GENOME-WIDE ASSOCIATION STUDIES

Rare and common variants: twenty arguments

Greg Gibson

Abstract | Genome-wide association studies have greatly improved our understanding of the genetic basis of disease risk. The fact that they tend not to identify more than a fraction of the specific causal loci has led to divergence of opinion over whether most of the variance is hidden as numerous rare variants of large effect or as common variants of very small effect. Here I review 20 arguments for and against each of these models of the genetic basis of complex traits and conclude that both classes of effect can be readily reconciled.

Common disease–common variant hypothesis
(CDCV hypothesis). The model that complex disease is largely attributable to a moderate number of common variants, each of which explains several per cent of the risk in a population.

Heritability
The proportion of the phenotypic variance in a population that is due to genotypic differences among individuals.

Genetic variance
The contribution of genotypic differences among individuals to phenotypic variation.

Narrow sense variance
The additive component of the genetic variance: namely, the average effect of substituting one allele for another at a locus.

School of Biology and Center for Integrative Genomics, 770 State Street, Georgia Institute of Technology, Atlanta, Georgia 30332, USA.
e-mail: greg.gibson@biology.gatech.edu
doi:10.1038/nrg3118

Soon after the ascension of genome-wide association studies (GWASs) as the pre-eminent tool for discovering polymorphic genes that influence disease susceptibility and quantitative traits, the field of genetics developed three major outlooks on the architecture of complex traits. When GWASs began, the field was dominated by the simple common disease–common variant hypothesis (CDCV hypothesis)[1–4]. This model has now been refuted in light of the so-called 'missing heritability problem': the observation that loci detected by GWASs explain almost without exception a small minority of the inferred genetic variance[5,6]. It is simply not the case that a few dozen loci of moderate effect and intermediate frequency each explain several per cent of disease risk in a population, as is typically observed in crosses or pedigrees. Since then, the genetic component has been attributed instead to one of three causes: a large number of small-effect common variants across the entire allele frequency spectrum (the infinitesimal model)[7,8], a large number of large-effect rare variants (the rare allele model)[9] or some combination of genotypic, environmental and epigenetic interactions (the broad sense heritability model)[10,11]. FIGURE 1 shows the expected distribution of genome-wide association profiles under each of these three models and the CDCV model.

GWASs are neither powered nor designed to detect variation under any of these models on a consistent basis, so there is as of yet insufficient empirical data to resolve the debate. In all likelihood, each of the three genetic architectures contributes, possibly to different degrees, to different diseases or traits. However, because the heritability of risk is generally estimated as narrow sense variance, interaction effects should be considered to be secondary to the main effects of common and rare variants. The purpose of this article is therefore to review 20 arguments that support or refute the infinitesimal and rare allele models. This is not a comprehensive survey but is rather an overview of five arguments in favour and five arguments against each of these two models that have been drawn from theory and data from diverse categories of disease. BOX 1 lists each of the arguments approximately according to their relative strength.

The models

Infinitesimal model: many variants of small effect. By 'infinitesimal model', I mean the proposition that common variants are among the major source of genetic variance for disease susceptibility and continuous traits, where hundreds or thousands of loci contribute in each case. The loci detected by GWASs are merely the largest effect sizes drawn from a Poisson or similar distribution. If half a dozen common variants explain 10% of risk in the population, the remainder is attributable to a myriad of variants that each explain considerably less than 1% of risk and have a genotype relative risk (GRR) of less than 1.1 (REF. 12). FIGURE 2a shows that affected individuals will tend to carry a slight excess of risk variants, as the overall distribution of the number of risk alleles per affected individual is skewed relative to unaffected individuals. If risk alleles follow the same distribution of allele frequencies as neutral variation, then they will include a large number of rare variants as well. Ultimately, every gene contributes to every trait, but with effect sizes that are so small that it would take samples greater than the population size of the species to detect them. In practice, as shown by the massive meta-analyses of GWASs for height and body mass index (BMI)[13,14], each involving several hundred thousand people, it is unlikely that more than a few hundred loci will ever be confirmed for most diseases, and these will not necessarily explain even half

