype	Species differences in physiology and in patterns of genetic variation	*ApoAV* and hypertriglyceridaemia[30]
Monogenic or syndromic forms of disease	Genes in which rare mutations lead to monogenic or syndromal forms of disease also show common genetic variation that predisposes to polygenic disease	Rare monogenic disorders establish the gene has a critical function and indicates lack of compensation or plasticity	Lack of an appropriate monogenic disease	Monogenic diabetes and type 2 diabetes (review[31])
Positional information	Genome-wide scans for linkage or association could indicate regions with a high probability of containing a susceptibility gene	Genes that account for reproducible linkage peaks in genome scans probably represent major susceptibility genes	Regions of interest defined by linkage remain large. Linkage studies are usually underpowered and will not detect all susceptibility genes	*APOE* and Alzheimer's disease[15]
Prior association data	Previous studies showing association with a gene or a meta-analysis of previous studies indicate that variation in the gene probably has an aetiological role	Previous positive studies can give some of the strongest prior knowledge for examining a gene	Caution is needed: initial association studies generally overestimate the effect size of the variant tested	*PPARG, KCNJ11, CAPN10* in type 2 diabetes[12,24,32]

Table 2: **Criteria for selection of candidate genes**

linkage disequilibrium in many parts of the genome. By typing a subset of variants that captures a disproportionate amount of the information in common regional haplotypes, it should be possible to maintain power while making considerable savings in genotyping[34] (see earlier paper in this series[9]). In its most conservative form, tagging simply avoids redundant typing of sets of variants that are in complete linkage disequilibrium with each other. Strategies for more subtle selection of tags are slowly emerging.[35] Whether such tag single nucleotide polymorphisms (SNPs) also have intrinsic biological merit as markers for complex trait susceptibility variants remains unresolved.[36–38] Advocates of the common disease, common variant hypothesis argue on theoretical[38] and empirical[2] grounds that many of the alleles affecting susceptibility to common complex traits will themselves be common. If so, then the typing of regional tag-SNPs that are selected specifically to capture such common genomic variation should provide an efficient approach for detecting complex trait susceptibility alleles.[6,39]

The second approach incorporates assessments of the likely functional effect of variation within a gene or region of interest. Those variants judged most likely to influence gene expression or function (irrespective of their correlation with local haplotype structure)[40] are then prioritised. Unfortunately, predicting the functional credentials of most variants remains extremely difficult. Non-synonymous coding variants—those that alter the aminoacid sequence in the gene product—are obvious targets, but assessment of the potential regulatory effect of intronic variants or those lying several kb upstream of a gene remains poor.[41] One critical issue is the need to define the extent of the regulatory elements influencing a given gene. Reports of association between haplotypes surrounding the beta-cell promoter of the *HNF4A* gene and type 2 diabetes highlight the difficulties.[42,43] This promoter, which lies 46 kb upstream of the coding region, was identified a decade after delineation of the coding sequence.[44,45] Cross-species sequence comparisons to identify conserved non-coding sequences that are likely to represent regulatory elements[46] and high-throughput experimental methods for defining sites of gene regulation[47] promise to aid both variant selection and subsequent interpretation.

Until such methods are established, much weight will be placed on the detailed functional assessment of those variants for which there is some preliminary evidence for association. It is certainly reasonable to argue that a functional effect provides biological corroboration, enhancing the probability that the statistical association is genuine, but such findings need careful interpretation. First, in-vitro functional studies have intrinsic limitations: it can be hard to know how to interpret results that suggest that the functional effects of a given variant are only evident in particular cell-lines, under particular experimental conditions, by particular investigators. Second, allelic variation in expression is likely to be present in most genes.[48] Thus, evidence that a particular variant affects the expression of a gene will add little to a questionable association finding, especially without any evidence that differential expression of the gene can be causally related to the trait of interest.

Concerns about SNP selection become more relevant as researchers begin genome-wide scanning for association.[49] For a comprehensive survey of the genome, hundreds of thousands of SNPs will need to be typed. Studies on this scale are increasingly feasible[50] but serious questions remain over SNP selection, marker density, and study design. One point of debate is the merit of focusing whole genome coverage around exonic SNPs.[50] Such a strategy should capture common variation in and around transcribed sequences well but variants within the regulatory sequence will be poorly represented. Until the relative importance of variation in transcribed and regulatory sequences in multifactorial trait susceptibility becomes clearer, a full assessment of the costs and benefits of exon-centric scanning strategies is not possible. Of course, with further advances in genotyping, difficult decisions about SNP selection will no longer have to be made.

How appropriate are the samples typed?

