Human genetic variation and its contribution to complex traits

Kelly A. Frazer, Sarah S. Murray, Nicholas J. Schork and Eric J. Topol

Abstract | The last few years have seen extensive efforts to catalogue human genetic variation and correlate it with phenotypic differences. Most common SNPs have now been assessed in genome-wide studies for statistical associations with many complex traits, including many important common diseases. Although these studies have provided new biological insights, only a limited amount of the heritable component of any complex trait has been identified and it remains a challenge to elucidate the functional link between associated variants and phenotypic traits. Technological advances, such as the ability to detect rare and structural variants, and a clear understanding of the challenges in linking different types of variation with phenotype, will be essential for future progress.

Structural variants
Broadly defined, these are all variants that are not single nucleotide variants. They include insertion–deletions, block substitutions, inversions of DNA sequences and copy number differences.

Genome-wide association (GWA) study
An investigation of the association between common genetic variation and disease. This type of analysis requires a dense set of markers (for example, SNPs) that capture a substantial proportion of common variation across the genome, and large numbers of study subjects.

Elucidating the inherited basis of genetic variation in human health and disease is one of the major scientific challenges of the twenty-first century. In 2001 two reference versions of the human genome were published. One was released by the Human Genome Sequencing Consortium and reflected the assembly of sequences derived from numerous donors[1], whereas the other, released by Celera Genomics, was a consensus sequence derived from five individuals[2]. Importantly, both versions represented the human genome as a haploid sequence and genetic variation was not annotated. In order to study how genetic variants contribute to phenotypic diversity, large-scale studies were initiated to identify and catalogue nucleotides that differ among individuals. Initial studies focused largely on understanding the range of patterns and frequencies of SNPs[3–5]. As the prevalence and contribution of structural variants to human biology was realized[6,7], consortia were formed and systematic studies were conducted to improve our understanding of this class of variants[8–10].

In 2007, the first complete genome sequence of an individual, J. Craig Venter[11], was published, followed shortly thereafter by the publication of a second individual's genome, that of James D. Watson[12]. Subsequently, two additional genomes from anonymous individuals were sequenced: one Han Chinese (Asian)[13] and one Nigerian (African)[14]. In aggregate, these studies — published after the release of the human genome reference sequence — have rapidly increased our knowledge of the various forms of human genetic variation, their evolutionary histories and the correlations between them. However, our understanding of the locations and frequencies of structural variants across the genome is still limited, and cataloguing these classes of alterations is a high priority.

Genome-wide association (GWA) studies are the most widely used contemporary approach to relate genetic variation to phenotypic diversity[15]. Over the past 2 years these studies have identified statistical association between hundreds of loci across the genome and common complex traits. The results of these studies have substantially increased our understanding of the diverse molecular pathways underlying specific human diseases. However, GWA studies have several limitations. First, there is great difficulty moving beyond mere statistical associations to identifying the functional basis of the link between a genomic interval and a given complex trait. Second, SNP associations identified in one population frequently are not transferable to members of other populations. Third, the bulk of the heritable fraction of complex traits has not been accounted for in recent GWA studies. This last point is probably explained by the fact that GWA studies do not capture information about rare variants and have limited statistical power to detect small gene–gene and gene–environment interactions.

The use of new technologies for assaying DNA sequences has provided important insights and raised new questions about the roles that different types of genetic variants have in human health and disease. Here, for each type of genetic variant we discuss their probable contribution to overall genetic variation, the approaches taken to assess their contribution to phenotypic variation and the successes achieved so far. There have been several excellent reviews on structural variation[16,17] as well as reviews describing the findings of GWA studies[15,18–20].

Scripps Genomic Medicine, Scripps Translational Science Institute and The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA.
Correspondence to K.A.F.
e-mail: kfrazer@scripps.edu
doi:10.1038/nrg2554

© 2009 Macmillan Publishers Limited. All rights reserved

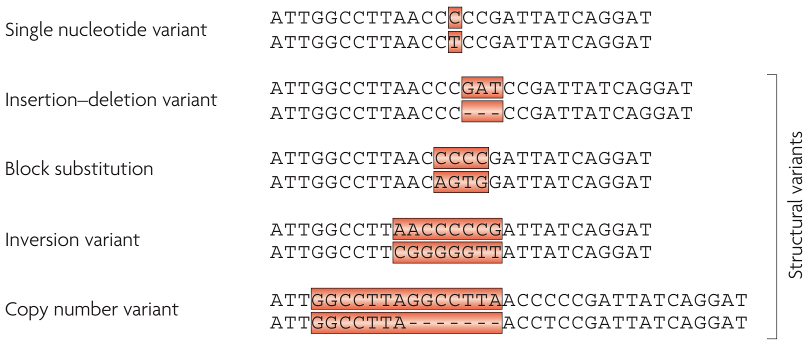

Single nucleotide variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACCTCCGATTATCAGGAT
Insertion–deletion variant
ATTGGCCTTAACCCGATCCGATTATCAGGAT
ATTGGCCTTAACCC---CCGATTATCAGGAT
Block substitution
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTAACAGTGGATTATCAGGAT
Inversion variant
ATTGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTCGGGGGTTATTATCAGGAT
Copy number variant
ATTGGCCTTAGGCCTTAACCCCCGATTATCAGGAT
ATTGGCCTTA-------ACCTCCGATTATCAGGAT
Structural variants

Figure 1 | **Classes of human genetic variants.** The nomenclature used to describe the various types of structural variants is not yet standard[121]. Here, the terminology used aims to describe the nucleotide composition of the variant and distinguish it from other types of variants. Single nucleotide variants are DNA sequence variations in which a single nucleotide (A, T, G or C) is altered. Insertion–deletion variants (indels) occur when one or more base pairs are present in some genomes but absent in others. They are generally composed of only a few bases but can be greater than 80 kb in length[11]. Block substitutions describe cases in which a string of adjacent nucleotides varies between two genomes. An inversion variant is one in which the order of the base pairs is reversed in a defined section of a chromosome. A well-characterized inversion variant that has been described in humans involves a section of chromosome 17 in which a ~900 kb interval is in the reverse order in approximately 20% of individuals with Northern European ancestry[122]. Copy number variants occur when identical or nearly identical sequences are repeated in some chromosomes but not others. The largest copy number variant identified in the Venter genome[11] was almost 2 Mb in length.

Here we unify the exciting discoveries of these two disciplines into a single Review to provide a comprehensive overview of our current knowledge of human genetic variation and where the key challenges lie for future research aimed at understanding the genetic architecture of complex traits.

Classes of human genetic variation

Human genetic variants are typically referred to as either common or rare, to denote the frequency of the minor allele in the human population. Common variants are synonymous with polymorphisms, defined as genetic variants with a minor allele frequency (MAF) of at least one percent in the population, whereas rare variants have a MAF of less than 1%. Genetic variants are also discussed in terms of their nucleotide composition. In the broadest sense, variants in the human genome can be divided into two different nucleotide composition classes: single nucleotide variants and structural variants[10] (FIG. 1). The vast majority of genetic variants are hypothesized to be neutral[21] (that is, they do not contribute to phenotypic variation), achieving significant frequencies in the human population simply by chance. However, the relative percentage of neutral, near-neutral[22] and non-neutral variants remains to be empirically determined.

Single nucleotide variants. SNPs are the most prevalent class of genetic variation among individuals. On the basis of survey sequencing results it has been estimated that the human genome contains at least 11 million SNPs, with ~7 million of these occurring with a MAF of over 5%[23] and the remaining having MAFs between 1 and 5%. Analysis of the four fully sequenced individual genomes suggests that these original estimates are fairly accurate and that most SNPs have been identified and information about them deposited in the Single Nucleotide Polymorphism database (dbSNP) (BOX 1). In addition to SNPs there are innumerable rare and novel or '*de novo*' single nucleotide variants, in some cases segregating only in a nuclear family or a single individual. For instance, any base pair that, when altered, is compatible with life is likely to be found in at least one of the ~6.7 billion people on Earth. However, it is important to note that in any given individual the majority of variants are those that are common in the population as a whole (BOX 1). Furthermore, when the genomes of two individuals are compared, the majority of the base pairs that differ are at positions with variants that are common in the population.

The alleles of SNPs located in the same genomic interval are often correlated with one another. This correlation structure, or linkage disequilibrium (LD)[24], varies in a complex and unpredictable manner across the genome and between different populations. The efforts of Phase I of the InternationalHapMapProject[3], along with those of Perlegen Sciences[5], paved the way for breaking the genome down into groups of highly correlated SNPs that are generally inherited together (known as LD bins). From Phase II of the International HapMap Project[4] it was determined that the vast majority of SNPs with a MAF of at least 5% could be reduced to ~550,000 LD bins for individuals of European or Asian ancestry and to 1,100,000 LD bins for individuals of African ancestry ($r^2 \geq 0.8$). By genotyping the DNA sample of an individual with a 'tagging' SNP from each LD bin, knowledge regarding over 80% of SNPs present at a frequency above 5% across the genome is gained[25–28].