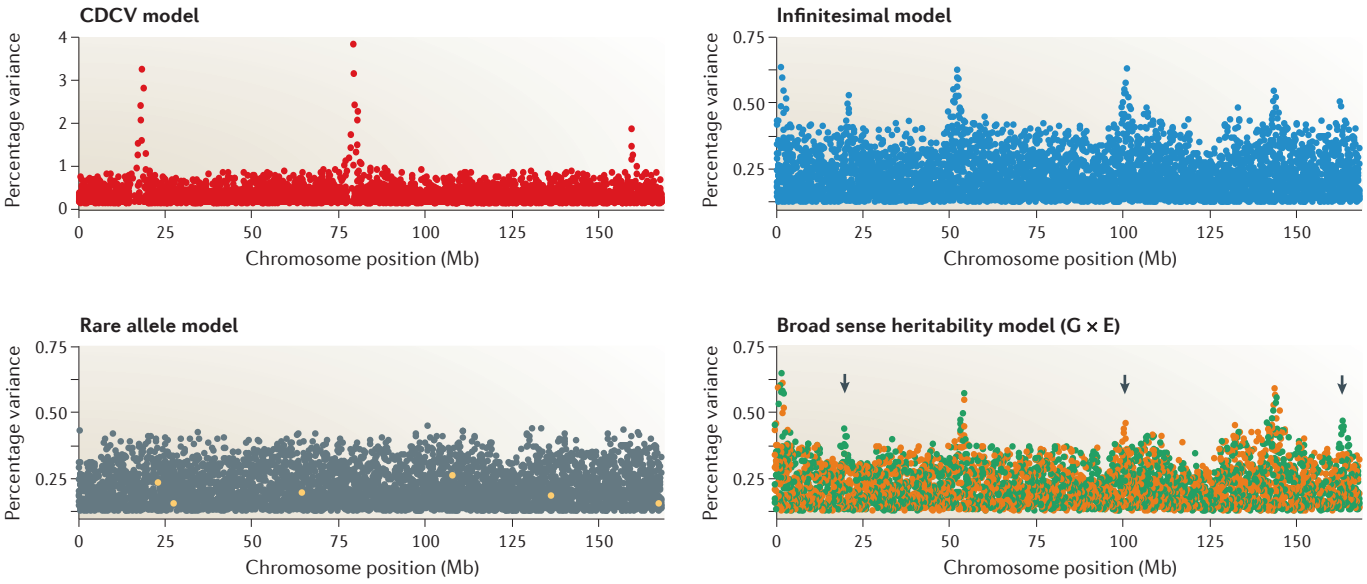

Figure 1 | **Different expected signatures from genome-wide association studies for four models of disease.** Each plot shows the approximate expected distribution of SNP effects for a modest study of 2,000 cases and controls. The *y* axis is the percentage of the variance for a trait or disease liability in the population explained by each SNP (note that standard Manhattan plots typically show the significance instead, which is represented as the negative $\log_{10}$ of the *P* value), and the *x* axis is the location of tens of thousands of SNPs along the chromosome. In the common disease–common variant (CDCV) model, a small number of moderate-effect loci would produce very strong signals, each of which explains several per cent of the genetic variance. Note the expanded scale of the *y* axis here relative to the other plots. In the rare allele model, causal variant effects (yellow dots) may be large in a few individuals but are not common enough to explain much variance or result in genome-wide significance. The infinitesimal model, by contrast, does produce some significant peaks owing to small effects of common variants, and in each case several SNPs within a linkage disequilibrium (LD) block associate with the trait. Finally, it can be argued that if associations are only seen in some environments (green and orange signals, bottom right), then in a mixed population the overall effect will be reduced at such loci (as indicated by the arrows), and fewer associations will be detected, explaining less of the variance.

Genotype relative risk
(GRR). The ratios of the risk of disease between individuals with and without the genotype. A ratio of 1.1 equates to a 10% increase in risk.

Penetrance
Describes the proportion of individuals with a mutation or risk variant who have the disease.

Expressivity
The severity of the disease in individuals who have the risk variant and disease.

Genotype-by-genotype interactions
(G × G interactions). Otherwise known as epistasis, this refers to the situation in which the effect of one genotype is conditional on genotypes at one or more other unlinked loci.

Genotype-by-environment interactions
(G × E interactions). Refers to the situation in which the effect of the genotype is conditional on the environment, which may include abiotic (temperature), biotic (viral exposure) and cultural/behavioural influences.

Parent-of-origin genetic contributions
Genetic effects that are only seen when the allele is transmitted either from the mother or from the father.

of the genetic variance. The term 'infinitesimal' borrows from the initial formulation of quantitative genetic theory by Fisher[7]. Here it simply signifies the idea that the heritability is not so much missing as it is hidden beneath the significance thresholds used to define risk alleles with high confidence[15].

Rare allele model: many rare alleles of large effect. The alternative view is that most of the variance for certain complex diseases is due to moderately highly penetrant rare variants, the allele frequency of which is typically <1%, most of which are recently derived alleles in the human population. Under this model, expressivity may be modified by other loci or by the environment[16,17], but the notion is that the rare susceptibility genotype is largely responsible for disease. The rare allele model generally refers to dominant effects owing either to haploinsufficient or gain-of-function alleles, where risk is elevated twofold or more over the background. Under these conditions, penetrance does not need to be anywhere near 100%. In fact, as shown in FIG. 2b, the vast majority of unaffected individuals are expected to carry one or more risk alleles. The notion is that a disease such as schizophrenia is actually a collection of hundreds, or possibly even thousands, of similar conditions that are attributable to rare variants at individual loci[18]. If each of these variants explains most of the risk in just a handful of people, their effects will not explain enough of the variance in a total population to be detected by standard GWAS procedures. The total number of loci that may contribute to a disease of a given prevalence is a function of the baseline disease incidence, the number and frequency of rare variants per locus and their GRRs (namely, effect size). For a disease with a high heritability, under a multiplicative model, relative risks rise steeply as the number of contributing rare alleles in an individual increases, but only a very small fraction of individuals have a sufficiently large number of alleles to ensure high sibling relative risks[19].

Broad sense heritability model: non-additive G×G and G×E interactions and epigenetic effects. The broad sense heritability model posits that additive contributions of common variants and large effects of rare variants are insufficient to explain the missing heritability. Proponents of this model point to a long history of detection of genotype-by-genotype interactions (G×G interactions; also known as epistasis) and genotype-by-environment interactions (G×E interactions) in model organism quantitative genetic research[20,21], and note the increasing number of studies documenting epigenetic effects[22], notably parent-of-origin genetic contributions[23,24] and inheritance of DNA methylation patterns[25]. The notion here is that as GWASs only measure the average effects of alleles across

Box 1 | **The twenty arguments**

Arguments for rare alleles
- Evolutionary theory predicts that disease alleles should be rare.
- Empirical population genetic data show that deleterious variants are rare.
- Rare copy number variants contribute to several complex psychological disorders.
- Many rare familial disorders are due to rare alleles of large effect.
- Synthetic associations may explain common variant effects.

