APPLICATIONS OF NEXT-GENERATION SEQUENCING

Uncovering the roles of rare variants in common disease through whole-genome sequencing

Elizabeth T. Cirulli and David B. Goldstein

Abstract | Although genome-wide association (GWA) studies for common variants have thus far succeeded in explaining only a modest fraction of the genetic components of human common diseases, recent advances in next-generation sequencing technologies could rapidly facilitate substantial progress. This outcome is expected if much of the missing genetic control is due to gene variants that are too rare to be picked up by GWA studies and have relatively large effects on risk. Here, we evaluate the evidence for an important role of rare gene variants of major effect in common diseases and outline discovery strategies for their identification.

Minor allele frequency
Ranging from 0 to 50%, this is the proportion of alleles at a locus that consists of the less frequent allele. This number does not take genotype into account.

Effect size
The increase in risk (or proportion of population variation) that is conferred by a given causal variant.

Heritability
The proportion of phenotypic variation in a trait that is due to underlying genetic variation. In studies of humans, this value is usually calculated by comparing trait correlations in individuals of varying degrees of relatedness.

Center for Human Genome Variation, Duke University Medical School, Durham, North Carolina 27708, USA.
Correspondence to D.B.G.
e-mail: d.goldstein@duke.edu
doi:10.1038/nrg2779

Genome-wide association (GWA) studies, which have thus far focused on very common SNPs, have been completed for most common human diseases and many related traits. These studies have found most of the very common gene variants (minor allele frequency (MAF) > ~5%) in the human genome and have identified over 500 independent strong SNP associations ($p < 1 \times 10^{-8}$) (see the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies). However, most of these associated SNPs have very small effect sizes, and the proportion of heritability explained is at best modest for most traits[1]. Furthermore, GWA signals have rarely been tracked to causal polymorphisms, which has led many to assume that the causal variants must have subtle regulatory effects. Partially in response to this assumption, genome-wide patterns of gene expression have been characterized in multiple tissue types, and many common variants that influence gene expression have been identified[2,3]. Surprisingly, variants associated with gene expression show little overlap with those associated with disease[4]. Although the systematic identification of rare variants associated with common diseases has not yet been feasible, several rare variants have nevertheless been identified that confer a substantial risk of disease. For example, autism, mental retardation, epilepsy and schizophrenia have been shown to be influenced by rare structural variants that affect genes[5]. Additionally, it seems possible that some, perhaps even many, of the current GWA signals could reflect the effect of multiple rare variants that have been credited to common variants[6]. Whatever underlies the GWA signals for common diseases, it is clear that GWA studies of very common variants rarely succeed in securely implicating specific genes in specific common diseases, which greatly limits their importance in applications such as drug development.

Taken together, these observations support the long-established idea that rare variants could be the primary drivers of common diseases[7–9]. Over the past several years, this view has become less popular, at least with some in the research community, than the theory that common diseases are more strongly influenced by many common gene variants with small effects on risk[10–13]. Although the relative impact of common and rare variants on common diseases remains an unanswered empirical question, we think that the results of the past several years make a strong case for common diseases being more similar to Mendelian diseases than is postulated by the common disease–common variant model[14]. In particular, it seems possible that much of the genetic control of common diseases is due to rare and generally deleterious variants that have a strong impact on the risk of disease in individual patients. It is also likely that the variants with the largest effect sizes will be those that have obvious functional consequences. This rare, functional variant model is not inconsistent with the absence of secure linkage evidence for most common diseases (for example, nearly every chromosome arm has been 'linked' to schizophrenia and bipolar disorder[15–17]). Variants that increase the risk of disease by less than about

Box 1 | **Variants identified through GWA for an example of a common human disease**

Type 2 diabetes has been studied using some of the largest genome-wide association (GWA) sample sizes to date. Three large case–control studies were recently subjected to a meta-analysis with a total sample size of 4,549 cases and 5,579 controls[22]: 1,924 cases and 2,938 controls came from the Wellcome Trust Case Control Consortium[78], 1,464 cases and 1,467 controls came from the Diabetes Genetics Initiative[79], and 1,161 cases and 1,174 controls came from the Finland–United States Investigation of Non-Insulin-Dependent Diabetes Mellitus Genetics[80]. The first two studies were performed using a different chip (Affymetrix 500k) from that used in the last study (Illumina 317k), but the imputation of missing genotypes allowed for a combined analysis of 2.2 million SNPs for all 10,128 participants. Twelve SNPs had already been established as associated with type 2 diabetes in these earlier studies, and the meta-analysis confirmed six additional associated SNPs. In total, these 18 variants only explained 6% of the heritability of type 2 diabetes[21]. Furthermore, the underlying causal variant has not definitively been found for any of these significant associations, although potassium inwardly-rectifying channel, subfamily J, member 11 (*KCNJ11*) and Wolfram syndrome 1 (*WFS1*) were already known to contain rare variants that influence the disease, and there is evidence supporting potentially causal common variants in transcription factor 7-like 2 (*TCF7L2*), solute carrier family 30, member 8 (*SLC30A8*) and *KCNJ11* (REFS 81–83). The table shows type 2 diabetes-associated variants reported by Zeggini *et al.*[22]; the maximum odds ratio reported was 1.37.

SNP	Odds ratio	SNP location
rs7903146	1.37	Intron of *TCF7L2*
rs13071168	1.35	Between *SYN2* and *PPARG*
rs7020996	1.26	Near *CDKN2A* and *CDKN2B*
rs6931514	1.25	Intron of *CDKAL1*
rs1801282	1.18	Exon of *PPARG*
rs4402960	1.17	Intron of *IGF2BP2*
rs5015480	1.17	Near *HHEX*
rs5215	1.16	Exon of *KCNJ11*
rs10282940	1.15	Exon of *SLC30A8*
rs8050136	1.15	Intron of *FTO*
rs7578597	1.15	Exon of *THADA*
rs10923931	1.13	Intron of *NOTCH2*
rs4580722	1.11	Near *WFS1*
rs12779790	1.11	Between *CDC123* and *CAMK1D*
rs17705177	1.1	Near *TCF2*
rs864745	1.1	Intron of *JAZF1*
rs7961581	1.09	Between *TSPAN8* and *LGR5*
rs4607103	1.09	Near *ADAMTS9*

Mendelian disease
A disease that is carried in families in either a dominant or recessive manner and that is typically controlled by variants of large effect in a single gene.

Imputation
Based on the known linkage disequilibrium structure in fully genotyped individuals, the genotype of untyped variants can be inferred in individuals who are genotyped for a smaller number of variants.

Exome
The exome is the collection of known exons in our genome: this is the portion of the genome that is translated into proteins. As exons comprise only 1% of the genome and contain the most easily understood, functionally relevant information, sequencing of only the exome is a cheaper method of identifying most of the variants that are most likely to affect a trait.

Linkage disequilibrium
A nonrandom association between alleles at different loci.

fourfold are expected to generate inconsistent linkage evidence[18], leaving scope for the presence of many rare variants with a strong impact on disease risk. The available linkage evidence for many common diseases securely confirms high locus heterogeneity; otherwise, the weak linkage evidence would tend to point towards the same genome regions in different families, which is generally not the case. This is not to say that common variants play no part in common diseases. Indeed, studies of some Mendelian diseases indicate that common variants may often have a key role as modifiers of the effects of rarer, more highly penetrant contributors to disease risk[19,20], and it seems reasonable to expect that this also holds true for common diseases. This view could have fundamental implications for the discovery of the genetic bases of common diseases and for therapeutic applications.

In this Review, we first discuss the insights provided by GWA studies into the genetic architecture underlying common diseases. We then present evidence supporting the potentially important role of rare variants in common diseases. Next, we outline strategies, namely whole-genome and whole-exome sequencing, for the discovery of rare causal variants, with a focus on family-based designs and extreme-trait designs. Finally, we suggest that the underlying causes of Mendelian and common diseases may be more similar than previously supposed, and we offer a perspective of what the future holds for these types of studies.

