Genome-wide association studies for complex traits: consensus, uncertainty and challenges

Mark I. McCarthy[*‡], *Gonçalo R. Abecasis*[§], *Lon R. Cardon*[*||], *David B. Goldstein*[¶], *Julian Little*[#], *John P. A. Ioannidis*[**‡‡] *and Joel N. Hirschhorn*[§§||||¶¶]

Abstract | The past year has witnessed substantial advances in understanding the genetic basis of many common phenotypes of biomedical importance. These advances have been the result of systematic, well-powered, genome-wide surveys exploring the relationships between common sequence variation and disease predisposition. This approach has revealed over 50 disease-susceptibility loci and has provided insights into the allelic architecture of multifactorial traits. At the same time, much has been learned about the successful prosecution of association studies on such a scale. This Review highlights the knowledge gained, defines areas of emerging consensus, and describes the challenges that remain as researchers seek to obtain more complete descriptions of the susceptibility architecture of biomedical traits of interest and to translate the information gathered into improvements in clinical management.

Genome-wide association (GWA) studies
Studies in which a dense array of genetic markers, which captures a substantial proportion of common variation in genome sequence, is typed in a set of DNA samples that are informative for a trait of interest. The aim is to map susceptibility effects through the detection of associations between genotype frequency and trait status.

*Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, UK.
Correspondence to M.I.M
e-mail: mark.mccarthy@drl.ox.ac.uk
doi:10.1038/nrg2344
Published online 9 April 2008

The first wave of large-scale, high-density genome-wide association (GWA) studies has improved our understanding of the genetic basis of many complex traits[1]. For several diseases, including type 1 (REFS 2,3) and type 2 diabetes[4–9], inflammatory bowel disease[10–14], prostate cancer[15–20] and breast cancer[21–23], there has been rapid expansion in the numbers of loci implicated in predisposition. For others, such as asthma[24], coronary heart disease[25–27] and atrial fibrillation[28], fewer novel loci have been found, although opportunities for mechanistic insights are equally promising. Several common variants influencing important continuous traits, such as lipids[7,29–31], height[32–35] and fat mass[36–38], have also been found. An updated list of published GWA studies can be found at the National Cancer Institute (NCI)-National Human Genome Research Institute (NHGRI)'s catalog of published genome-wide association studies.

These findings are providing valuable clues to the allelic architecture of complex traits in general. At the same time, many methodological and technical issues that are relevant to the successful prosecution of large-scale association studies have been addressed. However, despite understandable celebration of these achievements, sober reflection reveals many challenges ahead. Compelling signals have been found, often highlighting previously unsuspected biology, but, for most of the traits studied, known variants explain only a fraction of observed familial aggregation[39], limiting the potential for early application to determine individual disease risk. Because current technology surveys only a limited subset of potentially relevant sequence variation, this should come as no surprise. Much work remains to obtain a complete inventory of the variants at each locus that contribute to disease risk and to define the molecular mechanisms through which these variants operate. The ultimate objectives — full descriptions of the susceptibility architecture of major biomedical traits and translation of the findings into clinical practice — remain distant.

With completion of the initial wave of GWA scans, it is timely to consider the status of the field. This Review considers each major step in the implementation of a GWA scan, highlighting areas where there is an emerging consensus over the ingredients for success, and those aspects for which considerable challenges remain.

Subject ascertainment and design

Although there is a growing focus on the application of GWA methodologies to population-based cohorts, most published GWA studies have featured case–control designs, which raise issues related to the optimal selection of both case and control samples.

© 2008 Nature Publishing Group

Case–control design
An association study design in which the primary comparison is between a group of individuals (cases), ascertained for the phenotype of interest and that are presumed to have a high prevalence of susceptibility alleles for that trait, and a second group (controls), not ascertained for the phenotype and considered likely to have a lower prevalence of such alleles.

Selection bias
Bias arising from the fact that the samples ascertained for the study (particularly controls) might not be representative of the wider population that they are purported to represent.

Misclassification bias
Bias resulting from the failure to correctly assign individuals to the relevant group in a case–control study; for example, the presence of some individuals who meet the criteria for being cases in a population-based control sample.

Population stratification
The presence in study samples of individuals with different ancestral and demographic histories: if cases and controls differ with respect to these features, markers that are informative for them might be confounded with disease status and lead to spurious associations.

Case selection. The principal issues with regard to case ascertainment revolve around the extent to which selection should be driven by manoeuvres that are designed to improve study power through enrichment for specific disease-predisposing alleles. These include efforts to minimize phenotypic heterogeneity or to focus on extreme and/or familial cases (defined, for example, by early age of onset or ascertainment from multiplex pedigrees). Because the genetic architecture of most complex traits remains poorly understood, the value of such efforts is hard to predict. In most circumstances, and particularly when the total GWA sample size has financial or operational constraints, efforts to enrich case selection are likely to improve power. However, there are situations in which selection of familial cases or extreme individuals might have the opposite effect[40,41].

Control selection. Optimal selection of control samples remains more controversial, although the accumulating empirical data indicate that many commonly expressed concerns have been overstated. The Wellcome Trust Case Control Consortium (WTCCC) study was able to demonstrate the effectiveness of a 'common control' design in which 3,000 UK controls were compared with 2,000 cases from each of 7 different diseases[1]. The WTCCC also assuaged concerns about the potential for selection bias when using non-population-based controls[1]. Comparison of the genome-wide genotypic distributions from the two constituents of the WTCCC common-control resource (one derived from a population-based birth cohort, the other from opportunistic sampling of blood donors) revealed no excess of significant associations, indicating that ascertainment, selection and survival biases were, in this situation at least, having minimal impact on genotype distributions. Although each prospective control sample must be critically evaluated, these findings suggest that a broad range of ascertainment schemes are compatible with GWA analysis.

One consequence of the common-control design is the potential loss of power that is associated with the inability to exclude latent diagnoses of the phenotype of interest through intensive screening of controls. Fortunately, the consequences of misclassification bias are modest unless the trait is common, and any loss of power is recoverable by increasing the sample size (BOX 1). For common traits, such as obesity and hypertension, in which the effect of misclassification on power is greatest[1], one remedy involves adopting a more stringent case definition, for example, based on early age of onset or ascertainment of a more extreme phenotype, while still excluding monogenic cases. Although the most powerful strategy for a given fixed sample size involves a 'hypernormal' control group, it might be difficult to identify such individuals without introducing inadvertent selection effects. For instance, selecting extremely low-weight individuals as controls for a case–control study of obesity could result in overrepresentation of alleles primarily associated with chronic medical disease or nicotine addiction rather than weight regulation *per se*.

Other case–control design issues. Four other issues loom large in the design of case–control studies. The first is sample size, and with this issue the consensus view is clear: the more samples the better[1,34,35,38]. The initial wave of GWA studies has shown that, with rare exceptions, the effect sizes resulting from common SNP associations are modest, and that sample sizes in the thousands are essential[1].

The second issue relates to the propensity for latent population substructure (population stratification and cryptic relatedness) to inflate the type 1 error rate and generate spurious claims of association around variants that are informative for that substructure[42,43]. The evidence emerging from GWA studies is reassuring: as long as cases and controls are well matched for broad ethnic background, and measures are taken to identify and exclude individuals whose GWA data reveal substantial differences in genetic background, the impact of residual substructure on type 1 error seems modest[1]. Several statistical tools exist to detect and adjust for residual stratification[42,44], and inventories of markers that are informative for the detection of ethnic substructure are a useful by-product of current scans[1,45–47]. These approaches can be used to adjust for substructure even in populations with quite diverse antecedents (such as European-descent populations in North America)[46,47] and with negligible impact on power[48]. Analysis in African-descent populations is complicated by their greater haplotypic diversity and fine-scale geographical structure[49], and by the extensive admixture demonstrated by African-descent populations that are resident in Europe and North America. Furthermore, it is important to note that the tools mentioned above (particularly genomic-control approaches[44]) correct for 'average' genome-wide measures of ethnic admixture, and will not always eliminate spurious associations immediately adjacent to markers that are strongly informative about ancestry.

Author addresses

[‡]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Oxford, UK.
[§]Center for Statistical Genetics, Department of Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
[‖]Fred Hutchinson Cancer Research Center, Seattle, Washington 98109, USA.
[¶]Center for Population Genomics and Pharmacogenetics, Duke Institute for Genomic Sciences and Policy, Duke University Medical Center, Duke University, Durham, North Carolina 27708, USA.
[#]Department of Epidemiology and Community Medicine, University of Ottawa, Ottawa, Ontario, Canada.
[**]Clinical and Molecular Epidemiology Unit, Department of Hygiene and Epidemiology, University of Ioannina School of Medicine, and Biomedical Research Institute, Foundation for Research and Technology-Hellas, Ioannina 45110 Greece.
[‡‡]Department of Medicine, Tufts University School of Medicine, Boston, Massachusetts 02111, USA.
[§§]Division of Genetics and Endocrinology and Program in Genomics, Children's Hospital, Boston, Massachusetts 02115, USA.
[‖‖]Department of Genetics, Harvard Medical School, Boston, Massachusetts 02115, USA.
[¶¶]Broad Institute at MIT and Harvard, Cambridge, Massachusetts 02142, USA.

