Guilt beyond a reasonable doubt

David Altshuler & Mark Daly

Genome-wide association studies, exemplified by the Wellcome Trust Case Control Consortium and follow-up studies, have identified dozens of common variants robustly associated with common diseases, providing new clues about genetic architecture in humans. Finding all such loci, and fully defining genotype-phenotype correlation, will be a key to translating initial clues into pathophysiological understanding and clinical prediction.

Genetic screens are used to explore biological mechanisms *in vivo*, unbiased by prior assumptions about the DNA alterations responsible for phenotypic variation. In model systems, genome-wide, phenotype-driven screens typically identify many genes of unknown function, ultimately leading to a broad and deep understanding of mechanism.

In humans, success with phenotype-driven, genome-wide screening for inherited disease mutations has been limited to mendelian traits. Human phenotypic variation is largely polygenic rather than monogenic, however, and thus the vast majority of heritable factors for common human diseases remain unknown. Genome-wide association studies (GWASs) have been proposed as a new approach to 'forward genetics' in humans, but until recently they were untested for gene discovery.

The Wellcome Trust Case Control Consortium (WTCCC) now reports in *Nature* the largest GWAS thus far[1], scanning 17,000 individuals for seven diseases, with two follow-up studies reported in this issue, Todd *et al.* on type 1 diabetes (page 857) and Parkes *et al.* on Crohn's disease (page 830), and another on type 2 diabetes published elsewhere[2–4]. Together with other publications, statistically compelling associations have been identified this year by GWASs across a variety of diseases, including Crohn's disease, obesity, type 1 and type 2 diabetes, coronary heart disease and prostate and breast cancer (see **Supplementary Note** for additional references). In multiple diseases, five to ten independent genomic regions have been identified and confirmed. After years as 'Keystone Cops', complex trait geneticists can now find culprits not previously suspected and establish guilt beyond a reasonable doubt.

The current crop of successful studies shares five key features. First, they all use high-density SNP genotyping arrays (based on the Human Genome Project, the SNP Consortium and the HapMap Project) and analytical methods built on the synthesis of population genetics, statistical genetics and epidemiology. Second, the clinical investigators had the foresight to collect large patient samples that included detailed phenotype information, DNA samples and informed consent for genetic research. Third, they have paid careful attention in their design and analysis to minimizing bias (coming from, for example, population substructure, genotyping errors or variability in DNA quality and laboratory processing). Fourth, they have applied statistical thresholds appropriate to genome-wide searches. With ~10 million common SNPs to be tested genome-wide, and few true associations for which power is adequate, the prior probability of a true association is low—and the *P* value required to declare significance is correspondingly stringent (for further discussion of power in the WTCCC, see pages 815–816 in this issue). Finally, they have validated putative 'positives' in independent samples (preferably using independent genotyping technologies). Here, 'replication' refers to association of the same allele to the same trait under the same genetic model[5].

What has been learned?

The most important outcome of these studies is the discovery of new biological associations in genes or regions previously unrecognized to have a role in each disease. In some cases, links have been newly established between diseases and well-studied pathways (such as age-related macular degeneration and the complement pathway, Crohn's disease and autophagy). In many cases, however, associated regions contain genes of unknown function or do not contain annotated genes. Typical of genetic screens in model systems and mendelian genetics, an unbiased genetic approach highlights genes not previously identified.

Second, new mechanistic connections have been uncovered between diseases. Examples include SNPs in *IL23R* with Crohn's disease[6] and psoriasis[7], *PTPN2* with Crohn's disease and type 1 diabetes[1], *PTPN22* and *IL2RA* with type 1 diabetes and rheumatoid arthritis[1], 8q24 with prostate cancer and breast cancer[8] (see also Stacey *et al.* (page 865) and Hunter *et al.* (page 870), in this issue) and nearby SNPs in a noncoding region of 9p near *CDKN2B* and *CDKN2A* with type 2 diabetes[4,9,10] and coronary heart disease[1,11,12].

Third, the studies have found a substantial fraction of associations outside of transcription units. This is unsurprising, as coding sequences make up less than half of the evolutionarily conserved DNA in the human genome. Investigation of functional noncoding associations will be critical to unraveling molecular and cellular roles of noncoding functional DNA in humans.