The genetic architecture of common diseases

Although most of the complex traits that have been studied using the GWA approach have reasonable levels of heritability, the proportion of that heritability that has been explained by very common variants is surprisingly low[1]. This limited impact is well-illustrated by two diseases that have been subject to exceptionally large studies. Despite a discovery sample size of 10,128 and a replication sample size of 53,975, the 18 common variants that were found to be significantly associated with type 2 diabetes seem to explain only about 6% of the increased risk of disease among relatives[21,22] (BOX 1). Likewise, a meta-analysis of schizophrenia GWA studies that included a total of 8,008 cases and 19,077 controls identified only 7 significant SNPs, some in high linkage disequilibrium with each other and each with an odds ratio below 1.3, despite a heritability of 80–85% for schizophrenia[23]. Pharmacogenetic traits may be an exception: they have not been extensively studied, but associated variants of large effect have been identified by GWA studies[24–28]. This difference in effect size may reflect the lack of long-term selection on many such traits (BOX 2).

Several explanations for the missing heritability, and solutions for finding it, have been offered. One popular view had been that because disease status is determined by many underlying factors, intermediate phenotypes related to the condition would provide a more tractable target in GWA studies than the

Box 2 | **Pharmacogenetic traits**

Although the effect sizes of SNPs identified by genome-wide association (GWA) studies have been very modest for most complex traits, this has not been true for the small number of pharmacogenetic traits that have been studied in this manner. Of the nearly 450 (not screened for independence) GWA signals from case–control studies that were reported to have a *p*-value below 1×10^{-8} in the National Human Genome Research Institute (NHGRI) Catalog of Published Genome-Wide Association Studies, only five were studies of pharmacogenetic traits (two of these are replications of one of the others). Despite this imbalance, these traits have some of the strongest associations found: all five have an odds ratio above 1.95, which is true of only 13.8% of associated SNPs for non-pharmacogenetic traits, and two of them have odds ratios above 25.0, an effect size seen for only one non-pharmacogenetic association. These results may simply reflect the low-hanging fruit, and as more pharmacogenetic GWA studies are performed, it may be found that most pharmacogenetic variants do not have effects this large. However, thus far they support the long-standing theory that the genetic variants that contribute to pharmacogenetic traits will often have little phenotypic consequence before the administration of a drug[84] (note that this is not to say that variants that influence drug response have never been under selection for other reasons — for example, it has been postulated that genetic variants that affect cytochrome P450, family 2, subfamily D, polypeptide 6 (*CYP2D6*)[85] may have been influenced by alkaloid toxins in the diet). Because humans have spent most of their evolutionary history unexposed to drugs, many pharmacogenetic traits have largely escaped the attention of natural selection. This is in contrast to other complex traits, which have been under comparatively heavy selection unless their effects on fitness come very late in life. Because of this lack of constraint on the evolution of variants that affect pharmacogenetic traits, they have had the opportunity to rise in frequency through genetic drift. Although this does not mean that such variants will always be of a sufficiently high frequency to be detected by GWA studies, they are more likely to have a high frequency than the variants that have been under selection due to their effect on common diseases and other complex traits.

disease itself. Available data, however, provide little support for this view. For example, cognitive traits are perhaps the leading endophenotype in schizophrenia, but GWA studies have proved no more successful for uncovering the genetic basis of cognitive traits in normal individuals than for studying the causes of schizophrenia itself[29,30].

Another popular theory is that the variation contributing to common diseases has subtle effects and is context-dependent. For example, some have argued that specific combinations of common variants may lead to substantially greater effects on risk than is apparent for any individual common variant. Although this explanation is theoretically possible, there are currently few clear examples of common variants identified in GWA studies that interact strongly with each other or with the environment to affect a complex trait. Nevertheless, methods for detecting interactions among the many common variants in the human genome are still immature[31], and it remains possible that novel analytic and experimental approaches will identify new risk profiles that emerge from multi-way interactions.

Finally, a recent study suggests that parent-of-origin effects may be important for common disease[32], and that the incorporation of such information into GWA studies may increase the estimated contribution of associated common variants to disease risk.

Endophenotype
An intermediate phenotype that is heritable and associated with a disease but is not itself a symptom of the disease. Although there is little evidence to support the theory, it has been argued that endophenotypes would be a more tractable target for genetic analysis than the relevant disease state itself.

Haploinsufficiency
This occurs when a diploid organism only has one copy of a gene and both copies are required for correct function. This is one way that a protein-truncating mutation can influence predisposition to a disease.

Rare variants and common diseases

For common variants that have been definitively shown to be associated with a trait, the underlying causal variant has rarely been found, although there is supporting evidence for certain traits; examples include a prostate-cancer-associated SNP that is upstream of a gene and is associated with expression level in cell lines, and breast-cancer-associated intronic SNPs that are associated with expression and binding of transcription factors[33,34]. Because pinning down common causal variants with subtle effects is difficult, many could eventually be confirmed over time. However, even for apparently clear examples, such as age-related macular degeneration (AMD) — a now classic example of a disease-associated common variant — there is still no general agreement on whether the identified coding variant explains the observed associations[35–37]. The common disease–common variant hypothesis[14] led to the general assumption that some as-yet-unidentified common variant is responsible for the associations seen in most GWA studies[7,22,38,39]. This assumption has controlled much of the follow-up research, in which the focus has generally been on resequencing in the linkage disequilibrium block defined by common variants[40–43]. Even when rare variants in the same regions as GWA signals have been identified[44,45], supporting the view that they could influence disease risk, these seem to be viewed as independent of the original association.

What are rare variants? Although the GWA studies performed so far have been designed to fully characterize direct effects of very common variants, they have provided little information about whether the rest of the genetic control for each trait is only slightly more rare than the general detection threshold in these GWA studies (about 5%) or substantially more rare. This question is crucial because the answer determines the optimal research strategies. In TABLE 1, we outline a division of potential causal variant frequencies and their implications for analysis methods.

Rare copy-number variants. The same gene chips used to evaluate common SNPs can also be used to analyse the roles of common and rare large copy-number variants (CNVs), and this approach, along with others, has identified secure connections between CNVs and common diseases[5]. Although systematic work on large CNVs in disease is in its infancy, specific CNVs are associated with an effect on the risk of disease that dramatically exceeds the effects of most common variants associated with a disease. For example, three rare deletions were recently shown to be associated with schizophrenia with odds ratios estimated at 2.7, 11.5 and 14.8 (REF. 46), which are substantially higher than the typical values for common variants that are associated with a disease: of those GWA study case–control associations with *p*-values below 1×10^{-8}, only 13% had odds ratios above 2, and only 1% had odds ratios above 10 (see the NHGRI Catalog of Published Genome-Wide Association Studies). There are two obvious possibilities for the pathogenicity of these deletions: haploinsufficiency and the revealing of recessive

Table 1 | **Potential frequencies of causal variants in complex traits**

Variant class	Minor allele frequency	Implications for analysis
Very common	Between 5 and 50%	Amenable to association analysis using current genome-wide association methods
Less common	Between 1 and 5%	Amenable to association analysis using variants catalogued in the 1000 Genomes Project
Rare (but not private)	Less than 1% but still polymorphic in one or more major human populations	Amenable to framework of extreme phenotype resequencing, as well as co-segregation in families
Private	Restricted to probands and immediate relatives	Difficult to analyse except through co-segregation in families. As linkage evidence will (by definition) be modest, discovery would be limited to the most recognizable of variants

mutations on the homologous chromosome (FIG. 1). Rare CNVs are simply one class of rare variants that may contribute to common diseases, and the example of Mendelian diseases (see below) strongly suggests that rare CNVs will be a minority of all the rare variants that contribute to disease risk.