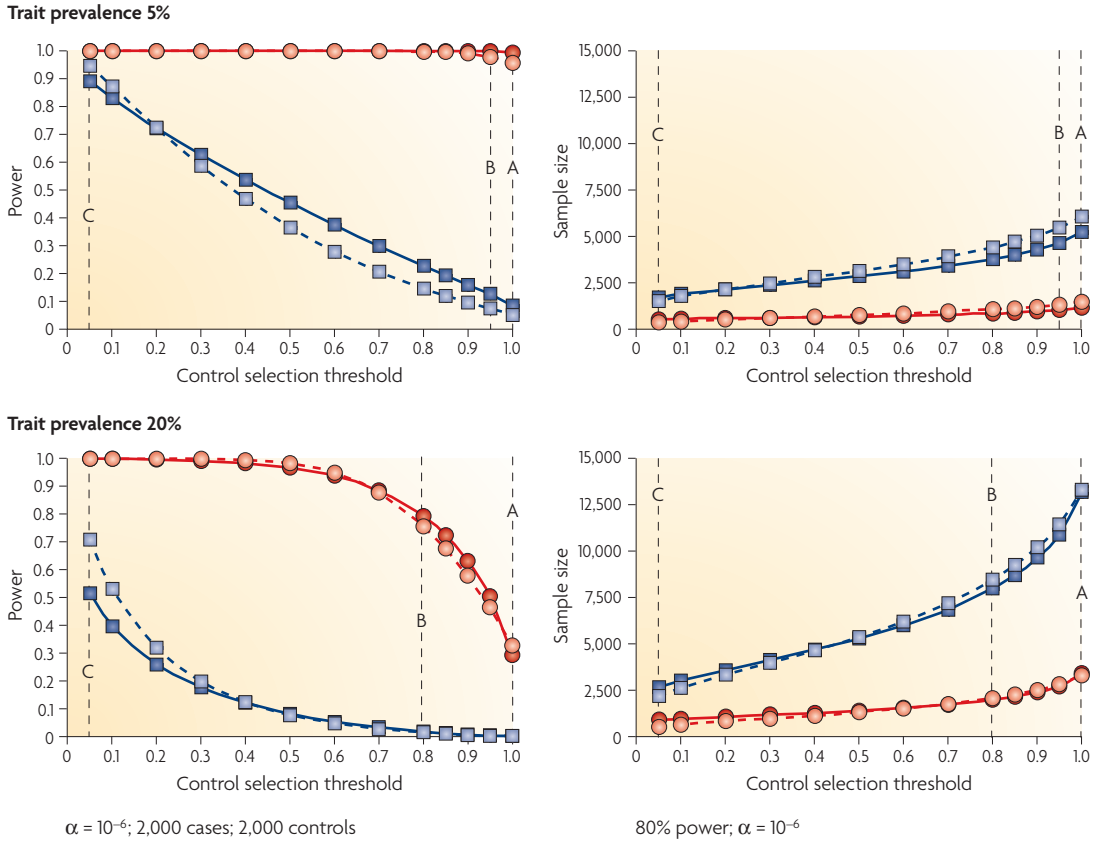
© 2008 Nature Publishing Group

Cryptic relatedness
Evidence — typically gained from analysis of GWA data — that, despite allowance for known family relationships, individuals in the study sample have residual, non-trivial degrees of relatedness, which can violate the independence assumptions of standard statistical techniques.

Family-based association methods
A suite of analytical approaches in which association testing is performed within families: such approaches offer protection from population substructure effects but at the price of reduced power.

The third issue concerns the relative merits of family-based and case–control association methods. Although family-based association methods provide a robust strategy for dealing with stratification, this typically comes at the cost of reduced power[50]. Given the ease with which GWA data enable the detection of, and correction for, population substructure[42,44], this particular justification has become less persuasive. Nevertheless, there are many valuable clinical resources (for example, isolates) for which pedigree information can be usefully exploited. One option for the efficient use of family data in such a setting is to restrict high-density scanning to a subset of pedigree members and then use information on patterns of chromosomal segregation derived from low-density

Box 1 | **The impact of selection by phenotype among controls on power and sample size**



In a case–control study, the manner in which the controls are ascertained (with respect to the phenotype of interest) has implications for the power of the study and for sample size. The panels on the left show estimates of power for a sample size of 2,000 cases and 2,000 controls and α (p value) = 10^{-6}. Those on the right show the sample sizes (that is, the number of case–control pairs) that are required for 80% power at the same threshold. In the upper panels, the disease of interest has a population prevalence of 5% (so that cases are ascertained purely from the top 5% of the population distribution); in the lower panels, the population prevalence is 20%. In each panel, power or sample size estimates are shown for a range of control selection thresholds, that is, the trait-distribution threshold that is used to define the controls. Under scenario A, controls are ascertained from the full distribution (that is, population-based controls): a proportion (5% or 20%) will meet the criteria for being cases. Under scenario B, controls are ascertained only if they cannot be cases: they come from the residual part (bottom 80% or 95%) of the distribution. Under scenario C, hypernormal controls have been selected exclusively from the lowest 5% of the distribution. Each panel considers four potential susceptibility loci. Tracks in blue denote loci that account for 0.25% of overall trait variance, tracks in red denote loci that account for 1%. Light red and light blue symbols denote that the variant responsible is common (overall allele frequency 30%), red and dark blue symbols denote that the variant is rare (1%).

As expected, in all settings, scenario C is the most powerful strategy for given overall case–control sample size, and scenario A is the least powerful strategy. When the disease prevalence is modest (5%; upper panels), the distinctions between scenarios A and B are not large, and it will often be easier to increase sample size than to undertake detailed phenotypic examination of the controls to exclude latent cases. When the disease prevalence is higher (20%; lower panels), misclassification is more prevalent under scenario A, the adverse consequences of using population-based controls are more marked, and the advantages of using hypernormal controls (scenario C), if available, are most obvious.

© 2008 Nature Publishing Group

genotyping in the remaining members to propagate genotypes through the family[51].

The fourth question relates to the potential to use historical control genotypes to substitute for, or supplement, newly typed controls in future GWA studies. The risks associated with this (particularly inflation of the type 1 error) will clearly depend on the extent to which there are disparities between the new cases and historical controls with respect to population origins, DNA format (whole-genome amplified DNA versus native DNA as well as storage conditions)[52,53] and genotyping implementation (platform, genotyping centre, generation of chip or allele-calling software). The limits of acceptable divergence are not yet known, but it seems safest that studies intending to use historical control data also type a sample of ethnically matched controls (including a subset of the historical control samples, if available) using the same assay as for the newly defined cases. This should allow the detection of systematic effects that are attributable to non-disease-related differences between the historical and new data. An alternative approach would involve the re-genotyping of any interesting GWA signals in all samples using a dedicated assay.

From case–control to cohort studies. Increasingly, the GWA approach is being extended from analysis of case–control samples to population-based cohorts[54–56]. Although typically underpowered for dichotomous phenotypes (given limited cases for any given disease), such cohorts often offer a rich tapestry of longitudinal measures for a wide range of quantitative traits, and lifestyle and exposure data can enable an evaluation of the joint effects of genes and environment. These studies promise new insights into the genetic basis of continuous traits and enhanced opportunities for revealing pleiotropy[57], although low power remains an issue — especially for the detection of non-additive gene–environment interactions[58,59]. Whereas GWA data meta-analysis is the obvious solution to overcome restrictions of sample size, such procedures are often complicated by the lack of standardization that characterizes the measurement of many key continuous biological traits and environmental exposures[60].

Implementation

Marker selection and assay design. The debates about marker selection that dominated early discussions of GWA approaches have now boiled down to choices between a limited range of commodity genome-wide chips[61,62]. This is not the place for a detailed discussion of the relative merits of specific array-designs other than to point out that, genotype for genotype, designs that take linkage disequilibrium (LD) structure into account when defining content will achieve greater coverage, in the index population at least. However, they will also be more vulnerable to loss of coverage when assays fail, or when typing samples whose genetic ancestry differs from that of the reference panel or panels used to guide marker selection[61,62]. Although greater feature density, which allows more variants to be typed in a single array, increases coverage, this does not necessarily equate to greater power. When funding is limited, but the availability of samples is not, overall power might be maximized by typing more individuals with a less dense and less costly array. In populations of non-African ancestry, some of the drive towards ever-increasing array density has been blunted by the capability to impute genotypes at untyped loci (discussed later)[63,64].

A new option in array selection comes with the inclusion on contemporary commodity arrays of additional probes that are designed to type copy number variants (CNVs)[65]. However, because the global inventory of CNVs remains incomplete[66], and with limited empirical data currently available, the extent to which the current round of products (such as the Affymetrix 6.0 and Illumina Human1M arrays) captures the structural variome remains unclear. However, their use should provide early insights into the contribution such variants make to common phenotypes of biomedical importance[67].

For any given study, the final choice of array platform is often a pragmatic one, based not only on the number and ethnic origin of the samples, and the overall research objectives, but also on factors such as cost, array delivery schedules and available genotyping capacity.

DNA pooling. The costs of well-powered GWA studies have reignited interest in the value of DNA pooling approaches as a means to conduct more economical genome-wide surveys for association[68]. However, even though several studies have been completed[69], the falling costs of commodity genotyping and the intrinsic limitations of the pooling approach (reduced power, loss of individual genotype data and difficulties ensuring equimolar representation of samples) mean that the future of this approach, for GWA studies at least, is uncertain.

Obtaining robust genotype data. Experience from the first wave of GWA studies has demonstrated that scrupulous attention to detail is required throughout because each stage is fraught with the potential for error and bias[1,52,53,70,71]. Many of these errors and biases have the potential to generate extreme values for the association test statistic; if uncorrected, these can dominate the tails of the distribution, such that interesting true associations become lost in a sea of spurious signals. Efforts to prevent, detect and eradicate sources of bias and error therefore remain a high priority in GWA studies[1], despite continuing improvements in genotyping performance.

These efforts start with careful attention to the quality and accurate quantification of the starting DNA. Differences in extraction methods between cases and controls can be an important source of bias[52,53]. Implementation of genotyping-performance metrics allows poorly performing arrays to be targeted for re-analysis and deficient samples to be selected for replacement[1].

Given the scale of data generation, conversion of raw experimental data into genotypes has necessitated the development of automated methods. Indeed, the very idea of assigning a discrete genotype call has increasingly been replaced by measures of the posterior probability of each possible genotype, given the observed data[1]. Several algorithms have been developed for defining the three

Pleiotropy
The phenomenon whereby a single allele can affect several distinct aspects of the phenotype of an organism, often traits not previously thought to be mechanistically related.

Linkage disequilibrium
(LD). The nonrandom allocation of alleles at nearby variants to individual chromosomes as a result of recent mutation, genetic drift or selection, manifest as correlations between genotypes at closely linked markers.

Copy number variant
(CNV). A class of DNA sequence variant (including deletions and duplications) in which the result is a departure from the expected diploid representation of DNA sequence.

DNA pooling approaches
Association studies that are conducted using estimates of allele frequencies derived from pools of DNA compiled from multiple subjects rather than individual DNA samples.

© 2008 Nature Publishing Group

genotype clusters at a given SNP, and for assigning genotype calls to each[1,53,72–74], with each generation of software heralding steady improvements in both accuracy and call rate. This last point is crucial because there is no room for complacency regarding SNPs that display low call rates[1,53]. Case–control studies, in particular, are vulnerable to informative missingness, whereby spurious associations arise through differences in the patterns of missing data with respect to genotype.

Considerations such as these have highlighted a tension between stringency and call rate, which has prevented the promulgation of quality-control thresholds with a universal value for defining the set of 'clean' genotypes for analysis. If researchers aim to maximize accuracy by setting the threshold for calling genotypes too high, the consequence for many SNPs will be a low call rate (such that some true signals are discarded) and/or a rise in type 1 error (due to informative missingness). However, if a lower threshold is favoured, as some groups have done[1], call rates will be preserved (and informative missingness minimized) at the expense of accuracy. Those who adopt this more liberal strategy accept that a non-trivial number of poorly performing SNPs will survive the quality-control process and might be disproportionately represented among the most extreme association signals. If resources are not to be wasted in fruitless validation and replication studies, it becomes essential to subject interesting signals to individual reviews of quality-control parameters (especially visualization of the signal intensity (cluster) plots) (BOX 2) before proceeding. One corollary is that GWA data-sharing efforts should extend to the provision of raw signal data as well as 'finished' genotypes.