Structural variants. Structural variation, broadly defined, refers to all base pairs that differ between individuals and that are not single nucleotide variants. Such variation includes insertion–deletions (indels), block substitutions, inversions of DNA sequences and copy number differences (FIG. 1). Compared with single nucleotide variants, the technological ability to detect structural variants in the human genome has only recently emerged[8,10,29–32]. Hence our understanding of the locations and frequencies of structural variants, and our ability to assay their association with complex traits, is still maturing[33–38]. Analysis of the four fully sequenced human genomes (BOX 1) combined with targeted sequencing of structural variants greater than 8 kb in length in eight human genomes[9] has provided tremendous insight. These studies suggest that structural variation accounts for at least 20% of all genetic variants in humans and underlies greater than 70% of the variant bases. Altogether, for any given individual, structural variants constitute between 9 and 25 Mb of the genome (~0.5 to 1%), underscoring the important roles of this class of variation in genome evolution and in human health and disease.

LD patterns of common structural variants

There has been conflicting initial evidence regarding whether the alleles of structural polymorphisms are in LD with SNPs, and are therefore assayed by proxy

Complex traits
Continuously distributed phenotypes that are classically believed to result from the independent action of many genes, environmental factors and gene-by-environment interactions.

Minor allele
The less common allele of a polymorphism.

Linkage disequilibrium
(LD). In population genetics, LD is the nonrandom association of alleles. For example, alleles of SNPs that reside near one another on a chromosome often occur in nonrandom combinations owing to infrequent recombination.

© 2009 Macmillan Publishers Limited. All rights reserved

Box 1 | **Sequenced genomes provide insights into genetic variation**

Single nucleotide variants in four human genomes		
	(n)	**In dbSNP (%)**
J. Craig Venter's genome	3,213,401	91.0
James D. Watson's genome	3,322,093	81.7
Asian genome	3,074,097	86.4
Yoruban genome	4,139,196	73.6
Structural variants in the Venter genome		
	(n)	**length (bp)**
Block substitutions	53,823	2–206
Indels (heterozygous)	851,575	1–82,711
Inversions	90	7–670,345
Copy number variants	62	8,855–1,925,949

So far, four individual genomes have been fully sequenced. Two of these have been from Caucasian individuals[11,12] (J. Craig Venter and James D. Watson) — both are well-known scientists. The other two have been from anonymous individuals, one Han Chinese (Asian)[13] and one Nigerian (African)[14]. The genome of J. Craig Venter was sequenced using Sanger dideoxy technology, whereas the other three genomes were generated using newer DNA sequencing technologies that are characterized by shorter read lengths. Although single nucleotide variants are accurately detected in all four genomes the longer reads generated for the Venter genome allowed for better assembly and more accurate detection of structural variants (FIG. 1).

The two Caucasian genomes have roughly similar numbers of single nucleotide variants (~3.3 million) with the vast majority of these sites previously identified as variants in the Single Nucleotide Polymorphism database (dbSNP; see the table). There are fewer novel single nucleotide variants in J. Craig Venter's genome owing to that fact that his genome was partially represented in the Celera human genome assembly[2] and variants in that assembly were subsequently mined and deposited into dbSNP[117]. The Asian genome has slightly fewer single nucleotide variants than the Caucasian genomes but approximately similar fractions are novel variants. The Yoruban genome has ~1.25-fold more single-base variants than the Caucasian genomes and a greater percentage is novel, which is reflective of the overall increased amount of diversity in genomes of individuals with African origins (BOX 2). The fact that in all four genomes the majority of single nucleotide variants are present in dbSNP suggests that most human high-frequency SNPs (minor allele frequency ≥ 10%) have been discovered.

Looking at the number of single nucleotide variants that are shared between the three 'out of Africa' genomes, ~1.2 million (67%) are shared by all three, ~1.7 million (52%) are shared between any set of two genomes, and each has ~1.0 million (30%) that are unique to their own genome[13]. Overall, ~5.2 million single nucleotide variants were identified in the three genomes, the majority being present in dbSNP. As additional genomes are sequenced the number of SNPs present in humans will become more apparent, but at this time previous estimates of ~11 million are reasonable[23]. Interestingly, these data indicate that most single nucleotide variants present in an individual are common rather than rare. The corollary to this is that when two human genomes are compared the majority of the bases that differ will be due to common variants.

On the basis of the Venter genome, Caucasians contain ~4.1 million genetic variants, of which ~22% are structural variants that account for 74% of all variant bases (see the table). This is likely to be an underestimate of the true contribution of structural variants to genetic diversity between individuals. In the Venter and Watson genomes, 10 to 30 Mb of novel sequences that are not present in the reference genome assembly were generated. Furthermore, in the Asian and African genomes that were sequenced, more than half of the structural variants identified were not present in the reference genome. Interestingly, a study of ~1,300 structural variants in the 270 HapMap individuals showed that when two genomes are compared 92% of the bases that vary are accounted for by common structural variants[41]. In total these data provide two important insights into structural variants: the majority of common structural variants are yet to be discovered; and common structural variants constitute the vast majority of base pairs that differ between any two individuals.

using tagging SNPs. Several studies have demonstrated that common short indels (1–5 bp)[33,39,40], as well as larger common structural polymorphisms in unique regions[8,41] of the genome, are in LD with tagging SNPs. Except for a potential skew towards lower MAFs, structural variants seem to behave similarly to SNPs in terms of both genomic and population distribution, indicating a similar evolutionary history: both types of variants are 'ancestral', having arisen once in human history and shared among individuals by descent rather than occurring as the result of recurrent mutations[39,41,42].

The evolutionary history and LD pattern of structural polymorphisms in segmental duplications has been more difficult to determine. Segmental duplications are composed of repeated sequences over 5 kb in length with >90% sequence identity[43]. Structural polymorphisms are highly enriched in regions of the genome that have recently undergone duplication. Indeed, 25 to 50% of all nucleotides in large structural variants map in segmental duplications, which constitute only 5.3% of genomic sequences[29]. This strong relationship between structural variants and segmental duplications is reflective of their similar natures. Recent data suggest that up to 25% of the intervals annotated as segmental duplications in the reference human genome sequence actually represent copy number variants between individuals rather than fixed duplication events[11]. Structural polymorphisms in segmental duplications exhibit low LD with tagging SNPs[44]. Recent studies indicate that this observed lower LD is the result of a paucity of validated SNPs[41] that can potentially serve as tags in segmental duplications compared with the rest of the genome. These studies indicate that common structural variants in segmental duplications share a similar evolutionary history with those in unique regions of the genome and are in LD with neighbouring SNPs.

© 2009 Macmillan Publishers Limited. All rights reserved

Contribution of variants to phenotypes

In humans, hundreds of complex phenotypic traits determine how we look and behave, and our propensity to develop certain diseases. Each complex phenotype is governed by a combination of inherited factors, which are largely believed to be genetic variants, and environmental influences. Full sequencing of human genomes has shown that in any given individual there are, on average, ~4 million genetic variants encompassing ~12 Mb of sequence (BOX 1). The challenge is to determine which of these variants underlies or is responsible for the inherited components of phenotypes. Over the last decade or so the human genetics field has debated[45,46] the common disease–common variant hypothesis, which posits that common complex traits are largely due to common variants with small to modest effect sizes[47–49]. The opposing theory, the rare variant hypothesis, posits that common complex traits are the summation of low-frequency, high-penetrance variants[50,51]. Overall the field is making earnest attempts to determine the relative importance of common and rare variants in common complex phenotypic traits.

GWA studies

Linking common genetic variants to common complex traits. Concurrent with the efforts of the scientific community to dissect the human genome into LD blocks were extraordinary technological advances in assaying SNPs. From 1997 to 2007, technological advances moved the field from testing one SNP at a time to the assessment of a million SNPs per individual. These two fronts of progress — one on the empirical determination of the LD structure of SNPs across the genome[3–5] and the other a new-found capacity to perform ultra-high-throughput genotyping — set the foundation for a veritable avalanche of discoveries of common genetic variants associated with various common traits and diseases through GWA studies.

There are several excellent reviews of GWA study designs and analysis that discuss selection of cases and controls, and statistical analyses — including dealing with population stratification and replication[15,52,53]. As these topics have been previously covered we do not discuss them here. Rather, we focus on the important insights about the genetic basis of human complex traits gained from GWA studies, as we believe these findings have immediate relevance to basic scientists as well as the medical research community. We also discuss the limitations of the GWA approach in identifying genetic variants underlying complex traits, providing insights into the key lines of experimentation for the future.