As in conventional epidemiology, the prospective cohort study is often regarded as the gold standard. Although cohort studies can measure risk at the population level, and therefore remain essential for genetic epidemiologists, they are usually not efficient for the initial stages of gene discovery. Unless the disease is very common, the study samples generated will have far fewer individuals with disease than without, and the nested case-control samples that emerge will often be small. Furthermore, the unselected nature of the cases could compromise power, especially when compared with samples that are deliberately enriched for genetic aetiology and disease homogeneity. A meta-analysis of the relation between variation at *IRS1* (encoding insulin receptor substrate-1, a key intermediate in insulin signalling) and type 2 diabetes reported that the association was restricted to people from hospital clinics and was absent in population cohorts.[51] Given these limitations, the case-control study remains the mainstay of genetic association studies, and the most important issues relate to choice of the two study groups.

In many ways, selection of cases is the easier task, because clinical presentation facilitates recruitment. However, researchers have to make several crucial decisions, explicitly or implicitly, that could influence study outcome. One decision relates to characterisation of cases: is it better to have many less well-characterised cases, accepting that some misclassification and heterogeneity is inevitable, or to aim for phenotypic homogeneity at the cost of reduced sample size?

	Frequency of susceptibility allele in controls					
γ (allelic odds ratio)	1%	5%	10%	20%	30%	40%
1·1	221 927	46 434	24 626	13 987	10 759	9505
1·2	58 177	12 217	6509	3730	2896	2581
1·3	27 055	5702	3051	1763	1380	1240
1·5	10 604	2249	1213	712	566	516
2·0	3193	687	377	229	188	177
4·0	598	134	78	52	46	47

Calculations assume multiplicative effect on disease risk (ie, homozygous susceptibility genotype has penetrance that exceeds that of heterozygote by factor γ, the genotype relative risk, and that of wild-type homozygote by γ^2). Under such model, each allele has independent effects on disease risk, and allelic odds ratio is also equal to γ. Sample sizes presented are total number of cases needed in case control study where controls are present in equal numbers. These sample size derivations assume best-case scenario in which susceptibility variant itself (or a perfect proxy) has been typed.

Table 3: **Approximate sample sizes necessary to detect significant association (power=90%, two-sided α=0·001) by effect size and allele frequency for predisposing allele**

Unfortunately, no general answer is possible: the correct strategy depends on disease-specific and study-specific factors, which are both known (eg, cost and specificity of the phenotyping tools used) and unknown (the genetic architecture of the condition). More useful might be selection of cases that are likely to be enriched for genetic susceptibility. There are good reasons to expect selection based on strong family history[52] or early age of onset[53] to increase the difference in frequency of susceptibility alleles between cases and controls and, for a given sample size, to improve power. When searching for associations within regions known to be linked to the disease of interest, this difference can be further heightened by selecting on the family-specific evidence for linkage and by use of allele-sharing information to select family members showing maximum sharing with other affected relatives.[54] However, although careful selection of cases (and controls) should provide a valid test of association with improved power, reliable estimation of population-based parameters (relative risk, population attributable risk) is not possible with extreme samples and will need population-based cohorts. Thus, although there is no universally correct recipe for case selection, the decisions taken by investigators can affect study power, and often in unpredictable ways. This fact could help explain some inconsistencies in outcome between studies.[51]

Selection of controls has been extensively discussed.[6,7,10] Controls must be selected from the same population as cases, or genomic control methods must be used to correct for any latent population stratification.[9,55] Selection by phenotype or by age remains controversial. Compared with population controls (a sample judged to be representative of the population from which the cases were drawn), use of hypernormal controls (such as older people known to be free of the disease of interest), would be expected to improve power by increasing the difference in susceptibility allele frequency between cases and controls. However, this benefit is usually not substantial,[7] and can easily be outweighed by the costs of defining a hypernormal population (the need to phenotype people for disease status, the potential for inadvertent selection for other phenotypes such as survivor effects). One favoured option is the use of large control cohorts of clear provenance, often nationally representative. One such example is the UK 1958 birth cohort, composed of individuals born in a single week during March 1958, for which extensive longitudinal phenotype data are available. Such cohorts can provide controls for studies in a wide range of diseases, though survivor bias becomes a concern where the mean age of cases differs appreciably from that of the control cohort. There are important benefits associated with building up a large body of genotype, phenotype, and quality-control data for a single control cohort. A later paper in this series considers family-based association resources as an alternative source of controls.[56]

Is the study size large enough?