Once armed with a set of called genotypes, the final phase of quality control beckons. Experience has shown that most SNPs showing extreme departures from Hardy–Weinberg equilibrium (HWE) in controls can be safely discarded[1], although lesser (but nevertheless quite marked) departures are to be expected under the null hypothesis, given the number of tests performed. Appropriate thresholds for any given study will depend on the sample size and overall data quality, and might best be defined by using the observed distribution of HWE statistics (the WTCCC used this approach to set a threshold at an exact $p < 5.7 \times 10^{-7}$ in controls)[1]. In any event, HWE is an imprecise tool for quality control purposes. Tests of departure from HWE are underpowered for the detection of genotyping error[75,76], whereas overenthusiastic use of HWE as a quality criterion can prove to be counterproductive given that modest disequilibrium (in cases particularly) can be a signature of true association.

Recent studies have emphasized the importance of detecting individuals whose GWA data reveal an ancestry that is discrepant with self-described labels, allowing them to be removed from consideration[1] or analysed separately. Similarly, sample integrity needs to be confirmed using recorded gender or previously obtained genotypes. Inadvertent sample duplication and swaps, cross-contamination and cryptic relatedness[43] are frequently revealed by analysis of GWA data, and, if unresolved, would violate assumptions of statistical independence and introduce misclassification effects. It is straightforward to identify first- and second-degree relatives at least and exclude one member of each pair from analysis. In the presence of more remote relationships, options include explicit modelling of those relationships, or adjustment through genomic-control approaches[43,44].

This sequence of manoeuvres has proved challenging enough for experienced groups given the size of the data sets concerned. However, all groups working with such data, whether generated themselves or downloaded from the web, need to appreciate the intricacies of data quality control if they are to avoid potential misinterpretation. The comments above refer to the implementation of SNP-based scans, but the task of converting intensity traces and SNP-based data into multi-allelic CNV genotypes is far more demanding and is only now being tackled[65].

Analysis and interpretation

Diagnostic plots. A limited number of formats have emerged as standard tools for representing the data emerging from a GWA scan, with the quantile-quantile plot (Q-Q plot) among the most widely used[52,77] (BOX 2). These plots help to indicate whether the study has generated more significant results than expected by chance and to put such findings in context. Undetected population stratification or cryptic relatedness result in deviation from the null across the entire distribution, whereas large-effect susceptibility loci generate deviations at the highly significant end of the range.

Single-point analyses. In most situations, the most powerful tool for the analysis of GWA data has been a single-point, one degree of freedom test of association, such as the Cochran–Armitage test. Such tests allow comparison of the genotype distributions of cases and controls at each SNP in turn, and can be conducted with or without adjustment for relevant covariates, such as the principal components of population substructure[42]. Although the Cochran–Armitage method directly tests only one of several possible genetic models, it has the merit of being robust to modest deviations from additivity on the logistic scale (at least to those most likely to be biologically relevant). Furthermore, in situations in which the true model at the causal variant is non-additive, even modest departures from perfect LD will result in greatly reduced power to detect that non-additivity at nearby variants: in the GWA context therefore, in situations when few causal variants will be directly typed, the additive model is likely to perform well. Whereas the use of alternative models (general, dominant or recessive) could result in enhanced detection of some signals[78], the use of multiple correlated tests also complicates computation of type 1 error rates and can reduce the efficiency of subsequent follow-up efforts.

Historically, interpretation of genetic association findings has adopted the standard frequentist approach to the evaluation of significance. From such a view-point, GWA results are compared against a single criterion of genome-wide significance. Although several benchmarks have been proposed, in European-descent GWA studies opinion is coalescing around the need to adjust for 1–2 million independent tests, which results in a target α (p value) of $\sim 5 \times 10^{-8}$ (REFS 49,79,80). However, such an approach fails

Informative missingness
If patterns of missing data are nonrandom with respect to both genotype and trait status, then analysis of the available genotypes can result in misleading associations where none truly exists.

Signal intensity (cluster) plots
Plots of raw intensity data for individual variants that are generated by the genotyping platform and represent the extent to which the various genotypes can be discriminated: these provide a useful visual diagnostic for the genotyping data quality.

Hardy–Weinberg equilibrium
(HWE). A theoretical description of the relationship between genotype and allele frequencies that is based on expectation in a stable population undergoing random mating in the absence of selection, new mutations and gene flow: in the context of genetic studies, departures from equilibrium can be used to highlight genotyping errors.

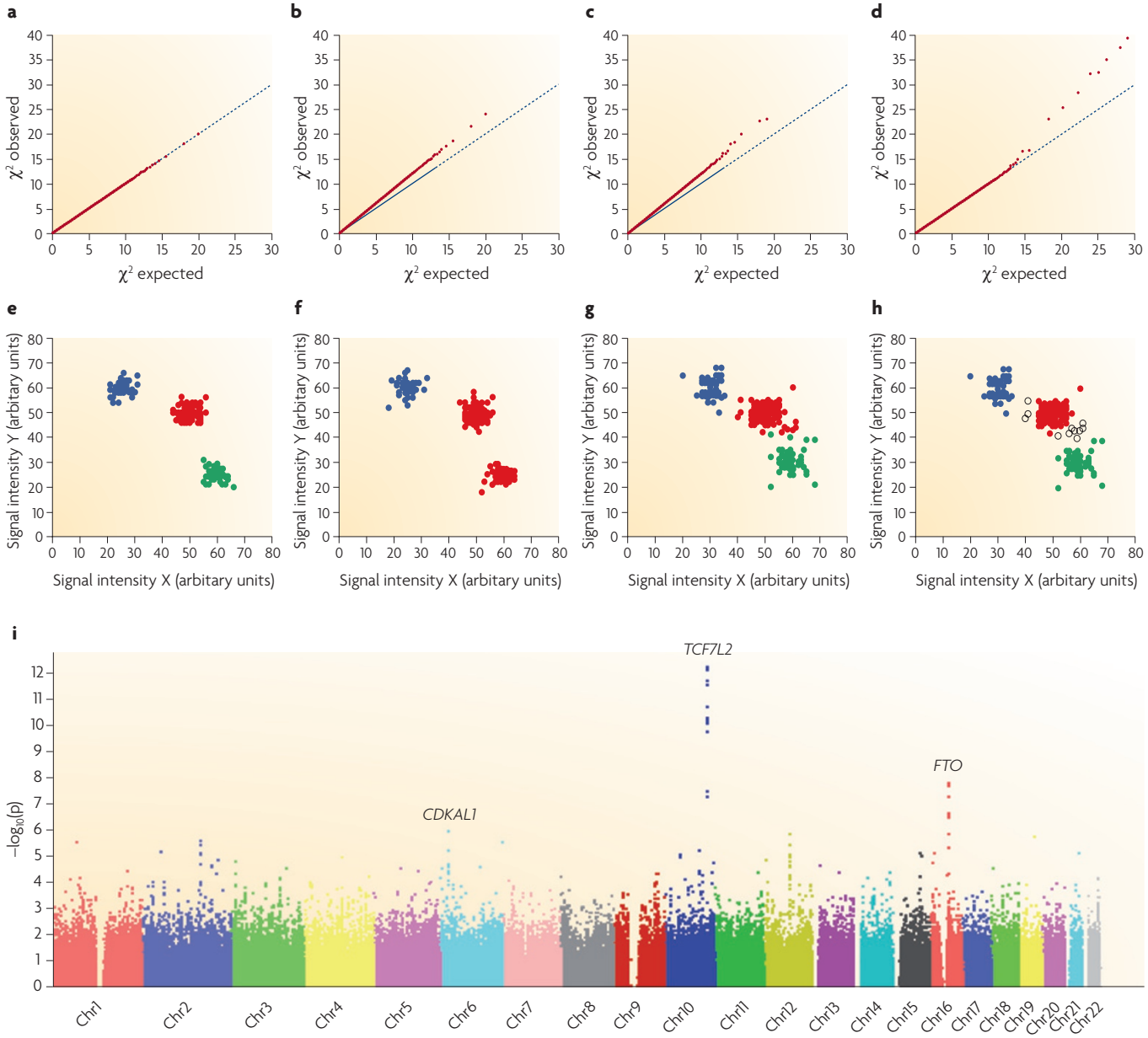
Quantile-quantile plot
(Q-Q plot). In the context of GWA studies, a Q-Q plot is a diagnostic plot that compares the distribution of observed test statistics with the distribution expected under the null.

Cochran–Armitage test
A genotype-based contingency-table test for association that is well suited to the detection of trends across ordinal categories (in this case, genotypes).

Frequentist
A school of statistics that uses p values and combines them with hypothesis testing to make inferences.

© 2008 Nature Publishing Group

Box 2 | **Visualization of genome-wide association data**



Quantile-quantile (Q-Q) plots provide a visual summary of the distribution of the observed test statistics generated by a genome-wide association (GWA) study[52,77]. Typically, a single test statistic (for case–control studies, a chi-squared (χ^2) comparison of absolute genotype counts) is calculated for each variant passing quality control. In panels **a**–**d**, the blue line denotes expectation under the null and red circles indicate idealized test results from hypothetical GWA data, generated under four scenarios: in panel **a** the observed data conforms closely to expectation indicating little evidence for association; in panel **b** inflation of the observed findings across the distribution is seen, indicative of population stratification or cryptic relatedness; in panel **c** there is similar evidence of population substructure, but some suggestion of an excess of strong associations; in panel **d** there is little evidence of substructure, but compelling evidence for an excess of disease associations.

Signal intensity (cluster) plots provide diagnostics at the level of individual SNPs. Typically, raw data from the genotyping platform is plotted along two axes (one for each allele) to define clusters of data corresponding to the three genotype groups. Panels **e**–**h** display idealized plots based on ~200 genotypes. In panel **e** the three clusters are well defined and individual genotypes are accurately called (as shown by the three colours). In panel **f** the clusters are well defined, but an error in allele calling has led to two clusters being assigned the same genotype. In panel **g** significant overlap between clusters is likely to result in failure to call certain genotypes (shown in open symbols in panel **h**). In this example, all failed genotypes are either heterozygotes or homozygotes for the green allele: this generates biased estimates of genotype frequencies, which can result in spurious association signals owing to informative missingness.

Finally, genome-wide Manhattan plots display GWA findings with respect to their genomic positions, highlighting signals of particular interest. In panel **i**, an example from the type 2 diabetes component of the Welcome Trust Case Control Consortium study[1,5], the strongest associations are seen on chromosomes 10 (transcription factor 7-like 2; *TCF7L2*), 16 (fat mass and obesity associated; *FTO*) and 6 (CDK5 regulatory subunit associated protein 1-like 1; *CDKAL1*). Additional strong signals on chromosomes 1, 2 and 12 did not replicate.