Synthetic associations. It has long been recognized that a common variant could record a diluted signal of risk due to a rarer causal variant of large effect. An early paper from the HapMap consortium made this point in 2003, noting: "...even a relatively uncommon disease-associated variant can potentially be discovered using this approach. Reflecting its historical origins, the uncommon variant will be travelling on a chromosome that carries a characteristic pattern of nearby sequence variants. In a group of people affected by a disease, the rare variant will be enriched in frequency compared with its frequency in a group of unaffected controls."[10]

Nevertheless, the properties of such associations had not been systematically explored until a recent study by Dickson *et al.*[6]. They considered a model in which one or more 'rare' variants (defined as below the threshold of reliable representation in GWA studies) were the only contributors to disease risk. They then asked how often a signal of risk conferred by such variants would be picked up and credited to common variants in a typical GWA study (FIG. 2). They found that in typical GWA sample sizes, rare variants can easily create genome-wide significant associations, and the properties of such associations can be dramatically different from what occurs when the causal variant is common. For example, the causal variants can be megabases away from the common variants that carry the signal of the association, and the real risk effects can be several-fold stronger than what is credited to a common variant. In addition, and counter-intuitively, the likelihood of a genome-wide significant signal being credited to common variants can increase with the number of rare causal variants in a region (because multiple rare variants can load up onto the same haplotype defined by a common variant). Dickson *et al.* referred to these associations as 'synthetic associations', a special case of indirect association. This term was not meant to imply that such associations are spurious but was intended to emphasize the differences between the properties of these associations and what is expected when the causal variant is common.

It is unknown how many GWA signals may be due to synthetic associations, but the ease with which they could occur does question some of the inferences that have been drawn about the results of GWA studies and also mandates a more open-minded approach to following up GWA signals. For example, when associated SNPs are far from the nearest gene, it is often assumed that a regulatory variant of subtle effect must be the cause, such as for the SNPs in 8q24 that are associated with colorectal cancer[47–49] and located hundreds of kilobases from the oncogene *MYC*. Although there is evidence supporting *MYC* regulation by this region, there has been no definitive link between the associated variants and this regulation[50,51]. However, as a synthetic association can be driven by rare variants that are megabases from the GWA signal[6] (FIG. 2), the 'locus' implicated by a GWA study might be much larger than has typically been assumed. The effect sizes of the causal variants may also be substantially larger than that of the common variant that tags them. Accepting the possibility of synthetic associations means that follow-up sequencing efforts for GWA study results should extend well beyond the block of linkage disequilibrium surrounding a discovery variant that is defined by common variants.

Although there has been little systematic effort devoted to finding synthetic associations, it has been shown that rare or less common variants in some of the same genomic regions as GWA signals influence disease risk. For example, one region that has been associated with height through GWA studies contains the gene growth differentiation factor 5 (*GDF5*), which was already known to contain rare variants that result in skeletal disorders[45,52]. In a study of type 1 diabetes that sequenced the area around a GWA signal in the gene interferon induced with helicase C domain 1 (*IFIH1*), four rare or less common functional variants were identified that were protective in addition to the originally associated common variant[44]. More recently, a common variant (MAF of 0.194 in European-Americans) in C20orf194 was found in a GWA study to be associated with haemolytic anaemia in response to interferon-α and ribavirin treatment (for chronic hepatitis C infection) with a p-value of 1.1×10^{-45} (REF. 27). However, follow-up genotyping of two functional variants with lower frequencies (MAFs of 0.076 and 0.123) in the nearby gene inosine triphosphatase (*ITPA*) showed that these variants entirely accounted for the original GWA signal.

Haplotype
A combination of alleles that are inherited together.

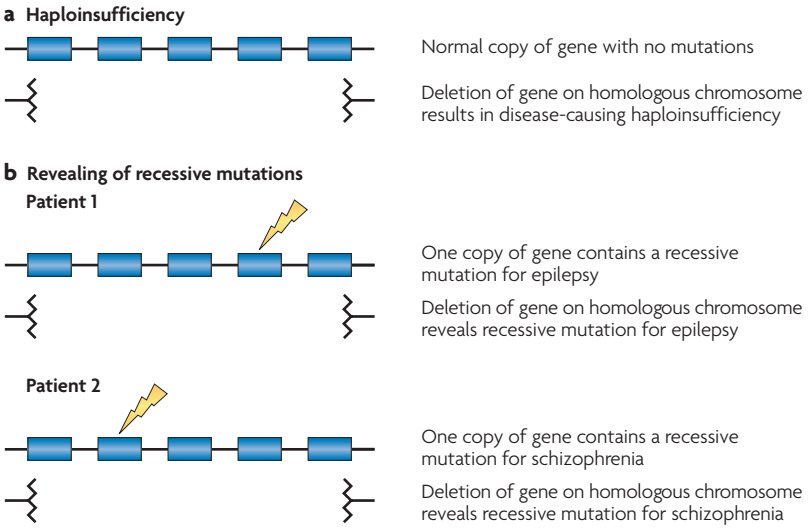

Figure 1 | **Role of deletions in disease predisposition.** The figure shows a heterozygous deletion (brackets) and five genes (blue boxes) on the homologous chromosome, with lightning bolts representing mutations. Deletions can be linked to a disease either by causing a deleterious haploinsufficiency (**a**) or by revealing deleterious recessive mutations (**b**). In **b**, the recessive mutations that are revealed could predispose to different conditions. Because individuals bearing a recessive mutation predisposing to one condition or another would effectively become homozygotes for those mutations in the presence of a deletion, the same deletion can become associated with multiple conditions. For example, deletion of a region at 2p16.3 has been shown to be associated with schizophrenia, autism and mental retardation[54,101–103]. The easiest way to resolve these possibilities will be through sequencing of the homologous chromosome in individuals with the relevant deletions.

1000 Genomes Project
An international research consortium that will sequence the genomes of 1,200 individuals of various ethnicities. Most individuals will be sequenced to low coverage, or in exons only. The goals are to catalogue human variation with minor allele frequencies of ~1% or greater and to refine and optimize strategies for sequencing large numbers of genomes.

Coverage
The number of sequence reads that have alignments that overlap a certain position. Because current sequencing strategies produce random reads, resulting in an uneven distribution of reads across the genome, a high average coverage is required to assure that most bases in the genome are covered by multiple reads.

Furthermore, using a model in which each minor allele, for either lower frequency variant, was counted once, the *p*-value dropped to 2.2×10^{-92}. This is one of the few examples in which the causal variants from a GWA signal have been definitively identified, and in some ways the findings are consistent with the characteristics of a synthetic association, in which multiple functional variants load up in the same direction on a haplotype defined by a more common variant included on the gene chip. In this case, the causal variants explaining the association are common, perhaps because the trait relates to a drug response[53] (BOX 2).

Strategies for identifying disease-causing variants

The basic paradigm used in the recently completed phase of GWA studies was to catalogue very common variants and genotype them either directly or indirectly (through linkage disequilibrium) in cases and controls. These GWA studies successfully represented variants with a frequency of ~5% or higher in the general population. If variants that are slightly less common than this, in the range of 1–5%, contribute significantly to common diseases, then a simple extension of this paradigm is appropriate. This line of thought implicitly underlies the 1000 Genomes Project, which is extending the catalogue of known human variants down to a frequency near 1% (see the 1000 Genomes website). Chips for a new wave of GWA studies, which will account for these less common variants, are already being designed by companies such as Illumina and Affymetrix. A second analysis of phenotypes using the millions of genotypes available on these chips is likely to identify some new associations. If, however, much of the causation is due to rare variants, such an approach is unlikely to uncover much more of the genetic control of diseases than has already been revealed by the current GWA efforts. The CNVs that have recently been implicated in many common diseases generally have population frequencies that are considerably below this threshold[46,54,55].