The key determinant of quality in an association study is sample size.[3] With the remote chance of finding common genes with large effects, studies must be powered to detect variants that are common—but have low relative risk—or rarer—but with higher relative risk—which means samples sizes of thousands (table 3). Rare variants with low relative risks are largely beyond the reach of genetic epidemiology because of the massive sample size that would be needed.[57] These calculations assume that the susceptibility variant itself (or a marker in complete linkage disequilibrium) has been typed, which is a best-case scenario. The apparent success in identifying interesting associations with studies much smaller than would be implied by table 3 might suggest that these calculations are too pessimistic. However, small initial studies rarely find the correct result,[3] and even when they do are likely to overestimate the true effect size.[2,58]

What are the practical implications of these sobering calculations? It is hard to make any case for doing even preliminary studies on small samples. If such an underpowered study generates positive findings, those will need to be replicated in a second, ideally larger, sample. The need for large studies[12] is fuelling the emphasis on national and international collaborations. Ideally, to avoid concerns that investigators have rushed to press once promising results have been identified, groups with access to multiple cohorts should type and report on all available samples when publishing positive results. However, large sample size alone is no guarantee of validity. Increasing size reduces sampling error but carries an increased danger of false positives as hypothesis testing becomes highly sensitive to the consequences of small bias effects due, for example, to population stratification.[59]

How good is the genotyping?

Most association studies assume implicitly that the genotypes are accurate. However, even with the best

methods, some assays will be unreliable; and the accuracy of earlier genotyping methods (on which much of the current published work is based) will have been even worse. Historically, few journals have required authors to report the steps taken to limit and assess genotyping error rates or the performance characteristics of their assays (by contrast with data routinely demanded for biochemical assays). Authors rarely admit to residual errors in their data, so readers have little information on which to assess the effect of genotyping error. Even well-regarded laboratories that use powerful new technologies assessed under ideal conditions report error rates close to 1% (and SNP-specific error rates up to 3%),[60] so error rates of a few percent are not atypical.[61]

Most interest has focused on the extent to which random genotyping error reduces the power to detect true case-control differences.[62,63] For example, each 1% rise in genotyping error might require sample size to increase by 2–8% to maintain constant type I and type II error rates.[63] Random error rates of a few percent therefore have an appreciable, though not calamitous, effect on power. Of greater concern are situations in which systematic genotyping error increases the danger of a wrongly attributed association. In view of the low prior probability of association expected of most variants, and the effect of publication and other biases that favour positive findings,[64] studies affected by genotyping error of this type are almost certainly disproportionately represented in the published work. Some study designs, such as family-based association tests with parent-offspring triads, are inherently prone to bias away from the null hypothesis, whether the errors are detected (and the relevant genotypes, individuals or families removed from analysis) or not.[65–67] In case-control studies, batch-related differences in genotyping performance can be a problem. Many laboratories store, and subsequently genotype, different sample sets (eg, cases and controls) on separate plates, so any variation between batches in genotyping accuracy or tendency to preferentially reject particular genotypes (usually heterozygotes) as assay performance falls will translate into between-group differences in genotype distribution. The consequences can be serious.[68]

Reducing the effect of genotyping error requires more accurate genotyping platforms. Such methods are now becoming available,[69] but not all researchers yet have access. Also, genotyping error needs to be taken more seriously by monitoring of assay performance (panel) and reducing sources of error and bias. Even with the best methods, some assays will be surreptitiously inaccurate.[70] Information about genotyping performance should be published for any association study,[8] providing journals (and their reviewers and readers) with the information necessary for critical appraisal, and facilitating meta-analyses (which should incorporate some understanding of error estimates).

How appropriate is the analysis?

All types of association analysis contain traps for the unwary. Ready access to so many analytical programs can easily lead to inappropriate use if the underlying assumptions are not clearly stated or appreciated. Haplotype reconstruction provides a good example. There are many ways of inferring haplotypes from unphased genotype data (as obtained in a sample of unrelated individuals).[71–75] However, the conceptual and computational differences between these methods translate into substantial differences in performance and, sometimes, in outcome.[73,74] Similarly, when inferring haplotypes within pedigrees, substantial errors can arise with methods that inappropriately assume that the typed markers are in linkage equilibrium.[76]

The inexperienced (or even experienced) reader might find it almost impossible to establish whether appropriate methods have been used, and the extent to which the results are robust. One recommendation is to urge authors to make their raw genotype data freely available[77] but doing so could raise confidentiality issues. The best guarantor of probity is likely to remain a significant input from experts in statistical genetics during study design, data analysis, and peer-review. Additional reassurance comes when a finding remains consistent when several complementary methods are used.

How appropriate is the interpretation?