© 2008 Nature Publishing Group

to account for factors such as the power of a study and the number of likely true positives, which are important components of any comprehensive evaluation of GWA findings[81]. Consider a small case–control GWA performed for a condition for which there is only weak evidence for genetic involvement: any associations found are unlikely to be genuine, whatever the p value obtained. Such limitations have excited interest in Bayesian approaches, which incorporate information on the likely number of true associations, and the power of a given study to detect associations of a given magnitude, into estimates of credibility[81,82]. These methods, which generate measures such as Bayes' factors, or the false-positive report probability rather than p values, avoid a single threshold for genome-wide significance but depend on the ability to assign plausible probabilities to each of the alternate hypotheses (see REF. 1 for further discussion). Nevertheless, both approaches conclude that only very low p values equate to strong evidence for association (that is, are associated with a low false-positive rate). Crucially, most GWA signals that have attained such significance levels have subsequently been confirmed by replication[1].

Multi-marker analyses. As expected, given the incomplete coverage of common variation that is provided by contemporary GWA platforms[61,62], a modest boost to power can be provided by computational approaches that improve the detection of associations that are attributable to variants that have not themselves been directly typed[83]. Haplotype-based methods and imputation methods are related but complementary approaches to achieve this. In situations when haplotype-based analyses[83,84] reveal evidence for association that exceeds that of any directly typed SNP in the vicinity (after allowance for the increased dimensionality), one can invoke either an effect that is directly attributable to the haplotype (that is, independent causal *cis* effects at multiple SNPs) or the explanation that the haplotype tags more efficiently than any individual genotyped SNP, an as yet untyped aetiological variant. Imputation methods[63,64] rely on information from sets of resequenced and/or densely genotyped individuals to infer missing genotypes at untyped variants. Because data from the International HapMap Consortium[49] are typically used as the reference, imputation methods have proved most powerful in recovering associations and causal effects attributable to HapMap SNPs that are not included on commercial arrays. Importantly, the use of such methods is not restricted to samples drawn from HapMap reference populations[85].

Challenges. Immediate challenges in this area are numerous. The imminent arrival of large-scale genome-wide CNV data[65] has focused attention on the development of methods that are suited to the specific features of such data (multiallelic, semi-quantitative and probably more error prone) and methods that facilitate the integration of CNV and SNP information. There is work to be done to understand how best to incorporate the intrinsic uncertainty that is associated with genotype calls derived from both direct and, in particular, imputed data[63,64], and in the development of analysis tools that take account of, and/or estimate, information on population history, which will prove especially valuable for studies in genetic isolates[86]. Evaluation of the contribution of rare variants to common disease susceptibility raises issues related to detection (rare variants are poorly captured by the standard GWA arrays[87]) and functional assessment (the sheer number of such variants and the limited power to test them for association[88]). Finally, there is a need for improved methods to estimate the joint effects of multiple genes and/or environmental exposures on disease predisposition. Such analyses raise both computational and statistical issues, related to the scale and complexity of the data and the large number of hypotheses that could be addressed[89].

Validation and replication

The importance of replication. The use of small samples, which are underpowered to detect loci of realistic effect size, and over-liberal declarations of association are the main reasons why so few of the complex-trait associations that were claimed in the pre-GWA era proved genuine[81,90]. This history, together with the high dimensionality of GWA studies, their vulnerability to a range of errors and biases, and the modest effect sizes to be anticipated for most complex-trait susceptibility alleles, help to explain the pre-eminent role of replication in the evaluation of GWA findings[91,92].

Terminology in this area can be confusing. Here, we use 'technical validation' to refer specifically to the reanalysis of original GWA samples using a second genotyping platform. Technical validation allows early detection of technical errors in typing or imputation that might have generated a spurious association signal, and is an important prelude to large-scale efforts to evaluate selected signals in additional independent samples (referred to here as 'replication').

Best practice. The aim of validation and replication is to determine which of the findings arising from the primary GWA reflect true reproducible associations. Accordingly, the focus is not merely to provide additional evidence to support or refute the original association, but also the systematic appraisal of potential sources of error and bias that could have been responsible. Hence, it is important to use independent replication samples, and to use distinct genotyping assays to expose technical artefacts. Credibility is increased when multiple investigative groups find the same association in independent samples.

Claims of replication should be reserved for findings involving the same allele or haplotype (or an established proxy thereof), the same phenotype and the same genetic model as the original signal. Otherwise the risk of spurious claims of association is increased by the testing of multiple hypotheses. This has an important bearing on decisions about the extent to which early replication efforts at a given signal should focus exclusively on the index variant, as opposed to including additional adjacent SNPs[93]. The inclusion of adjacent SNPs runs the risk of generating a profusion of apparent associations around spurious GWA signals — complicating interpretation of the evidence for replication — and can involve a substantial waste of genotyping effort around the many false-positive signals.

Bayes' factors
The Bayesian alternative to classical frequentist approaches to hypothesis testing, essentially equivalent to likelihood ratio tests: prior and posterior information are combined in a ratio that measures the strength of the evidence in favour of one model rather than the other.

False-positive report probability
The probability that a reported association between a genetic variant and a trait of interest is not true.

Haplotype-based methods
Association methods that rely on the relationship between the distribution of estimated haplotype frequencies and trait status, rather than each individual variant in turn.

Imputation methods
A set of approaches for filling in missing genotype data using a sparse set of genotypes (for example, from a GWA scan) and a scaffold of linkage disequilibrium relationships (as provided by the HapMap).

© 2008 Nature Publishing Group

Box 3 | **Informative heterogeneity**

Replication of results is an essential step in establishing that the associations revealed by initial genome-wide association (GWA) studies are genuine. Failure to replicate (in an otherwise well-performed and well-powered study) is usually interpreted as indicating that the initial finding was spurious. However, failure to replicate can also result from substantive differences between the discovery and replication studies, for example, with respect to sample ascertainment.

A recent example is provided by the highly significant association between variants in the fat mass and obesity associated (*FTO*) gene and type 2 diabetes that was first detected in a UK GWA scan (odds ratio for diabetes ~1.27, $p = 2 \times 10^{-8}$) and subsequently strongly replicated in other UK samples (odds ratio for diabetes ~1.22, $p = 5 \times 10^{-7}$)[1,5,36]. However, this association could not be replicated in several well-powered diabetes GWA scans[4,6–8]. The explanation for these divergent findings derives from the fact that *FTO* was shown to influence diabetes risk through a primary effect on weight regulation, such that the diabetes risk allele is associated with higher fat mass, weight and body mass index[36,37]. In the UK studies, marked differences in weight between diabetic cases and non-diabetic controls meant that differences in *FTO* genotype frequencies were observed in diabetes case–control analyses. Several other diabetes GWA scans had explicitly targeted case selection on relatively lean individuals[4,7] (to remove the confounding effects of obesity), thereby abolishing the differences in weight between the diabetic cases and controls, and with it the between-group differences in *FTO* genotype distributions. Rather than dismissing the *FTO* association with type 2 diabetes as a failure of replication, identification of the source of the heterogeneity (what might be termed informative heterogeneity) provided an explanation for the discrepant findings and highlighted the likely mechanism of its action.

In most instances, therefore, it seems wise first to obtain definitive evidence of association at the index variant. So that failure to replicate can be meaningfully interpreted, the samples used in early rounds of replication should be broadly similar (with respect to ethnicity and ascertainment) to those of the original study. Of course, as loci are validated as truly causal, extension into multiple ethnicities is highly desirable to test the generalizability and consistency of the proposed association, although such efforts will probably entail additional efforts at variant discovery given ethnic differences in LD.

Multi-stage designs. Several recent reports have emphasized the potential advantages of multi-stage designs, in which signals from an initial, first-stage GWA are used to define a subset of SNPs that are retyped in additional second-stage samples[94–96]. Such designs have been seen as an effective way of retaining power while reducing genotyping costs. However, the substantial price differential between commodity and custom genotyping means that those cost benefits can be less dramatic than comparisons of genotype numbers alone would suggest. In circumstances in which the second-stage and GWA samples have similar provenance, so that the prospects for appreciable genetic heterogeneity between the stages are low, the best-powered analytical strategy involves a joint analysis (in which the distribution of test statistics across the data from both stages combined is considered), rather than the conventional replication design (which considers the second-stage results in isolation)[94]. Such considerations blur the boundaries of where exactly replication starts, but whichever analytical approach is taken, confirmation in many independent samples is important and it is the overall strength of the evidence of association that matters.

Power, sample size and heterogeneity. An appreciation of power and sample size is central to the design and interpretation of appropriate replication studies. Studies that lack the power to offer convincing support or refutation of the original finding can generate misleading inferences when considered in isolation, although combinations of such studies might be of value provided that all suitable studies have been included. Calculations of replication sample size need to consider the so-called 'winner's curse' effect, whereby the original study will typically overestimate the true effect size[97]. Replication efforts that fail to make such allowance will probably be underpowered[98,99].

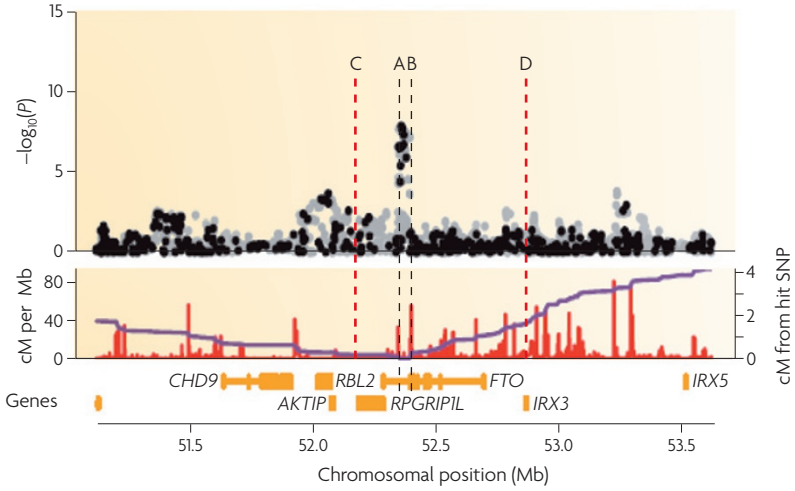
If well-performed replication studies confirm the original findings, then the evidence in favour of association is enhanced (unless of course both the original and replication studies have succumbed to the same errors). Interpretation of a failure to replicate is more difficult. If it is clear that the replication studies were well powered and well performed, and that there is genuine divergence between the effect-size estimates (for example, no overlap of 95% confidence intervals), then there are two possible explanations. Either the original finding was wrong, or the difference in findings is attributable to some source of heterogeneity[100,101]. The list of potential causes of heterogeneity is long: it includes variable patterns of LD between the genotyped SNP and untyped causal alleles (although this is unlikely if the samples are of similar ancestry); differences in the distribution, frequency or effect size of the causal alleles at a given locus (due to, for example, drift or selection, or differences in case ascertainment); and the impact of non-additive interactions with other genetic variants or environmental exposures.