To capture such rare variants, it will be necessary to sequence entire genomes of cases instead of genotyping a catalogue of variants. Up until recently, this was a daunting task. However, new machines[56] can currently sequence 55 billion bases in about 10 days; although this rate has been increasing with remarkable rapidity, the cost per accurate base call is still the most important parameter and needs improving. This impressive amount of data provides sufficient 'coverage' to identify most variants present in a given genome, and the increase in sequencing capacity is allowing more laboratories to sequence whole genomes. Although sample sizes would have to be large to identify rare variants of small effect, there are many examples of variants (or genotypes) with large enough effects to be strongly enriched in specific populations. For example, individuals who are homozygous for the C-C chemokine receptor 5 Δ32 (*CCR5Δ32*) deletion, who comprise around 1% of the general European population, are virtually immune to HIV infection[57–59]. This protection increases the frequency of such homozygotes by up to 20-fold in individuals who have been highly exposed to HIV and yet remain uninfected[60]. In another example, the frequency of the risk factor human leukocyte antigen-B*5701 (*HLA-B*5701*) is increased as much as 30-fold in individuals who are hypersensitive to the antiretroviral drug abacavir compared with the general population[61–63]. Although neither of these variants is rare, they show that the frequency of variants influencing complex traits can be so enriched in carefully defined populations that they could be readily identified even in a data set as large as would be generated in a whole-genome sequencing study.

Whole-genome and whole-exome sequencing. The most comprehensive study of the role of inherited variation in diseases will involve whole-genome sequencing of all enrolled subjects. Such studies will eventually be carried out in a manner similar to GWA studies, with very large sample sizes that will provide sufficient statistical evidence to implicate variants on the basis of association evidence alone. In the interim, it will be important to focus on designs that are optimized to detect the role of causal variants in smaller samples.

Proof-of-concept examples of the identification of rare, disease-causing variants are now available for whole-genome and whole-exome sequencing strategies. For example, Ng *et al.*[64] sequenced the exomes of four unrelated cases with Freeman–Sheldon syndrome and eight controls. Although the cause of this disease was already known, the authors showed that the causal

variants were evident after whole-exome sequencing, as the causal gene was the only one that contained at least one coding indel or non-synonymous or splice-site variant in all four cases that was not present in any of the controls nor in dbSNP.

Choi *et al.*[65] used whole-exome sequencing to discover the cause of disease in an individual with an unclear diagnosis. They identified a small number of homozygous missense mutations in positions that were highly conserved from invertebrates to humans. One such variant was in a gene known to cause congenital chloride-losing diarrhoea, consistent with the patient's symptoms.

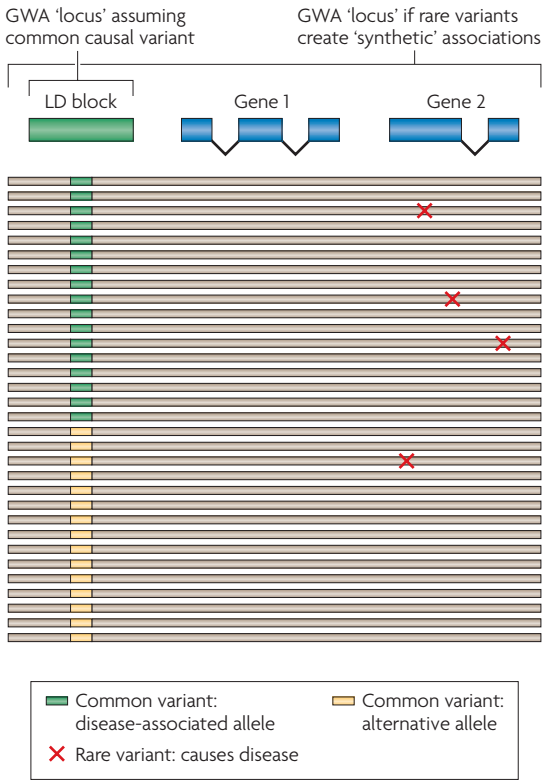

Figure 2 | **Synthetic associations.** Shown is a set of chromosomes (grey boxes) containing a common variant (green and yellow boxes) and four rare variants (red Xs). At the top, the area of linkage disequilibrium (LD) surrounding the common variant is shown, as are nearby genes. Each rare variant shown can cause the disease, and collectively they are more common on the haplotype containing the green version of the common variant. In a genome-wide association (GWA) study for this disease, the rare variants will not be directly assessed but will often create a signal that is credited to one or more common variants. In most cases (as illustrated), the signal credited to the common variant would be far weaker than the real effects of the causal variants. Furthermore, because the LD block containing the common variants does not extend to the gene containing the causal variants, one might assume that the causal variant must be regulatory. However, in synthetic associations, causal variants can lie within a much larger region than the LD block surrounding the associated, common variant.

Most importantly, whole-genome and whole-exome sequencing have been used to identify the previously unknown causes of diseases. For example, Ng *et al.*[66] discovered the unknown cause of Miller syndrome by performing whole-exome sequencing in four cases from three kindreds and eight controls. They located the causal gene by identifying genes that had coding indels, non-synonymous variants or splice-site variants in all of the cases but in none of the controls. Although the causes of such Mendelian diseases will be easier to discover than those of more complex diseases, these successes indicate that whole-genome sequencing of even a small number of individuals can identify causal variants.

Until complete genomic sequencing is inexpensive enough to use in the large sample sizes that would be needed to perform whole-genome association studies with no *a priori* weighting of putative functional variants, two designs are likely to be the primary engines of discovery. The first design involves selecting families that have multiple affected individuals (family-based sequencing), and the second approach involves selecting individuals that are at the extreme ends of a trait distribution (extreme-trait designs) (FIG. 3). Extreme-trait designs will be particularly important for identifying variants that are rare but not private and that have modest to high effect sizes.

Sequencing affected individuals in families. Under this design, a family with multiple individuals affected with a common disease would be selected. One economical design could involve the initial sequencing of the most distantly related, co-affected family members and the identification of overlapping variants, as the more distantly related the co-affected individuals, the fewer genetic variants they will share. However, even distant relatives will share too many variants to allow easy identification of the causal variants, even when relatives share the same rare causal variants (which will not always be the case). The list of variants will therefore need to be further screened by function, zygosity, population frequency and/or the type of gene affected to focus on the most likely candidates. Promising variants would then be checked for co-segregation in the family. However, for most common diseases, the evidence emerging from co-segregation analyses will not be sufficient to distinguish among a large set of candidate pathogenic variants. For example, consider a pedigree involving 16 individuals who are affected with bipolar disorder and 35 who are unaffected[67]; the probability of co-segregation by chance would be 0.0003. This statistic means that co-segregation analysis would be a possible approach for prioritizing among a small set of key functional variants, such as protein-truncating variants, but would be insufficient for distinguishing among more general classes, such as all coding variants, let alone all identified variants. With such a small discovery sample size, the successful narrowing down of possible causal variants will be heavily dependent on the causal variant being rare or private and of obvious functional consequence. Most likely, the discovery paradigm will require combining co-segregation evidence with evidence from additional

Indel
A small insertion or deletion of nucleotides. If it occurs in an exon and is not a multiple of three in length, it results in a frameshift and usually the loss of gene function.

Splice-site variant
A variant, usually found at the intron–exon boundary, that alters the splicing of an exon to its surrounding exons.

Non-synonymous variant
A genetic variant that changes a codon for one amino acid to another amino acid. Many non-synonymous variants are well-tolerated, but others can cause a disease.

Co-segregation
In the pedigree of a family with a condition, the segregation pattern shows how often the putative causal variant is found to coincide with the condition. When a variant coincides with the condition in a family, the condition and the variant are said to co-segregate.

families and association evidence that is based on the typing of candidate variants in large case–control cohorts. The development of appropriate test statistics that combine these different lines of evidence is a current priority for the field.

This suggests that modest linkage evidence may be a promising way to narrow down the region (and variants) of interest. For example, Sobreira *et al.* discovered the cause of metachondromatosis, a dominant Mendelian disease with incomplete penetrance, by first running a linkage analysis on a small pedigree with DNA for only 12 family members[68]. The analysis implicated six regions, with logarithm of the odds (LOD) scores of 2.5, 1.8 and 1.07, that comprised a total of 42 Mb. The authors then sequenced the entire genome of one affected patient from this pedigree and eight unrelated controls, and focused analysis on private, good-quality variants with functional consequences for protein-coding genes in the 42 linked megabases. This allowed them to identify the causative protein-truncating variant, which was found in a linkage peak with the second-highest LOD score; this variant was confirmed by locating a similar variant in a second, unrelated family with the disease. Although these results are for a Mendelian disease, and similar studies for common diseases are unlikely to be this simple, they show that a combination of weak linkage evidence and functional prioritization can be used to identify the causative variant by sequencing only one patient genome.

