Human Molecular Genetics, 2008, Vol. 17, Review Issue 2 ***R156–R165***
doi:10.1093/hmg/ddn289

Genome-wide association studies: potential next steps on a genetic journey

Mark I. McCarthy[1,2,3,*] and Joel N. Hirschhorn[4,5,6,*]

[1]Oxford Centre for Diabetes, Endocrinology and Metabolism, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK, [2]Wellcome Trust Centre for Human Genetics, University of Oxford, Old Road, Headington, Oxford OX3 7BN, UK, [3]Oxford NIHR Biomedical Research Centre, Churchill Hospital, Oxford OX3 7LJ, UK, [4]Program in Genomics and Divisions of Genetics and Endocrinology, Children's Hospital, Boston, MA 02115, USA, [5]Broad Institute of Harvard and MIT, Cambridge, MA 02142, USA and [6]Department of Genetics, Harvard Medical School, Boston, MA 02115, USA

Received September 2, 2008; Revised and Accepted September 5, 2008

Genome-wide association studies have successfully identified numerous loci at which common variants influence disease risk or quantitative traits. Despite these successes, the variants identified by these studies have generally explained only a small fraction of the heritable component of disease risk, and have not pinpointed with certainty the causal variant(s) at the associated loci. Furthermore, the mechanisms of action by which associated loci influence disease or quantitative phenotypes are often unclear, because we do not know through which gene(s) the associated variants exert their effects or because these gene(s) are of unknown function or have no clear connection to known disease biology. Thus, the initial set of genome-wide association studies serve as a starting point for future genetic and functional studies. We outline possible next steps that may help accelerate progress from genetic studies to the biological knowledge that can guide the development of predictive, preventive, or therapeutic measures.

INTRODUCTION

The first successful wave of genome-wide association (GWA) studies has, over the last year, identified common variants associated with numerous common polygenic diseases and quantitative traits. As illustrated in several reviews in this special issue, these GWA studies have mapped many novel, convincingly associated loci [for example, at least 32 for Crohn's disease (1), 14 for prostate cancer (2), 15 for type 2 diabetes (3) and 40 for height (4–6)]. These discoveries have more often than not implicated previously unsuspected genes, highlighting the power of unbiased genetic screens to uncover novel biology. However, these loci in combination typically explain only a fraction of the inherited contribution to risk, raising the question of how best to find the variation responsible for the remainder. As discussed in several of the articles in this issue, the successful GWA studies were often poorly powered to discover most of the loci that they identified, indicating that more loci with equivalent effect sizes can be discovered simply by increasing sample size. In addition, most GWA studies have been completed in European-derived populations, suggesting that performing GWA studies in non-European samples will also be important (7). However, it is not clear how much of the remaining relevant variation will be uncovered by these steps.

In most cases, the associated loci themselves demand additional exploration. The causal variant is usually not identified by GWA studies (see Finding Additional Associated Loci and Variants section) and in some cases the causal variant may be more strongly associated (and explain more of the risk) than the marker detected in the initial GWA. Furthermore, the associated loci remain essentially unexplored for independent causal alleles, which may account for additional genetic risk. Finally, it is possible that new approaches (such as genome-wide resequencing) designed at detecting variants not well assayed in current GWA studies will be required to define more fully the inherited basis of common disease.

*To whom correspondence should be addressed. Tel: +44 1865 857298; Fax: +44 1865 857299; Email: mark.mccarthy@drl.ox.ac.uk (M.I.M.); Children's Hospital, Enders 561, 300 Longwood Avenue, Boston, MA 02115, USA. Tel: +1 617 9192129; Fax: +1 617 7300253; Email: joelh@broad.mit.edu (J.N.H.)

One of the main goals of genetic studies of complex traits is to flag pathways relevant to disease that could reveal novel therapeutic targets. However, for most associated loci, there is substantial ignorance regarding the mechanisms by which genetic variation could influence phenotype: the identity of the gene(s) affected by the susceptibility variant(s) at each locus is often uncertain, and the mechanisms by which the causal variants (also often unknown) influence phenotype is usually unclear. This lack of knowledge is a substantial impediment to the understanding needed to make progress towards new therapies or preventive measures. This obstacle highlights the need to pinpoint the causal variants and the genes affected by those variants, as well as for informative functional and computational studies to move from gene identification to possible mechanisms that could guide translational progress.

Clearly, this first wave of GWA studies represents a starting point on the journey to elucidating and understanding the genetic basis of complex traits and common disease and translating this knowledge into clinically useful insights. In this review, we describe some possible next steps. Specifically, we outline possible approaches in three areas: finding additional loci that contain causal variants, refining the location and phenotypic consequences of causal variants and progressing from known loci and variants to functional mechanisms. We do not address the ability to translate genetic discoveries into predictive tests, as finding additional causal variation is a necessary precursor to making genotype-based predictors more accurate and useful.

FINDING ADDITIONAL ASSOCIATED LOCI AND VARIANTS

Keeping with what's working: the prospects for more GWA studies

Most common diseases and quantitative traits have heritabilities between 30 and 90% (8). With a few exceptions (1,9–12), the loci discovered by association studies individually account for a small fraction ($<1\%$) of population variation, and, even when considered in combination, most of the inherited component of disease predisposition remains unexplained. It is important to remember that the loci that have been identified have not yet been explored for additional variation (common or rare) that could contribute to inherited variation in risk. However, given the extent of the 'missing heritability' for most diseases and quantitative traits, as yet undiscovered loci are also likely to be important. One immediate question that arises is whether, given the small amount of variation explained by the loci found in the first wave of GWA studies, additional similar studies may be productive.

The total amount of inherited variation that will be discovered with any particular sample size depends on the underlying genetic architecture (the distribution of effect sizes): this is extremely difficult to predict from initial association studies. In general, diseases where individual loci have relatively large effects (such as type 1 diabetes) have yielded additional loci of moderate effect as sample sizes have increased (1), but GWA studies have discovered multiple loci even for diseases without major genes, such as type 2 diabetes (13–18). Despite this unavoidable uncertainty, completed GWA studies can give some guidance as to the likelihood that additional GWA studies will identify novel loci. Most GWA studies have actually had low power to discover the loci that have emerged: for many of the loci identified, chance (in the form of sampling error) played a useful role in boosting their detectability (the so-called 'winner's curse') (19,20). Therefore, many loci of equivalent effect size are likely to be present, and can be unearthed by additional studies. For example, if a GWA has 5% power to detect each of 40 loci, one would expect any single study to discover two of these on average, leaving the remaining 38 to be identified by future studies.

Populations of different ancestry may also be helpful in discovering new loci. Some genetic variation is private to populations with particular continental ancestry, preventing its discovery in other populations. Effect sizes may be larger in certain populations, thereby increasing power. Even if a causal allele is present in multiple populations and has consistent effect sizes, allele frequencies may vary across populations, leading to different power in different groups (21). There are already several examples of variants that, despite the emphasis on European-derived populations for GWA studies, were discovered first in non-European-derived populations (22–24).

Moving beyond the main effects

Almost all GWA studies to date have concentrated on the detection and characterization of main effects. While opinion remains divided as to the extent to which non-additive effects [often described in terms of gene–gene ($G\times G$) and gene–environment ($G\times E$) interaction] will explain the missing inherited risk not attributable to the variants so-far uncovered, the truth is that there are very few empirical data to guide us.