We should be wary of appealing to heterogeneity as a rationale for failure to replicate, as over-eagerness to deploy such an explanation would mean that no report of association could ever be refuted. Nevertheless, there are established instances in which the effects of proven associations can vary substantially across studies, so clearly circumstances exist where it can be justified.

The role of variants in the fat mass and obesity associated (*FTO*) gene on the risk of diabetes and obesity (BOX 3) provides one such example, illustrating how a clear case for heterogeneity can be made, especially when the initial association finding has been robustly replicated and the source of the heterogeneity is apparent[1,5,36,37]. Some sources of heterogeneity can be directly evaluated, including differences in LD structure between populations (particularly once the true causal allele has been identified), clear variation in case ascertainment that can be correlated with effect size or a highly significant interaction with a well-measured covariate. Other sources of heterogeneity, such as an unmeasured or poorly defined environmental exposure, will be difficult or impossible to demonstrate. The number of associations for which heterogeneity has an important role cannot be reliably estimated from the evidence to date: because our inventory of 'proven' complex-trait associations is heavily biased towards variants that have shown consistent replication across studies (this being such an important factor in determining proof), there is likely to be a substantial under-representation of causal loci that feature appreciable between-study

© 2008 Nature Publishing Group

Box 4 | **Strategies for resequencing within genome-wide association signals**



Because genome-wide association (GWA) studies directly genotype only a small proportion of the variants that segregate within the population examined, it is unlikely that the causal variant(s) will be among those for which genotype data are available. Imputation methods[63,64] allow association analysis to be extended to a larger set of variants (HapMap SNPs for which reliable imputation is possible), but this still comprises a minority of all common SNPs, and representation of rare variants is poorer still[87]. Targeted resequencing allows recovery of a more complete inventory of sequence variation within regions of interest, and enables systematic fine-mapping efforts to identify those putatively causal variants with the strongest effects on disease susceptibility.

For any given region, the extent and scope of resequencing efforts depends on the strategic goal. Consider the example of the fat mass and obesity associated (*FTO*) region that is implicated in obesity and type 2 diabetes (see the figure)[1,5,36]. Association analysis using directly typed (black) and imputed (grey) SNPs localized the association signal to a 47 kb region (A–B) defined by flanking recombination hot spots (red trace), within intron 1 of the *FTO* gene. If the goal is to identify common causal variants that could underlie this index signal, it should suffice to resequence this ~50 kb interval in ~200 individuals (a total of ~10 Mb). If the aim is, instead, to identify all common variants contributing to regional susceptibility (including those independent of the index signal, so-far missed owing to incomplete coverage) then all sequence that is relevant to gene expression or function (arbitrarily, C–D) needs to be considered (~600 kb, hence a total of ~120 Mb). Should the aim extend to identification of rare variants with independent effects on disease risk, recovery of the full allelic spectrum of sequence variants will require deep resequencing in ~500–1000 individuals: in the first instance, such deep resequencing efforts might be preferentially targeted to exonic and conserved non-coding sequence. *AKTIP*, AKT interacting protein; *CHD9*, chromodomain helicase DNA binding protein 9; *IRX3*, iroquois homeobox 3; *IRX5*, iroquois homeobox 5; *RBL2*, retinoblastoma-like 2; *RPGRIP1L*, retinitis pigmentosa GTPase regulator interacting protein 1-like.

heterogeneity[102]. Moreover, estimation of between-study heterogeneity is highly imprecise when the number of studies is limited (less than 10) and/or their individual power is low. Systematic efforts to define the impact of gene–gene and gene–environment interactions, well-powered GWA studies in multiple ethnicities and comprehensive exploration of variation around strong signals of association should all help to address this point.

Finding additional susceptibility genes

Meta-analysis. Individual GWAs are underpowered to detect all but the biggest effects, and the susceptibility variants identified so far are probably only a subset of the loci that would be detectable using this approach if power were increased[39]. Joint (meta) analysis of data from comparable GWA scans[9,34,35,38,103] provides a low-cost approach to enhance power for both main and joint (gene–gene and gene–environment) effects, obtain *in silico* replication, inform SNP selection for subsequent replication efforts and explore potential sources of heterogeneity. In type 2 diabetes, joint analysis of three GWA scans (4,700 cases and 5,700 controls when combined)[5–7,9] was central to the robust identification of several novel susceptibility variants, and allowed confirmation of the role of previously known susceptibility variants of modest effect size, such as those in *KCNJ11* (potassium inwardly-rectifying channel, subfamily J, member 11) and *PPARG* (peroxisome proliferator-activated receptor gamma).

As increasingly large data sets are deployed to find signals for which smaller samples were underpowered, the average effect size of the novel variants that are discovered will decrease. For some purposes, such as predictive diagnostics, this has implications for their likely translational impact. Even so, it is worth remembering that owing to imperfect LD, a modest finding from the initial GWA might ultimately lead to fine mapping of variants with considerably larger effect sizes. When it comes to insights into disease pathogenesis, however, locus effect size is almost immaterial: even loci with modest effects, once they are established as genuine, can reveal novel causal mechanisms. Thus, discovery efforts are likely to continue to bear fruit as long as the clinical, logistical and financial resources are available to support them, and, as genotyping costs fall, ever larger studies will become possible.

The technical aspects of successful data combination have been widely discussed in the meta-analysis literature[100,104]. Clearly, researchers seeking to obtain an unbiased estimate of the significance of a given association are duty-bound to be as inclusive as possible[90]. Although it might be tempting to exclude particular studies on the basis of some perceived failure of quality or potential heterogeneity, any such decisions must be carefully justified. In the presence of heterogeneity, random-effects models provide more appropriate estimates of overall effect size, because heterogeneity violates the basic assumption of fixed-effects analysis[100]. Although summary-level data will suffice for many meta-analysis purposes, access to primary individual-level data allows for more sophisticated reanalysis, including the capacity to undertake haplotype and conditional analyses, to perform imputation, to examine the joint effects of genes and environment, and to explore phenotypic heterogeneity.

Reintroducing biology. So far, most efforts at replication have concentrated, not unreasonably, on the signals for which the statistical evidence is strongest. However, the efficient identification of additional susceptibility loci with more modest effect sizes might benefit from the integration of statistical evidence with some assessment of functional candidacy. Several prioritization strategies for defining variant subsets with increased 'prior' odds for association can be envisaged, embracing aspects of biological candidacy (for example, variants mapping to genes that interact with, or lie within the same pathways as, those

© 2008 Nature Publishing Group

Box 5 | **Challenges in following-up confirmed associations**

Dramatic advances in identifying gene variants that influence human complex traits have yet to be accompanied by consistent progress in understanding the mechanisms by which these variants influence disease. For most associations, systematic efforts to identify the underlying causal variant or variants have yet to be reported. There is a clear need to establish tools — both bioinformatic and experimental — to support efforts to map confirmed associations and thereby maximize the biological information gathered from GWA studies. On the bioinformatics side, for example, tools are required to display association data in the context of the increasingly rich functional annotation of the genome (for example, the WGAViewer[126]).

On the experimental side, the availability of genome-wide profiles of gene expression and alternate splicing across a range of human tissues and/or cell lines, alongside dense genotype data from the same samples, would be a valuable resource. Even for major cell types these do not currently exist[108–110]. The paucity of such information for multiple brain tissues and defined populations of primary lymphocytes, for example, impedes progress in defining causal mechanisms in neuropsychiatric, infectious and autoimmune diseases. In some situations, we might expect causal variants to have quite marked effects on expression phenotypes; in other situations, their molecular consequences will be subtle, or restricted in terms of space (for example, tissue specific) or time (for example, to particular periods of development). Although it is straightforward to resequence the exons of a gene of interest to screen for non-synonymous coding variants, establishing whether a variant some hundreds of kilobases from the gene could be exerting modest effects on the expression of specific transcripts in particular tissues is far more challenging, and far less amenable to large-scale genome-wide analysis.

that were previously implicated in susceptibility to the disease of interest) and genomic annotation (for example, all variants capturing variation in microRNA target sites).

Other populations. Among the GWA scans that have already been performed there has been a strong bias towards samples of North European origin[1–38]. There are excellent reasons for extending analyses to samples from populations with differing mutational and demographical histories, including other major ethnic groups[49] and population isolates[86]. Each such sample offers new opportunities in terms of detectable susceptibility variants, generating ethnic-specific patterns with respect to the identity of the loci themselves, the frequency and LD relationships of disease-susceptibility alleles, and the presence of additional genetic or environmental factors with which they might interact. Studies in other populations are therefore capable of revealing novel susceptibility loci and aetiological pathways, as well as assisting with the fine-mapping of causal variants within those loci already confirmed[105]. Given the large sample sizes that are needed to achieve these ends, the ascertainment of GWA and replication sample sets from diverse ethnic groups is a priority.

Other sequence variants. Finally, it is worth reiterating that, so far, GWA scans have focused almost exclusively on the detection of effects that are attributable to common SNPs. The first wave of GWA arrays offered limited power to capture structural variants[106] and rare variants of any type[87]. The arrival of tools that are better-suited to the direct evaluation of these variant types is likely to provide the most immediate source of novel susceptibility loci, although each brings its own set of technical and analytical challenges[65].

Following-up confirmed signals

The confirmed signals emerging from GWA scans and subsequent replication efforts are just that — association signals. The causal variants will only occasionally be among those that are directly typed in GWA scans, and the interval within which the aetiological variant(s) are expected to lie (typically defined in terms of flanking recombination hot spots) can be sizeable, often containing several genes[1,24]. Even worse, because it seems likely that many complex-trait susceptibility effects are mediated through remote regulatory elements[107], the coding exons of the susceptibility gene could lie well beyond the interval of maximal association.

Resequencing and fine mapping. The task of moving from confirmed association signal to complete enumeration of the pattern of causal variants at a given locus poses significant challenges. On occasion, useful shortcuts to exhaustive fine mapping might be available. The set of associated variants might include a number with particularly strong biological credentials — a non-synonymous coding SNP in a compelling biological candidate, for example. Alternatively, clues can be gathered from expression studies[108–110]: the cluster of associated variants might display strong *cis* associations with expression of one of the nearby genes, transcript levels of which are themselves associated with the phenotype of interest[13,24]. Although caution is warranted in placing weight on *in silico* or *in vitro* functional data (the concordance between epidemiological associations and measurable functional effects has historically been poor[111]), such findings can provide a rapid route towards direct functional confirmation of the implicated molecular mechanisms.