Extreme-trait sequencing. A basic extreme-trait design would be to sequence a small, carefully selected population at one or both ends of the extremes of a phenotype (this idea has been discussed previously[69] and an exploration of power in this context is available[70]). Obvious examples include individuals who are known to be highly exposed to HIV but remain uninfected or individuals at the extremes of the distributions for blood pressure. Because variants that contribute to the trait will be enriched in frequency in such a population, even small sample sizes may suggest many candidate variants that can then be genotyped for confirmation in a much larger group of samples. Consider a case in which a variant is enriched 30-fold in such an extreme population, as with hypersensitivity to abacavir. If this variant were rare in the general population, say with a MAF of 0.1%, one would need to sequence about 30 extreme individuals to see it once. Although identifying enrichment for this particular variant would probably involve the sequencing of at least 100 extreme individuals, this would not

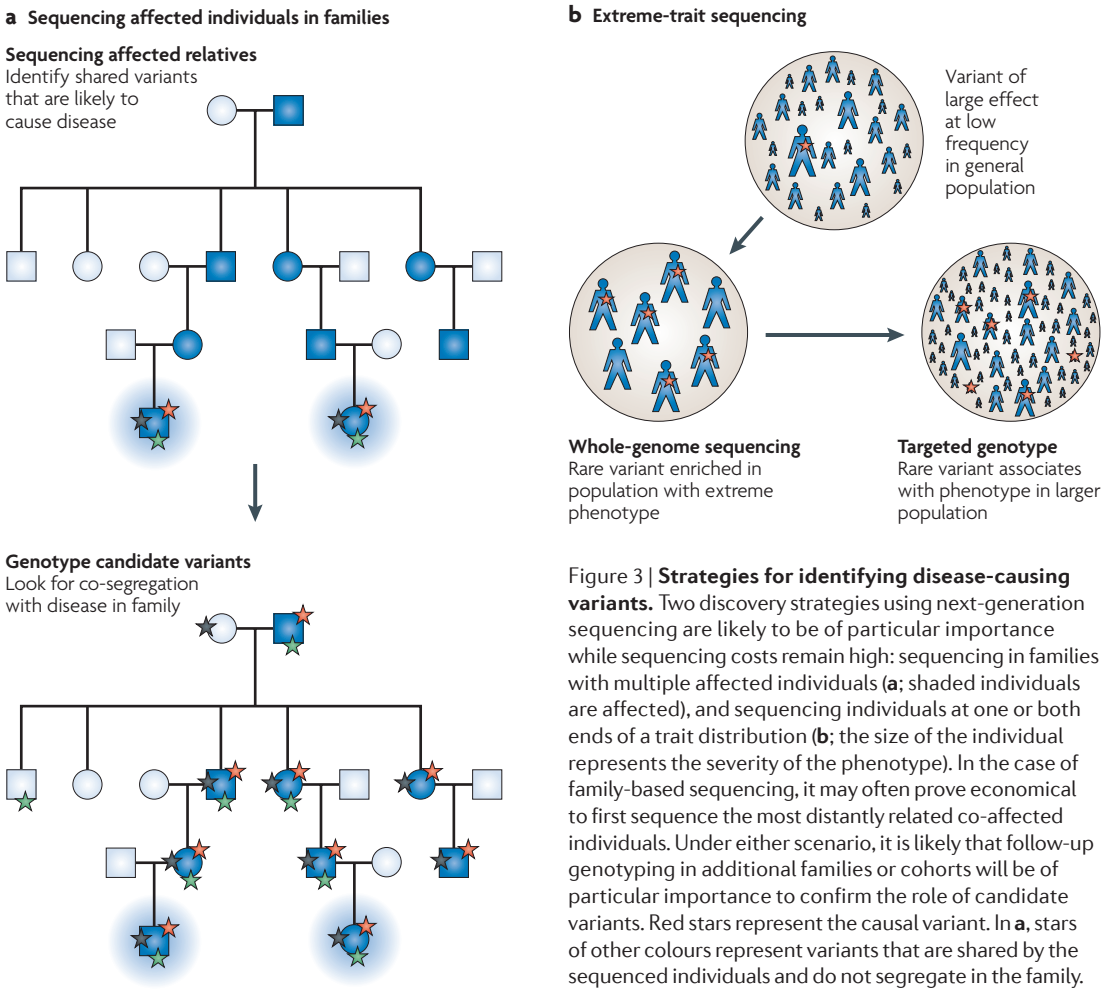

Figure 3 | **Strategies for identifying disease-causing variants.** Two discovery strategies using next-generation sequencing are likely to be of particular importance while sequencing costs remain high: sequencing in families with multiple affected individuals (**a**; shaded individuals are affected), and sequencing individuals at one or both ends of a trait distribution (**b**; the size of the individual represents the severity of the phenotype). In the case of family-based sequencing, it may often prove economical to first sequence the most distantly related co-affected individuals. Under either scenario, it is likely that follow-up genotyping in additional families or cohorts will be of particular importance to confirm the role of candidate variants. Red stars represent the causal variant. In **a**, stars of other colours represent variants that are shared by the sequenced individuals and do not segregate in the family.

be the case if multiple rare variants in the same gene affect the phenotype. Searching for genes with an enrichment of rare variants in extreme individuals will implicate this gene even at a low number of sequenced genomes, although again this will be heavily dependent on the variants being of obvious function. Although currently the only examples of the identification of causal variants by whole-genome or whole-exome sequencing come from Mendelian diseases, Ng *et al.*[64,66] used this technique to identify genes that were enriched for rare variants in a very small number of cases versus controls: the same variant did not need to be present in all cases. For the follow-up of variants identified in extreme-trait sequencing, family members of the extreme individuals will be invaluable for confirming potentially causal variants through co-segregation analysis. This strategy will be an effective complement to extreme-trait sequencing in nearly every situation in which family members who show variability in the trait are available for such a follow-up.

These two strategies will be some of the most effective methods for using whole-genome sequence data from a small number of individuals. It seems likely that for many diseases that affect only a small proportion of the population (for example, schizophrenia, epilepsy and amyotrophic lateral sclerosis), both designs will be used and will complement one another. However, such studies will also face many challenges. The success of family-based sequencing will depend on the rarity of causal variants in the population at large, and the genetic bases of conditions that affect a relatively small proportion of the population are likely to be more easily identified than conditions, such as type 2 diabetes, that affect a much higher proportion of the population. In the case of such common diseases, even families that carry relatively high impact rare variants will show imperfect co-segregation, and the enrichment of specific causal variants among affected, unrelated individuals will be much lower. For extreme-trait sequencing, accurate phenotyping will be of vital importance; as only a small number of individuals will be sequenced, the inclusion of even a small proportion of misclassified individuals because of factors such as measurement error could affect the analysis. At small sample sizes, both methods will also depend on the causal variants being of obvious functional consequence. Furthermore, the accurate calling of variants and sufficient coverage to identify the presence of causal variants in all individuals will be crucial. Finally, the sheer number of rare and private variants that will be identified will make identification of the causal variants difficult; this problem will be compounded by both locus and allelic heterogeneity.

The importance of recognizability

In both family-based and extreme-trait discovery paradigms, a key determinant of statistical power will be whether causal sites tend to be 'recognizable' — that is, whether the variants have an obvious function. Examples of recognizable variants would be ones that delete some or all of a gene, introduce a premature stop site into a protein or result in a non-conservative amino acid substitution. The general failure of GWA studies of very common variants to identify causal sites has led to the widespread belief that most signals must reflect subtle regulatory variants that are not easily recognized[69–71]. However, this view is based on assumption, not observation, and it is possible that the variants that contribute to common diseases behave more like those that contribute to Mendelian diseases, in which the variants tend to have obvious effects[72,73]. In contrast to this lack of information about the underlying causes of complex diseases, tens of thousands of variants that cause Mendelian diseases have been catalogued and characterized. Most of these variants are readily recognizable[72,73]. If common diseases have genetic causes that are, in this way, similar to those of Mendelian diseases, we can expect most of them to be missense, nonsense or splicing variants, or obviously functional deletions or insertions.