Arguments against rare alleles
- Simulation of the allele frequency distribution of data from genome-wide association studies (GWASs) is not consistent with rare variant explanations.
- Genome-wide associations are consistent across populations.
- Sibling recurrence rates are greater than the postulated effect sizes of rare variants.
- Epidemiological transitions cannot be attributed to rare variants.
- Rare variants do not have obviously additive effects.

Arguments for common alleles
- GWASs have successfully identified thousands of common variants.
- Model organism research supports common variant contributions to complex phenotypes.
- Variation in endophenotypes is almost certainly due to common variants.
- The infinitesimal model is standard quantitative genetic theory.
- Common variants collectively capture most of the genetic variance in GWASs.

Arguments against common alleles
- The missing heritability has not been accounted for.
- Demographic phenomena suggest more than a simple common variant model.
- The quantitative trait locus (QTL) paradox: QTLs that are consistently detected in pedigrees and in experimental crosses are not observed in outbred populations.
- Absence of blending inheritance.
- Very few common variants for disease have been functionally validated.

thousands of individuals, they are incapable of capturing heterogeneity of effect sizes at the family level that would be the hallmark of these broader components of the genetic architecture. Although broad sense heritability is not considered any further in this article (but see BOX 2), I do not disregard its potential contribution. Rather, my purpose here is to contrast the two narrow sense models, as resolution of their contributions lays the foundation for consideration of other genetic mechanisms.

Arguments in favour of the rare allele model

Evolutionary theory predicts that disease alleles should be rare. Perhaps the strongest argument for the rare allele model comes from evolutionary theory. As disease is deleterious to fitness, variants that promote disease should be selected against; disease-promoting variants should therefore not be common[3,26,27]. The existence of disease-promoting variants reflects the balance between mutation creating new susceptibility variants and selection preventing them from drifting to a higher frequency in the population[2,28]. Mutation rates are sufficiently large that purifying selection cannot remove all deleterious variants, and those variants that have a modest effect on fitness (for example, if they influence late-onset diseases[29]) may rise to allele frequencies of 1% or occasionally more, particularly if the effect is recessive. But selection is sufficiently efficient that even a fitness reduction of a fraction of a per cent will keep allele frequencies from reaching common levels. The argument has been made that relaxed selection in modern humans may favour the accumulation of deleterious rare alleles, greatly increasing the prevalence of disease hundreds of generations into the future[30].

Empirical population genetic data shows that deleterious variants are rare. It has been appreciated for some time that the distribution of minor allele frequencies (MAFs) is strongly skewed towards an excess of rare variants: over one-third of all polymorphisms have frequencies below 5%[31]. Multiple factors contribute to this skewed distribution, but the finding from whole-exome sequence data that nonsynonymous substitutions are even more significantly skewed towards low frequencies almost certainly reflects the operation of purifying selection[32–34]. As a class, amino acid substitutions appear to be deleterious. It does not need to follow from the observations above that the reduction in fitness is due to promotion of chronic disease or that all rare nonsynonymous variants are deleterious. However, these findings are consistent with the theory that selection keeps fitness-reducing alleles at a large proportion of genes at low frequency[35]. It remains to be seen whether the same is true of regulatory polymorphisms[36,37], because (despite the considerable technological achievements of the ENCODE project) we still lack efficient procedures for identifying enhancers and other regulatory regions that polymorphisms could disrupt[38].

Many rare familial disorders are due to rare alleles of large effect. This statement does not apply solely to conditions that are caused by rare, high-penetrance Mendelian mutations, such as cystic fibrosis and muscular dystrophy. There are numerous chronic conditions that have familial analogues: well-known examples include rare variants promoting atherosclerosis through hypercholesterolemia[39], the lesions that are responsible for maturity onset diabetes of the young (MODY)[40] and the *BRCA1* and *BRCA2* breast cancer susceptibility mutations[41]. In fact, perusal of the Online Mendelian Inheritance in Man (OMIM) database provides examples of near-Mendelian cases of many common disorders[42]. Probably the most comprehensive survey of this model is the demonstration that one-quarter of the cases of X-chromosome-linked intellectual disability can be ascribed to rare protein-coding mutations, which were discovered by comprehensive sequencing of X-chromosomal exons[43]. There is thus extensive precedent for rare variants contributing substantially to special cases of complex disease, including to risk of infection[44].

Rare copy number variants contribute to several complex psychological disorders. Copy number variants (CNVs) are either hemizygous deletions or local duplications that result in three or even four copies of a locus[45]. Five per cent of cases of schizophrenia and of autism have each been attributed to CNVs at fewer than half a dozen genomic locations[46–48]. These effects are less highly penetrant than Mendelian mutations, implying modification by the genetic background. There is no evidence to support

Purifying selection
Selection against genetic variants that reduce fitness. Purifying selection generally keeps deleterious alleles at a low frequency or removes them from the population.

Chronic disease
Medical conditions that develop slowly and persist, generally with a strong genetic component.