GWA studies published to date have used various commercial genotyping platforms containing approximately 300,000 to 500,000 common SNPs to detect differences in allele frequencies between cases and controls[25–28]. Such studies are hypothesis-free, as there is no bias or presumptive list of candidate genes that are being tested[54]. However, the term 'genome-wide' is a misnomer, because approximately 20% of common SNPs are only partially tagged or not tagged at all, and rare variants are generally not tagged. For over 80 phenotypes — including diseases and biological measurements — GWA studies have provided remarkably compelling statistical associations for a total of over 300 different loci in the human genome[55]. The results have been reported on almost a weekly basis from April 2007, with over 220 studies reported to date. Almost all disease categories have been addressed, including cardiovascular, neurodegenerative, neuropsychiatric, metabolic, autoimmune and musculoskeletal diseases, and several types of cancer.

Enhanced understanding of human diseases. The most impressive outcome of this knowledge base, which connects genomic intervals with complex traits, is a new understanding of the molecular underpinnings and pathways of many diseases[56]. Notably, most of the genes or genomic loci that have been identified through GWA studies have not previously been known to be related to the complex trait under investigation. For a substantial number of common diseases the newly identified pathways suggest that molecular subphenotypes may exist; that is, although a number of different pathways might potentially be involved in the development of a particular disease when all cases are considered, in any individual with the disease only one or a subset of these pathways might be involved. For example, the genetic propensity to develop type 2 diabetes (T2D) seems to involve genes in several different pathways that affect pancreatic β-cell formation and function, as well as pathways affecting fasting glucose levels and obesity[57–59] (FIG. 2). Likewise, many of the loci associated with multiple sclerosis involve immune function — including the interleukin receptor genes *IL2RA* and *IL7RA*, and the *HLA-DRA* locus — but a gene encoding a protein involved in axonal function, kinesin family member 1B (*KIF1B*), is also associated with this disease[60,61]. Clinicians previously considered these conditions as simple phenotypes, with all patients with the diagnosis having the same underlying biological disorder.

Surprisingly, there have been several instances in which one genomic interval has been associated with two or more seemingly distinct diseases. This convergence of genes associated with multiple diseases has led to the concept of the 'diseaseome'[62], which maps a network of how different genes and pathways connect to various diseases (FIG. 3). Examples include different interleukin receptor genes that are associated with Crohn's disease, multiple sclerosis, systemic lupus erythematosus and rheumatoid arthritis[56,63]. Such diseases had already been thought of as sharing a common immune-mediated aetiology, but now there is discrete evidence for a common genetic underpinning. Another example is the common SNP on chromosome 9p21 that is associated with three vascular phenotypes — myocardial infarction[64–66], abdominal aortic aneurysm and intracranial aneurysm[67]. Such conditions would not previously have been thought to have a common pathogenic thread. The recent exceptional advances in associating genes with many diseases have led some to suggest that the textbooks of medicine need to rewritten to account for our enhanced understanding of the interconnectivity of the molecular basis underlying distinct diseases.

Population stratification
Subdivision of a population into different ethnic groups with potentially different marker allele frequencies and different disease prevalences.

© 2009 Macmillan Publishers Limited. All rights reserved

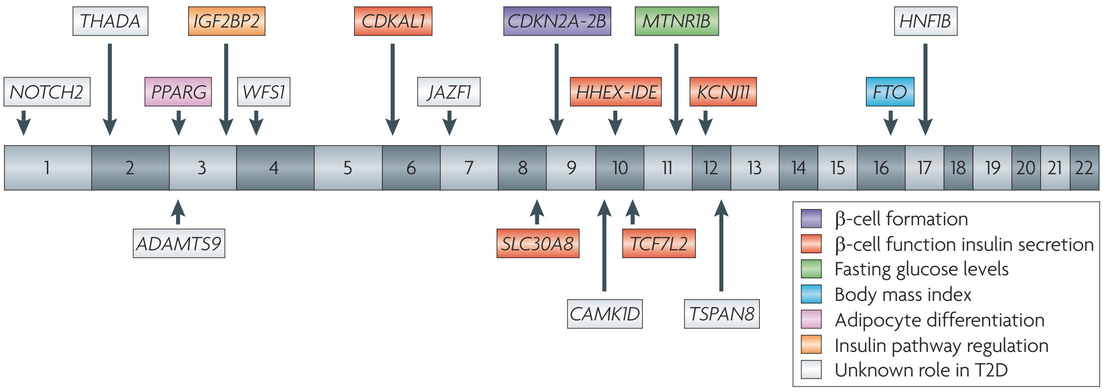


Figure 2 | **Insights into the genetic basis of type 2 diabetes (T2D).** Genome-wide association (GWA) studies have identified 18 genomic intervals that confer increased risk to T2D in Caucasians[58,59,72–75,123–127]. Four of these contain previously known candidate genes, based on the involvement of rare mutations in monogenic forms of diabetes. However, the remaining 14 intervals contain genes that were previously unsuspected in playing a part in the genetic basis of T2D. Traditional risk factors for T2D include obesity (defined as increased body mass index), elevated fasting glucose levels and impaired β-cell function, which results in reduced insulin secretion[57]. GWA studies have revealed new loci that are associated with these phenotypes in T2D cases. For example, the association of melatonin receptor 1B (*MTNR1B*) shows the involvement of the circadian rhythm pathway in fasting glucose levels and T2D[58,59,128]. The functions of the genes suspected of playing a part in β-cell dysfunction are diverse, with functions that include pancreatic islet proliferation, insulin secretion and cell signalling. Six additional genes contain variants that are statically associated with T2D, but their role in the disorder has not yet been elucidated. The functional diversity of T2D genes and the multitude of pathways in which they are members were not imagined before the results of the GWA studies. Note that although insulin-like growth factor 2 mRNA binding protein 2 (IGF2BP2) is known to regulate insulin signalling through its binding to insulin-like growth factor 2 (IGF2), there is no data indicating its role in diabetes. *ADAMTS9*, ADAM metallopeptidase with thrombospondin type 1 motif, 9; *CAMK1D*, calcium/calmodulin-dependent protein kinase ID; *CDKAL1*, CDK5 regulatory subunit associated protein 1-like 1; *CDKN*, cyclin-dependent kinase inhibitor; *FTO*, fat mass and obesity associated; *HHEX*, hematopoietically expressed homeobox; *HNF1B*, HNF1 homeobox B (also known as *TCF2*); *IDE*, insulin-degrading enzyme; *JAZF1*, JAZF zinc finger 1; *KCNJ11*, potassium inwardly rectifying channel, subfamily J, member 11; *NOTCH2*, Notch homologue 2; *PPARG*, peroxisome proliferator-activated receptor gamma; *SLC30A8*, carrier family 30 (zinc transporter), member 8; *TCF7L2*, transcription factor 7-like 2 (T-cell specific, HMG-box); *THADA*, thyroid adenoma associated; *TSPAN8*, tetraspanin 8; *WFS1*, Wolfram syndrome 1.

Limitations of GWA studies in identifying causative variants. Despite this exceptional progress, there are substantial limitations to the GWA study approach. Although statistically compelling associations have been identified, there is an enormous gap in the ability to provide the biological explanation for why a genomic interval tracks with a complex trait. For the most part, all we know is that a tag SNP for an LD bin is statistically associated with a trait, but we have no idea of the precise variants in the bin that have a causal role in contributing to variation in the trait. It is important to emphasize that tag SNPs are in LD not only with other SNPs but also with common structural variants, the majority of which have not yet been identified. The best way to move from a statistical association to knowledge of the causative variant is unclear. In most cases it will be straightforward to identify causative variants that are in LD with a tagging SNP and that are located in exons that truncate or otherwise alter the gene product. However, the causative variants underlying GWA study associations are likely to be regulatory rather than coding. For instance, many of the associations so far are not even localized to intervals that include a gene. For example, the variant at 9p21 that associates with myocardial infarction is 150 kb from the nearest gene[64–66], and for the variants on 8q24 that are associated with susceptibility to multiple solid tumours this distance is 300 kb[68,69]. Experiments are being conducted that simultaneously assay global gene expression and genome-wide variation in a large number of individuals to map genetic factors underlying differences in expression levels[70]. These data sets may be valuable tools for identifying the causative variants and biological bases for many loci associated with a complex trait through GWA studies.

Transferring GWA study results to other populations. With rare exceptions, the GWA studies carried out so far have focused on populations of European ancestry for the primary, high-throughput genotyping and have only interrogated other ancestries using limited replication genotyping. Unless a particular functional variant has been unambiguously identified, testing a tag SNP that is associated with a disease or trait in one population for risk assessment in an individual from another population can be problematic. The problem stems from both allele frequency differences between populations[71] and the fact that LD patterns across loci that mark or co-segregate with a putative causally associated genetic variant may be different from population to population.