To illustrate the importance of recognizability, consider the number of candidate variants in different functional groups. For example, Shianna *et al.* compared the frequencies of different classes of variations in ten case and ten control genomes (K. V. Shianna *et al.*, unpublished data). There were 383,913 variants (single-nucleotide variants and indels) present in at least two cases and no controls. Testing such a large number of variants would require a p-value of 1.3×10^{-7} to declare a significant association. However, if testing is restricted to only variants that affect the coding sequence, this number drops to 2,354, which requires a p-value of only 2.2×10^{-5}. If testing is restricted to only protein-truncating variants, the number drops further to 152, which requires a p-value of 3.3×10^{-4}. It will therefore be essential to develop methods that prioritize variants based on the likelihood that they contribute to disease (BOX 3).

Future directions

Both the professional and lay communities have adjusted to the apparently limited impact of genetic differences on common diseases. For example, there seems to have been a dampening of enthusiasm in industry for the use of genetic association data in prioritizing drug targets, whereas the provision of genetic risk profiles has become commonplace and even allows social engagement at 'spit parties', at which party-goers produce saliva samples for genotyping. This attitude is in stark contrast to the more than 25 million Americans with rare genetic diseases (see the US National Institutes of Health Office of Rare Diseases Research website), many of which can be diagnosed with highly accurate genetic tests (see the GeneTests website). However, these reactions to the characteristics of common genetic variation may be at odds with rare genetic contributions to common diseases. The identification of rare variants that influence diseases is suddenly feasible thanks to next-generation sequencing, and it is possible that discovery will move faster than generally anticipated. If rare, obviously functional variants have a big influence on common diseases, whole-genome sequencing will allow definitive connections to be rapidly established between specific genes and many important common diseases. Definitive connections — for example, a clearly functional mutation

Box 3 | **Bioinformatic approaches and challenges for whole-genome sequencing studies**

The analysis of sequence data presents a fundamentally different challenge from the analysis of a targeted set of polymorphisms, as was the paradigm for genome-wide association (GWA) studies. The most fundamental distinction is that in a sequencing study, it is necessary to assess all variants that have been identified, and many of these variants will not have yet been included in polymorphism databases. To deal with such data, it is necessary to have an integrated bioinformatic environment that begins with the accurate calling of variants based on the sequence data and that includes the categorization of all identified variants, novel or known, into functional categories.

The challenges begin with variant calling. Even after the sequence reads for a whole genome have been aligned, which takes a few days using programs such as BWA[86] or Bowtie[87], single-nucleotide variants and indels must be identified, a process that takes programs such as SAMTools[88] or ssahaSNP a few more days. The identification of other classes of variants, such as copy-number variants or inversions, requires the use of variant calling methods that are based on a combination of read depth and read pair methods[89–91]. Following these variant calling procedures, it is crucial that the called variants are annotated, distinguished in terms of their most likely functional consequences and presented to researchers in a format that can be used for analysis. Many existing visualization tools can conveniently display the aligned short reads from sequence data in the context of annotation features (for example, genes and known variants); however, they were not designed to actually annotate the variants[92–95]. To perform annotation and analysis, many groups have constructed local annotation servers[64], used their own in-house annotation pipelines[96–99] or sometimes developed new genome browsers[100]. However, there is still a need for an efficient and publicly available software tool that can annotate, organize and interpret the millions of variants that are called in each genome.

Our group has developed a unique software environment called SequenceVariantAnalyzer that takes as input the variants called for each genome sequenced, as well as the quality scores and coverage statistics, and annotates and stores them for subsequent analysis. Individual genomes are classified as either cases or controls, and variants can be screened by function, location, zygosity and quality. The software comes with a number of functions for prioritizing genes or variants, such as identifying variants that are found in the homozygous state only in cases, searching for genes that are enriched with truncating variants only in cases and not in controls, and identifying compound heterozygotes. Software that performs these functions makes the analysis of whole-genome sequence data possible, and will be essential for human geneticists in the coming years.

in a single gene conferring a strongly elevated risk of a disease — would provide validated therapeutic targets for the pharmaceutical industry. Indeed, it may not be an overstatement to suggest that genetic discovery could be the most likely avenue for ameliorating the ongoing crisis in global drug development[74]. However, this prediction assumes that rare variants will be found that have large influences on common diseases, that their biological functions will be obvious and that locus and allelic heterogeneity will not obfuscate insights into the mechanisms of disease. How often these assumptions will hold is currently unknown and will largely determine the rate of discovery in the coming years.

Before discovery genetics focused on rare variants can be systematized in a manner analogous to what was successfully accomplished in the study of common variants, many technical challenges will need to be overcome. The development of analysis techniques to cope with the millions of variants called per genome will be a high priority, as will the development of techniques that can combine data about different rare variants into one analysis. New software for accurately identifying CNVs and inversions will also be important, as there is currently no highly accurate method for identifying even a high proportion of all such variants present in a genome. Even methods for calling simpler variants, such as those affecting a single nucleotide, will need to be improved, as currently there are tens of thousands of false-positive variant calls per genome[75], which further complicates the search for causal variants. Even more effort must be applied to accurately calling small insertions and deletions, which are currently called much less accurately than single site changes. Furthermore, it will be essential for researchers to carefully choose cases and controls for each study: as costs restrict sample size, the success of each project will be highly dependent on the careful selection of individuals to sequence. If rare variants of large effect are found, a transition towards the study of particular pathogenic variants that strongly influence diseases will greatly increase the importance of designs that allow for the evaluation of family members carrying putative causal variants. The availability of large control cohorts who can be recalled and phenotypically evaluated will also be crucial, especially if unscreened controls from the general population are used for multiple studies involving different phenotypes. Unlike variants with weak influences, which could be expected to appear in the population at large without much phenotypic effect, a variant of strong effect is less likely to appear in an individual without any phenotypic consequence. Confirmation of such potentially causal variants will therefore often require the careful evaluation of the phenotype in any controls who are carriers.

Until the cost of whole-genome sequencing drops to a level that is similar to whole-exome sequencing (including the cost of the exome capture step), whole-exome sequencing will be a good alternative strategy for many studies. Because the most obvious disease-influencing variants will be the clearly functional ones, whole-exome sequencing should pack most of the variation of interest into a much smaller, more cost-effective and more easily interpreted bundle. However, there are drawbacks to this technique: most importantly, it almost entirely misses structural variation. Whole-exome sequencing is also restricted to a certain set of exons: if causal variants lie in exons that are not targeted, they will not be identified.

Compound heterozygote
When an individual inherits two different recessive mutations, one from each parent, in the same gene that cause the same phenotype. An example would be a single-nucleotide variant causing a codon for an amino acid to be changed into a stop codon in one allele and a 4-bp deletion in the other allele: each of these variants knock out their respective allele, resulting in neither copy functioning.

Additionally, the capture methods currently used are nonspecific enough that they require the sequencing of a far greater number of bases than expected based on the size of the exome, which makes whole-exome sequencing prices comparable to those of low-coverage whole-genome sequencing. However, low-coverage sequencing will miss many of the variants present in only a single individual. For these reasons, high-coverage whole-genome sequencing will be the method of choice once it becomes more affordable, and the rapid increase in the sequencing capacity of existing platforms, as well as the development of new, less-expensive platforms[76], suggests this will not be more than a few years away.

We predict that, within the next 5 years, rare and private variants with moderate to large effects on many complex traits will be discovered. With regard to personal genomics, it is possible that genome sequencing could provide specific enough information about the risk of common diseases to influence lifestyle choices and the use of relatively non-invasive monitoring programs (for example, imaging); it might even lead to an expansion of interest in couples-based and fetal screening. This outlook is tempered by past experience that has shown that risk factors for disease, even when well-characterized, are sometimes not particularly useful in the clinical setting[77]; this may well continue to be true even if rare variants of large effect are found to influence common diseases. However, we may be cautiously optimistic that if a large enough set of definitive connections can be established between specific genes and diseases, a subset of these discoveries may lead to clear clinical deliverables.