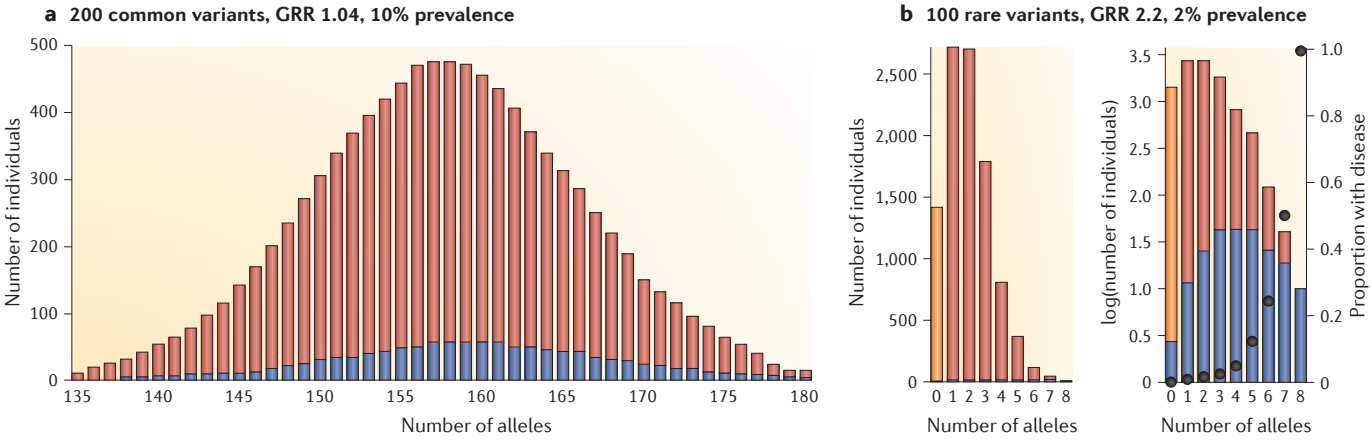

Figure 2 | **Expected distribution of risk variants.** The approximate frequency distribution of risk alleles in cases (shown in blue) and controls (shown in red) under the infinitesimal model for a disease with a high heritability and 10% prevalence (**a**) and under a multiplicative rare allele model for a disease with a high heritability and 2% prevalence (**b**). **a** | This parameterization assumes 200 loci with risk allele frequencies from 0.1 to 0.9 but is skewed towards lower frequencies. Each risk allele is assumed to increase the probability of disease additively by 1.04 relative to the overall risk of 10%. The frequency distribution in cases is skewed to the right, but note that the median number of risk alleles in affected individuals is only slightly greater than it is in unaffected individuals. **b** | The multiplicative risk for a disease with prevalence K is often represented as $K = f_0(1+p(\tau-1))2n$, where f_0 is the baseline risk in the absence of disease alleles, τ is the genotype relative risk (GRR) of each of n alleles at frequency p (REFS 130,131). This parameterization (shown in the left-hand panel) assumes 100 loci, each with a risk allele frequency of 1%, such that each risk allele multiplies a background risk of 0.2% by a factor of 2.2. The vast majority of unaffected individuals carry at least one allele; the orange bars show the expected number of individuals without any risk alleles. The right-hand panel shows the same figure on the logarithmic scale, emphasizing how relative risk increases with the number of variants carried. Note that the measured per-allele GRR across the population in the presence of 100 other alleles is ~1.15, which is much smaller than the 2.2-fold multiplicative risk due to a single variant. For higher risks (say, for example, fivefold) and 100 alleles, the frequencies must be very low (~0.1%) for a disease prevalence of 1%, and affected individuals will only carry one or two risk alleles.

the hypothesis that rare single-nucleotide variants at the same loci are the major source of genetic variance[9], but ongoing deep-sequencing studies will quantify such effects. In the case of the ciliopathies, there is evidence from functional assays that rare variants can both promote disease and modify its severity[49]. It is thus incontrovertible that rare variants contribute to disease risk and that identification of such alleles is a powerful mode of genetic analysis. The question is whether they can account for more than a minor fraction of complex disease risk.

Synthetic associations may explain common variant effects. 'Synthetic association' describes the situation in which the association of a common variant with a disease is actually due to linkage disequilibrium (LD) between the common variant and several disease-promoting rare variants that happen to segregate on the same haplotype[50]. Thus, a common variant that is present in 20% of cases and that mathematically explains 1% of disease susceptibility may actually simply report the activity of two or three rare variants that each substantially elevate risk in one or two percent of the cases. Until this year, rare variants have been excluded from the major whole-genome genotyping platforms, so there has been no way to document their contribution systematically. Synthetic association is not expected to account for most of the missing heritability[19,51], but this type of effect must always be considered as an explanation for apparent common variant effects[52].

Ciliopathies
A class of diseases due to disruption of the cilium, a cellular organelle.

Linkage disequilibrium
(LD). Nonrandom association between genotypes, generally discussed in relation to loci that are closely located on a chromosome: for example, within a gene.

Haplotype
A set of alleles that commonly segregate together and are defined as regions of extended linkage disequilibrium, which in humans is often up to 100 kb in length.