For instance, the tagging SNP rs10757278, which is found on chromosome 9p21 and is associated with myocardial infarction in Caucasians[64–66], is in strong LD with multiple SNPs in this population (BOX 2); however, in

© 2009 Macmillan Publishers Limited. All rights reserved

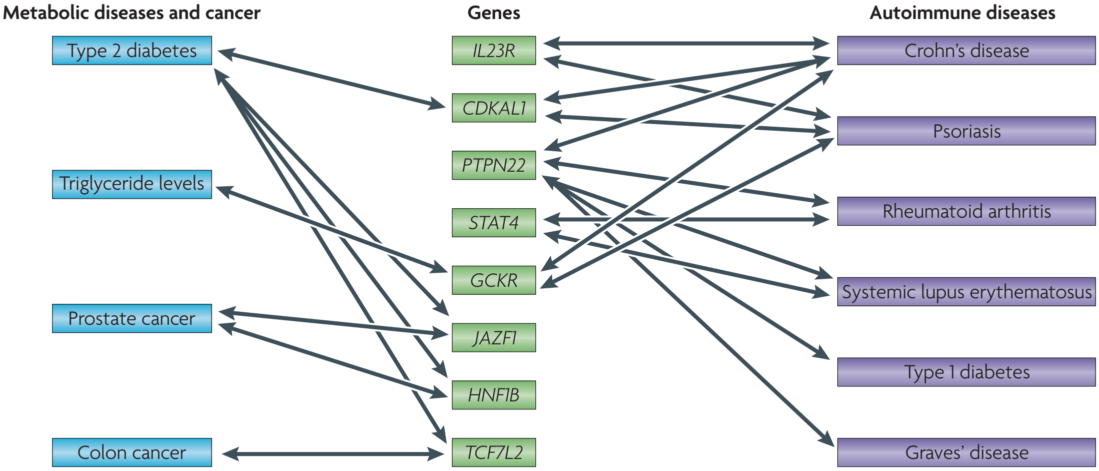


Figure 3 | **Overlap of genetic risk factor loci for common diseases.** A surprising finding of genome-wide association (GWA) studies is that over 15 loci are associated with the risk of developing two or more diseases; eight are shown here for illustrative purposes. Some alleles may be protective for one disease but confer susceptibility to another; for example, the SNP R602W in *PTPN22* (protein tyrosine phosphatase, non-receptor type 22 (lymphoid)) protects against Crohn's disease but predisposes to several other autoimmune diseases[129,130]. For some genes, distinct risk alleles are associated with different diseases; for example, JAZF zinc finger 1 (*JAZF1*) has a role in prostate cancer and in type 2 diabetes[75,131,132]. Thus, GWA study results indicate that many diseases that were previously viewed as having distinct aetiologies probably share common molecular causes. In some cases, the 'sharing' of associated genetic variants across diseases may be expected owing to shared clinical features of these disorders; for example, among the autoimmune diseases. In other cases this sharing is more surprising; for example, the involvement of glucokinase regulatory protein (*GCKR*) in both triglyceride levels and autoimmune diseases. *CDKAL1*, CDK5 regulatory subunit associated protein 1-like 1; *HNF1B*, HNF1 homeobox B (also known as *TCF2*); *IL23R*, interleukin 23 receptor; *STAT4*, signal transducer and activator of transcription 4; *TCF7L2*, transcription factor 7-like 2 (T-cell specific, HMG-box).

Asians this SNP is in a singleton block, and in Africans it is in LD with only a subset of the same SNPs present in Caucasians. Thus, rs10757278 probably tags so far undiscovered variants differently in the three populations. By contrast, the tagging SNP rs13266634 on 8q24, which has been associated with T2D in Caucasians[72–75], is in LD with the same set of SNPs in all three populations, suggesting that it may tag so far undiscovered variants similarly in the three populations. Interestingly, the structure of the LD bin is similar but the frequency of the variants is different in the populations. Thus, although panels of markers that capture as much variation as possible across the genome have been devised to facilitate association studies in different populations[25–28,76], markers that are found to be associated with a particular trait or disease in any given population will often not be transferable for risk prediction in individuals from a different population.

GWA studies of structural variants. The fact that structural variants underlie greater than 70% of the bases that vary in humans suggests that they will play a profound part in phenotypic diversity between individuals. Thus, there is tremendous interest among many researchers to test structural variants for association with specific complex traits. It is important that association studies involving structural variants are subjected to the same standards of quality control and replication that have been developed for SNP-based studies[77].

Interestingly, recent studies that have looked for associations between rare structural variants and autism and schizophrenia have identified specific deletions involved in both of these diseases. Notable among these is the association between rare recurrent deletions and duplications of a 600 kb interval at 16p11.2 that was observed in multiple unrelated individuals with autism and was estimated to account for 1% of the cases[78–80]. Additionally, large deletions (>3 Mb) on chromosomes 22q11.2, 1q21.1 and 15q13.3, each with high estimated odd ratios (>17), show significant association with schizophrenia[81,82]. In contrast to these associations with specific structural associations, several studies have presented evidence that individuals with schizophrenia have a slightly increased (1.15-fold) overall load of large (>100 kb) structural variants in their genomes compared with control individuals[82,83]. It is currently unclear what this slight increase in rare genome-wide structural variants in schizophrenia patients compared with controls means for the aetiology of this disease. First, normal individuals harbour many large structural variants (>8 kb). Second, the current framework for understanding the inherited basis of phenotypic traits is that specific genetic loci will be associated. Further studies examining larger sample numbers will hopefully provide insights into the mechanisms underlying these associations.

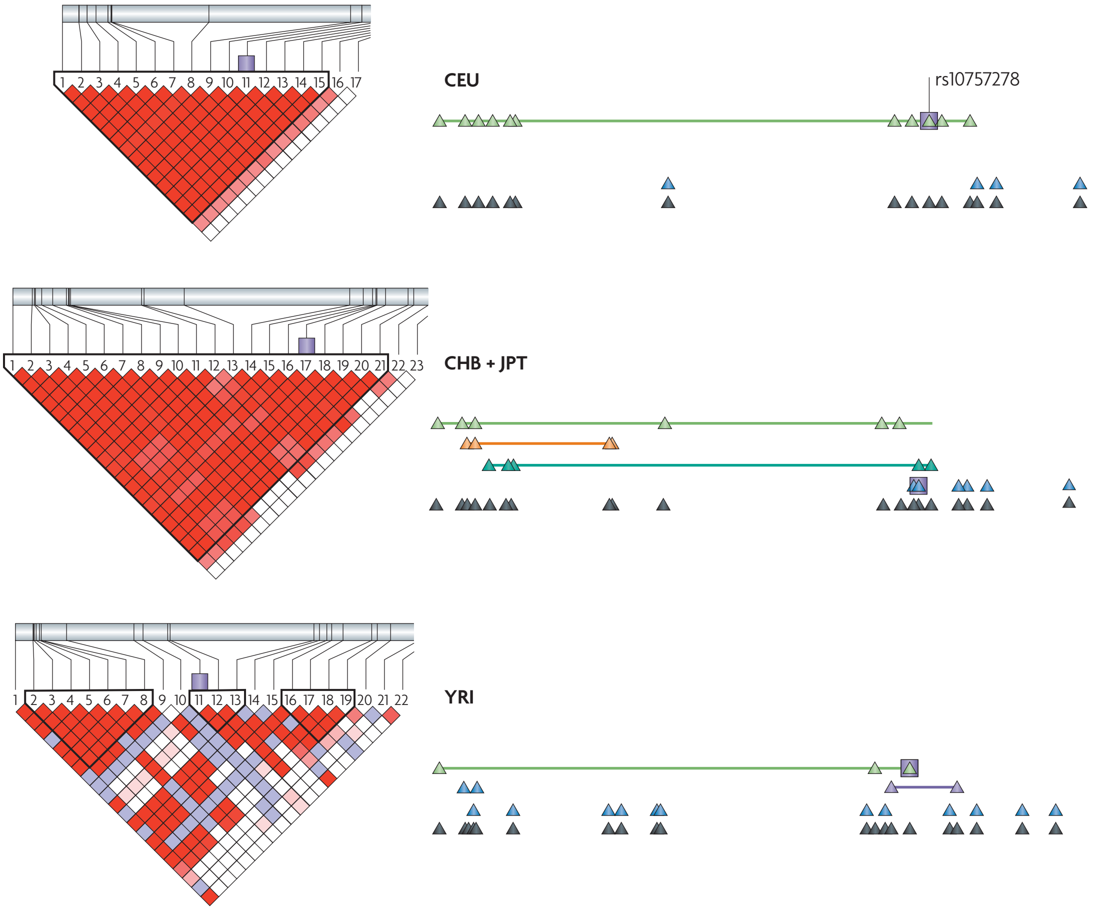
Beyond current GWA studies

An unforeseen limitation of GWA studies is that the genomic markers that are found to be associated with any given complex trait each have less impact on susceptibility than was anticipated. The small magnitude of susceptibility risk (or protection from the condition of

Odds ratio
A measurement of association that is commonly used in case–control studies. It is defined as the odds of exposure to the susceptible genetic variant in cases compared with that in controls. If the odds ratio is significantly greater than one, then the genetic variant is associated with the disease.