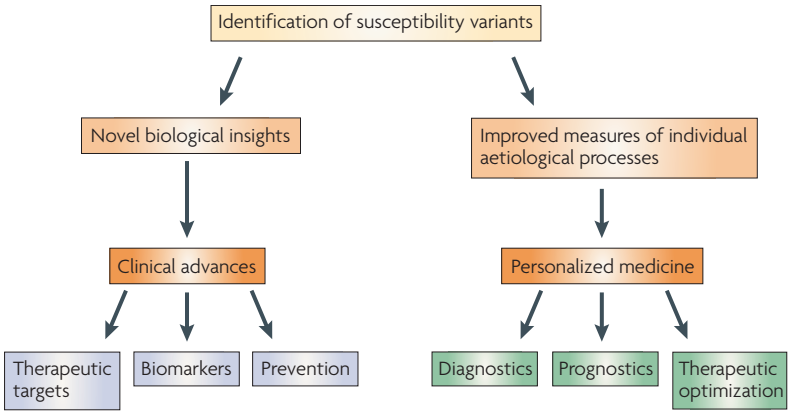
In the absence of such insights, further progress will generally require exhaustive examination of the region, first, to generate a comprehensive inventory of regional variation, and second, to use fine-mapping approaches to define the signals with the strongest statistical claims[112]. Despite recent advances in sequencing and genotyping, both are daunting tasks.

Resequencing plans need to be tailored to the desired objective (BOX 4). Implementation raises a host of unanswered issues related to optimal study design and data interpretation. There are numerous choices to be made: first, about the samples to be resequenced, that is, the balance of HapMap individuals versus disease cases and whether to favour cases carrying the known susceptibility variant or haplotype; second, the depth of resequencing to be undertaken, which defines the allele spectrum recovered; and finally, the merits of extending the search for rare variants beyond well-annotated sequence, given the difficulties associated with obtaining robust evidence for a statistical or functional effect for rare variants in unannotated sequence (BOX 4).

Related challenges lie in the development of efficient strategies for fine mapping that take into account the desire both to discriminate between highly correlated variants (to determine which variants are causal) and to search for additional, independent signals. Given the relatively high price of custom genotyping,

© 2008 Nature Publishing Group

Box 6 | **Clinical translation of findings from GWA studies**



Recent successes in the identification of susceptibility variants that underlie many important biomedical phenotypes has increased confidence that this information can be translated into clinically beneficial improvements in management. There are two principal routes through which such translation might be effected (see the figure).

In the first, identification of novel causal pathways provides new opportunities for clinical advances of generic benefit to all those suffering from (or at risk of) the disease concerned. This might involve identification of therapeutic targets within causal pathways, leading to novel therapeutic agents for treatment and/or prevention. Identification of causal pathways should bolster efforts to identify biomarkers, allowing improved disease prediction and monitoring of disease progression and treatment response. Sometimes, genetic discoveries can highlight important environmental contributors to disease, enabling public-health-based disease prevention measures. Note that even modest genotype–phenotype associations (provided they are confirmed as genuine through extensive replication and functional studies) can offer significant new translational opportunities through the identification of novel modifiable pathways.

The second translational route lies through using knowledge of individual patterns of disease predisposition (for example, through genetic profiling) to develop more personalized approaches to disease management. The major limitation here, for most complex traits, is that the variants so far identified explain only a small proportion of individual variation in disease risk[39]. Consequently, for most individuals (all but the small proportion who have inherited extremely high or low numbers of susceptibility alleles for a given disease), genetic profiling using currently available markers provides limited information on disease risk beyond that available from conventional risk factors. If genetic profiling is to become widely applied in clinical practice, we need to: improve the accuracy of risk prediction through identification of additional susceptibility variants; demonstrate, by prospective studies, that profiling results in beneficial modifications of medical care and/or personal responsibility; and establish an appropriate regulatory environment for the use of such tests.

exhaustive fine-mapping efforts across multiple signals can cost far more than the original GWA scan. Multi-stage approaches to fine mapping (homing in on causal variants through successive rounds of genotyping, whereby ever fewer SNPs are typed in increasingly large samples) might seem efficient but they can involve costly, time-consuming investment in successive assay designs. Given inter-ethnic differences in patterns of LD and mutational histories (see above), samples from other ethnicities (especially those of African origin) should be extremely valuable tools for fine mapping, and considerable effort is being expended to identify samples on the scale required for robust inference.

Mendelian randomization
An analytical approach that allows one to test for a causal relationship between two phenotypes that show observational associations, but are subject to confounding: Mendelian randomization makes use of the random segregation of susceptibility alleles at meiosis to explore causality in a model that is freed from most sources of confounding.

Function and translation. Once all the statistical evidence has been extracted and putative causal variants are identified, substantial challenges remain (BOX 5). Prominent among these is the need to obtain functional confirmation that the variants implicated are truly causal, and to reconstruct the molecular and physiological mechanisms whereby they have an impact on the phenotype of interest. Because many of the complex-trait susceptibility variants so far identified map to sequence of unknown function that is some distance from the nearest coding sequence[5–9,15,17,25–27], the design of contextually appropriate functional assays is far from straightforward. Improved tools for the functional annotation of the genome (for example, from the ENCODE project)[107] will be essential, but generating such annotations for all tissues and/or cells of biomedical relevance remains a monumental task.

It is also important to move beyond the selected samples used in the discovery phase and evaluate the impact of the variants on unbiased population samples. From an epidemiological perspective, consensus guidelines have been proposed to grade the strength of the epidemiological evidence for a given association[113]. These take into account the extent of the evidence, the consistency of the replication and the protection from bias in the accumulated data. Population-based studies will enable research to establish whether, using studies of the joint effects of genes and environment and Mendelian randomization approaches[114], genetic discoveries can be used to pinpoint modifiable environmental exposures. Such analyses will require very large study samples for which both genetic and environmental exposures are accurately measured[58,59].

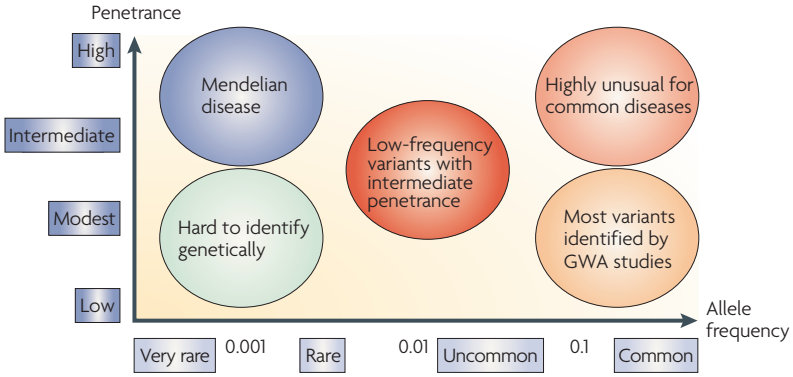
Finally, the ultimate objective of genetic research lies in the translation of the findings into advances in clinical care (BOX 6). The mechanistic insights generated by gene discovery might identify new therapeutic targets and lead to novel pharmaceutical and preventative approaches. In addition, there is growing expectation that individual patterns of genetic predisposition will be of value in health-care delivery (personalized medicine). For many, but not all[115], diseases the modest effect sizes of the variants emerging from GWA studies limits the degree to which this is possible (BOX 7). Higher penetrance, lower frequency variants that are not detected by current GWA approaches but are amenable to new high-throughput sequencing efforts might prove more valuable in this respect[116].

Reporting and deposition

Because wide availability of data is central to the success of many of these endeavours, there has been substantial investment in structures to support data-sharing between investigators, such as the National Institutes of Health (NIH)-supported dbGAP programme[117], and the European Bioinformatics Institute-based European Genotyping Archive. Moreover, the NIH recently announced a policy for sharing of data obtained in NIH supported or conducted GWA studies. Although aggregate or summary data (genotype frequencies and association p values by group as provided by the WTCCC

© 2008 Nature Publishing Group

Box 7 | **Low-frequency variants and disease susceptibility**



Genome wide association (GWA) studies are proving adept at identifying common variants contributing to the inherited component of common diseases. Almost all such variants seem to have modest effect sizes and, even when combined, their impact on overall population variance and predictive power is limited[127]. There is a marked disparity between the extent of overall familial aggregation observed for many common diseases and that attributable to variants identified to date. In type 2 diabetes, known variants collectively account for a sibling relative risk of ~1.07 in Europeans, way below the overall figure (~3) from epidemiological studies[4–9].

Although the identification of additional common risk variants (at the already identified loci, and the others that are yet to be found) will explain some of this deficit, one emerging hypothesis anticipates that a significant proportion of this 'missing heritability' will be attributable to low-frequency variants with intermediate penetrance effects, which have been largely refractory to conventional gene-discovery approaches[116].

Consider a hypothetical variant with a minor allele frequency of 1% and an allelic odds ratio of 3. Given a disease prevalence of 5%, the penetrance of the risk homozygote (~45%) is too low to support Mendelian segregation and detection by traditional linkage approaches. At the same time, the low risk-allele frequency means low detectability by GWA[1,87]. Yet this variant has a stronger effect on familial risk than most known common susceptibility variants: a locus-specific sibling relative risk of 1.038 comfortably exceeds that of the strongest diabetes-susceptibility effects — that of transcription factor 7-like 2 (*TCF7L2*), which is strongly associated to diabetes, is approximately 1.025. As few as thirty such variants across the genome would jointly generate a sibling relative risk >3, and offer impressive predictive power (a discriminative accuracy of 77%)[127]. Novel resequencing technologies, allied to large-scale association testing, provide the potential to identify and characterize variants with these properties and evaluate their contribution to disease risk. In the first instance, such efforts are likely to be targeted to genes already implicated in disease susceptibility.

and the National Cancer Institute Cancer Genetic Markers of Susceptibility (CGEMS) studies, for example[1,118]) provide an excellent and convenient substrate for many data-sharing purposes, individual-level data (including raw signal-intensity information) provide a more valuable resource that supports a wider range of analyses (see above).

Lack of transparency and incomplete reporting of both data and study methods have raised concerns in many health research fields[119–121], and poor reporting has been associated with biased estimates of effect[122]. The importance of transparency was emphasized by the NCI-NHGRI Working Group on Replication in Association Studies[91]. To help remedy this problem, some groups have advocated the development of evidence-based reporting guidelines similar to the consolidated standards of reporting trials (CONSORT) guidelines for the reporting of randomized clinical trials[123]. In particular, the epidemiology community has recently developed a reporting guidance (strengthening the reporting of observational studies in epidemiology; STROBE) for cross-sectional, case–control and cohort studies[124]: extension of this guidance to genetic association studies is in an advanced stage of development (see the Human Genome Epidemiology Network (HuGeNet) website).

In addition to formal deposition mechanisms, there has been a proliferation of investigator networks, designed to exploit the potential for novel signal discovery through integration of data from large numbers of GWA scans that are informative for traits of interest[1,9,34,35,38,103,125]. For phenotypes such as height and body mass index, which are widely and reliably recorded in many cohorts undergoing GWA analysis, aggregate data from several tens of thousands of scans has facilitated detection of variants with small effects[34,35,38]. Such data-sharing efforts will be vital for the success of the coming wave of cohort-based GWAs, each of which will allow the analysis of large numbers of phenotypes. In this situation, focused replication of the best signals for each trait becomes increasingly unattractive, and enlarging sample size through accumulation of additional GWA data seems to be the most feasible option, at least for initial discovery and validation efforts. Management of these large, dynamic and overlapping consortia can be challenging, particularly with respect to ensuring that all those involved (particularly junior investigators) obtain appropriate credit for their contributions.