Arguments against the rare allele model

Analysis of GWAS data is not consistent with rare variant explanations. A major argument against rare variants as the predominant source of missing heritability comes from the analysis of the allele frequency distributions of GWAS data[19]. Considerations of LD using standard quantitative genetic theory[19] strongly limit the number of rare variants and the range of their effect sizes that would be compatible with them making a large contribution to disease risk yet remaining undetected in GWASs (FIG. 3). Furthermore, analyses of the distribution of risk allele frequencies across 8 traits argue that, if anything, MAFs are skewed to be >0.2, providing strong empirical evidence that rare alleles are not alone responsible[53]. Rare allele proponents point out that it is difficult to model the true distribution of rare variants and that the under-representation of rare variants on genotyping arrays complicates the interpretation. No-one doubts that some fraction of the total risk for any complex disease is due to rare alleles, but these studies argue that it is not the majority.

Sibling recurrence rates are greater than would be expected by the postulated effect sizes of rare variants. Heritability of disease is often inferred from elevated sibling recurrence rates relative to incidence in unrelated individuals. Intuitively, if a disease has a prevalence of 1% in the general population, but 50% of siblings are affected, then the odds are elevated 50-fold in the family, whereas a single segregating rare variant with a fivefold effect would

Box 2 | **Broad sense heritability models**

Phenotypic variance is traditionally decomposed into genetic and environmental components, and heritability is defined as the ratio of the genetic to the total phenotypic variance in a population[8]. The genetic component of phenotypic variance can be further decomposed into additive, dominance and interaction effects. The additive component is the average effect of substituting one allele for the other, irrespective of whether dominance is present: namely, whether the heterozygotes are closer in phenotype to one class of homozygote.

'Broad sense heritability' refers to the genetic effect that includes non-additive components, such as genotype-by-genotype (G×G) interactions, also known as epistasis, genotype-by-environment (G×E) interactions and epigenetics. Epigenetics refers to the impact of chromatin modification on the effect of a genotype through DNA methylation, alteration of the histone code and, some would argue, microRNA expression.

Genome-wide association studies (GWASs) are not generally powered to detect epistasis or G×E interactions[121], and they are not designed to detect epigenetic influences[122] (although recent sampling designs to detect parent-of-origin associations do provide evidence for them[23]). There are two major obstacles to detecting epistasis: very large samples are required to find sufficient individuals of each genotype combination to measure small effects accurately, and the number of comparisons scales exponentially with the number of interactions, so the testing burden is enormous. Additionally, epistatic effects are generally thought to probably be small relative to main effects, and the GWAS literature provides few examples of departure from additivity. Similarly, there is as of yet little evidence for environmental modification of genotype effects: a recent example relating to the protective effect of coffee consumption on Parkinson's disease risk provides an exception[108]. One of the difficulties may be that the environment is difficult to define and to measure, and it may be unique individual exposures that are more important than exposures that are shared by thousands of people.

A particular class of genetic interaction that has yet to attract full attention is deleterious intergenic compound heterozygosity: namely, interactions between multiple rare variants that lead to disease. Whether and how often such interactions extend to heterozygous combinations of alleles has considerable implications for the potential of broad sense heritability to be a major component of disease susceptibility. If hundreds of mutations that are each at a frequency of <1% all affect a particular disease, then many more people will be doubly or triply heterozygous for various combinations than are homozygous for any one variant. It is not known how often heterozygotes that do not individually associate with disease would be deleterious in combination. Synthetic genetic interaction screens in yeast[123,124] have demonstrated that thousands of combinations of individually viable mutations can jointly be lethal, and similar examples have been documented in *Drosophila melanogaster* and *Caenorhabditis elegans*[125,126].

only be expected to result in disease for 5% of carrier individuals. Sibling recurrence rates are generally much higher than can be attributed to rare variants with the postulated effect sizes[17]. If rare variants are contributing, then they do so in the context of many other genetic variants in the pedigrees, as shown by a recent analysis of schizophrenia genetics[54]. Under multiplicative models, diseases may cluster in families owing to segregation of multiple rare variants that happen to be brought together in the pedigree. Nevertheless, the rare allele model must not only explain association data in populations, but it must also fit demographic distributions of disease within and among families, and more theory is required to support the claim that they can do both[55–57].

Rare variants do not obviously have additive effects. On the face of it, the widely documented additivity of genetic associations is inconsistent either with the dominance of rare variant effects or with the assumption that they interact multiplicatively within individuals to influence disease. However, it turns out that rare variants can easily induce apparently additive effects statistically, because the homozygotes for the tagging variant are twice as likely as the heterozygotes to carry the rare variant. Multiplicative interactions between rare variants are additive on the logarithmic scale but cannot be measured in GWASs because of low power. It will be important to establish mechanistically whether combinations of two or more such mutations increase risk in a linear or synergistic manner[58,59]. Compound heterozygosity for two different rare variants at one locus is well documented in diseases such as cystic fibrosis, haemochromatosis and sickle cell syndromes; extension of the concept to include intergenic multiple heterozygosity could represent a large source of genetic variance that is almost impossible to detect with current methods.

Epidemiological transitions cannot be attributed to rare variants. The fourth argument against rare variant effects is a demographic one — namely, the changing prevalence of so many chronic diseases in a span of just two or three generations and the known impact of environmental variables on risk. For example, diabetes and heart disease have greatly increased in incidence in India and China in the past 10 years[60,61]: an epidemiological transition that at best implies a change in penetrance of genetic effects that are attributable to any class of variant, whether rare or common, in the contemporary environment. Schizophrenia, a disease with a very high heritability (as inferred from twin studies) and for which very few replicated hits have been identified by GWASs, despite extensive scans, nevertheless shows such demographic influences as whether the parents live in rural or urban areas (disease rates are elevated in children born after migration to cities)[62]. Paternal age effects on psychological disease[63] might be attributed to elevated mutation rates in sperm, but other hypotheses are equally compatible with the data, and maternal age effects operate in the opposite direction, as younger mothers have higher likelihoods of having affected children[64]. In other words, rare variants alone cannot explain the demographic distribution of disease incidence.