The perceived unreliability of association studies has generated much discussion about the level of evidence needed before an association can be regarded as proven. The emphasis has generally been on the need for greater stringency.[1,7] If there are around 10^6 variants within the genome that can, in principle, affect any given human trait, an appropriate threshold might be a p value of 5×10^{-8} (that is the standard 0·05 corrected for 10^6 tests).[1] Arguments from a Bayesian perspective suggest that 5×10^{-5} should be sufficient to constrain the false discovery rate.[7]

	Susceptibility allele frequency in controls					
	1%	5%	10%	20%	30%	40%
α						
0·05	13 599	2866	1533	886	694	623
0·01	19 258	4058	2171	1255	982	883
0·001	27 055	5702	3051	1763	1380	1240
5×10–5	36 869	7770	4157	2403	1881	1690
5×10–8	58 678	12 366	6617	3825	2994	2690

Numbers indicate sample size needed to detect significant association (power=90%) for different values of α, assuming allelic odds ratio of 1·3, given differing allele frequencies for predisposing allele or haplotype. Assumptions are same as for table 3.

Table 4: **Effect of differing statistical significance levels on sample size**

***Panel:* Detection and reduction of genotyping error in association studies**

Ways of detecting error

- Duplicate genotype assignment: discordance rate between two operators scoring same genotypes
- Duplicate genotyping on same assay: discrepancy rate when a proportion (more than 10%) of samples genotyped again with same method
- Duplicate genotyping on another assay: discrepancy rate when a proportion (more than 10%) of samples genotyped again with a different method
- Blank control wells: is there contamination of the PCR reagents? Is the plate oriented correctly?
- Expected allele frequencies: are allele (or haplotype) frequencies in line with those expected from previous data from similar ethnic groups?
- Hardy-Weinberg equilibrium: are genotype frequencies consistent with Hardy-Weinberg equilibrium? (Modest departures, especially in case samples, could be evidence of association, but see also an earlier paper in this series[9])
- Between-run consistency: are there outlier batches with extreme allele or haplotype distributions that need to be targeted for repeat?
- Mendelian consistency: where related individuals have been typed, are genotypes consistent with mendelian expectation? (Note that in simple family structures typed for biallelic markers, not all genotyping errors will generate mendelian inconsistency)
- Assay performance: where the method provides these, are quality scores for individual genotypes and/or batches satisfactory?
- Linkage disequilibrium relations and haplotype structure: where several adjacent SNPs have been typed in a region of strong linkage disequilibrium, do haplotype relations suggest any fluctuations in genotyping performance? Appearance of haplotypes that have never been identified before can be a very sensitive measure of error

Ways of minimising error

- Assay design: robust assay design incorporating, where possible, quality checks (eg, obligate restriction sites in a PCR-RFLP assay). Confirm that assay detects all genotypes accurately by comparison with reference samples
- Monitor assay QC on assay-specific basis: derive and monitor quality control metrics (above) for each assay
- Complete genotyping: batch-related biases are most likely when genotyping call rates are low, so beware of any assay with poor call rate, or appreciable difference in call rate between cases and controls
- Mix sample sets within batches: ensure that cases and controls are mixed on the same plate (eg, by intercalating 96-well plates onto a single 384-well plate)
- If in doubt, retype: no single measure will detect all available errors and correcting all detectable errors will not reduce the error rate to zero. If the quality control metrics indicate assay failure, redesign the assay and retype

As with linkage studies,[78] such thresholds provide useful benchmarks, whether for a single study or a meta-analysis. However, a distinction needs to be made between thresholds and standards for proof beyond reasonable doubt (which should clearly tend towards the stringent) and those for publication (just that the study has been designed and done well). To limit publication to studies that prove association (or that achieve the difficult task of showing conclusively that variation within a gene does not make a material contribution) would paralyse research. Although the increases in sample size necessary to satisfy more stringent criteria are surprisingly modest (table 4), they can still exceed the scope of many current collections. Moreover, any policy insisting on such stringent criteria for publication would further complicate the task of obtaining an unbiased overview of all the data, positive or negative. Many good (largely negative, presumably) studies would not be disseminated, leaving the published work even more prone to bias.

The headline p value of an association study has little to do with experimental quality. Well run studies with large, appropriate samples, robust low-error genotyping, and clear primary hypotheses will advance knowledge even if none of the association tests reaches significance at the threshold chosen. By contrast, few associations first reported from small sample sizes prove reliable in larger datasets.[3,4] Studies of the role of the insertion/deletion variant in the gene encoding angiotensin-converting-enzyme in myocardial infarction[79] and of the Pro12Ala SNP in *PPARG* in type 2 diabetes[12] illustrate this point. Larger studies of the angiotensin-converting-enzyme insertion/deletion variant made clear that small case-control samples had overestimated the true effect, probably as a result of publication bias.[79] In the other example, many of the early case-control studies were too small to provide any useful estimate of the true effect size.[12]

However, experimental quality cannot be the sole consideration for publication. Journals have to take account of the scientific interest of the findings; and, since only a few variants are likely to show association to any given phenotype, it is legitimate to regard positive findings as having intrinsically greater importance. The solution to this problem has to be rigorous criteria for study quality along with acceptable alternative mechanisms for the deposition and dissemination of all good-quality

association data. In the meantime, peer-review needs to weigh study quality and the apparent biological interest of the headline findings more equitably.