© 2009 Macmillan Publishers Limited. All rights reserved

Box 2 | The LD of common variants in the human genome differs between populations



The linkage disequilibrium (LD) structure of SNPs in a 13 kb interval of chromosome 9p21 is shown for the three HapMap populations: CEU (European ancestry), JPT + CHB (Asian ancestry) and YRI (African ancestry) (see the figure). There are two commonly used definitions of LD, D' and r^2, that capture different aspects of nonrandom association. On the left of the figure, the LD structures of the interval are shown quantified using D'[24]. SNPs that were ascertained in Phase II of the HapMap Project with a minor allele frequency (MAF) ≥ 5% are shown in their respective map positions. There are population differences in the numbers of SNPs that meet the ≥ 5% criterion. The pairwise correlation of SNPs — which are shown as vertical lines — is shown as red and white boxes, with red indicating high correlation ($D' \approx 1$) and white indicating no correlation ($D' \approx 0$) (other colours represent intermediate values). On the right of the figure, all SNPs are shown on the bottom row as black triangles. Above this, SNPs are grouped together into bins at an $r^2 > 0.8$ (using the ldSelect algorithm)[118]. SNPs that are efficiently tagged by each other ($r^2 > 0.8$) are shown in the same colour and are connected by a line. Singleton bins that do not tag any other SNPs are shown as individual blue triangles.

Using both r^2 and D' it is clear that LD is less strong in the African population than in the Caucasian and Asian populations. This reflects the fact that some haplotype patterns across the genome were lost in population bottlenecks associated with human migration out of Africa. Using the D' statistic fewer SNPs are correlated, as indicated by a lower number of red boxes and more white boxes. Using the r^2 statistic Africans have a greater number of singleton SNPs, and the LD bins have fewer numbers of variants and span shorter lengths. The r^2 statistic shows greater differences in pairwise correlations between SNPs in the populations, which is due to the fact that allele frequencies vary substantially between the three groups. For example, in Caucasians, SNP rs10757278 (purple box), which has been associated with myocardial infarction[64–66], lies in a bin composed of eleven SNPs; in Africans it is in LD with three of these SNPs, and in Asians it is in a singleton LD bin. For rs10757278 the MAF is the same in Caucasians and Asians (50%) but considerably less in Africans (5%).

interest) for each genomic marker needs to be emphasized (BOX 3). Most of the odds ratios for the heterozygote genotypes of the associated variants that have been identified so far are approximately 1.1, a figure that can increase to 1.5–1.6 for homozygote genotypes. Fewer than 12 common genetic variants (excluding those in the human leukocyte antigen (HLA) region) have high odds ratios between 2 and 10 for association with a particular trait[84–93]. Therefore, only a limited amount of the genetic variance underlying the heritable component of any of the ~80 complex traits that have been examined has been identified (BOX 3). Even for disease traits for which a large number of common genetic variants have been identified, only a small fraction of the inherited risk

© 2009 Macmillan Publishers Limited. All rights reserved

Epistasis
In statistical genetics, this term refers to an interaction of multiple genetic variants (usually at different loci) such that the net phenotypic effect of carrying more than one variant is different than would be predicted by simply combining the effects of each individual variant.

has been explained. For example, in the case of Crohn's disease, although over 30 associated genomic markers have been validated, these account for less than 10% of the cumulative genetic variance[87], and the 44 loci associated with height account for ~5%[94–99]. Currently, there are almost no complex traits for which there is much greater than 10% of the genetic variance explained, leaving the bulk of heritability unexplained by the common variants identified so far. Thus, we are left wondering: where is the rest of the genetic variation underlying these heritable traits, and how do we capture it?

One possibility is that the missing variation is accounted for by common genetic variants with small effect sizes that have not yet been identified. Many GWA studies have been conducted using sample sizes of 2,000 to 5,000 individuals and have sufficient statistical power to confidently identify common variants with odds ratio of 1.5 or greater[18]. Therefore, it is likely that only a few, if any, common variants with moderate to large effect sizes remain to be discovered for most complex traits investigated to date. Sample sizes of 60,000 are required to provide sufficient power to identify the majority of variants with odds ratios of 1.1 (REF. 18). This raises the possibility that when higher-powered GWA studies are ultimately performed the number of common variants with small effect sizes — but unequivocal statistical association — may substantially increase. It is important to note that GWA studies and meta-analyses of combined GWA studies have been conducted using 20,000 to 40,000 samples for lipid phenotypes (low- and high-density lipoprotein, total cholesterol and triglyceride levels)[100–102] but still only a small proportion of trait variance (5 to 10%) has been identified, leaving much of the heritability of these traits unexplained. Similarly, a meta-analysis of GWA studies for height that used an effective sample size of 27,000 (REFS 98,99) found less than 5% of the genetic variance underlying the trait. It is unknown how much more of the genetic variance underlying complex traits will be accounted for by increasing sample sizes in GWA studies. However, given the results of the studies conducted so far using large numbers of individuals it is reasonable to expect that GWA studies using 60,000 to 100,000 individuals will probably capture at most 10–15% of the genetic variance underlying any given phenotype.

Some of the missing heritability is likely to be accounted for by rare and novel variants (BOX 4). Because rare variants are not in LD with tagging SNPs the GWA studies currently being conducted are not able to capture their contribution to complex traits. The sequencing of candidate genes involved in the development of colorectal adenomas[103] and in the metabolism of lipid[104–106] and folate[107] suggests that rare variants with moderate to high penetrances contribute to the genetic components of common complex traits. To investigate the role of rare variants in complex traits, a technological advance enabling rare variants to be directly assayed is required. This breakthrough will probably come from new DNA sequencing technologies, which may, in the next 3 to 5 years, be capable of generating genomic sequences of thousands of individuals in a cost-effective way. The methods being proposed to analyse such future data sets to identify rare variants associated with complex traits are in theory straightforward (BOX 4), but they will be complex to implement as they largely rely on looking for frequency differences of functional rare variants in cases versus controls. Similar to the identification of functional common variants, the alleles of rare variants that cluster in coding sequences will be easy to find but the methods for efficiently annotating variants in other functional elements are not yet on the horizon. To develop such methods it is necessary to first identify and catalogue all functional elements in the human genome[108], and then determine which nucleotides in these elements, if altered, would have functional consequences.

Finally, there are statistical limitations of the GWA approach in identifying gene–gene and gene–environment interactions, which are likely to be profoundly important. The effects of genetic background on the impact that a particular genetic variant has on a phenotype are well documented in the model organism literature. For example, it is well known that the results of knockout, transgene, chromosome substitution and QTL mapping studies in mice are all influenced by the choice of strain[109–113]. Initial attempts to identify epistasis in GWA data have been unfruitful[114–116] and it is unclear how to proceed. Conceptual advances in our understanding of the mechanisms underlying gene–gene and gene–environment interactions are required before we can accurately model and measure their effects in complex traits in humans.

Box 3 | **How many genetic variants do we expect to find for complex traits?**

The 18 genetic variants that have been associated with type 2 diabetes (FIG. 2) have minor allele frequencies (MAFs) ranging from 0.073 to 0.50 and odds ratios (ORs) ranging from 1.05 to 1.15, except for the *TCF7L2* gene, which has an OR of 1.37. These MAFs and ORs are typical of what is observed for the genetic variants discovered in genome-wide association (GWA) studies for other diseases and complex phenotypic traits. Altogether, these 18 variants explain less than 4% of the total liability of the trait, which is only a small fraction of the estimated heritability. This implies that there are many more genes to be identified that contribute to the genetic components of the disease. Assuming that the undiscovered genetic variants have similar MAFs and ORs as those that have been identified, and estimating 40% heritability, more than 800 genetic variants are required (Y. Pawitan, personal communication). If we assume that the undiscovered genetic variants are largely rare (BOX 4) with MAFs that are ~10 times smaller than those identified to date (0.0073 to 0.05) and ORs that are ~10 times larger (1.63 to 4.05), then ~85 variants are required (Y. Pawitan, personal communication).