David Altshuler and Mark Daly are at Massachusetts General Hospital, Harvard Medical School and the Broad Institute of Harvard and MIT, Boston, Massachusetts, USA. e-mail: altshuler@molbio.mgh.harvard.edu and mjdaly@chgr.mgh.harvard.edu

Table 1 Power of GWASs to discover several recently defined associations

		Power in a 'typical' GWAS (1,000 cases/1,000 controls)			Power in WTCCC (2,000 cases/3,000 controls)			Sample size required for 90% power, $P < 10^{-8}$		
Gene	Disease	1.0×10^{-2}	1.0×10^{-4}	1.0×10^{-8}	1.0×10^{-2}	1.0×10^{-4}	1.0×10^{-8}		RAF	RR
ATG16L1	CD	>0.99	>0.99	0.74	>0.99	>0.99	>0.99	2,430	0.5	1.5
IRGM	CD	0.67	0.19	<0.01	0.98	0.8	0.16	10,902	0.075	1.4
PTPN2	T1D, CD	0.37	0.05	<0.01	0.82	0.34	<0.01	19,754	0.17	1.2
IL2	T1D	0.11	<0.01	<0.01	0.31	0.04	<0.01	54,600	0.26	1.1
9p21	MI	0.97	0.87	0.09	>0.99	>0.99	0.86	5,066	0.47	1.25
9p21	T2D	0.36	0.05	<0.01	0.79	0.31	<0.01	20,220	0.83	1.2
CDKAL1	T2D	0.35	0.04	<0.01	0.79	0.31	<0.01	20,700	0.31	1.15

Approximate risk models estimated from published replication studies and power computed using the Genetic Power Calculator[15] (http://pngu.mgh.harvard.edu/~purcell/gpc/). Sample size calculation assumes equal numbers of cases and controls. RAF, risk allele frequency; RR, relative risk; CD, Crohn's disease; T1D, type 1 diabetes; MI, myocardial infarction; T2D, type 2 diabetes; WTCCC, Wellcome Trust Case Control Consortium.

Fourth, the results indicate that individual SNPs have very modest effects in the population: associated SNPs rarely show odds ratios of >2.0 (CFH in age-related macular degeneration), and more typically, odds ratios are <1.5. Undiscovered common variants are likely to have similar or smaller effects (or are in low linkage disequilibrium with SNPs on arrays).

Fifth, strong evidence is lacking for epistasis among associated SNPs, despite joint analysis in large cohorts. Similarly, little evidence has been obtained for strong association of disease-associated SNPs to homogeneous disease subtypes, or quantitative 'endo-phenotypes' (such as glycemic and obesity traits in type 2 diabetes). Sixth, despite substantial progress, the vast majority of heritability remains unexplained. To some extent, the magnitude of the associations discovered is currently underestimated, because the full spectrum of causal variation at each locus has yet to be defined by deep sequencing.

A less obvious but still important implication is that many more such loci must remain to be found. Even for the confirmed associations identified, statistical power was limited in the genome-wide scans that found them (**Table 1**). Even in the large WTCCC study (which included 2,000 cases and 3,000 controls)[1], the power to obtain a genome-wide $P < 10^{-8}$ was <1% for many of the confirmed associations discovered by comparison across studies and by replication studies. This explains the tendency of different GWASs to find partially overlapping sets of associations and makes it implausible that most regions harboring relevant associations have been identified.

Where to from here?

These papers provide proof-of-concept that GWASs can identify previously unknown causal loci. The next steps are to obtain a full picture of genotype-phenotype correlation at these loci and to find remaining loci. A more complete picture will be critical to understanding the disease mechanisms underlying the associations and to assess SNPs for clinical management.

Rarely will the SNPs used to discover each locus prove causal; exhaustive sequencing of each region will be needed to discover all causal mutations and fully define genotype-phenotype correlation. In many cases, multiple independent common variants[13] and rare variants[14] will be found at the same locus. Sequencing of exons in each associated region may identify coding mutations of stronger effect, which may be easier to study *in vitro* and in individual subjects. In addition, identification of 'smoking gun' causal coding mutations may help prove which gene at each locus is responsible for the association and may, in aggregate, increase the overall predictive value of genotype.