A related issue is the need for more complete statements of prior hypotheses and the range of analyses undertaken. Too often, in the effort to produce at least one p value that reaches nominal significance, exploratory analyses arising from post-hoc subdivision and stratification of the data are presented as major findings. One aspect of such multiple testing is easy to solve—by reporting the number of variants typed[80] and adjusting for this by standard methods based on the Bonferroni correction or alternatives such as the false discovery rate.[81] These adjustments can readily take account of correlations between the typed markers caused by linkage disequilibrium.[82] Allowing for the proliferation of analyses that might be done at each locus is not so easy. For example, association could be sought at the level of the allele or the genotype: if the latter, several genetic models (eg, dominant, recessive) can be tested.[83] If several markers have been typed, analyses of haplotype or diplotype frequency become possible.[84] Analyses might account for the effects of stratification by, or adjustment for, factors such as sex or disease subtype or age; additional phenotypes (age at disease onset, or another phenotype related to disease) could be included as outcome variables. Evidence of interaction effects (gene-gene or gene-environment) might be sought.[85] Where parent-offspring triads are typed, it becomes possible to test for parent-of-origin and a variety of epigenetic effects.[86] Any of these analytical manoeuvres might be entirely justified in any given study on the basis of prior biological hypotheses or the desire to replicate specific findings arising from previous studies. However, when there is no such justification, such repeated analyses will inflate the type 1 error of the study if no correction is made. In practice, the high degree of correlation between these analyses can make determination of the extent of any correction extremely difficult.

The challenge is to avoid making false attribution of associations based on post-hoc analyses without ruining the capacity to undertake potentially informative data exploration in what might be expensively acquired samples of richly-characterised individuals. Part of the answer must be a clearer statement of prior hypotheses (including, potentially, some form of prestudy registration of intent), and a more complete exposition of all the different analyses undertaken. Most importantly, showing that hypotheses generated from post-hoc analyses in one sample can be confirmed in directed analyses in others is essential for establishing their credibility.

The future

It is time for a mature and balanced view of the merits of association studies. Individual association studies can be viewed as stages on a journey that starts with ignorance and (hopefully) ends with a clear and robust assessment of whether or not variation at a given locus contributes to disease susceptibility. Very rarely will conclusive evidence come from a single study. As Page and colleagues[77] point out, the major purpose of replication in association studies is not to improve the statistical significance of the findings. Replication studies provide insurance against errors and biases that can unavoidably afflict any individual study, and amplify confidence that any associations uncovered reflect processes that are biologically interesting (ie, a variant that truly influences disease susceptibility), rather than methodological inadequacies (inappropriate control groups, genotyping error, investigator biases, over-elaborate data exploration).[77,87]

The theoretical promise of association methodology in the analysis of multifactorial traits remains largely unproven. However, armed with burgeoning knowledge of the human genome, the availability of larger sample sets, and the increasing ability to type those samples rapidly and cheaply, the framework for progress is in place. Improved adherence to principles of good study design will help to ensure that this new capacity contributes to our understanding of the cause of common disease.

Conflict of interest statement

We declare that we have no conflict of interest.

Acknowledgments

We thank many colleagues in the genetics of type 2 diabetes and other polygenic disease for their useful discussion on these topics, especially Steven Wiltshire, Eleftheria Zeggini, Chris Groves, Tim Frayling, Michael Weedon, Kirsten Ward. ATH is a Wellcome Trust clinical research leave fellow.

References

1 Risch N, Merikangas K. The future of genetic studies of complex human diseases. *Science* 1996; **273:** 1516–17.

2 Lohmueller KE, Pearce CL, Pike M, Lander ES, Hirschhorn JN. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat Genet* 2003; **33:** 177–82.

3 Ioannidis JPA, Trikalinos TA, Ntzani EE, Contopoulos-Ioannidis DG. Genetic associations in large versus small studies: an empirical assessment. *Lancet* 2003; **361:** 567–71.

4 Ioannidis JP, Ntzani EE, Trikalinos TA, Contopoulos-Ioannidis DG. Replication validity of genetic association studies. *Nat Genet* 2001; **29:** 306–09.

5 Hirschhorn JN, Lohmueller K, Byrne E, Hirschhorn K. A comprehensive review of genetic association studies. *Genet Med* 2002; **4:** 45–61.

6 Zondervan KT, Cardon LR. The complex interplay among factors that influence allelic association. *Nat Rev Genet* 2004; **5:** 89–100.