GWA studies provide a finite data set for exploring the joint effects of genes. Why then, given that epistasis is so prevalent in animal models (25), and with modifier genes implicated in many ostensibly Mendelian diseases (26,27), have $G\times G$ effects been so hard to detect? One obvious reason is that the understandable reliance on main-effect testing and replication in validating association signals has biased discoveries towards signals that are not subject to $G\times G$ contingencies. As far as unbiased genome-scale $G\times G$ discovery is concerned, the computational burden imposed by any comprehensive search for higher-order effects has been an important limitation (28). Another relates to sample size and power: individual studies are only likely to detect interaction effects substantially larger than the main effect-sizes that have so far emerged. This follows, in part, from the very nature of GWA scans: while incomplete linkage disequilibrium (LD) (e.g. an r^2 of 0.8) between the causal variant and a typed proxy may be perfectly adequate for detecting main effects, the power to detect non-additivity is severely dented when the causal variant has not been directly assayed (29,30). Exhaustive studies for $G\times G$ effects are therefore going to be dependent on knowledge of causal variants (or failing that, ever more dense GWA data), larger data-sets and more efficient computational approaches.

The challenges with respect to the detection of $G\times E$ interaction (or, more generally, understanding of the joint effects of

G and E) are, if anything, greater. The overall parameter space is effectively unlimited (given all the possible exposures one could conceive of) and many of the exposures most likely to be relevant to disease predisposition (for example, diet and physical activity in the case of diabetes) are hard to measure in detailed, standardized fashion in large sample collections (29,30). Such issues impinge on the power to detect G×E effects in the first place, but also mean that differences between study samples (with respect to the exposures themselves, and the measurements thereof) may make it extremely difficult to differentiate between failure of replication and genuine heterogeneity. This is particularly worrying since evidence for heterogeneity of genetic effect attributable to variation in a given exposure is precisely what one needs for identifying modifiable risk factors amenable to public health intervention. Progress in this area is going to be dependent on harmonization of biobank measures and outcomes on a scale not yet attempted.

Assessing variation that is not captured in current GWA studies

The commercial arrays used for GWA scans are designed to provide excellent coverage of common SNPs, but have only limited potential to capture rare and low frequency variants (i.e. those with a minor allele frequency below 5%) (31). [The specific case of detection of copy number variants is discussed in (32).] Indeed, the extent to which low-frequency, intermediate penetrance variants contribute to disease predisposition, and explain the large proportion of the inherited risk yet to be localized, represents one of the major unanswered questions in human genetics. Such variants will have 'flown below the radar' of available genome-wide technologies, neither penetrant enough to show Mendelian segregation and to be detected through traditional linkage approaches, nor frequent enough to be captured by GWA approaches. Yet, such variants could, individually and collectively, have greater impact in terms of explaining familial risk, and providing individual prediction of disease risk, than the common variants emerging from GWA studies (33). For instance, the locus-specific sibling relative risk attributable to a variant with control MAF of 1% and a per-allele odds ratio of 3 exceeds that of the strongest common T2D-susceptibility variant currently known (*TCF7L2*) and around 30 such variants distributed across the genome could explain all the residual missing inherited risk for this disease.

Identification of such variants remains a substantial challenge, though advances in high-throughput resequencing technologies and the efforts of the Thousand Genomes Project should enable rapid progress. Initial efforts are likely to be focused around genes already (by virtue of a role in monogenic or multifactorial forms of disease) implicated in disease pathogenesis, since functional variants in such genes represent particularly impressive candidates. Genome-wide surveys for such variants are likely to become practical first in recent isolates and 'self-contained' populations (such as Iceland and Finland) where homozygosity mapping (34) and long-range haplotype phasing (35) will prove valuable tools for the detection of rare disease-associated haplotypes. Because low frequency variants are likely to reflect relatively recent mutational events, evaluation of signals emerging from any given study will likely be complicated by substantial allelic heterogeneity and ethnic differences: evidence that the variants identified are causal may well need to be built up on a 'locus-wide' basis by studies conducted across multiple samples.

More unusual sources of inherited variation

Large samples and currently available technologies, or those that will likely be available in the near future, should enable assessment of the various types of genetic variation described earlier. It is possible that such comprehensive efforts, plus consideration of gene–gene and gene–environment interactions, will explain the bulk of inherited variance. However, epigenetic effects, such as methylation or histone modifications, offer an additional possible contribution to heritability. Note that epigenetic effects that track with underlying DNA sequence variation (36) should be detectable in the usual fashion: in this case, epigenetic effects are part of the explanation of mechanism by which DNA variation affects phenotype. In addition, while non-sequence-dependent epigenetic effects may also have strong influences on gene expression and phenotype, they are usually not transmitted across generations, so they are unlikely to contribute to estimates of heritability (at least in the usual scenario where twins and parent/offspring trios provide reasonably consistent estimates of heritability). However, one potential source of heritable variation not captured by traditional approaches is epigenetic changes that are transmitted across generations but where the inherited epigenetic state does not track with underlying DNA sequence variation. Under certain circumstances, changes in DNA methylation in mice can be induced purely by dietary modification and then transmitted across generations even after the dietary modification has been removed (37). As yet, similar heritable, non-sequence-dependent changes in methylation or other epigenetic modifications of DNA have not been documented in humans but, if these phenomena were prevalent in humans, new approaches would be required to detect these effects.

GENETIC AND PHENOTYPIC CHARACTERIZATION OF ASSOCIATED LOCI AND CAUSAL VARIANTS

Genetic refinement of association signals identified by GWA studies

With current imputation methods (38,39), about 3 million SNPs [those in HapMap (40)] can be tested for association, either because they have been directly genotyped or because their genotypes can be imputed from data at nearby variants. However, because there are many more than 3 million common variants in the genome (41), not all common causal variants will be represented in HapMap and not all will have been tested, either directly or indirectly through imputation. The current generation of successful GWA studies will instead provide a list of common variants that show convincing evidence of association because they are correlated with nearby causal variants. Due to strong LD throughout most of the genome (41), there will often be several variants that show more or less equivalent evidence of association for any given signal of association. There may also

be multiple independent signals of association at a locus, such as at 8q24 for prostate cancer (2), and the chromosome 9 locus for type 2 diabetes (16).

If the goal were to identify markers of disease risk for the purposes of prediction, validating associations without further genetic localization of the source of the signal would be a sufficient endpoint for GWA studies (assuming the associated variant is reasonably tightly correlated to the causal variant and thus accurately reflects the contribution to inherited risk). However, to facilitate follow-up functional studies or generate hypotheses regarding mechanism, it is essential to refine the location of causal variants as sharply as possible. [We note that fine mapping may also increase the strength of association, although the gains in significance may not be dramatic if the GWA study evaluated a reasonably dense set of SNPs (42).] In the next few paragraphs, we discuss possible steps to localize the variant(s) responsible for a signal of association.

To generate a more comprehensive list of potential causal variants that could explain an association signal, resequencing across the entire region of association (at least out to the point at which LD has substantially decayed) and confirmatory genotyping efforts will generally be required. Improved sequencing methodologies (43) currently make this task much more practical than previously. Furthermore, the 1000 genomes project, once completed, should lessen the necessity of the sequencing step. We note that comprehensively searching for additional causal variants that are independent from the association signal and may be of lower frequency will require much more substantial sequencing and genotyping efforts, both in depth and breadth, to fully interrogate nearby genes for possible additional susceptibility alleles.