1. Maher, B. Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008).
One of the first articles to explicitly recognize that GWA studies explain a small part of the genetic components of many diseases.
2. Kasowski, M. *et al.* Variation in transcription factor binding among humans. *Science* **328**, 232–235 (2010).
3. Pickrell, J. K. *et al.* Understanding mechanisms underlying human gene expression variation with RNA sequencing. *Nature* **464**, 768–772 (2010).
4. Heinzen, E. L. *et al.* Tissue-specific genetic control of splicing: implications for the study of complex traits. *PLoS Biol.* **6**, e1 (2008).
5. Stankiewicz, P. & Lupski, J. R. Structural variation in the human genome and its role in disease. *Annu. Rev. Med.* **61**, 437–455 (2010).
6. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
7. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genet.* **40**, 695–701 (2008).
8. Schork, N. J., Murray, S. S., Frazer, K. A. & Topol, E. J. Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).
9. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
10. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
11. Pritchard, J. K. & Cox, N. J. The allelic architecture of human disease genes: common disease-common variant.or not? *Hum. Mol. Genet.* **11**, 2417–2423 (2002).
12. Stephens, J. W. & Humphries, S. E. The molecular genetics of cardiovascular disease: clinical implications. *J. Intern. Med.* **253**, 120–127 (2003).
13. Plomin, R., Haworth, C. M. & Davis, O. S. Common disorders are quantitative traits. *Nature Rev. Genet.* **10**, 872–878 (2009).
14. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
15. Tandon, R., Keshavan, M. S. & Nasrallah, H. A. Schizophrenia, 'just the facts' what we know in 2008. 2. Epidemiology and etiology. *Schizophr. Res.* **102**, 1–18 (2008).
16. Crow, T. J. How and why genetic linkage has not solved the problem of psychosis: review and hypothesis. *Am. J. Psychiatry* **164**, 13–21 (2007).
17. Serretti, A. & Mandelli, L. The genetics of bipolar disorder: genome 'hot regions', genes, new potential candidates and future directions. *Mol. Psychiatry* **13**, 742–771 (2008).
18. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
19. Steinberg, M. H. & Adewoye, A. H. Modifier genes and sickle cell anemia. *Curr. Opin. Hematol.* **13**, 131–136 (2006).
20. Thein, S. L. & Menzel, S. Discovering the genetics underlying foetal haemoglobin production in adults. *Br. J. Haematol.* **145**, 455–467 (2009).
21. Manolio, T. A. *et al.* Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
22. Zeggini, E. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nature Genet.* **40**, 638–645 (2008).
23. Shi, J. *et al.* Common variants on chromosome 6p22.1 are associated with schizophrenia. *Nature* **460**, 753–757 (2009).
24. SEARCH Collaborative Group *et al. SLCO1B1* variants and statin-induced myopathy — a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
25. Tanaka, Y. *et al.* Genome-wide association of *IL28B* with response to pegylated interferon-α and ribavirin therapy for chronic hepatitis C. *Nature Genet.* **41**, 1105–1109 (2009).
26. Daly, A. K. *et al. HLA-B*5701* genotype is a major determinant of drug-induced liver injury due to flucloxacillin. *Nature Genet.* **41**, 816–819 (2009).
27. Fellay, J. *et al. ITPA* gene variants protect against anemia in patients treated for chronic hepatitis C. *Nature* **464**, 405–408 (2010).
28. Ge, D. *et al.* Genetic variation in *IL28B* predicts hepatitis C treatment-induced viral clearance. *Nature* **461**, 399–401 (2009).
29. Need, A. C. *et al.* A genome-wide study of common SNPs and CNVs in cognitive performance in the CANTAB. *Hum. Mol. Genet.* **18**, 4650–4661 (2009).
30. Cirulli, E. T. *et al.* Common genetic variation and performance on standardized cognitive tests. *Eur. J. Hum. Genet.* 3 Feb 2010 (doi:10.1038/ejhg.2010.2).
31. Bhattacharjee, S. *et al.* Using principal components of genetic variation for robust and powerful detection of gene–gene interactions in case–control and case-only studies. *Am. J. Hum. Genet.* **86**, 331–342 (2010).
32. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
33. Meyer, K. B. *et al.* Allele-specific up-regulation of *FGFR2* increases susceptibility to breast cancer. *PLoS Biol.* **6**, e108 (2008).
34. Chang, B. L. *et al.* Fine mapping association study and functional analysis implicate a SNP in *MSMB* at 10q11 as a causal variant for prostate cancer risk. *Hum. Mol. Genet.* **18**, 1368–1375 (2009).
35. Hughes, A. E. *et al.* A common *CFH* haplotype, with deletion of *CFHR1* and *CFHR3*, is associated with lower risk of age-related macular degeneration. *Nature Genet.* **38**, 1173–1177 (2006).
36. Hageman, G. S. *et al.* Extended haplotypes in the complement factor H (*CFH*) and CFH-related (*CFHR*) family of genes protect against age-related macular degeneration: characterization, ethnic distribution and evolutionary implications. *Ann. Med.* **38**, 592–604 (2006).
37. Spencer, K. L. *et al.* Deletion of *CFHR3* and *CFHR1* genes in age-related macular degeneration. *Hum. Mol. Genet.* **17**, 971–977 (2008).
38. Frayling, T. M. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Rev. Genet.* **8**, 657–662 (2007).
39. McCarthy, M. I. & Hirschhorn, J. N. Genome-wide association studies: potential next steps on a genetic journey. *Hum. Mol. Genet.* **17**, R156–R165 (2008).
40. Bouatia-Naji, N. *et al.* A variant near *MTNR1B* is associated with increased fasting plasma glucose levels and type 2 diabetes risk. *Nature Genet.* **41**, 89–94 (2009).
41. 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).
42. 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).
43. Pillai, S. G. *et al.* A genome-wide association study in chronic obstructive pulmonary disease (COPD): identification of two major susceptibility loci. *PLoS Genet.* **5**, e1000421 (2009).
44. Nejentsev, S., Walker, N., Riches, D., Egholm, M. & Todd, J. A. Rare variants of *IFIH1*, a gene implicated in antiviral responses, protect against type 1 diabetes. *Science* **324**, 387–389 (2009).
This study showed that rare variants in the same region as a GWA signal for diabetes were associated with the disease.
45. Sanna, S. *et al.* Common variants in the *GDF5–UQCC* region are associated with variation in human height. *Nature Genet.* **40**, 198–203 (2008).
46. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
One of the first studies to identify rare CNVs associated with a common disease.
47. Gruber, S. B. *et al.* Genetic variation in 8q24 associated with risk of colorectal cancer. *Cancer Biol. Ther.* **6**, 1143–1147 (2007).
48. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nature Genet.* **39**, 984–988 (2007).
49. Zanke, B. W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nature Genet.* **39**, 989–994 (2007).
50. Prokunina-Olsson, L. & Hall, J. L. No effect of cancer-associated SNP rs6983267 in the 8q24 region on co-expression of *MYC* and *TCF7L2* in normal colon tissue. *Mol. Cancer* **8**, 96 (2009).
51. Sotelo, J. *et al.* Long-range enhancers on 8q24 regulate c-Myc. *Proc. Natl Acad. Sci. USA* (2010).
52. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40**, 575–583 (2008).
53. Goldstein, D. B. Common genetic variation and human traits. *N. Engl. J. Med.* **360**, 1696–1698 (2009).
54. Need, A. C. *et al.* A genome-wide investigation of SNPs and CNVs in schizophrenia. *PLoS Genet.* **5**, e1000373 (2009).
55. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
56. Metzker, M. L. Sequencing technologies — the next generation. *Nature Rev. Genet.* **11**, 31–46 (2010).