Summary and conclusions

Genetic architecture of complex traits. During the past few years there have been tremendous advances in our knowledge of genome-wide LD patterns of SNPs, the relative contributions of single nucleotide versus structural variants in overall genetic diversity, and the range of effect sizes for common variants. In spite of these advances we have a limited understanding of the genetic architecture of complex traits, including the number of genetic variants that influence any one trait, their allele frequencies, effect sizes and modes of interactions. The results of GWA studies over the past 2 years have cast doubt over the validity of the common disease–common variant hypothesis. This is largely because the low odds ratios of common single nucleotide variants (BOX 3), even assuming

© 2009 Macmillan Publishers Limited. All rights reserved

additive penetrances, preclude them from being responsible for the familial clustering of most complex traits. What about common structural variants? Based on our current understanding the majority of common structural variants should be in LD with SNPs and thus may have already been assayed by proxy in GWA studies[18,41]. Although a major challenge with current technologies, it is a priority to catalogue the locations and frequencies of common structural variants and empirically determine their LD patterns across the genome. Only after this is accomplished will we have a firm understanding of how well common structural variants are assayed in GWA studies based on tagging SNPs.

Another pressing issue is the importance of rare and novel variants in the familial aggregation of complex traits. On the basis of the analysis of the Venter genome, Caucasians are likely to carry 200 to 500 non-synonymous rare (MAF < 5%) and/or novel single nucleotide variants that affect protein function[117]. Gene-centric, genome-wide rare variant sequencing programmes are underway, and therefore the extent to which these variants contribute to the familial aggregation of any given complex trait should be determined in the next 3 to 4 years. However, we will not be able to fully address the role of rare variants, including non-coding and structural variants, until rapid cost-effective methods for sequencing entire genomes are available. Given that the new DNA sequencing technologies have difficulties in accurately identifying and characterizing structural variants, it is hard to predict when this will become technically feasible.

Over the next 5 to 10 years, systematic exploration of the universe of variants and epistasis, and of epigenomics, will undoubtedly provide tremendous insights into the genetic architecture of complex traits. With our current

Box 4 | **Linking rare genetic variants to common complex traits**

Much remains to be determined about the relative contribution of rare variants to common complex traits, but the findings of several studies to date provide some insights. Although the rigid definition of a rare variant is one present with a minor allele frequency (MAF) of less than 1%, the frequency boundaries used in the literature vary. Here, we define variants with MAFs between 0.1% and 3% as rare variants and MAFs of less than 0.1% as novel (for high-frequency common variants that are in linkage disequilibrium (LD) with one another the MAF is greater than 5%).

Rare variants identified in GWA studies

In general, rare variants are not in LD with common variants and therefore will not be detected in genome-wide association (GWA) studies[51]. However, when the cohorts in a GWA study have a substantial number of individuals that share distant ancestors, 10 to 20 generations ago, it is sometimes possible to identify rare, highly penetrate variants. In a recent GWA study[119] involving 809 Old Order Amish individuals, an SNP on chromosome 11q23 with a minor allele frequency (MAF) of 0.028 was associated with markedly lower fasting serum triglycerides levels, higher levels of high-density lipoprotein (HDL) cholesterol and lower levels of low-density lipoprotein (LDL) cholesterol. With much study and a keen understanding of lipid metabolism the investigators demonstrated that the associated SNP tags a loss of function variant located 823 kb away in apolipoprotein C3 (*APOC3*). Consistent with a favourable lipid profile, carriers of this variant (*APOC3* R19X) are significantly less likely than non-carriers to have coronary artery calcification. Importantly, the origin of the *APOC3* R19X variant was shown to be from a founding couple of the Lancaster Amish born in the early 1800s, with all carriers being descended from this couple. In a second GWA study[120] involving 4,763 individuals in the Northern Finland Birth Cohort 1966, a variant on chromosome X with a MAF of 0.017 was identified as associated with markedly increased LDL cholesterol levels in the 38 males that carried it. The variant is located in an intron of *AR* — a ligand-dependent transcription factor controlling circulating androgen levels, which are in part responsible for sex-specific dyslipidaemias. It is likely that this variant was present in a founding member of the Finnish population and was inherited in the 38 males by descent.

Rare variant identification and association testing in sequencing studies

The discovery and association testing of rare variants with the propensity to develop colorectal adenomas has been performed through a candidate gene sequencing study[103]. The investigators analysed 124 UK patients with multiple polyps and 483 random controls for germ line variants in five genes, three that are involved in the Wnt signalling pathway (*APC*, *AXIN1* and *CTNNB1*) and two in mismatch repair (*MLH1* and *MSH2*) by DNA sequencing. Overall, 24% of the individuals with adenoma had a rare potentially pathogenic variant in one of the five genes compared with 11.5% of the controls. The rare variants aggregated as a class and their combined frequency differences between cases and controls are significantly different with an odds ratio of 2.2. Interestingly, several of the rare variants identified in this study with odds ratios of ~2.0 were shown by examining surrounding sequences to have a common origin and thus, in effect, be founder alleles.

Rare variant characteristics

These studies demonstrate several important characteristics of rare variants. First, rare variants will often have a different population history to common variants. Common variants are ancient and are frequently present in all human populations (BOX 2), whereas rare variants are likely to be population specific[51], having originated from founder effects 10 to 20 generations ago. Second, rare variants that are associated with complex phenotypes are likely to have effect sizes larger than those of common variants. The penetrance of rare variants will vary and in some cases be high — such as *APOC3* R19X, in which all carriers have a favourable lipid profile[119] — and in other cases the penetrance will be considerably lower — such as those found underlying colorectal adenomas[103]. Finally, as technological advances allow rare variants to be directly assayed, some individual alleles will be significant when tested for association with a complex trait[119,120]. In other cases, rare variants will have to be aggregated as a class and compared between cases and controls[103] — this will always be the case for novel variants. Similar studies examining the roles of rare and novel variants in triglyceride and cholesterol serum levels overall support these conclusions[104–106].

© 2009 Macmillan Publishers Limited. All rights reserved

knowledge, what do we predict that the future will tell us? Given the low effect sizes of common variants identified through GWA studies and the fact that a surprising fraction are associated with multiple diseases, it will probably be shown that they do not account for familial concentration of phenotypic traits but rather that they modify the penetrance of casual rare variants with large effect sizes.

The impact of human genetics on medicine. The knowledge gained through human genetic studies will have a major impact on medical sciences. In the short term our increased understanding of the molecular pathways involved in disease provides new potential drug targets. In the long term the ability to predict disease susceptibility, as well as classify diseases into subphenotypes from genotypic information, will result in improved treatment and an expanded use of pharmacogenetics. The ability to stratify individuals according to genotype has the potential to make clinical trials more cost-effective and time-efficient by enrolling a much smaller number of patients with an anticipated larger treatment effect when the intervention is more precisely matched with the underlying altered biology. The majority of existing cohorts have been collected for case–control study designs and therefore can only provide a snapshot assessment of the association of a genetic variant and a particular trait. However, the natural progression of a disease cannot be adequately probed through such studies. Therefore, we call for the collection and analysis of carefully phenotyped prospective cohorts, which will be essential to develop accurate risk and disease course prediction from genotypic information.

1. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
2. Venter, J. C. *et al.* The sequence of the human genome. *Science* **291**, 1304–1351 (2001).
3. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
4. Frazer, K. A. *et al.* A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
Publication of the HapMap Phase II results genotyping over 3.1 million SNPs in 270 individuals from four geographically diverse populations.
5. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079 (2005).
6. Sebat, J. *et al.* Large-scale copy number polymorphism in the human genome. *Science* **305**, 525–528 (2004).
7. Iafrate, A. J. *et al.* Detection of large-scale variation in the human genome. *Nature Genet.* **36**, 949–951 (2004).
8. Redon, R. *et al.* Global variation in copy number in the human genome. *Nature* **444**, 444–454 (2006).
9. Kidd, J. M. *et al.* Mapping and sequencing of structural variation from eight human genomes. *Nature* **453**, 56–64 (2008).
Demonstrates the prevalence and importance of structural variation in the human genome, which historically had not been given much attention.
10. Eichler, E. E. *et al.* Completing the map of human genetic variation. *Nature* **447**, 161–165 (2007).
11. Levy, S. *et al.* The diploid genome sequence of an individual human. *PLoS Biol.* **5**, e254 (2007).
The first publication of a genome sequence of a single individual (J. Craig Venter).
12. Wheeler, D. A. *et al.* The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
The first paper to demonstrate how technological advances will enable the rapid sequencing of individual human genomes in the near future. Interestingly, the individual sequenced here is Jim Watson, who won the nobel prize for discovery of the DNA double helix.
13. Wang, J. *et al.* The diploid genome sequence of an Asian individual. *Nature* **456**, 60–65 (2008).
14. Bentley, D. R. *et al.* Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**, 53–59 (2008).
15. McCarthy, M. I. *et al.* Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nature Rev. Genet.* **9**, 356–369 (2008).
A useful review of appropriate study design, analysis, and interpretation of human GWA studies.
16. Feuk, L., Carson, A. R. & Scherer, S. W. Structural variation in the human genome. *Nature Rev. Genet.* **7**, 85–97 (2006).
17. Conrad, D. F. & Hurles, M. E. The population genetics of structural variation. *Nature Genet.* **39**, S30–S36 (2007).
18. Altshuler, D., Daly, M. J. & Lander, E. S. Genetic mapping in human disease. *Science* **322**, 881–888 (2008).
A good recent review of the results of human GWA studies. Interestingly, the authors compare sample size requirements for genetic association studies of common and rare variants.
19. Donnelly, P. Progress and challenges in genome-wide association studies in humans. *Nature* **456**, 728–731 (2008).
20. Kruglyak, L. The road to genome-wide association studies. *Nature Rev. Genet.* **9**, 314–318 (2008).
21. Kimura, M. Evolutionary rate at the molecular level. *Nature* **217**, 624–626 (1968).
22. Ohta, T. Near-neutrality in evolution of genes and gene regulation. *Proc. Natl Acad. Sci. USA* **99**, 16134–16137 (2002).
23. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
24. Slatkin, M. Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. *Nature Rev. Genet.* **9**, 477–485 (2008).
25. Barrett, J. C. & Cardon, L. R. Evaluating coverage of genome-wide association studies. *Nature Genet.* **38**, 659–662 (2006).
26. Eberle, M. A. *et al.* Power to detect risk alleles using genome-wide tag SNP panels. *PLoS Genet.* **3**, e170 (2007).
27. Pe'er, I. *et al.* Evaluating and improving power in whole-genome association studies using fixed marker sets. *Nature Genet.* **38**, 663–667 (2006).
28. Clark, A. G. & Li, J. Conjuring SNPs to detect associations. *Nature Genet.* **39**, 815–816 (2007).
29. Tuzun, E. *et al.* Fine-scale structural variation of the human genome. *Nature Genet.* **37**, 727–732 (2005).
30. Cooper, G. M., Zerr, T., Kidd, J. M., Eichler, E. E. & Nickerson, D. A. Systematic assessment of copy number variant detection via genome-wide SNP genotyping. *Nature Genet.* **40**, 1199–1203 (2008).
31. Korbel, J. O. & al, e. Paired-end mapping reveals extensive structural variation in the human genome. *Science* **318**, 420–426 (2007).
32. Khaja, R. *et al.* Genome assembly comparison identifies structural variants in the human genome. *Nature Genet.* **38**, 1413–1418 (2006).
33. Conrad, D. F., Andrews, T. D., Carter, N. P., Hurles, M. E. & Pritchard, J. K. A high-resolution survey of deletion polymorphism in the human genome. *Nature Genet.* **38**, 75–81 (2006).
34. Barnes, C. *et al.* A robust statistical method for case-control association testing with copy number variation. *Nature Genet.* **40**, 1245–1252 (2008).
35. McCarroll, S. A. & Altshuler, D. M. Copy-number variation and association studies of human disease. *Nature Genet.* **39**, S37–S42 (2007).
36. Sebat, J. Major changes in our DNA lead to major changes in our thinking. *Nature Genet.* **39**, S3–S5 (2007).
37. Korn, J. M. *et al.* Integrated genotype calling and association analysis of SNPs, common copy number polymorphisms and rare CNVs. *Nature Genet.* **40**, 1253–1260 (2008).
38. Cooper, G. M., Nickerson, D. A. & Eichler, E. E. Mutational and selective effects on copy-number variants in the human genome. *Nature Genet.* **39**, S22–S29 (2007).
39. Hinds, D. A., Kloek, A. P., Jen, M., Chen, X. & Frazer, K. A. Common deletions and SNPs are in linkage disequilibrium in the human genome. *Nature Genet.* **38**, 9–11 (2006).
40. McCarroll, S. A. *et al.* Common deletion polymorphisms in the human genome. *Nature Genet.* **38**, 86–92 (2006).
41. McCarroll, S. A. *et al.* Integrated detection and population-genetic analysis of SNPs and copy number variation. *Nature Genet.* **40**, 1166–1174 (2008).
Demonstrates that common structural variants are in LD with common SNPs in the human genome.
42. Jakobsson, M. *et al.* Genotype, haplotype and copy-number variation in worldwide human populations. *Nature* **451**, 998–1003 (2008).
43. Bailey, J. A. *et al.* Recent segmental duplications in the human genome. *Science* **297**, 1003–1007 (2002).
44. Locke, D. P. *et al.* Linkage disequilibrium and heritability of copy-number polymorphisms within duplicated regions of the human genome. *Am. J. Hum. Genet.* **79**, 275–290 (2006).
45. 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).
46. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
47. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
48. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
49. Chakravarti, A. Population genetics — making sense out of sequence. *Nature Genet.* **21**, 56–60 (1999).
50. Fearnhead, N. S., Winney, B. & Bodmer, W. F. Rare variant hypothesis for multifactorial inheritance: susceptibility to colorectal adenomas as a model. *Cell Cycle* **4**, 521–525 (2005).
51. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genet.* **40**, 695–701 (2008).
The authors discuss the concepts behind the common disease common–variant hypothesis and contrast them to the basic ideas that underlie the rare variant hypothesis.
52. Pearson, T. A. & Manolio, T. A. How to interpret a genome-wide association study. *JAMA* **299**, 1335–1344 (2008).
53. Iles, M. M. What can genome-wide association studies tell us about the genetics of common disease? *PLoS Genet.* **4**, e33 (2008).
54. Altshuler, D. & Daly, M. J. Guilt beyond a reasonable doubt. *Nature Genet.* **39**, 813–815 (2007).
55. Hindorff, L. A., Junkins, H. A., Mehta, J. P. and Manolio, T. A. A Catalog of Published Genome-Wide Association Studies. *National Human Genome Research Institute* [online], <www.genome.gov/26525384> (accessed 1 Jan 2009).
56. Xavier, R. J. & Rioux, J. D. Genome-wide association studies: a new window into immune-mediated diseases. *Nature Rev. Immunol.* **8**, 631–643 (2008).
57. Frayling, T. M. Genome-wide association studies provide new insights into type 2 diabetes aetiology. *Nature Rev. Genet.* **8**, 657–662 (2007).