Once a set of potential causal variants has been assembled (from the variants showing the best evidence for association in the GWA data, and any strongly-correlated newly identified variants), various methods can be in theory be used to test which are most likely to explain the signal of association (44,45). One such method, stepwise logistic regression, can be used to test whether one of a set of variants is necessary and sufficient to explain the association signal—if variant A remains significant with variant B in the regression model, but variant B is not significant with variant A in the model, than variant A is more likely to be causal. In practice, the combination of strong LD (many pairs of variants have correlation coefficients that approach 1) and modest effect sizes means that samples of hundreds of thousands of individuals may be required for this or any other method to distinguish between multiple nearly equivalent variants. Indeed, the level of pairwise correlation between variants may approach the genotyping accuracy rate (often between 0.99 and 1 for directly genotyped SNPs, and lower for imputed SNPs), in which case even minimal genotyping error rates can still confound these analyses. Thus, fine mapping will be challenging even in large, densely genotyped data sets.

One method of lessening the obstacle of strong LD is to perform fine mapping in populations of different ancestries, in the hope that pairwise correlation coefficients will not be equally high in all populations (and/or that additional causal alleles with more attractive fine-mapping potential are revealed). Individuals of recent African ancestry may be particularly helpful because of the lower levels and often distinct patterns of LD (7). By genotyping all of the equivalently associated variants in multiple populations, it is possible that a subset of variants may emerge that show a more consistent pattern of association across populations, making these more likely candidates for being causal.

By taking associations discovered in one population and testing them in populations of different ancestries, a new challenge may arise. It is not yet known how often loci discovered in one population (until now, usually European-derived populations) will also show association in other ethnic groups, and it is possible that none of the associated variants from, say, a European population will show association when genotyped in a population of recent African ancestry, making fine mapping impossible. Of primary concern is that power could be limiting, either due to the unavailability of sufficiently large samples in populations of African ancestry, and/or a smaller effect size in these populations. Thus, the success of this approach will depend on the collection of large samples from multiple ethnic groups.

It is not known how often effect sizes will vary substantially across populations, because, as of yet, few truly validated associations have been exhaustively examined across multiple ethnicities. However, some well-validated associations where the causal variant has almost certainly been identified, such as ApoE4 and Alzheimer's disease, do show smaller effects in African-derived populations than in the European-derived populations in which the association was discovered (46). However, comparisons of effect sizes for type 2 diabetes susceptibility loci between European and Asian populations have shown an encouraging consistency across these groups at least (47–50).

Variable effect sizes across populations of different ancestries could arise for many different reasons. As one possibility, the strengths of joint gene–gene or gene–environment effects may vary across populations and thereby modify the observed main effects of causal variants. For example, if a variant only influences the risk of cancer in individuals exposed to an unmeasured environmental variable, the power to detect the association would track with the prevalence of the environmental exposure. Similarly, gene–gene interactions, if strong, could modify effect sizes through variable allele frequencies across populations. Differences in the phenotype itself could influence power: if a variant only increases risk for a certain clinical subtype [such as has been seen for estrogen receptor positive versus negative breast cancer (2)], power will be greater in those populations where that clinical subtype represents a greater fraction of ascertained cases. Finally, the very differences in LD that inspire fine mapping in multiple ethnicities may also complicate the interpretation of a negative result, as it may be possible that the causal variant is in LD with known variants in some populations but not in others (Fig. 1). This last possibility highlights the potential importance of comprehensive resequencing in regions of association in diverse populations.

Finally, many of the more straightforward methods used for fine mapping assume that a single causal variant is responsible for the association signal. However, given the large number of loci that influence common disease, and the examples of multiple independent causal variants at a single locus, it is possible that some signals could result from two or more

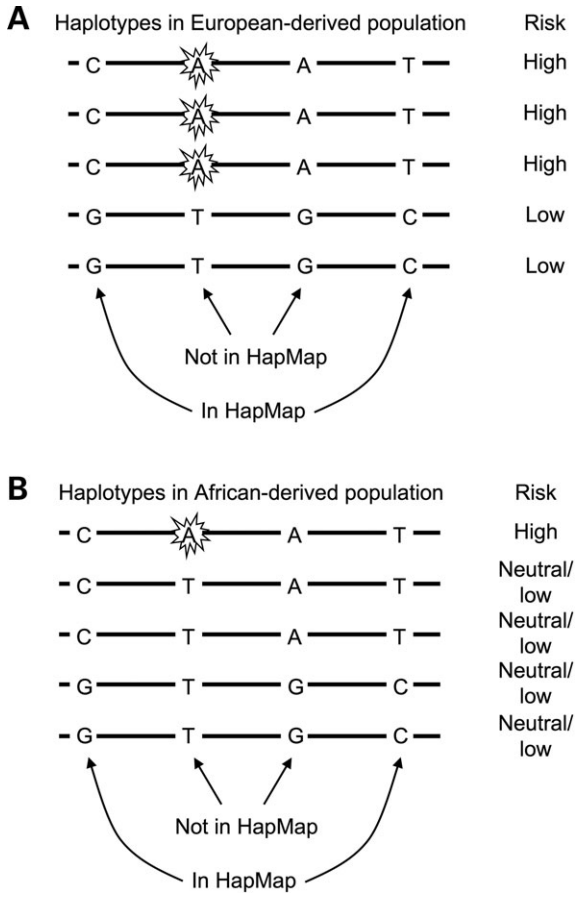

Figure 1. Different LD patterns can yield different patterns of association. Hypothetical haplotypes in an associated region and their effects on disease risk are shown for a European-derived population (**A**) and an African-derived population (**B**). In European-derived populations, several SNPs show equivalent signals of association, including the causal SNP (marked by jagged lines). Two of these are in HapMap, and have been tested via genotyping or imputation, permitting the effect of the causal SNP (which is not in HapMap) to be detected indirectly. In African-derived populations, the causal SNP is rarer and is no longer strongly correlated with the surrounding SNPs in HapMap, so the surrounding SNPs will not show strong association. Thus, a fine-mapping approach based only on HapMap SNPs but without additional resequencing may fail to detect a signal in the African-derived population.

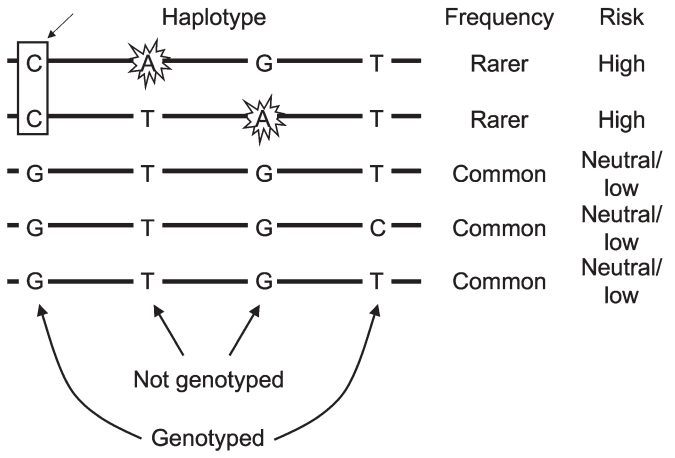

Figure 2. A common SNP may be strongly associated because it tags multiple rarer causal variants. In this hypothetical example, the C allele of the genotyped SNP on the left (indicated by the box) is strongly associated with disease risk because it tags a combination of two rarer causal variants which are themselves only weakly correlated with the associated SNP. Sequencing in affected individuals carrying high-risk haplotypes might be required to uncover the actual causal variants, which in this example have not been genotyped.

causal variants that are in strong LD with each other. In this case, a SNP that happens to tag both high-risk alleles or both low-risk alleles may show the strongest statistical evidence of association but not be a causal variant (or even tightly correlated with either causal variant; Fig. 2).