A testable hypothesis suggested by the power calculations in **Table 1** is that a more extensive set of loci that influence each disease may be found by GWASs of greater power (or by combining existing GWASs). Common sense dictates that a complete set of susceptibility loci will provide greater biological insight than an incomplete set. Moreover, the biological insight provided by any locus is not necessarily related to the size of the effect of common variants used to discover it, nor is it predictive of the combined effect of all rare and common variants at that locus. Thus, the discovery of additional causal loci should be pursued, followed by exhaustive sequencing to fully define genotype-phenotype correlation.

Some loci may be missed by well-powered GWASs because none of the causal variants are in linkage disequilibrium with SNPs on the genotyping arrays. Some of these may be found by genome-wide measurement of copy number variation. Thus, these GWASs are the first in a series of genome-wide, phenotype-driven approaches in humans, which, when integrated, will provide a more complete picture of human phenotype variation and inborn susceptibility to disease.

Ultimately, the value of this endeavor must be measured in the resulting clinical and biological advances. Predictive testing will have value in cases in which effective preventative interventions exist, and when modest changes in risk improve clinical decision-making. Achieving a clinical benefit will be challenged by the modest magnitude of SNP effects and by the likelihood that genetic tests will be made available (and aggressively promoted) before or instead of mounting clinical trials to evaluate the value of genetically enabled decision-making.

New tools and frameworks will be required to translate genetic insights into knowledge of disease pathogenesis and new therapeutics: there is little precedent for functional analysis based on genes discovered by polygenic inheritance, noncoding DNA changes and quantitative alteration of gene function. This quest is worth mounting, however, as it is in pursuit of culprits whose guilt in human disease has been established beyond a reasonable doubt.

Note: Supplementary information is available on the Nature Genetics website.

COMPETING INTERESTS STATEMENT
The authors declare no competing financial interests.

1. Wellcome Trust Case Control Consortium. *Nature* **447**, 661–678 (2007).
2. Parkes, M. *et al. Nat. Genet.* **39**, 830–832 (2007).
3. Todd, J. *et al. Nat. Genet.* **39**, 857–864 (2007)
4. Zeggini, E. *et al. Science* **316**, 1336–1341 (2007).
5. NCI-NHGRI Working Group on Replication in Association Studies *et al. Nature* **447**, 655–660 (2007).
6. Duerr, R.H. *et al. Science* **314**, 1461–1463 (2006).
7. Cargill, M. *et al. Am. J. Hum. Genet.* **80**, 273–290 (2007).
8. Easton, D.F. *et al. Nature*, advance online publication 27 May 2007 (doi:10.1038/nature05887).
9. Diabetes Genetics Initiative of Broad Institute of Harvard and MIT, Lund University, and Novartis Institutes of BioMedical Research *et al. Science* **316**, 1331–1336 (2007).
10. Scott, L.J. *et al. Science* **316**, 1341–1345 (2007).

11. McPherson, R. *et al. Science*, **316**, 1488–1491 (2007).
12. Helgadottir, A. *et al. Science*, **316**, 1491–1493 (2007).
13. Haiman, C. *et al. Nat. Genet.* **39**, 638–644 (2007).
14. Kotowski, I.K. *et al. Am. J. Hum. Genet.* **78**, 410–422 (2006).
15. Purcell, S. *et al. Bioinformatics* **19**, 149–150 (2003).

Conjuring SNPs to detect associations

Andrew G Clark & Jian Li

Human genome-wide association studies pose a challenge in identifying significant disease associations from nearly half a million statistical tests. A new report describes an especially promising approach, recently applied to the Wellcome Trust Case Control Consortium data sets, that uses the correlated structure of genomic variation to impute genotypes at missing sites and to test association with both observed and imputed SNPs.