7 Colhoun HM, McKeigue PM, Davey Smith G. Problems of reporting genetic associations with complex outcomes. *Lancet* 2003; **361:** 865–72.

8 Little J, Bradley L, Bray MS, et al. Reporting, appraising, and integrating data on genotype prevalence and gene-disease associations. *Am J Epidemiol* 2002; **156:** 300–10.

9 Cordell HJ, and Clayton DG. Genetic association studies. *Lancet* 2005; 366: 1121–31.

10 Cardon LR, Palmer LJ. Population stratification and spurious allelic association. *Lancet* 2003; **361:** 598–604.

11 Altmuller J, Palmer LJ, Fischer G, Scherb H, Wjst M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am J Hum Genet* 2001; **69:** 936–50.

12 Altshuler D, Hirschhorn JN, Klannemark M, et al. The common PPARg Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nat Genet* 2000; **26:** 76–80.

13 Bertina RM, Koeleman BP, Koster T, et al. Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* 1994; **369:** 64–67.

14 Hugot JP, Chamaillard M, Zouali H, et al. Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* 2001; **411:** 599–603.

15 Corder EH, Saunders AM, Strittmatter WJ, et al. Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* 1993; **261:** 921–23.

16 Rubinsztein DC, Easton DF. Apolipoprotein E genetic variation and Alzheimer's disease. a meta-analysis. *Dement Geriatr Cogn Disord* 1999; **10:** 199–209.

17 Mann V, Ralston SH. Meta-analysis of COL1A1 Sp1 polymorphism in relation to bone mineral density and osteoporotic fracture. *Bone* 2003; **32:** 711–17.

18 Mann V, Hobson EE, Li B, et al. A COL1A1 Sp1 binding site polymorphism predisposes to osteoporotic fracture by affecting bone density and quality. *J Clin Invest* 2001; **107:** 899–907.

19 Zareparsi S, Branham KE, Li M, et al. Strong association of the Y402H variant in complement factor H at 1q32 with susceptibility to age-related macular degeneration. *Am J Hum Genet* 2005: **77:** 149–53.

20 Hageman GS, Anderson DH, Johnson LV, et al. A common haplotype in the complement regulatory gene factor H (HF1/CFH) predisposes individuals to age-related macular degeneration. *Proc Natl Acad Sci USA* 2005; **102:** 7227–32.

21 Haines JL, Hauser MA, Schmidt S, et al. Complement factor H variant increases the risk of age-related macular degeneration. *Science* 2005; **308:** 419–21.

22 Klein RJ, Zeiss C, Chew EY, et al. Complement factor H polymorphism in age-related macular degeneration. *Science* 2005; **308:** 385–89.

23 Edwards AO, Ritter R 3rd, Abel KJ, Manning A, Panhuysen C, Farrer LA. Complement factor H polymorphism and age-related macular degeneration. *Science* 2005; **308:** 421–24.

24 Gloyn AL, Weedon MN, Owen KR, et al. Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes. *Diabetes* 2003; **52:** 568–72.

25 Ueda H, Howson JM, Esposito L, et al. Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature* 2003; **423:** 506–11.

26 Marron MP, Raffel LJ, Garchon HJ, et al. Insulin-dependent diabetes mellitus (IDDM) is associated with CTLA4 polymorphisms in multiple ethnic groups. *Hum Mol Genet* 1997; **6:** 1275–82.

27 Chistiakov DA, Turakulov RI. CTLA-4 and its role in autoimmune thyroid disease. *J Mol Endocrinol* 2003; **31:** 21–36.

28 Bennett ST, Todd JA. Human type 1 diabetes and the insulin gene: principles of mapping polygenes. *Annu Rev Genet* 1996; **30:** 343–70.

29 Engel LS, Taioli E, Pfeiffer R, et al. Pooled analysis and meta-analysis of glutathione S-transferase M1 and bladder cancer: a HuGE review. *Am J Epidemiol* 2002; **156:** 95–109.

30 Pennacchio LA, Olivier M, Hubacek JA, et al. An apolipoprotein influencing triglycerides in humans and mice revealed by comparative sequencing. *Science* 2001; **294:** 169–73.

31 McCarthy MI. Progress in defining the molecular basis of type 2 diabetes mellitus through susceptibility-gene identification. *Hum Mol Genet* 2004; **13** (suppl 1): R33–41.

32 Weedon MN, Schwarz PE, Horikawa Y, et al. Meta-analysis and a large association study confirm a role for calpain-10 variation in type 2 diabetes susceptibility. *Am J Hum Genet* 2003; **73:** 1208–12.

33 McCarthy MI, Smedley D, Hide W. New methods for finding disease-susceptibility genes: impact and potential. *Genome Biol* 2003; **4:** 119.