Phenotypic refinement of association signals identified by GWA studies

GWA association studies generally focus on single phenotypes, but associated variants may actually influence multiple traits. In some cases, the phenotypes are correlated, such as the association of *FTO* variation with diabetes, obesity and other obesity-related phenotypes (3,51). In others, the phenotypes have a clear connection, such as the association of variants with different combinations of autoimmune diseases (1) or cancers (2). But, in yet others, the phenotypes shared have no obvious connection, such as the association of *HNF1B* and *JAZF1* variants with both prostate cancer and type 2 diabetes (18,52,53), and the evidence for pleiotropy therefore suggests previously unsuspected mechanistic connections between two apparently unrelated diseases (54). Thus, it will often be productive to analyze validated associated variants with respect to a wide variety of phenotypes.

In theory, associated variants can also be used to reverse the usual experimental flow of GWA studies. Traditionally, GWA studies begin with a fixed phenotype (such as type 2 diabetes) and then search across a large range of possible genotypes for associations. A reverse experiment might begin with a fixed genotype at a validated variant (such as the diabetes-associated SNP at *TCF7L2*), and search across a range of possible phenotypes for new associations. One might extend this analogy to imputation as well: traditionally, genotypes at untyped markers are inferred using combinations of genotypes from correlated variants. The effects of unmeasured phenotypes might be inferred if combinations of phenotypes show stronger associations than the measured phenotypes themselves; in this case, the combination of phenotypes might represent a new 'unmeasured' phenotype. In this way, associations from GWA studies might lead to the discovery of new phenotypes, and might suggest possible mechanisms of action of the associated variants.

COMBINING FUNCTION AND GENETICS

Although statistical ('reverse genetics') approaches can help move a robust association closer to an improved understanding of disease processes, 'functional' data can certainly support and inform such efforts. Of course, functional inferences are already widely used to identify the leading positional candidates which often (even before definitive evidence of their

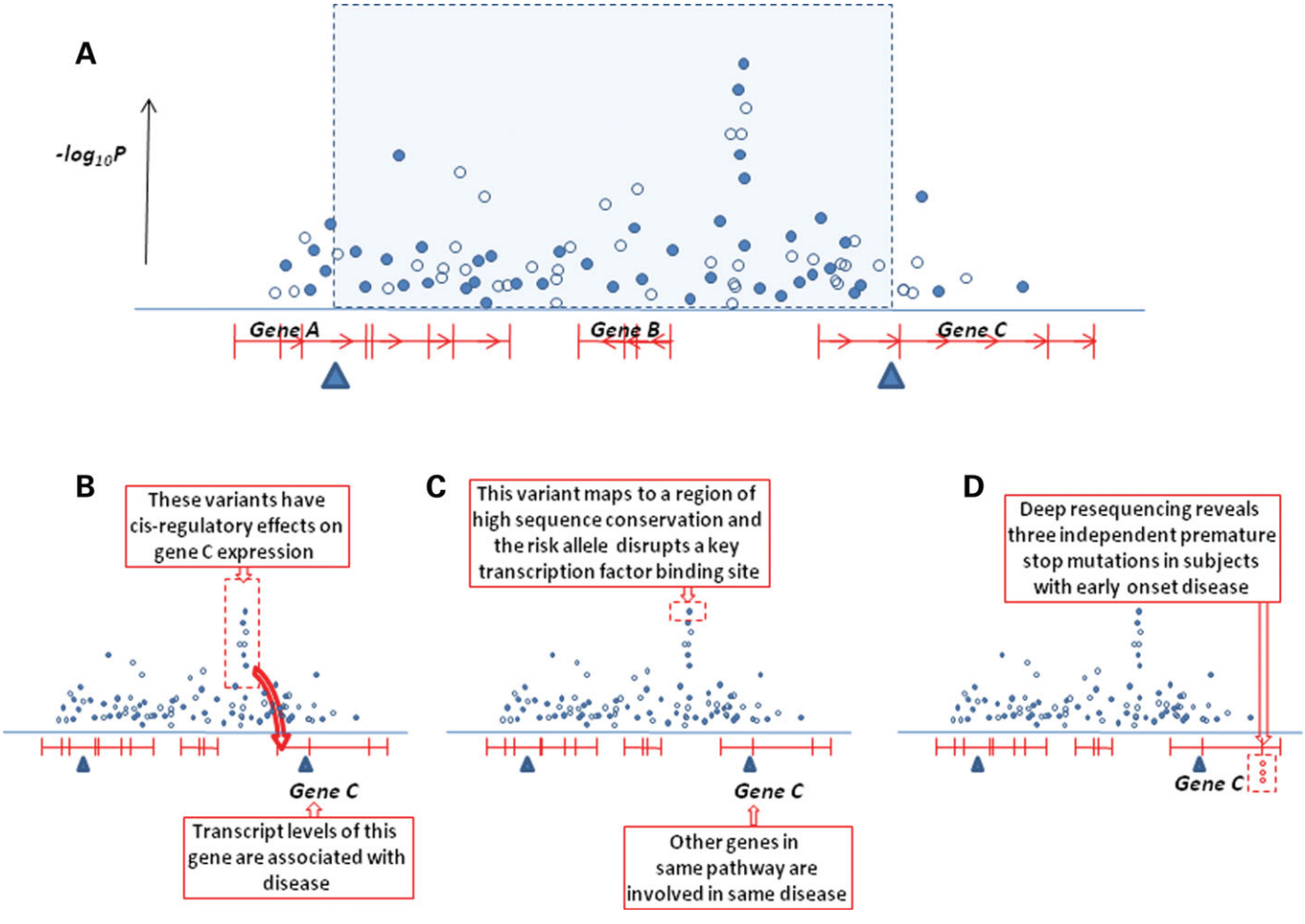

Figure 3. Strategies for using functional data to support causal variant and causal gene identification. The figure illustrates ways in which fine-mapping efforts can be supported by clues from functional data: (**A**) consider a locus at which GWA analysis (complemented by replication data—not shown) has revealed a highly significant association mapping between the coding regions of genes B and C. Directly typed SNPs are shown in the filled symbols, imputed SNPs in open symbols. Flanking recombination hotspots (blue triangles) define an interval within which the variant causal for that signal is most likely to reside. This interval contains the entire coding sequence of gene B, and portions of genes A and C. For the purposes of this cartoon, the causal variant turns out to be the typed SNP with the strongest association, and it exerts its effect on disease through altering expression of gene C; (**B**) clues to the identity of the causal gene are derived by expression QTL studies in a tissue relevant to disease: not only is the expression of gene C associated with the same cluster of variants which shows the disease association; but there are also directionally-consistent associations between gene C transcript levels and disease state; (**C**) clues to the identity of the causal gene are derived from analysis of genome annotations: not only does gene C code for a member of a pathway previously implicated in the disease, but the associated variants are predicted to have strong functional credibility; (**D**) clues to the identity of the causal gene are derived from deep exon resequencing of genes A–C: three independent premature stop-codon mutations in gene C (predicted to lead to generation of a truncated protein product with dominant-negative effects) are found in subjects with severe, early-onset forms of the disease of interest.

direct involvement is available) provide 'shorthand' notation for the association signals with which they co-localize (18). Sometimes, the combination of a credible biological candidate and putatively functional variant [e.g. the R325W variant in *SLC30A8* in type 2 diabetes (13)] will provide a very strong functional hypothesis that can be directly evaluated. However, poor understanding of disease mechanisms (frustrating efforts to define candidacy within associated loci), and the sheer challenge of evaluating the potential impact of many possible candidate variants on the expression and/or function of nearby genes, will often leave researchers dependent on statistical approaches to do the 'heavy-lifting', turning to direct experimental functional evaluation once the list of potential candidate variants has been reduced to manageable proportions. Opportunities for more effective integration of functional and statistical approaches are discussed below (and summarized in Fig. 3).