GWAS associations are consistent across populations. An empirical argument against pervasive rare variant effects is that common variants are often consistent across populations — such as between Caucasians and Asians — despite differences in allele frequencies[65,66]. If rare variants are recently derived relative to the common variants, then they should be at different frequencies in Caucasians and Asians, and we would expect that they would only induce synthetic common variant associations in one of these populations or at least that they would not tend to have the same magnitude of effect. This would be especially true of common variants that differ in frequency between the two populations. Under a mutation–selection balance model, it may be possible for different novel rare variants to have effects in each population, and fine-mapping studies sometimes reveal subtle differences in the patterns of association (for example, REFS 67,68). Nevertheless, the simplest interpretation of

Mutation–selection balance
An evolutionary model that accounts for the maintenance of genetic variation as a balance between mutation generating variance and purifying selection removing it.

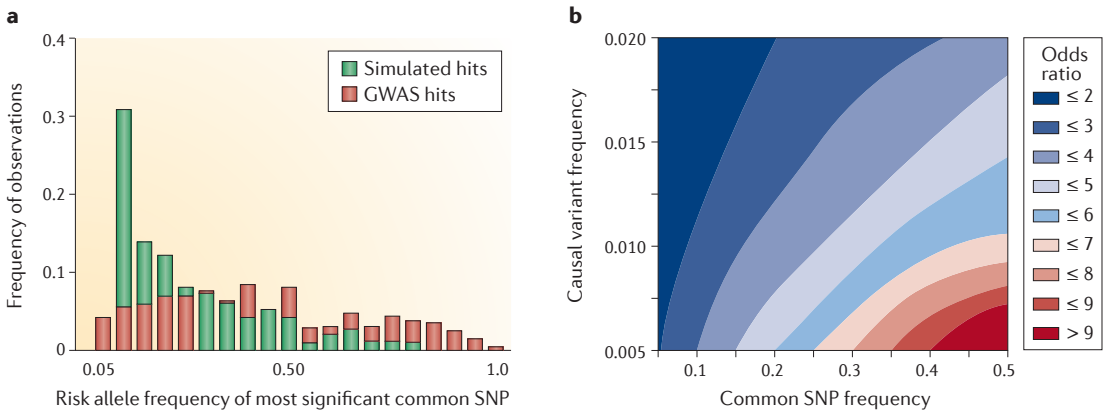

Figure 3 | **Inconsistency between genome-wide association study results and rare variant expectations.** **a** | The frequency distribution of risk allele frequencies (shown in light red) for 414 common variant associations with 17 diseases is only slightly skewed towards lower-frequency variants. By contrast, simulations — in this case, assuming up to nine rare causal variants inducing the common variant association with SNPs at the same frequency as observed on common genotyping platforms (light green bars) — result in a marked left-skew with a peak for common variants whose frequency is less than 10%. (The skew is even stronger if only a single causal variant is responsible.) The observed data are thus not immediately consistent with the rare variant model. **b** | Part of the problem with synthetic associations is that they would explain too much heritability if they were pervasively responsible for common variant effects. This is due to the relationship between allele frequency, maximum possible linkage disequilibrium (LD) and the amount of variance explained[19]. The plot shows the expected odds ratio due to a rare variant of the indicated frequency (from 0.5% to 2%) if it increases the odds ratio at a common SNP (with which it is in maximum possible LD) by 1.1-fold. Intermediate effect sizes (2 < odds ratio < 5) require combined causal variant frequencies in excess of 1%. As the number of rare variants increases, the likelihood that they are in high LD with the common variant also drops, further reducing the probability that they can explain observed common variant association. Suppose that a disease has a prevalence of 1%. Then ten causal variants that are each at a frequency of 1% would result in 20% of people carrying a causal variant. If the penetrance is 5%, then 1% of people would have the disease, and these 10 variants would completely explain the genetic risk. Similarly, if 100 causal variants were each at 0.1% frequency, it would take ~10 such variants to induce each single common variant association with an observed odds ratio of 1.1. If large genome-wide association studies (GWASs) detect dozens of such common loci, and they were actually due to LD with rare variants, then the heritability would be explained several times over. Alternatively, if hundreds of very rare causal variants are not in LD with common variants, we do not expect to see significant GWAS associations. Data taken from REF. 19.

the consistency of common variant effects is that they are actually due to the common variants themselves or to unobserved common variants in high LD across all populations.

Arguments in favour of the infinitesimal model

The infinitesimal model underpins standard quantitative genetic theory. Just as evolutionary theory provides a strong argument in favour of rare variants, standard quantitative genetic theory provides ample support for the infinitesimal model[7,8]. Whatever the causes of the maintenance of genetic variance may be, the consistent observation is that all diseases have moderately high heritability, and so purifying selection has been unable to purge the population of disease-promoting variants[2]. At face value, the existence of dozens of susceptibility alleles for metabolic and immunological diseases with effect sizes that are just not detected for psychological diseases implies a difference in genetic architecture between the two categories of conditions. This may imply different intensities of purifying selection, although other models, including decanalization[69], are also compatible with the data. Because most of the genetic variance remains unexplained, it is *a priori* just as likely to exist in the form of rare or common alleles, and the fact is that there is nothing about GWAS findings that is inconsistent with the infinitesimal model of many variants of very small effect across the full allele frequency spectrum. This model has served applied quantitative geneticists as well as evolutionary biologists for close to a century and, in a sense, it can be regarded as the null model that needs to be disproved before it is abandoned.