Genetic mapping has always relied on statistical inference, but this enterprise has never been so utterly dependent on rigorous analytical methods as it is with genome-wide association studies (GWASs). For each of the nearly 500,000 SNPs in the human genome scored by widely used genotyping platforms for GWASs, it is possible to perform a simple statistical test of association with disease state. Even if the null hypothesis of no association were true for all SNPs, we would expect some of these tests to provide nominal *P* values on the order of 10^{-6}. In order to avoid false-positive calls, we need to identify SNPs for which the *P* values are even lower. We could increase the power to appropriately reject the null hypothesis (that is, to correctly infer that a SNP is truly associated with disease) by elevating the sample size or restricting attention to intermediate-frequency SNPs and by being judicious in our choice of test. In this issue, Marchini *et al.*[1] (page 906) show that thoughtful application of population genetic principles and use of HapMap data can provide an additional source of power for association tests. They have successfully applied these methods to the Wellcome Trust Case Control Consortium (WTCCC) data[2] and have identified a collection of new genes associated with seven complex medical disorders (see pages 813–815 of this issue for discussion of the WTCCC studies).

Imputation to boost power

The more genetic data that we have for each individual, the greater the chance of finding variants that influence disease risk directly. This is true even if some of those variants are statistically inferred or 'imputed' from the observed genetic data. To see how imputation can give a boost in the power of tests of association, consider the situation where a SNP that has a direct effect on disease risk is in the HapMap set of SNPs but is not on the 500K genotyping platform used in a given GWAS (**Fig. 1a**). In this case, if only the observed marker SNP were used, the association test would be weakened by any observed departure from perfect linkage disequilibrium between the observed SNP and the unobserved risk-enhancing SNP. This contrasts with the hypothetical case (**Fig. 1b**) in which the risk-enhancing SNP is observed directly. If no other genetic variation in this genomic region influences risk, then the test based on this SNP alone will be the most powerful. One can see that such a direct test provides a greater chance to detect a significant association. Because we often do not observe the risk-enhancing SNP directly, imputation can be used to close some of the gap between these two extremes. High linkage disequilibrium in the human genome means that we can impute the unobserved genotype of many of the missing SNPs with surprisingly high accuracy (>98% in many cases). This accuracy will be reduced in regions of the genome with unusually high recombination rates (for example, SNPs within hotspots). The example in **Figure 1c** is for an imputation accuracy of 99%, and it is clear that the probability of detecting the association is much greater than in **Figure 1a**, where we did not apply imputation. Marchini *et al.*[1] and Scott *et al.*[3] use multiple flanking SNPs to impute missing SNP genotypes, and they find that the *P* values for tests of association are often an order of magnitude lower with the imputed SNPs than with the observed SNP data only.

This may seem like sleight of hand, because there seems to be a gain in power without any additional information, as the missing SNPs are imputed from the observed marker SNPs. One might think that tests based only on haplotypes of the observed SNPs[4–6] would do just as well, because they, after all, are what allows prediction of the missing SNPs. But the method does incorporate haplotype information of observed SNPs along with the linkage disequilibrium structure of the full HapMap sample to perform the imputation. By leveraging the observed marker SNPs and by predicting missing data from the pattern of linkage disequilibrium in the HapMap data, we get the best of both worlds.

Testing association

In a GWAS, the meaning of a *P* value becomes challenged in the context of so many simultaneous tests. One solution to this problem is to calculate the false discovery rate[7,8]; however, this approach was developed for testing a single hypothesis, as opposed to simultaneously testing a battery of SNPs associated with a disease. Association testing can be done with standard frequentist methods like logistic regression, where the model may specify either allelic or genotypic effects. Likelihood methods can be used to deal with the uncertainty in the imputations of missing genotype data. Bayesian methods also allow inference of probability of association conditional on observed genotype data and can accommodate imputed genotypes easily. Marchini *et al.*[1] make use of one useful measure of the relative likelihood of association, the Bayes factor, a term closely related to likelihood ratio and defined in this case as the probability of the observed data, given that the association is real, divided by the probability of the observed data under

Andrew G. Clark and Jian Li are in the Department of Molecular Biology and Genetics, Cornell University, Ithaca, New York 14853, USA.
e-mail: ac347@cornell.edu