Using expression information

The most obvious shortcut from association signal to putative mechanism lies in the use of expression quantitative trait locus (eQTL) data (55,56). Publicly available eQTL data exist for a growing number of tissues (57–62), and there is growing interest in expanding the range of tissues and cell-lines for which equivalent information is available. With such data, it becomes possible to establish if any of the variants within the association signal have transcriptional effects (most particularly, on those genes that lie nearby). Overlap between the association patterns with respect to disease and gene

expression, particularly if supported by independent evidence that expression of the gene(s) concerned is correlated with disease state, has the potential to highlight putative mechanisms and enable a targeted approach to resequencing and fine-mapping (Fig. 3B). In principle, these 'genetical genomics' approaches can be extended to proteomic and metabolomic data (63). However, with so many variants to consider, and so many potential eQTL associations, it can be difficult to know precisely what weight to place on such observations, especially if the eQTL data come from a tissue or cell-line of limited pathophysiological relevance to the condition of interest (64). Nonetheless, such approaches at least offer mechanistic hypotheses amenable to early experimental evaluation. The eQTL approach is likely to be particularly valuable in circumstances where causal variants exert remote regulatory effects on genes whose coding regions lie outside the boundaries of the region of maximal association and therefore would not otherwise be considered as strong candidates for involvement in the disease process.

Using genome annotation

A visit to any genome browser will reveal the vast range of annotation which now adorns the human genome sequence (65). As more of this information becomes available, researchers have to consider how to exploit these rich data sources to support identification of causal variants (Fig. 3C).

In principle, one strategy would involve using such annotations to assign different prior odds (based, for example, on estimated functional impact) to the variants that map within a region of maximal association, allowing formal integration of functional and statistical data to inform fine-mapping analyses. This would, in effect, attempt to extend and formalize heuristics that currently assign the greatest causal credibility to non-synonymous coding SNPs that lead to non-conservative amino acids changes in critical parts of the protein product. The equivalent approach at the level of the gene makes use of mechanistic insights from previous rounds of disease-susceptibility gene discovery to inform evaluations of positional candidacy in new signals [as in the case of Crohn's disease and autophagy, for example (66,67)].

Of course, the challenge with such strategies (particularly those operating at the level of sequence variation) lies in defining those priors with any accuracy: the sheer breadth of annotation available can be intimidating, and it may be unclear which particular annotations are most relevant in any given setting. Nevertheless, such approaches may provide some rational basis for defining the order and nature of functional experimentation when statistical methods alone have failed to resolve completely the identity of the causal variants. Equally, *in silico* assessments of functional credibility are likely to become increasingly important as researchers target low-frequency variants that are individually too rare for definitive association testing (68). The most powerful association tests in such a setting may well involve 'locus-wide' comparisons of the overall mutational load in cases and controls, and such analyses would certainly be strengthened if it became possible to integrate estimates of the potential functional impact of each of the variants concerned.

Using functional experimentation

If *in silico* assignment of function is difficult, experimental evaluation of the functional impact of putative causal variants is even more challenging. Many GWA signals map some distance from the nearest coding sequence, and are likely to mediate disease predisposition through remote regulatory effects on transcription, or (in the case of microRNAs) translation. Such variants are notoriously resistant to functional enquiry: ensuring that assays developed are relevant to the disease situation will depend on making appropriate choices with respect to the tissue and cell-type of interest, the stage of cellular and organismal development, and the inclusion of pertinent environmental factors, decisions which are almost impossible to make with confidence. Similar obstacles are likely in the use of animal models to interrogate the function of putative causal variants. Functional studies can be troublesome enough for alleles causal for Mendelian disease: in comparison, causal alleles involved in complex trait susceptibility are likely to have much more subtle molecular and cellular effects. It is essential to bear in mind that measurable effects of an allele in a given functional assay does not, by itself, prove a causal role in disease pathogenesis. Unless the functional assay is particularly compelling as a model for the processes involved in disease pathogenesis in man, such functional data cannot substitute for convincing genetic evidence.

Using genetics again: finding new causal variants at associated loci

Allelic heterogeneity is a common feature of Mendelian disease, and many genes have been implicated in both rare and common forms of the same condition [one obvious example is *KCNJ11* variants in which are related to diabetes phenotypes ranging from syndromic neonatal diabetes to common type 2 diabetes (69,70)]. Once there is evidence that one variant in a given gene influences a particular disease phenotype, the probability that other functional variants in the same gene also modify disease risk is markedly enhanced.

Such considerations open up the prospect of deploying targeted resequencing efforts to identify independent causal variants within the genes mapping to an association signal, in the hope that such discoveries will reveal which gene is responsible for the index association (Fig. 3D). The explicit aim is to identify 'smoking gun' mutations (typically of low-frequency and modest- to high-penetrance) which avoid the problems of functional attribution that complicate mechanistic inferences for common, low-penetrance variants. A typical experiment might resequence several hundreds of individuals (selected to represent both extremes of the phenotypic trait distribution, to capture both protective and risk-increasing low-frequency alleles). Such experiments might reasonably target the most functionally important components (exons, promoters, UTRs, highly-conserved sequences) of each gene within or adjacent to the region of maximal association. The advent of high-throughput resequencing technologies, often combined with careful DNA pooling strategies, has simplified both the economics and logistics of such studies. While it might appear desirable (since some causal variants will

surely lie outside sequences of premium functional importance) to extend deep-resequencing efforts to the entire association signal, it remains unclear, given the limitations of interpreting and evaluating the likely functional impact of many intergenic and intronic variants, how sequence information from such regions could be used effectively. Indeed, the difficulty now is not so much the capacity to discover low frequency variants, but the ability to establish robust criteria for their evaluation. Those criteria can be based on *in silico* parameters [such as limiting analyses to rare, non-conservative, missense mutations (71)], large-scale association testing (when adequately powered), co-segregation in pedigrees, or where feasible, high-throughput functional assays.

If resequencing efforts are successful, they will identify independent causal variants (subsequently confirmed by association or co-segregation), or demonstrate differences in mutational load [such as a shift in the synonymous to non-synonymous ratio for coding variants (68)] between subject groups that clearly mirror the phenotype of interest. In this case, fine-mapping of the index association may no longer be as critical, because the relevant gene will have been identified, enabling subsequent translational progress. An additional and important benefit of the deep-resequencing approach is the potential to uncover alleles with more severe effects at the molecular and clinical levels: such variants are likely to be more attractive substrates for functional and clinical investigation than the common variant which was originally detected.