Common variants collectively capture the majority of the genetic variance in GWASs. Direct empirical support for the infinitesimal model comes from genomic variance analyses[70,71]. Animal breeders have been using genomic selection methods with great success for the past decade[72], basing their selection of sires and dams on the overall predicted breeding value, which is determined from the full set of genomic markers that capture variation distributed throughout the genome. Similarly, in humans, by taking all nominally significant SNPs rather than just the significant ones from GWASs, it is possible to capture much more of the genetic variance than is explained by the highly significant loci[73,74] (BOX 3). A multivariate version of this approach, which is implemented by regression of phenotypic similarity on genetic relatedness, also implies that common variants capture most of the genetic variants[71]. Furthermore, partitioning of the genetic variance

Decanalization
The notion that genetic systems evolved to be buffered but that large effect mutations or environmental change can overcome this buffering, thereby increasing the genetic variance.

Genomic selection
The use of genetic markers that are spread throughout the genome to select individuals with desired predicted breeding values.

Predicted breeding value
The estimated phenotype of progeny of individuals that have a particular genotype.

Box 3 | **Association tests in unrelated individuals**

Genome-wide association studies (GWASs) generally start with comprehensive SNP-wise evaluation of the significance of the correlation between the genotype and the trait. For the most part, it is assumed that genotypic effects are ordered: namely, that the heterozygotes will have intermediate risk or trait values to homozygotes.

As millions of tests are performed, the results must be adjusted to control for false positives (usually using the conservative Bonferroni correction[127]), resulting in the standard GWAS threshold of 5×10^{-8}. The significance of the test statistic is then plotted against SNP position along the chromosome (FIG. 1). For the most part, genotyped SNPs are thought to 'tag' the actual causal variant, so efforts are made to estimate (or 'impute') as many genotypes in the region of a GWAS hit as possible[128], increasing the chances that the common causal variant is represented or that the effects of less common ones are better captured. Owing to improved methods for imputation, most minor alleles down to 5% frequency can now be inferred with a high accuracy if the reference panel is appropriately matched for population structure[129]. Rare variants, particularly those at a frequency of < 1%, need to be directly sequenced.

Although association tests are reasonably powered for detecting common variants that have a genotypic relative risk of 1.2 or more in meta-analyses that include tens of thousands of individuals, detecting the effects of single rare variant remains problematic. Various score statistics are being developed that evaluate pools of rare variants in a part of a gene, or in a group of related genes, for over-representation in cases or controls[119]. These can be conditioned on prior estimation of the likelihood that a nucleotide substitution is deleterious. Functional tests are increasingly used to validate candidate causal polymorphisms.

Two additional strategies have been proposed to overcome the very high false-negative rate in GWASs that results from adopting the strict GWAS significance threshold. One strategy is to generate a weighted sum of the contributions of all variants beyond nominal statistical thresholds that are observed in a discovery sample and then to ask whether it is predictive of risk or phenotype in a second sample[53]. In several instances in which a SNP-wise GWAS has uncovered few loci, this approach has provided clear evidence for polygenic risk. However, it must be recognized that the high variance in estimation of effects results in noisy scores that, although highly significant in the replication sample, only capture over an order of magnitude less of the variance and are far from predictive. See REF. 74 for another related strategy. The second approach is to evaluate all SNPs simultaneously by multiple regression or to use the elegant equivalent of regressing the similarity between individuals on their overall genetic similarity[71]; this gives rise to an estimate of the genetic variance explained by all SNPs. Partitioning of the genome into chromosome-sized, or smaller, units[75] allows for estimation of the contribution of each region of the genome to the genotypic variance.

on a chromosome-by-chromosome basis for a diverse set of traits shows that the proportion of variance explained is consistently proportional to chromosome length[75,76]. Variance is distributed along all of the chromosomes and is therefore attributable to hundreds of loci. Because common variants are used for the partitioning, it is most parsimonious to conclude that they are responsible, and simulations of rare variants that are so distributed capture much less of the variance.

Variation in endophenotypes is almost certainly due to common variants. Threshold-dependent models[77] postulate that disease is more likely to arise in individuals who have extreme values of underlying endophenotypes[78,79]. In many cases, causal common variants that are associated with a continuous endophenotype have been associated with disease (for example, REFS 80–82), and in some cases these have been confirmed by *in vitro* biochemical assays for structural and regulatory effects (for example, REFS 83–85). Expression quantitative trait locus analysis (eQTL analysis) shows that gene expression and splicing are heavily influenced by common variants, possibly for most transcripts[86–88]. To ignore these data is to deny that such transcriptional variation and metabolic variation are relevant to disease. It is of course possible that disease represents a discrete phase shift (for example, in gene expression profiles) that takes the organism outside the normal range of continuous variation. Thus, tumour samples have discrete transcriptome profile differences[89,90], and it is not obvious that the normal variability is relevant to pathology. Similarly, at least some cortex samples from autistic brains converge on a common transcriptional profile[91]. By contrast, various blood disorders have been shown to correlate with extreme values for the major vectors of modules of gene expression[92]. More research on the relationship between endophenotypes and disease is needed, but most observers would consider it to be implausible that natural variation in physiology is irrelevant to variation in disease susceptibility and would maintain that common variants are most likely to be responsible for continuously distributed physiological variation.