34 Johnson GC, Esposito L, Barratt BJ, et al. Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29:** 233–37.

35 Carlson CS, Eberle MA, Rieder MJ et al. Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am J Hum Genet* 2004; **74:** 106–20.

36 Wright AF, Hastie ND. Complex genetic diseases: controversy over the Croesus code. *Genome Biol* 2001; **2:** 1–8.

37 Weiss KM, Clark AG. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18:** 19–24.

38 Reich DE, Lander ES. On the allelic spectrum of human disease. *Trends Genet* 2001; **17:** 502–10.

39 Gabriel SB, Schaffner SF, Nguyen H, et al. The structure of haplotype blocks in the human genome. *Science* 2002; **296:** 2225–29.

40 Tabor HK, Risch NJ, Myers RM. Candidate-gene approaches for studying complex genetic traits: practical considerations. *Nat Rev Genet* 2002; **3:** 391–97.

41 Hudson TJ. Wanted: regulatory SNPs. *Nat Genet* 2003; **33:** 439–40.

42 Silander K, Mohlke KL, Scott LJ, et al. Genetic variation near the hepatocyte nuclear factor-4 gene predicts susceptibility to type 2 diabetes. *Diabetes* 2004; **53:** 1141–49.

43 Love-Gregory L, Wasson J, Ma J, et al. A common polymorphism in the upstream promoter region of the hepatocyte nuclear factor4 gene on chromosome 20q is associated with type 2 diabetes and appears to contribute to the evidence for linkage in an Ashkenazi Jewish population. *Diabetes* 2004; **53:** 1134–40.

44 Thomas H, Jaschkowitz K, Bulman M, et al. A distant upstream promoter of the HNF-4alpha gene connects the transcription factors involved in maturity-onset diabetes of the young. *Hum Mol Genet* 2001; **10:** 2089–97.

45 Boj SF, Parrizas M, Maestro MA, Ferrer J. A transcription factor regulatory circuit in differentiated pancreatic cells. *Proc Natl Acad Sci USA* 2001; **98:** 14481–86.

46 Loots GG, Ovcharenko I, Pachter L, Dubchak I, Rubin EM. rVista for comparative sequence-based discovery of functional transcription factor binding sites. *Genome Res* 2002; **12:** 832–39.

47 Knight JC, Keating BJ, Rockett KA, Kwiatkowski DP. In vivo characterization of regulatory polymorphisms by allele-specific quantification of RNA polymerase loading. *Nat Genet* 2003; **33:** 469–75.

48 Lo HS, Wang Z, Hu Y, et al. Allelic variation in gene expression is common in the human genome. *Genome Res* 2003; **13:** 1855–62.

49 Carlson CS, Eberle MA, Kruglyak L, Nickerson DA. Mapping complex disease loci in whole-genome association studies. *Nature* 2004; **429:** 446–52.

50 Ozaki K, Ohnishi Y, Iida A, et al. Functional SNPs in the lymphotoxin-α gene that are associated with susceptibility to myocardial infarction. *Nat Genet* 2002; **32:** 650–54.

51 Jellema A, Zeegers MP, Feskens EJ, Dagnelie PC, Mensink RP. Gly972Arg variant in the insulin receptor substrate-1 gene and association with type 2 diabetes: a meta-analysis of 27 studies. *Diabetologia* 2003; **46:** 990–95.

52 Teng J, Risch N. The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases. II. Individual genotyping. *Genome Res* 1999; **9:** 234–41.

53 Frayling TM, Wiltshire S, Hitman GA, et al. Young-onset type 2 diabetes families are the major contributors to genetic loci in the Diabetes UK Warren 2 genome scan and identify putative novel loci on chromosomes 8q21, 21q22, and 22q11. *Diabetes* 2003; **52:** 1857–63.

54 Fingerlin TE, Boehnke M, Abecasis GR. Increasing the power and efficiency of disease-marker case-control association studies through use of allele-sharing information. *Am J Hum Genet* 2004; **74:** 432–43.

55 Bacanu SA, Devlin B, Roeder K. The power of genomic control. *Am J Hum Genet* 2000; **66:** 1933–44.

56 Hopper JL, Bishop DT, Easton DF. Population-based family studies in genetic epidemiology. *Lancet* (in press).

57 Campbell MJ, Julious SA, Altman DG. Estimating sample sizes for binary, ordered categorical, and continuous outcomes in two group comparisons. *BMJ* 1995; **311:** 1145–48.

58 Wacholder S, Chanock S, Garcia-Closas M, El Ghormli L, Rothman N. Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J Natl Canc Inst* 2004; 96: 434–42.

59 Marchini J, Cardon LR, Phillips MS, Donnelly P. The effects of human population structure on large genetic association studies. *Nat Genet* 2004; **36:** 512–17.