CONCLUSION

For many different common diseases and quantitative traits, GWA studies have successfully identified multiple loci with associated common variants. Because the loci rarely encompass previously-noted candidate genes, these discoveries have generally shown that previously unsuspected biology is important in leading to disease. These insights have the potential to open new routes to novel treatments and preventive measures. However, in most cases, only a small fraction of the known genetic contribution to phenotype has been accounted for by these associated variants. Additional GWA studies, in larger samples and multiple ethnicities, will almost certainly lead to new discoveries and incremental gains in the amount of risk accounted for by identified genetic variants. In addition, exploration of these novel loci will very likely uncover additional alleles, both common and rare, that explain additional variance in phenotype, help pinpoint which gene(s) are responsible for the association and provide better clinical and molecular tools for assessing function and mechanism of disease. Looking ahead, new methodologies and approaches may be needed to discover the remaining as yet unidentified genetic contributors to disease risk. At associated loci, fine mapping can help narrow down the list of possible causal variants and simplify future functional studies. Finally, the challenges of moving from associated variant to mechanism of action are substantial, especially where the identity of the relevant gene(s) is uncertain or genome annotation is not helpful. Thus, GWA studies have already generated significant advances, but much of the potential impact of these advances has yet to be felt. Fulfilling the promise of GWA studies to improve our understanding of human disease and biology will require additional tools and resources (Box 1), and a coordinated effort from not only geneticists but also a broader range of biologists.

Box 1 Resources needed to progress from current findings of GWA studies.

- Large samples in diverse populations for multiple diseases/traits.
- Complete knowledge of common variation across the genome in multiple populations.
- Methods to interrogate efficiently structural variation in large samples.
- Improved sequencing technology and/or other methods for interrogating low frequency variation.
- Computational methods to interpret sequence data from large samples.
- Expression data from densely genotyped human samples and covering diverse tissue types.
- Improved genome annotation, especially of non-coding regions.
- Relevant and validated functional assays for associated genes.
- Tractable animal models or highly relevant *in vitro* models in which human causal variants can be assessed.
- Coordinated assessment of environmental exposures and disease outcomes in large cohorts with DNA samples available.
- Computational tools for comprehensive assessment of $G \times G$ and $G \times E$ joint effects.
- Assessment of the role of epigenetics in the inherited risk of disease.

FUNDING

MIMcC is partly supported by the Oxford NIHR Biomedical Research Centre. JNH is partially supported by R01DK075875 from NIDDK and by a grant from the Sandler Program in Asthma Research.

Conflict of Interest statement. None declared.