Summary and conclusions

The past year has seen a remarkable shift in our capacity to dissect the genetic basis of common diseases and continuous traits of biomedical significance. The GWA approach has proven itself extremely well-suited to the identification of common SNP-based variants with modest to large effects on phenotype. Careful implementation and appropriate interpretation has resulted in discoveries that have proven more robust than many had anticipated. Growing numbers of novel susceptibility loci have been identified, shedding light on the fundamental mechanisms that influence disease predisposition, and much is being learned about the complex relationships between changes in genome sequence and phenotypic variation.

However, we are far from the end of this particular voyage, and recent discoveries are nothing more than initial forays into the *terra incognita* of our genomes. We remain unable to explain more than a small proportion of observed familial clustering for most multifactorial traits, a fact that emphasizes the need to extend analysis to a more complete range of potential susceptibility variants, and to support more explicit modelling of the joint effects of genes and environment. Many of the greatest challenges to be faced in the years ahead lie not so much in the identification of the association signals themselves, but in defining the molecular mechanisms through which they influence disease risk and/or phenotypic expression.

© 2008 Nature Publishing Group

1. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
In this study, high density, genome-wide association data on 17,000 individuals identified many novel complex-trait susceptibility loci and explored key methodological and technical issues relevant to the GWA approach.
2. Todd, J. A. *et al.* Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nature Genet.* **39**, 857–864 (2007).
3. Hakonarson, H. *et al.* A genome-wide association study identifies *KIAA0350* as a type 1 diabetes gene. *Nature* **448**, 591–594 (2007).
4. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
5. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
6. Scott, L. J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
7. Diabetes Genetics Initiative. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
8. Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nature Genet.* **39**, 770–775 (2007).
9. Zeggini, E., Scott, L. J., Saxena, R., Voight, B. & DIAGRAM Consortium. Meta-analysis of genome-wide association data and large-scale replication identifies several additional susceptibility loci for type 2 diabetes. *Nature Genet.* 30 Mar 2008 (doi:10.1038/nrg.120).
10. Parkes, M. *et al.* Sequence variants in the autophagy gene *IRGM* and multiple other replicating loci contribute to Crohn's disease susceptibility. *Nature Genet.* **39**, 830–832 (2007).
11. Duerr, R. H. *et al.* A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
12. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39**, 596–604 (2007).
13. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of *PTGER4*. *PLoS Genet.* **3**, e58 (2007).
14. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in *ATG16L1*. *Nature Genet.* **39**, 207–211 (2007).
15. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genet.* **39**, 631–637 (2007).
16. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nature Genet.* **39**, 977–983 (2007).
This paper is one of the clearest demonstrations so far of the potential for pleiotropy: the same variants in *TCF2* influence risk to both type 2 diabetes and prostate cancer.
17. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genet.* **39**, 645–649 (2007).
18. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genet.* **40**, 310–315 (2008).
19. Gudmundsson, J. *et al.* Common sequence variants on 2p15 and Xp11.22 confer susceptibility to prostate cancer. *Nature Genet.* **40**, 281–283 (2008).
20. Eeles, R. A. *et al.* Multiple newly identified loci associated with prostate cancer susceptibility. *Nature Genet.* **40**, 316–321 (2008).
21. Easton, D. F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
22. Hunter, D. J. *et al.* A genome-wide association study identifies alleles in *FGFR2* associated with risk of sporadic postmenopausal breast cancer. *Nature Genet.* **39**, 870–874 (2007).
23. Stacey, S. N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nature Genet.* **39**, 865–869 (2007).
24. Moffatt, M. F. *et al.* Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
25. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493 (2007).
26. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
27. Samani, N. J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
28. Gudbjartsson, D. F. *et al.* Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007).
29. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genet.* **40**, 161–169 (2008).
30. Kathiresan, S. *et al.* Six new loci associated with blood low-density lipoprotein cholesterol, high-density lipoprotein cholesterol or triglycerides in humans. *Nature Genet.* **40**, 189–197 (2008).
31. Kooner, J. S. *et al.* Genome-wide scan identifies variation in *MLXIPL* associated with plasma triglycerides. *Nature Genet.* **40**, 149–151 (2008).
32. Weedon, M. N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nature Genet.* **39**, 1245–1250 (2007).
This paper demonstrates the power of the GWA approach to identify genes influencing continuous biomedical phenotypes, in this case, height.
33. Sanna, S. *et al.* Common variants in the *GDF5-UQCC* region are associated with variation in human height. *Nature Genet.* **40**, 198–203 (2008).
34. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* (in the press).
35. Lettre, G. *et al.* Genome-wide association studies identify 10 novel loci for height and highlight new biological pathways in human growth. *Nature Genet.* (in the press).
36. Frayling, T. M. *et al.* A common variant in the *FTO* gene is associated with body mass index and predisposes to childhood and adult obesity. *Science* **316**, 889–894 (2007).
37. Scuteri, A *et al.* Genome-wide association scans shows genetic variants in the *FTO* gene are associated with obesity-related traits. *PLoS Genet.* **3**, e115 (2007).
38. Loos, R. J. F. *et al.* Association studies involving over 90,000 people demonstrate that common variants near to MC4R influence fat mass, weight and risk of obesity. *Nature Genet.* (in the press).
39. Altshuler, D. & Daly, M. Guilt beyond a reasonable doubt. *Nature Genet.* **39**, 813–815 (2007).
40. Li, M., Boehnke, M. & Abecasis, G. R. Efficient study designs for test of genetic association using sibship data and unrelated cases and controls. *Am. J. Hum. Genet.* **78**, 778–792 (2006).
41. Howson, J. M., Barratt, B.J., Todd, J. A. & Cordell, H. J. Comparison of population- and family-based methods for genetic association analysis in the presence of interacting loci. *Genet. Epidemiol.* **29**, 51–67 (2005).
42. Price, A. L. *et al.* Principal components analysis corrects for stratification in genome-wide association studies. *Nature Genet.* **38**, 904–909 (2006).
43. Voight, B. F. & Pritchard, J. K. Confounding from cryptic relatedness in case–control association studies. *PLoS Genet.* **1**, e32 (2005).
44. Zheng, G., Freidlin, B. & Gastwirth, J. L. Robust genomic control for association studies. *Am. J. Hum. Genet.* **78**, 350–356 (2006).
45. Paschou, P. *et al.* PCA-correlated SNPs for structure identification in worldwide human populations. *PLoS Genet.* **3**, e160 (2007).
46. Tian, C. *et al.* Analysis and application of European genetic substructure using 300K SNP information. *PLoS Genet.* **4**, e4 (2008).
47. Price, A. L. *et al.* Discerning the ancestry of European Americans in genetic association studies. *PLoS Genet.* **4**, e236 (2008).
48. Fellay, J. *et al.* A whole-genome association study of major determinants for host control of HIV-1. *Science* **317**, 944–947 (2007).
49. International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
50. Laird, N. M. & Lange, C. Family-based designs in the age of large-scale gene-association studies. *Nature Rev. Genet.* **7**, 385–394 (2006).
51. Chen, W. M. & Abecasis, G. R. Family-based association tests for genomewide association scans. *Am. J. Hum. Genet.* **81**, 913–926 (2007).
52. Clayton, D. G. *et al.* Population structure, differential bias and genomic control in a large-scale, case–control association study. *Nature Genet.* **37**, 1243–1246 (2005).
This paper presents a detailed description of the potential for bias and error to complicate the analysis of large-scale genetic association data.
53. Plagnol, V., Cooper, J. D., Todd, J. A. & Clayton D. G. A method to address differential bias in genotyping in large-scale association studies. *PLoS Genet.* **3**, e74 (2007).
54. Cupples, L. A. *et al.* The Framingham Heart Study 100k SNP genome-wide association study resource: overview of 17 phenotype working group reports. *BMC Med. Genet.* **8**, S1 (2007).
55. Ridker, P. M. *et al.* Rationale, design, and methodology of the Women's Genome Health Study: A genome-wide association study of more than 25,000 initially healthy American women. *Clin. Chem.* **54**, 249–255 (2008).
56. Li, S. *et al.* The *GLUT9* gene is associated with serum uric acid levels in Sardinia and Chianti cohorts. *PLoS Genet.* **3**, e194 (2007).
57. Cordell, H. J. & Clayton, D. G. Genetic association studies. *Lancet* **366**, 1121–1131 (2005).
58. Wong, M. Y., Day, N. E., Luan, J. A., Chan, K.P & Wareham, N. J. The detection of gene–environment interaction for continuous traits: should we deal with measurement error by bigger studies or better measurement? *Int. J. Epidemiol.* **32**, 51–57 (2003).
59. Wong, M. Y., Day, N. E., Luan, J. A. & Wareham, N. J. Estimation of magnitude in gene–environment interactions in the presence of measurement error. *Stat. Med.* **23**, 987–998 (2004).
60. Burke, W., Khoury, M. J., Stewart, A., Zimmern, R. L. & Bellagio Group. The path from genome-based research to population health: development of an international public health genomics network. *Genet. Med.* **8**, 451–458 (2006).
61. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38**, 659–662 (2006).
62. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genet.* **38**, 663–667 (2006).
63. Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genet.* **39**, 906–913 (2007).
64. Servin, B. & Stephens, M. Imputation-based analysis of association studies: candidate regions and quantitative traits. *PLoS Genet.* **3**, e114 (2007).
65. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature Genet.* **39**, S37–S42 (2007).
This paper gives an excellent summary of the challenges to be addressed if large-scale genetic association studies are to be extended to CNVs.
66. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature Genet.* **39**, S7–S15 (2007).
67. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
68. Sham, P., Bader, J. S., Craig, I., O'Donovan, M. & Owen, M. DNA pooling: a tool for large-scale association studies. *Nature Rev. Genet.* **3**, 862–871 (2002).
69. Cargill, M. *et al.* A large-scale genetic association study confirms *IL12B* and leads to the identification of *IL23R* as psoriasis-risk genes. *Am. J. Hum. Genet.* **80**, 273–290 (2007).
70. Wang, W. Y., Barratt, B. J., Clayton, D. G. & Todd, J. A. Genome-wide association studies: theoretical and practical concerns. *Nature Rev. Genet.* **6**, 109–118 (2005).
71. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
72. Nicolae, D.L,. Wu, X., Miyake, K. & Cox, N. J. GEL: a novel genotype calling algorithm using empirical likelihood. *Bioinformatics* **22**, 1942–1947 (2006).
73. Rabbee, N. & Speed, T. P. A genotype calling algorithm for affymetrix SNP arrays. *Bioinformatics* **22**, 7–12 (2006).
74. Xiao, Y., Segal, M. R., Yang, Y. H. & Yeh, R. F. A multi-array multi-SNP genotyping algorithm for Affymetrix SNP microarrays. *Bioinformatics* **23**, 1459–1467 (2007).