© 2009 Macmillan Publishers Limited. All rights reserved

58. Lyssenko, V. *et al.* Common variant in *MTNR1B* associated with increased risk of type 2 diabetes and impaired early insulin secretion. *Nature Genet.* **41**, 82–88 (2009).
59. 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).
60. Aulchenko, Y. S. *et al.* Genetic variation in the *KIF1B* locus influences susceptibility to multiple sclerosis. *Nature Genet.* **40**, 1402–1403 (2008).
61. Hafler, D. A. *et al.* Risk alleles for multiple sclerosis identified by a genomewide study. *N. Engl. J. Med.* **357**, 851–862 (2007).
62. Goh, K. I. *et al.* The human disease network. *Proc. Natl Acad. Sci. USA* **104**, 8685–8690 (2007).
63. Lettre, G. & Rioux, J. D. Autoimmune diseases: insights from genome-wide association studies. *Hum. Mol. Genet.* **17**, R116–R121 (2008).
64. McPherson, R. *et al.* A common allele on chromosome 9 associated with coronary heart disease. *Science* **316**, 1488–1491 (2007).
65. Samani, N. J. *et al.* Genomewide association analysis of coronary artery disease. *N. Engl. J. Med.* **357**, 443–453 (2007).
66. Helgadottir, A. *et al.* A common variant on chromosome 9p21 affects the risk of myocardial infarction. *Science* **316**, 1491–1493 (2007).
67. Helgadottir, A. *et al.* The same sequence variant on 9p21 associates with myocardial infarction, abdominal aortic aneurysm and intracranial aneurysm. *Nature Genet.* **40**, 217–224 (2008).
68. Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature Genet.* **38**, 652–658 (2006).
69. Freedman, M. L. *et al.* Admixture mapping identifies 8q24 as a prostate cancer risk locus in African-American men. *Proc. Natl Acad. Sci. USA* **103**, 14068–14073 (2006).
70. Cookson, W. *et al.* Mapping complex disease traits with global gene expression. *Nature Rev. Genet.* **10**, 184–194 (2009).
71. Myles, S., Davison, D., Barrett, J., Stoneking, M. & Timpson, N. Worldwide population differentiation at disease-associated SNPs. *BMC Med. Genomics* **1**, 22 (2008).
72. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
73. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
74. 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).
75. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
76. Xing, J. *et al.* HapMap tagSNP transferability in multiple populations: general guidelines. *Genomics* **92**, 41–51 (2008).
77. Chanock, S. J. *et al.* Replicating genotype–phenotype associations. *Nature* **447**, 655–660 (2007).
78. Weiss, L. A. *et al.* Association between microdeletion and microduplication at 16p11.2 and autism. *N. Engl. J. Med.* **358**, 667–675 (2008).
79. Kumar, R. A. *et al.* Recurrent 16p11.2 microdeletions in autism. *Hum. Mol. Genet.* **17**, 628–638 (2008).
80. Marshall, C. R. *et al.* Structural variation of chromosomes in autism spectrum disorder. *Am. J. Hum. Genet.* **82**, 477–488 (2008).
81. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
82. Consortium, I. S. Rare chromosomal deletions and duplications increase risk of schizophrenia. *Nature* **455**, 237–241 (2008).
83. Walsh, T. *et al.* Rare structural variants disrupt multiple genes in neurodevelopmental pathways in schizophrenia. *Science* **320**, 539–543 (2008).
84. Richards, J. B. *et al.* Male-pattern baldness susceptibility locus at 20p11. *Nature Genet.* **40**, 1282–1284 (2008).
85. Link, E. *et al.* *SLCO1B1* variants and statin-induced myopathy — a genomewide study. *N. Engl. J. Med.* **359**, 789–799 (2008).
86. Graham, R. R. *et al.* Genetic variants near *TNFAIP3* on 6q23 are associated with systemic lupus erythematosus. *Nature Genet.* **40**, 1059–1061 (2008).
87. 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).
One of the human traits for which a large number of loci has been identified; the majority have modest effect sizes and in sum explain only a minority of the overall heritability.
88. Sulem, P. *et al.* Two newly identified genetic determinants of pigmentation in Europeans. *Nature Genet.* **40**, 835–837 (2008).
89. Sulem, P. *et al.* Genetic determinants of hair, eye and skin pigmentation in Europeans. *Nature Genet.* **39**, 1443–1452 (2007).
90. Stokowski, R. P. *et al.* A genomewide association study of skin pigmentation in a South Asian population. *Am. J. Hum. Genet.* **81**, 1119–1132 (2007).
91. Buch, S. *et al.* A genome-wide association scan identifies the hepatic cholesterol transporter ABCG8 as a susceptibility factor for human gallstone disease. *Nature Genet.* **39**, 995–999 (2007).
92. Klein, R. J. *et al.* Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
93. Thorleifsson, G. *et al.* Common sequence variants in the *LOXL1* gene confer susceptibility to exfoliation glaucoma. *Science* **317**, 1397–1400 (2007).
94. Weedon, M. N. *et al.* A common variant of *HMGA2* is associated with adult and childhood height in the general population. *Nature Genet.* **39**, 1245–1250 (2007).
95. Sanna, S. *et al.* Common variants in the *GDF5-UQCC* region are associated with variation in human height. *Nature Genet.* **40**, 198–203 (2008).
96. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40**, 575–583 (2008).
97. Lettre, G. *et al.* Identification of ten loci associated with height highlights new biological pathways in human growth. *Nature Genet.* **40**, 584–591 (2008).
98. Gudbjartsson, D. F. *et al.* Many sequence variants affecting diversity of adult human height. *Nature Genet.* **40**, 609–615 (2008).
99. Weedon, M. N. & Frayling, T. M. Reaching new heights: insights into the genetics of human stature. *Trends Genet.* **24**, 595–603 (2008).
100. Aulchenko, Y. S. *et al.* Loci influencing lipid levels and coronary heart disease risk in 16 European population cohorts. *Nature Genet.* **41**, 47–55 (2008).
101. Willer, C. J. *et al.* Newly identified loci that influence lipid concentrations and risk of coronary artery disease. *Nature Genet.* **40**, 161–169 (2008).
102. Kathiresan, S. *et al.* Common variants at 30 loci contribute to polygenic dyslipidemia. *Nature Genet.* **41**, 56–65 (2009).
103. Fearnhead, N. S. *et al.* Multiple rare variants in different genes account for multifactorial inherited susceptibility to colorectal adenomas. *Proc. Natl Acad. Sci. USA* **101**, 15992–15997 (2004).
104. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
One of the first studies to demonstrate that multiple rare alleles with high penetrance collectively contribute to a common phenotype in the general population.
105. Kotowski, I. K. *et al.* A spectrum of *PCSK9* alleles contributes to plasma levels of low-density lipoprotein cholesterol. *Am. J. Hum. Genet.* **78**, 410–422 (2006).
106. Romeo, S. *et al.* Population-based resequencing of *ANGPTL4* uncovers variations that reduce triglycerides and increase HDL. *Nature Genet.* **39**, 513–516 (2007).
107. Marini, N. J. *et al.* The prevalence of folate-remedial MTHFR enzyme variants in humans. *Proc. Natl Acad. Sci. USA* **105**, 8055–8060 (2008).
108. Birney, E. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816 (2007).
The goal of this project was to develop efficient methods for functionally annotating human genomic sequences. The work yielded new understandings of transcription regulatory sequences and their relationships with features of chromatin accessibility and histone modification.
109. Wade, C. M. & Daly, M. J. Genetic variation in laboratory mice. *Nature Genet.* **37**, 1175–1180 (2005).
110. Erickson, R. P. Mouse models of human genetic disease: which mouse is more like a man? *Bioessays* **18**, 993–998 (1996).
111. Linder, C. C. The influence of genetic background on spontaneous and genetically engineered mouse models of complex diseases. *Lab. Anim. (NY)* **30**, 34–39 (2001).
112. Frankel, W. N. Taking stock of complex trait genetics in mice. *Trends Genet.* **11**, 471–477 (1995).
113. Shao, H. *et al.* Genetic architecture of complex traits: large phenotypic effects and pervasive epistasis. *Proc. Natl Acad. Sci. USA* **105**, 19910–19914 (2008).
114. Maller, J. *et al.* Common variation in three genes, including a noncoding variant in *CFH*, strongly influences risk of age-related macular degeneration. *Nature Genet.* **38**, 1055–1059 (2006).
115. Li, M. *et al.* *CFH* haplotypes without the Y402H coding variant show strong association with susceptibility to age-related macular degeneration. *Nature Genet.* **38**, 1049–1054 (2006).
116. Rioux, J. D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nature Genet.* **39**, 596–604 (2007).
117. Ng, P. C. *et al.* Genetic variation in an individual human exome. *PLoS Genet.* **4**, e1000160 (2008).
118. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
119. Pollin, T. I. *et al.* A null mutation in human *APOC3* confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008).
120. Sabatti, C. *et al.* Genome-wide association analysis of metabolic traits in a birth cohort from a founder population. *Nature Genet.* **41**, 35–46 (2008).
121. Scherer, S. W. *et al.* Challenges and standards in integrating surveys of structural variation. *Nature Genet.* **39**, S7–S15 (2007).
122. Stefansson, H. *et al.* A common inversion under selection in Europeans. *Nature Genet.* **37**, 129–137 (2005).
123. 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).
124. Steinthorsdottir, V. *et al.* A variant in *CDKAL1* influences insulin response and risk of type 2 diabetes. *Nature Genet.* **39**, 770–775 (2007).
125. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
126. 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).
127. Gerken, T. *et al.* The obesity-associated *FTO* gene encodes a 2-oxoglutarate-dependent nucleic acid demethylase. *Science* **318**, 1469–1472 (2007).
128. Prokopenko, I. *et al.* Variants in *MTNR1B* influence fasting glucose levels. *Nature Genet.* **41**, 77–81 (2009).
129. Bayat, A., Barton, A. & Ollier, W. E. Dissection of complex genetic disease: implications for orthopaedics. *Clin. Orthop. Relat. Res.* **419**, 297–305 (2004).
130. Vang, T. *et al.* Autoimmune-associated lymphoid tyrosine phosphatase is a gain-of-function variant. *Nature Genet.* **37**, 1317–1319 (2005).
131. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nature Genet.* **39**, 977–983 (2007).
132. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nature Genet.* **40**, 310–315 (2008).

Acknowledgements

The authors are supported by a National Institutes of Health grant (NIH 1U54RR025204-01).

FURTHER INFORMATION

The Scripps Research Institute: http://www.scripps.edu/e_index.html
dbSNP: http://www.ncbi.nlm.nih.gov/projects/SNP
International HapMap Project: http://www.hapmap.org
Perlegen Sciences: http://www.perlegen.com

ALL LINKS ARE ACTIVE IN THE ONLINE PDF

© 2009 Macmillan Publishers Limited. All rights reserved