REFERENCES

1. Lettre, G. and Rioux, J.D. (2008) Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.* (this issue).

2. Easton, D. (2008) Genome-wide association studies in cancer. *Hum. Mol. Genet.* (this issue).
3. Mohlke, K.L., Boehnke, M. and Abecasis, G.R. (2008) Metabolic and cardiovascular traits: an abundance of recently identified common genetic variants. *Hum. Mol. Genet.* (this issue).
4. Weedon, M.N., Lango, H., Lindgren, C.M., Wallace, C., Evans, D.M., Mangino, M., Freathy, R.M., Perry, J.R., Stevens, S., Hall, A.S. *et al.* (2008) Genome-wide association analysis identifies 20 loci that influence adult height. *Nat. Genet.*, **40**, 575–583.
5. Lettre, G., Jackson, A.U., Gieger, C., Schumacher, F.R., Berndt, S.I., Sanna, S., Eyheramendy, S., Voight, B.F., Butler, J.L., Guiducci, C. *et al.* (2008) Identification of ten loci associated with height highlights new biological pathways in human growth. *Nat. Genet.*, **40**, 584–591.
6. Gudbjartsson, D.F., Walters, G.B., Thorleifsson, G., Stefansson, H., Halldorsson, B.V., Zusmanovich, P., Sulem, P., Thorlacius, S., Gylfason, A., Steinberg, S. *et al.* (2008) Many sequence variants affecting diversity of adult human height. *Nat. Genet.*, **40**, 609–615.
7. Cooper, R.S., Tayo, B. and Zhu, X. (2008) Genome-wide association studies: implications for multiethnic samples. *Hum. Mol. Genet.* (this issue).
8. King, R., Rotter, J. and Motulsky, A. (eds) (2002) *The Genetic Basis of Common Diseases*, Oxford University Press, New York.
9. Menzel, S., Garner, C., Gut, I., Matsuda, F., Yamaguchi, M., Heath, S., Foglio, M., Zelenika, D., Boland, A., Rooks, H. *et al.* (2007) A QTL influencing F cell production maps to a gene encoding a zinc-finger protein on chromosome 2p15. *Nat. Genet.*, **39**, 1197–1199.
10. Uda, M., Galanello, R., Sanna, S., Lettre, G., Sankaran, V.G., Chen, W., Usala, G., Busonero, F., Maschio, A., Albai, G. *et al.* (2008) Genome-wide association study shows BCL11A associated with persistent fetal hemoglobin and amelioration of the phenotype of beta-thalassemia. *Proc. Natl Acad. Sci. USA*, **105**, 1620–1625.
11. Thorleifsson, G., Magnusson, K.P., Sulem, P., Walters, G.B., Gudbjartsson, D.F., Stefansson, H., Jonsson, T., Jonasdottir, A., Stefansdottir, G., Masson, G. *et al.* (2007) Common sequence variants in the LOXL1 gene confer susceptibility to exfoliation glaucoma. *Science*, **317**, 1397–1400.
12. Sulem, P., Gudbjartsson, D.F., Stacey, S.N., Helgason, A., Rafnar, T., Magnusson, K.P., Manolescu, A., Karason, A., Palsson, A., Thorleifsson, G. *et al.* (2007) Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nat. Genet.*, **39**, 1443–1452.
13. Sladek, R., Rocheleau, G., Rung, J., Dina, C., Shen, L., Serre, D., Boutin, P., Vincent, D., Belisle, A., Hadjadj, S. *et al.* (2007) A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature*, **445**, 881–885.
14. Saxena, R., Voight, B.F., Lyssenko, V., Burtt, N.P., de Bakker, P.I., Chen, H., Roix, J.J., Kathiresan, S., Hirschhorn, J.N., Daly, M.J. *et al.* (2007) Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science*, **316**, 1331–1336.
15. Scott, L.J., Mohlke, K.L., Bonnycastle, L.L., Willer, C.J., Li, Y., Duren, W.L., Erdos, M.R., Stringham, H.M., Chines, P.S., Jackson, A.U. *et al.* (2007) A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science*, **316**, 1341–1345.
16. Zeggini, E., Weedon, M.N., Lindgren, C.M., Frayling, T.M., Elliott, K.S., Lango, H., Timpson, N.J., Perry, J.R., Rayner, N.W., Freathy, R.M. *et al.* (2007) Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science*, **316**, 1336–1341.
17. Steinthorsdottir, V., Thorleifsson, G., Reynisdottir, I., Benediktsson, R., Jonsdottir, T., Walters, G.B., Styrkarsdottir, U., Gretarsdottir, S., Emilsson, V., Ghosh, S. *et al.* (2007) A variant in CDKAL1 influences insulin response and risk of type 2 diabetes. *Nat. Genet.*, **39**, 770–775.
18. Zeggini, E., Scott, L.J., Saxena, R., Voight, B.F., Marchini, J.L., Hu, T., de Bakker, P.I., Abecasis, G.R., Almgren, P., Andersen, G. *et al.* (2008) Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.*, **40**, 638–645.
19. Lohmueller, K.E., Pearce, C.L., Pike, M., Lander, E.S. and Hirschhorn, J.N. (2003) Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nat. Genet.*, **33**, 177–182.
20. Goring, H.H., Terwilliger, J.D. and Blangero, J. (2001) Large upward bias in estimation of locus-specific effects from genomewide scans. *Am. J. Hum. Genet.*, **69**, 1357–1369.
21. McCarthy, M.I. (2008) Casting a wider net for diabetes susceptibility genes. *Nat. Genet.*, **40**, 1039–1040.
22. Yasuda, K., Miyake, K., Horikawa, Y., Hara, K., Osawa, H., Furuta, H., Hirota, Y., Mori, H., Jonsson, A., Sato, Y. *et al.* (2008) Variants in KCNQ1 are associated with susceptibility to type 2 diabetes mellitus. *Nat. Genet.*, **40**, 1092–1097.
23. Unoki, H., Takahashi, A., Kawaguchi, T., Hara, K., Horikoshi, M., Andersen, G., Ng, D.P., Holmkvist, J., Borch-Johnsen, K., Jorgensen, T. *et al.* (2008) SNPs in KCNQ1 are associated with susceptibility to type 2 diabetes in East Asian and European populations. *Nat. Genet.*, **40**, 1098–1102.
24. Nalls, M.A., Wilson, J.G., Patterson, N.J., Tandon, A., Zmuda, J.M., Huntsman, S., Garcia, M., Hu, D., Li, R., Beamer, B.A. *et al.* (2008) Admixture mapping of white cell count: genetic locus responsible for lower white blood cell count in the Health ABC and Jackson Heart studies. *Am. J. Hum. Genet.*, **82**, 81–87.
25. Valdar, W., Solberg, L.C., Gauguier, D., Burnett, S., Klenerman, P., Cookson, W.O., Taylor, M.S., Rawlins, J.N., Mott, R. and Flint, J. (2006) Genome-wide genetic association of complex traits in heterogeneous stock mice. *Nat. Genet.*, **38**, 879–887.
26. Accurso, F.J. and Sontag, M.K. (2008) Gene modifiers in cystic fibrosis. *J. Clin. Invest.*, **118**, 839–841.
27. Lettre, G., Sankaran, V.G., Bezerra, M.A., Araújo, A.S., Uda, M., Sanna, S., Cao, A., Schlessinger, D., Costa, F.F., Hirschhorn, J.N. and Orkin, S.H. (2008) DNA polymorphisms at the *BCL11A, HBS1L-MYB*, and beta-globin loci associate with fetal hemoglobin levels and pain crises in sickle cell disease. *Proc. Natl Acad. Sci. USA*, **105**, 11869–11874.
28. Marchini, J., Donnelly, P. and Cardon, L.R. (2005) Genome-wide strategies for detecting multiple loci that influence complex diseases. *Nat. Genet.*, **37**, 413–417.
29. Wong, M.Y., Day, N.E., Luan, J.A., Chan, K.P. and Wareham, N.J. (2003) 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.
30. Wong, M.Y., Day, N.E., Luan, J.A. and Wareham, N.J. (2004) Estimation of magnitude in gene-environment interactions in the presence of measurement error. *Stat. Med.*, **23**, 987–998.
31. Zeggini, E., Rayner, W., Morris, A.P., Hattersley, A.T., Walker, M., Hitman, G.A., Deloukas, P., Cardon, L.R. and McCarthy, M.I. (2005) An evaluation of HapMap sample size and tagging SNP performance in large-scale empirical and simulated data sets. *Nat. Genet.*, **37**, 1320–1322.
32. McCarroll, S.A. (2008) Extending genome-wide association studies to copy-number variation. *Hum. Mol. Genet.* (this issue).
33. McCarthy, M.I., Abecasis, G., Cardon, L.R., Little, J., Ioannidis, J.P.A., Goldstein, D.B. and Hirschhorn, J.N. (2008) Genome wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.*, **9**, 356–369.
34. Morrow, E.M., Yoo, S.Y., Flavell, S.W., Kim, T.K., Lin, Y., Hill, R.S., Mukaddes, N.M., Balkhy, S., Gascon, G., Hashmi, A. *et al.* (2008) Identifying autism loci and genes by tracing recent shared ancestry. *Science*, **321**, 218–223.
35. Kong, A., Masson, G., Frigge, M.L., Gylfason, A., Zusmanovich, P., Thorleifsson, G., Olason, P.I., Ingason, A., Steinberg, S., Rafnar, T. *et al.* (2008) Detection of sharing by descent, long-range phasing and haplotype imputation. *Nat. Genet.*, **40**, 1068–1075.
36. Kerkel, K., Spadola, A., Yuan, E., Kosek, J., Jiang, L., Hod, E., Li, K., Murty, V.V., Schupf, N., Vilain, E. *et al.* (2008) Genomic surveys by methylation-sensitive SNP analysis identify sequence-dependent allele-specific DNA methylation. *Nat. Genet.*, **40**, 904–908.
37. Morgan, H.D., Sutherland, H.G., Martin, D.I. and Whitelaw, E. (1999) Epigenetic inheritance at the agouti locus in the mouse. *Nat. Genet.*, **23**, 314–318.
38. Marchini, J., Howie, B., Myers, S., McVean, G. and Donnelly, P. (2007) A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet.*, **39**, 906–913.
39. Scheet, P. and Stephens, M. (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase *Am. J. Hum. Genet.*, **78**, 629–644.
40. Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., Leal, S.M. *et al.* (2007) A second generation human haplotype map of over 3.1 million SNPs. *Nature*, **449**, 851–861.
41. International HapMap Consortium. (2005) A haplotype map of the human genome. *Nature*, **437**, 1299–1320.