© 2008 **Nature Publishing Group**

75. Wittke-Thompson, J. K., Pluzhnikov, A. & Cox, N. J. Rational inferences about departures from Hardy–Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 967–986 (2005).
76. Cox, D. G. & Kraft, P. Quantification of the power of Hardy–Weinberg equilibrium testing to detect genotyping error. *Hum. Hered.* **61**, 10–14 (2006).
77. Smyth, D. J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (*IFIH1*) region. *Nature Genet.* **38**, 617–619 (2006).
78. Lettre, G., Lange, C. & Hirschhorn, J. N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31**, 358–362 (2007).
79. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
80. Hoggart, C. J. *et al.* Genome-wide significance for dense SNP and resequencing data. *Genet. Epidemiol.* **32**, 179–185 (2008).
81. 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 Cancer Inst.* **96**, 434–442 (2004).
This is an influential paper setting out the rationale for a Bayesian interpretation of genetic association findings, focusing on methods for establishing the confidence with which any given positive association can be regarded.
82. Wakefield, J. A Bayesian measure of the probability of false discovery in genetic epidemiology studies. *Am. J. Hum. Genet.* **81**, 208–227 (2007).
83. De Bakker, P. I. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* **37**, 1217–1223 (2005).
84. Morris, A. P. A flexible Bayesian framework for modeling haplotype association with disease, allowing for dominance effects of the underlying causative variants. *Am. J. Hum. Genet.* **79**, 679–694 (2006).
85. De Bakker, P. I. *et al.* Transferability of tag SNPs in genetic association studies in multiple populations. *Nature Genet.* **38**, 1298–1303 (2006).
86. Service, S. *et al.* Magnitude and distribution of linkage disequilibrium in population isolates and implications for genome-wide association studies. *Nature Genet.* **38**, 556–560 (2006).
87. Zeggini, E. *et al.* An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nature Genet.* **37**, 1320–1322 (2005).
88. Easton, D. F. *et al.* A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the *BRCA1* and *BRCA2* breast cancer-predisposition genes. *Am. J. Hum. Genet.* **81**, 873–883 (2007).
89. Marchini, J., Donnelly, P. & Cardon, L. R. Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nature Genet.* **37**, 413–417 (2005).
90. Hirschhorn, J.N., Lohmueller, K., Byrne, E. & Hirschhorn, K. A comprehensive review of genetic association studies. *Genet. Med.* **4**, 45–61 (2002).
91. NCI-NHGRI Working Group on Replication in Association Studies. Replicating genotype–phenotype associations: what constitutes replication of a genotype–phenotype association, and how best can it be achieved? *Nature* **447**, 655–660 (2007).
This feature article is a thoughtful summary of the main issues relating to replication of genetic association studies.
92. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
93. Clarke, G. M., Carter, K. W., Palmer, L. J., Morris, A. P. & Cardon, L. R. Fine mapping versus replication in whole-genome association studies. *Am. J. Hum. Genet.* **81**, 995–1007 (2007).
94. Skol, A. D., Scott, L. J., Abecasis, G. R. & Boehnke M. Optimal designs for two-stage genome-wide association studies. *Genet. Epidemiol.* **31**, 766–788 (2007).
95. Wang, H., Thomas, D. C., Pe'er, I. & Stram, D. O. Optimal two-stage genotyping designs for genome-wide association scans. *Genet. Epidemiol.* **30**, 356–368 (2006).
96. Müller, H. H., Pahl, R. & Schäfer, H. Including sampling and phenotyping costs into the optimization of two stage designs for genome wide association studies. *Genet. Epidemiol.* **31**, 844–852 (2007).
97. Zollner, S. & Pritchard, J. K. Overcoming the winner's curse: estimating penetrance parameters from case–control data. *Am. J. Hum. Genet.* **80**, 605–615 (2007).
98. Yu, K *et al.* Flexible design for following up positive findings. *Am. J. Hum. Genet.* **81**, 540–551 (2007).
99. Gorrochurn, P., Hodge, S. E., Heiman, G. A., Durner, M. & Greenberg, D. A. Non-replication of association studies: 'pseudo-failures' to replicate? *Genet. Med.* **9**, 325–331 (2007).
100. Ioannidis J. P., Patsopoulos, N. A. & Evangelou, E. Heterogeneity in meta-analyses of genome-wide association investigations. *PLoS ONE* **2**, e841 (2007).
101. Ioannidis J. P. Non-replication and inconsistency in the genome-wide association setting. *Hum. Hered.* **64**, 203–213 (2007).
102. Moonesinghe, R., Khoury, M. J., Liu, T. & Ioannidis, J. P. Required sample size and nonreplicability thresholds for heterogeneous genetic associations. *Proc. Natl Acad. Sci. USA* **105**, 617–622 (2008).
103. The GAIN Collaborative Research Group. New models of collaboration in genome-wide association studies: the Genetic Association Information Network. *Nature Genet.* **39**, 1045–1051 (2007).
104. Egger, M., Schneider, M. & Davey Smith, G. Spurious precision? Meta-analysis of observational studies. *BMJ* **316**, 140–144 (1998).
105. Helgason, A. *et al.* Refining the impact of *TCF7L2* gene variants on type 2 diabetes and adaptive evolution. *Nature Genet.* **39**, 218–225 (2007).
106. Locke, D. P., *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
107. ENCODE Project Consortium. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
This is a detailed examination of the functional annotation of a subset of the human genome, which reveals the complexity of genomic organization.
108. Stranger, B. *et al.* Population genomics of human gene expression. *Nature Genet.* **39**, 1217–1224 (2007).
109. Dixon, A. L. *et al.* A genome-wide association study of global gene expression. *Nature Genet.* **39**, 1202–1207 (2007).
110. Goring, H. H. *et al.* Discovery of expression QTLs using large-scale transcriptional profiling in human lymphocytes. *Nature Genet.* **39**, 1208–1216 (2007).
111. Ioannidis, J. P. & Kavvoura, F. K. Concordance of functional *in vitro* data and epidemiological associations in complex disease genetics. *Genet. Med.* **8**, 583–593 (2006).
112. Lowe, C. E. *et al.* Large-scale genetic fine mapping and genotype–phenotype associations implicate polymorphism in the *IL2RA* region in type 1 diabetes. *Nature Genet.* **39**, 1074–1082 (2007).
113. Ioannidis, J. P. *et al.* Assessment of cumulative evidence on genetic associations: interim guidelines. *Int. J. Epidemiol.* **37**, 120–132 (2008).
114. Davey Smith, G. & Ebrahim, S. 'Mendelian randomization': can genetic epidemiology contribute to understanding environmental determinants of disease? *Int. J. Epidemiol.* **32**, 1–22 (2003).
115. Zheng, S. L. *et al.* Cumulative association of five genetic variants with prostate cancer. *N. Engl. J. Med.* **358**, 910–919 (2008).
116. Stratton, M. R. & Rahman, N. The emerging landscape of breast cancer susceptibility. *Nature Genet.* **40**, 17–22 (2008).
117. Mailman, M. D. *et al.* The NCBI dbGaP database of genotypes and phenotypes. *Nature Genet.* **39**, 1181–1186 (2007).
118. Zheng, S. L. *et al.* Association between two unlinked loci at 8q24 and prostate cancer risk among European Americans. *J. Natl Cancer Inst.* **99**, 1499–1501 (2007).
119. Von Elm, E. & Egger, M. The scandal of poor epidemiological research. *BMJ* **329**, 868–869 (2004).
120. Brazma, A. *et al.* Minimum information about a microarray experiment (MIAME) — toward standards for microarray data. *Nature Genet.* **29**, 356–371 (2001).
121. Altman, D. & Moher, D. Developing guidelines for reporting healthcare research: scientific rationale and procedures. *Med. Clin. (Barc).* **125**, 8–13 (2005).
122. Gludd, L. L. Bias in clinical intervention research. *Am. J. Epidemiol.* **163**, 493–501 (2006).
123. Altman, D. G. *et al.* The revised CONSORT statement for reporting randomized trials: explanation and elaboration. *Ann. Intern. Med.* **134**, 663–694 (2001).
124. Von Elm, E. *et al.* The strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. *Lancet* **370**, 1453–1457 (2007).
125. Seminara, D. *et al.* The emergence of networks in human genome epidemiology: challenges and opportunities. *Epidemiology* **18**, 1–8 (2007).
126. Ge, D. *et al.* WGAViewer: a software for genomic annotation of whole genome association studies. *Genome Res.* 3 Mar 2008 (doi:10.1101/gr.071571.107).
127. Janssens, A. C. J.W, Gwinn, M., Subramonia-Iyer, S. & Khoury, M. J. Does genetic testing really improve the prediction of future type 2 diabetes? *PLOS Med.* **3**, e114 (2006).

Acknowledgements
Preparation of this article was supported by funding from the European Commission to the MolPAGE Consortium (LSHG-CT-2004-512066: MMcC) and by research grants from the National Institutes for Health (NHGRI and NHLBI; GRA). We thank our colleagues — particularly P. Donnelly, J. Marchini, J. Barrett, E. Zeggini, C. Lindgren, M. Boehnke, F. Collins, C. Spencer and D. Altshuler for discussions and the reviewers for their comments.

Competing interests statement
The authors declare competing financial interests: see web version for details.

DATABASES
Entrez Gene: http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=gene
CDKAL1 | *FTO* | *KCNJ11* | *PPARG* | *TCF7L2*

FURTHER INFORMATION
The McCarthy Group homepage: http://www.well.ox.ac.uk/mccarthy
Catalog of published genome-wide association studies: http://www.genome.gov/26525384
Consolidated standards of reporting trials (CONSORT): http://www.consort-statement.org
dbGAP: http://www.ncbi.nlm.nih.gov/dbgap
ENCODE project: www.genome.gov/10005107
European Genotyping Archive (EGA): http://www.ebi.ac.uk/ega
Genetic Association Information Network (GAIN): http://www.genome.gov/19518664
Human Genome Epidemiology Network (HuGeNet): http://www.hugenet.org.uk
International HapMap Consortium: www.hapmap.org
National Cancer Institute's cancer genetic markers of susceptibility (CGEMS) study: http://cgems.cancer.gov
Policy for sharing of data obtained in NIH supported or conducted GWA studies: http://grants.nih.gov/grants/guide/notice-files/NOT-OD-07-088.html
Strengthening the reporting of observational studies in epidemiology (STROBE): http://www.strobe-statement.org
Wellcome Trust Case Control Consortium: www.wtccc.org.uk
WGAViewer: http://www.genome.duke.edu/centers/pg2/downloads/wgaviewer.php

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2008 Nature Publishing Group