57. Dean, M. *et al.* Genetic restriction of HIV-1 infection and progression to AIDS by a deletion allele of the *CKR5* structural gene. Hemophilia Growth and Development Study, Multicenter AIDS Cohort Study, Multicenter Hemophilia Cohort Study, San Francisco City Cohort, ALIVE Study. *Science* **273**, 1856–1862 (1996).
58. Liu, R. *et al.* Homozygous defect in HIV-1 coreceptor accounts for resistance of some multiply-exposed individuals to HIV-1 infection. *Cell* **86**, 367–377 (1996).
59. Samson, M. *et al.* Resistance to HIV-1 infection in caucasian individuals bearing mutant alleles of the *CCR-5* chemokine receptor gene. *Nature* **382**, 722–725 (1996).
60. Huang, Y. *et al.* The role of a mutant *CCR5* allele in HIV-1 transmission and disease progression. *Nature Med.* **2**, 1240–1243 (1996).
61. Mallal, S. *et al.* Association between presence of *HLA-B*5701*, *HLA-DR7*, and *HLA-DQ3* and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. *Lancet* **359**, 727–732 (2002).
62. Martin, A. M. *et al.* Predisposition to abacavir hypersensitivity conferred by *HLA-B*5701* and a haplotypic *Hsp70-Hom* variant. *Proc. Natl Acad. Sci. USA* **101**, 4180–4185 (2004).
63. Young, B. *et al.* First large, multicenter, open-label study utilizing *HLA-B*5701* screening for abacavir hypersensitivity in North America. *AIDS* **22**, 1673–1675 (2008).
64. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
The first study to show that next-generation sequencing can be used to identify disease-causing variants.
65. Choi, M. *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc. Natl Acad. Sci. USA* **106**, 19096–19101 (2009).
The first study to diagnose a disease using next-generation sequencing.
66. Ng, S. B. *et al.* Exome sequencing identifies the cause of a mendelian disorder. *Nature Genet.* **42**, 30–35 (2010).
67. Yang, S. *et al.* Genomic landscape of a three-generation pedigree segregating affective disorder. *PLoS ONE* **4**, e4474 (2009).
68. Sobreira, N. L. M. *et al.* Whole genome sequencing of a single proband together with linkage analysis identifies a Mendelian disease gene. *PLoS Genet.* (in the press).
69. Manolio, T. A., Brooks, L. D. & Collins, F. S. A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118**, 1590–1605 (2008).
70. Verlaan, D. J. *et al.* Targeted screening of *cis*-regulatory variation in human haplotypes. *Genome Res.* **19**, 118–127 (2009).
71. Barrett, J. C. *et al.* Genome-wide association defines more than 30 distinct susceptibility loci for Crohn's disease. *Nature Genet.* **40**, 955–962 (2008).
72. Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
73. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genet.* **33**, 228–237 (2003).
A thoughtful overview of the kinds of mutations responsible for Mendelian disease that provides many insights about appropriate designs for studying common disease.
74. Caskey, C. T. The drug development crisis: efficiency and safety. *Annu. Rev. Med.* **58**, 1–16 (2007).
75. Roach, J. C. *et al.* Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science* 10 Mar 2010 (doi:10.1126/science.1186802).
76. Drmanac, R. *et al.* Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* **327**, 78–81 (2010).
77. Clayton, D. G. Prediction and interaction in complex disease genetics: experience in type 1 diabetes. *PLoS Genet.* **5**, e1000540 (2009).
78. 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).
A technically important early study providing well-powered GWA tests for multiple conditions.
79. Diabetes Genetics Initiative. Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
80. 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).
81. Hamming, K. S. *et al.* Coexpression of the type 2 diabetes susceptibility gene variants *KCNJ11* E23K and *ABCC8* S1369A alter the ATP and sulfonylurea sensitivities of the ATP-sensitive K^+ channel. *Diabetes* **58**, 2419–2424 (2009).
82. Nicolson, T. J. *et al.* Insulin storage and glucose homeostasis in mice null for the granule zinc transporter ZnT8 and studies of the type 2 diabetes-associated variants. *Diabetes* **58**, 2070–2083 (2009).
83. Gaulton, K. J. *et al.* A map of open chromatin in human pancreatic islets. *Nature Genet.* **42**, 255–259 (2010).
84. Motulsky, A. G. Drug reactions enzymes, and biochemical genetics. *JAMA* **165**, 835–837 (1957).
85. Ingelman-Sundberg, M. Genetic polymorphisms of cytochrome P450 2D6 (*CYP2D6*): clinical consequences, evolutionary aspects and functional diversity. *Pharmacogenomics J.* **5**, 6–13 (2005).
86. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows–Wheeler transform. *Bioinformatics* **25**, 1754–1760 (2009).
87. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
88. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
89. Hormozdiari, F., Alkan, C., Eichler, E. E. & Sahinalp, S. C. Combinatorial algorithms for structural variation detection in high-throughput sequenced genomes. *Genome Res.* **19**, 1270–1278 (2009).
90. Yoon, S., Xuan, Z., Makarov, V., Ye, K. & Sebat, J. Sensitive and accurate detection of copy number variants using read depth of coverage. *Genome Res.* **19**, 1586–1592 (2009).
91. Simpson, J. T., McIntyre, R. E., Adams, D. J. & Durbin, R. Copy number variant detection in inbred strains from short read sequence data. *Bioinformatics* **26**, 565–567 (2010).
92. Milne, I. *et al.* Tablet — next generation sequence assembly visualization. *Bioinformatics* **26**, 401–402 (2010).
93. Bao, H. *et al.* MapView: visualization of short reads alignment on a desktop computer. *Bioinformatics* **25**, 1554–1555 (2009).
94. Manske, H. M. & Kwiatkowski, D. P. LookSeq: a browser-based viewer for deep sequencing data. *Genome Res.* **19**, 2125–2132 (2009).
95. Arner, E., Hayashizaki, Y. & Daub, C. O. NGSView: an extensible open source editor for next-generation sequencing data. *Bioinformatics* **26**, 125–126 (2010).
96. Schuster, S. C. *et al.* Complete Khoisan and Bantu genomes from southern Africa. *Nature* **463**, 943–947 (2010).
97. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
One of the first studies to sequence an entire human genome using next-generation sequencing.
98. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
99. Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet.* **4**, e1000160 (2008).
100. Axelrod, N. *et al.* The HuRef Browser: a web resource for individual human genomics. *Nucleic Acids Res.* **37**, D1018–D1024 (2009).
101. Kirov, G. *et al.* Comparative genome hybridization suggests a role for *NRXN1* and *APBA2* in schizophrenia. *Hum. Mol. Genet.* **17**, 458–465 (2008).
102. Friedman, J. M. *et al.* Oligonucleotide microarray analysis of genomic imbalance in children with mental retardation. *Am. J. Hum. Genet.* **79**, 500–513 (2006).
103. Autism Genome Project Consortium *et al.* Mapping autism risk loci using genetic linkage and chromosomal rearrangements. *Nature Genet.* **39**, 319–328 (2007).

Acknowledgements

We thank D. Ge, E. L. Heinzen, A. C. Need, J. C. Fellay, J. M. Maia, E. K. Ruzzo and H. F. Willard for helpful comments on the manuscript.

Competing interests statement

The authors declare no competing financial interests.

DATABASES

Entrez Gene: http://www.ncbi.nlm.nih.gov/gene
CYP2D6 | *GDF5* | *IFIH1* | *ITPA* | *KCNJ11* | *SLC30A8* | *TCF7L2* | *WFS1*
OMIM: http://www.ncbi.nlm.nih.gov/omim
Freeman–Sheldon syndrome | metachondromatosis

FURTHER INFORMATION

Authors' homepage: http://humangenome.duke.edu
1000 Genomes: http://www.1000genomes.org
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
GeneTests: http://genetests.org
NHGRI Catalog of Published Genome-Wide Association Studies: http://www.genome.gov/26525384
SequenceVariantAnalyzer: http://www.svaproject.org
ssahaSNP: http://www.sanger.ac.uk/resources/software/ssahasnp
US National Institutes of Health Office of Rare Diseases Research: http://rarediseases.info.nih.gov/AboutUs.aspx#GARD

ALL LINKS ARE ACTIVE IN THE ONLINE PDF