60 Mein CA, Barratt BJ, Dunn MG, et al. Evaluation of single nucleotide polymorphism typing with invader on PCR amplicons and its automation. *Genome Res* 2000; **10:** 330–43.

61 Bogardus ST Jr, Concato J, Feinstein AR. Clinical epidemiological quality in molecular genetic research: the need for methodological standards. *JAMA* 1999; **281:** 1919–26.

62 Kang SJ, Gordon D, Finch SJ. What SNP genotyping errors are most costly for genetic association studies? *Genet Epidemiol* 2004; **26:** 132–41.

63 Gordon D, Finch SJ, Nothnagel M, Ott J. Power and sample size calculations for case-control genetic association tests when errors are present: application to single nucleotide polymorphisms. *Hum Hered* 2002; **54:** 22–33.

64 Ioannidis JP. Effect of the statistical significance of results on the time to completion and publication of randomized efficacy trials. *JAMA* 1998; **279:** 281–86.

65 Curtis D, Sham PC. A note on the application of the transmission disequilibrium test when a parent is missing. *Am J Hum Genet* 1995; **56:** 811–12.

66 Knapp M, Becker T. Impact of genotyping errors on type I error rate of the haplotype-sharing transmission/disequilibrium test (HS-TDT). *Am J Hum Genet* 2004; **74:** 589–91.

67 Mitchell AA, Cutler DJ, Chakravarti A. Undetected genotyping errors cause apparent overtransmission of common alleles in the transmission/disequilibrium test. *Am J Hum Genet* 2003; **72:** 598–610.

68 Kirk KM, Cardon LR. The impact of genotyping error on haplotype reconstruction and frequency estimation. *Eur J Hum Genet* 2002; **10:** 616–22.

69 Oliphant A, Barker DL, Stuelpnagel JR, Chee MS. BeadArray technology: enabling an accurate, cost-effective approach to high-throughput genotyping. *Biotechniques* 2002; **32** (suppl): S56–61.

70 Ewen KR, Bahlo M, Treloar SA, et al. Identification and analysis of error types in high-throughput genotyping. *Am J Hum Genet* 2000; **67:** 727–36.

71 Clark AG. Inference of haplotypes from PCR-amplified samples of diploid populations. *Mol Biol Evol* 1990; **7:** 111–22.

72 Excoffier L, Slatkin M. Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* 1995; **12:** 921–27.

73 Stephens M, Donnelly P. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 2003; **73:** 1162–69.

74 Niu T, Qin ZS, Xu X, Liu JS. Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms. *Am J Hum Genet* 2002; **70:** 157–69.

75 Stephens M, Smith NJ, Donnelly P. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 2001; **68:** 978–89.

76 Schaid DJ, McDonnell SK, Wang L, Cunningham JM, Thibodeau SN. Caution on pedigree haplotype inference with software that assumes linkage equilibrium. *Am J Hum Genet* 2002; **71:** 992–95.

77 Page GP, George V, Go RC, Page PZ, Allison DB. "Are we there yet?": deciding when one has demonstrated specific genetic causation in complex diseases and quantitative traits. *Am J Hum Genet* 2003; **73:** 711–19.

78 Lander E, Kruglyak L. Genetic dissection of complex traits: guidelines for interpreting and reporting linkage results. *Nat Genet* 1995; **11:** 241–47.

79 Keavney B, McKenzie C, Parish S, et al. Large-scale test of hypothesised associations between the angiotensin-converting-enzyme insertion/deletion polymorphism and myocardial infarction in about 5000 cases and 6000 controls. *Lancet* 2000; **355:** 434–42.

80 Redden DT, Allison DB. Nonreplication in genetic association studies of obesity and diabetes research. *J Nutr* 2003; **133:** 3323–26.

81 Storey JD, Tibshirani R. Statistical significance for genomewide studies. *Proc Natl Acad Sci USA* 2003; **100:** 9440–45.

82 Nyholt DR. A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74:** 765–69.

83 Schaid DJ. General score tests for associations of genetic markers with disease using cases and their parents. *Genet Epidemiol* 1996; **13:** 423–49.

84 Epstein MP, Satten GA. Inference on haplotype effects in case-control studies using unphased genotype data. *Am J Hum Genet* 2003; **73:** 1316–29.

85 Kraft P. Multiple comparisons in studies of gene x gene and gene x environment interaction. *Am J Hum Genet* 2004; **74:** 582–84.

86 Weinberg CR. Methods for detection of parent-of-origin effects in genetic studies of case-parents triads. *Am J Hum Genet* 1999; **65:** 229–35.

87 Lykken DT. Statistical significance in psychological research. *Psychol Bull* 1968; **70:** 151–59.