42. Wiltshire, S., Morris, A.P. and Zeggini, E. (2008) Examining the statistical properties of fine-scale mapping in large-scale association studies. *Genet. Epidemiol.*, **32**, 204–214.
43. Mardis, E.R. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet.*, **24**, 133–141.
44. Lowe, C.E., Cooper, J.D., Brusko, T., Walker, N.M., Smyth, D.J., Bailey, R., Bourget, K., Plagnol, V., Field, S., Atkinson, M. *et al.* (2007) Large-scale genetic fine mapping and genotype-phenotype associations implicate polymorphism in the IL2RA region in type 1 diabetes. *Nat. Genet.*, **39**, 1074–1082.
45. Ueda, H., Howson, J.M., Esposito, L., Heward, J., Snook, H., Chamberlain, G., Rainbow, D.B., Hunter, K.M., Smith, A.N., Di Genova, G. *et al.* (2003) Association of the T-cell regulatory gene CTLA4 with susceptibility to autoimmune disease. *Nature*, **423**, 506–511.
46. Farrer, L.A., Cupples, L.A., Haines, J.L., Hyman, B., Kukull, W.A., Mayeux, R., Myers, R.H., Pericak-Vance, M.A., Risch, N. and van Duijn, C.M. (1997) Effects of age, sex, and ethnicity on the association between apolipoprotein E genotype and Alzheimer disease. A meta-analysis. APOE and Alzheimer Disease Meta Analysis Consortium. *JAMA*, **278**, 1349–1356.
47. Ng, M.C., Park, K.S., Oh, B., Tam, C.H., Cho, Y.M., Shin, H.D., Lam, V.K., Ma, R.C., So, W.Y., Cho, Y.S. *et al.* (2008) Implication of genetic variants near *TCF7L2, SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2*, and *FTO* in type 2 diabetes and obesity in 6,719 Asians. *Diabetes*, **57**, 2226–2233.
48. Wu, Y., Li, H., Loos, R.J., Yu, Z., Ye, X., Chen, L., Pan, A., Hu, F.B. and Lin, X. (2008) Common variants in *CDKAL1, CDKN2A/B, IGF2BP2, SLC30A8* and *HHEX/IDE* genes are associated with type 2 diabetes and impaired fasting glucose in a Chinese Han population. *Diabetes*, [Epub ahead of print Jul 15].
49. Sanghera, D.K., Ortega, L., Han, S., Singh, J., Ralhan, S.K., Wander, G.S., Mehra, N.K., Mulvihill, J.J., Ferrell, R.E., Nath, S.K. and Kamboh, M.I. (2008) Impact of nine common type 2 diabetes risk polymorphisms in Asian Indian Sikhs: *PPARG2* (Pro12Ala), *IGF2BP2*, *TCF7L2* and *FTO* variants confer a significant risk. *BMC Med. Genet.*, **9**, 59.
50. Omori, S., Tanaka, Y., Takahashi, A., Hirose, H., Kashiwagi, A., Kaku, K., Kawamori, R., Nakamura, Y. and Maeda, S. (2008) Association of *CDKAL1, IGF2BP2, CDKN2A/B, HHEX, SLC30A8*, and *KCNJ11* with susceptibility to type 2 diabetes in a Japanese population. *Diabetes*, **57**, 791–795.
51. Frayling, T.M., Timpson, N.J., Weedon, M.N., Zeggini, E., Freathy, R.M., Lindgren, C.M., Perry, J.R., Elliott, K.S., Lango, H., Rayner, N.W. *et al.* (2007) A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, **316**, 889–894.
52. Gudmundsson, J., Sulem, P., Steinthorsdottir, V., Bergthorsson, J.T., Thorleifsson, G., Manolescu, A., Rafnar, T., Gudbjartsson, D., Agnarsson, B.A., Baker, A. *et al.* (2007) Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.*, **39**, 977–983.
53. Thomas, G., Jacobs, K.B., Yeager, M., Kraft, P., Wacholder, S., Orr, N., Yu, K., Chatterjee, N., Welch, R., Hutchinson, A. *et al.* (2008) Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.*, **40**, 310–315.
54. Frayling, T.M., Colhoun, H. and Florez, J.C. (2008) A genetic link between type 2 diabetes and prostate cancer. *Diabetologia*, **51**, 1757–1760.
55. Libioulle, C., Louis, E., Hansoul, S., Sandor, C., Farnir, F., Franchimont, D., Vermeire, S., Dewit, O., de Vos, M., Dixon, A. *et al.* (2007) 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.
56. Moffatt, M.F., Kabesch, M., Liang, L., Dixon, A.L., Strachan, D., Heath, S., Depner, M., von Berg, A., Bufe, A., Rietschel, E. *et al.* (2007) Genetic variants regulating *ORMDL3* expression contribute to the risk of childhood asthma. *Nature*, **448**, 470–473.
57. Chen, Y., Zhu, J., Lum, P.Y., Yang, X., Pinto, S., MacNeil, D.J., Zhang, C., Lamb, J., Edwards, S., Sieberts, S.K. *et al.* (2008) Variations in DNA elucidate molecular networks that cause disease. *Nature*, **452**, 429–435.
58. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.
59. Emilsson, V., Thorleifsson, G., Zhang, B., Leonardson, A.S., Zink, F., Zhu, J., Carlson, S., Helgason, A., Walters, G.B., Gunnarsdottir, S. *et al.* (2008) Genetics of gene expression and its effect on disease. *Nature*, **452**, 423–428.
60. Myers, A.J., Gibbs, J.R., Webster, J.A., Rohrer, K., Zhao, A., Marlowe, L., Kaleem, M., Leung, D., Bryden, L., Nath, P. *et al.* (2007) A survey of genetic human cortical gene expression. *Nat. Genet.*, **39**, 1494–1499.
61. Schadt, E.E., Molony, C., Chudin, E., Hao, K., Yang, X., Lum, P.Y., Kasarskis, A., Zhang, B., Wang, S., Suver, C. *et al.* (2008) Mapping the genetic architecture of gene expression in human liver. *PLoS Biol.*, **6**, e107.
62. Stranger, B.E., Nica, A.C., Forrest, M.S., Dimas, A., Bird, C.P., Beazley, C., Ingle, C.E., Dunning, M., Flicek, P., Koller, D. *et al.* (2007) Population genomics of human gene expression. *Nat. Genet.*, **39**, 1217–1224.
63. Dumas, M.E., Wilder, S.P., Bihoreau, M.T., Barton, R.H., Fearnside, J.F., Argoud, K., D'Amato, L., Wallis, R.H., Blancher, C., Keun, H.C. *et al.* (2007) Direct quantitative trait locus mapping of mammalian metabolic phenotypes in diabetic and normoglycemic rat models. *Nat. Genet.*, **39**, 666–672.
64. Nica, A.C. and Dermitzakis, E.T. (2008) Using gene expression to investigate the genetic basis of complex disorders. *Hum. Mol. Genet.* (this issue).
65. ENCODE Project ConsortiumBirney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigó, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T. *et al.* (2007) Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, **447**, 799–816.
66. Barrett, J.C., Hansoul, S., Nicolae, D.L., Cho, J.H., Duerr, R.H., Rioux, J.D., Brant, S.R., Silverberg, M.S., Taylor, K.D., Barmada, M.M. *et al.* (2008) Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nat. Genet.*, **40**, 955–962.
67. Rioux, J.D., Xavier, R.J., Taylor, K.D., Silverberg, M.S., Goyette, P., Huett, A., Green, T., Kuballa, P., Barmada, M.M., Datta, L.W. *et al.* (2007) Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.*, **39**, 596–604.
68. Romeo, S., Pennacchio, L.A., Fu, Y., Boerwinkle, E., Tybjaerg-Hansen, A., Hobbs, H.H. and Cohen, J.C. (2007) Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. *Nat. Genet.*, **39**, 513–516.
69. Gloyn, A.L., Pearson, E.R., Antcliff, J.F., Proks, P., Bruining, G.J., Slingerland, A.S., Howard, N., Srinivasan, S., Silva, J.M., Molnes, J. *et al.* (2004) Activating mutations in the gene encoding the ATP-sensitive potassium-channel subunit Kir6.2 and permanent neonatal diabetes. *N Engl. J. Med.*, **350**, 1838–1849.
70. Gloyn, A.L., Weedon, M.N., Owen, K.R., Turner, M.J., Knight, B.A., Hitman, G., Walker, M., Levy, J.C., Sampson, M., Halford, S. *et al.* (2003) Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (*KCNJ11*) and SUR1 (*ABCC8*) confirm that the *KCNJ11* E23K variant is associated with type 2 diabetes. *Diabetes*, **52**, 568–572.
71. Ji, W., Foo, J.N., O'Roak, B.J., Zhao, H., Larson, M.G., Simon, D.B., Newton-Cheh, C., State, M.W., Levy, D. and Lifton, R.P. (2008) Rare independent mutations in renal salt handling genes contribute to blood pressure variation. *Nat. Genet.*, **40**, 592–599.