Model organism research supports common variant contributions to complex phenotypes. In model organism research, both pedigree analyses and genetic crosses in which linkage mapping is used to localize QTLs almost always lead to the identification of multiple variants influencing the quantitative trait of interest[20]. This is as true of threshold-dependent characters and cryptic variation[93] as it is of continuous variation. Furthermore, the phenomenon of transgressive segregation in mapping populations of mice and flies established from eight founder strains provides strong empirical support for the existence of common polygenes[94–96]. The overwhelming evidence from classical quantitative genetics is that traits are regulated by many loci with a wide range of effect sizes. A counterargument is that in any cross or pedigree, there is no information about the frequency of contributing QTL alleles in the population, so some fraction of the mapped factors are likely to be rare variants — and if the parents were selected from a base population, possibly most are unusual rare variants, including mutations that were unconsciously selected in the laboratory. In the past year, resequencing of evolved outbred populations of *Drosophila melanogaster* has provided strong support for selection on thousands of variants being responsible for changes in the highly complex traits of body size and fecundity[97,98].

GWASs have successfully identified thousands of common variants[99]. Although there has been a very public focus on the failure of GWASs to find the missing heritability[5,6], the simplest explanation for this is that the expectations were based on unrealistic prior assumptions of effect sizes. After it has been accepted that most alleles are associated with relative risks <1.2, it becomes clear that hundreds of thousands of individuals are required to identify more than a few dozen loci than to explain >20% of the variance. In fact, sample sizes can be extrapolated from the range of variant effects in initial discovery samples[12]. In the case of human height, the GIANT Consortium confirmed the inferences from 30,000 individuals[100] when they increased the study sample to 180,000 individuals,

Threshold-dependent models
A model that postulates that individuals who exceed some threshold value of a continuous physiological characteristic (called 'liability') have or are at high risk for disease.

Endophenotypes
Intermediate physiological or psychological traits, such as metabolite and transcript abundance or a specific neuronal function.

Expression quantitative trait locus analysis
(eQTL analysis). Studies of the association between genotypes and gene expression (transcript abundance), leading to the detection of eQTLs.

which led to the identification of an additional 180 or so loci[14]. Simply put, the data are generally consistent with the infinitesimal model to a first approximation, even where variants are yet to be identified. It is the version of the CDCV model involving common variants of moderate effect that needs to be discarded.

Arguments against the infinitesimal model

The QTL paradox. Where are all of the QTLs that are so consistently detected in pedigrees and in experimental crosses when we transition to outbred populations? Why is it that 10 loci can each explain 50% of the genetic variance and most of the heritability in a cross between two strains, but in no cases have GWASs found more than one locus with an effect size that large? Two explanations can be forwarded immediately: first, the QTLs are actually rare variants that only contribute in that cross, and so are precisely the rare variants predicted by the rare allele theory; or second, each cross captures just a small fraction of the genetic variance in a population, so QTLs with large effects in one cross will be expected to be diluted in their contribution relative to other QTLs when measured in the entire population. Also, in many cases, the effect size estimated in the cross will be an overestimate owing to the Beavis effect (or 'winner's curse')[101,102], whereas in an unknown proportion of cases, QTLs will turn out to be due to multiple linked variants that coarse mapping fails to resolve into individual SNPs[103]. Nevertheless, there remains a general paradox that GWASs have found so few variants of moderate effect.

Absence of blending inheritance. A more pointed argument against the infinitesimal nature of effects is that it predicts less granularity in the distribution of risk and phenotypic trait variation than is often observed, although I am not aware of a quantitative assessment of this claim. In the infinitesimal model, disease risk ought to blend smoothly when unrelated people have children. The larger the number of alleles that affect a trait, the lower the among-individual variance should be under random mating as most individuals will share similar numbers of risk alleles. However, disease incidence and complex phenotypes generally cluster in families. A possible resolution of this conundrum is that the observed clustering of disease in families could be explained by stochastic variation in the number of susceptibility alleles: if two people happen to have more than the average number of small variants, so will their children. Furthermore, homophily — which is the tendency for couples to pair on the basis of shared attributes (including subclinical disease indicators) — will tend to enrich for variants that promote those attributes[104]. Granularity of traits is difficult to document, but facial features provide a good example of a suite of traits that do not simply follow either blending or Mendelian inheritance[105]. Certain features, such as the shape of the nose, location of the cheek bones or curve of the lips, run strongly in families, appearing in distant relatives in patterns that are suggestive of large genetic effects. If such clustering is also true of endophenotypes, then the infinitesimal model will be incomplete.

Demographic phenomena suggest more than a simple common-variant model. As with the rare allele model, infinitesimal effects are also not consistent with a wide range of demographic effects that are indicative of G×E interactions and complex genetic interactions. Prime among these are: the pervasiveness of differences in disease risk between geographic areas that are not obviously explained by genetic differentiation; increasing burden of complex disease in the span of one or two generations (both of these phenomena are obvious on browsing the US Centers for Disease Control and Prevention (CDC) website for incidence data); and conditioning of the risk for one disease on another disease in the same individual[106,107]. This does not counter the existence of thousands of small-effect loci that affect each trait or disease risk profile, but it suggests that the narrow sense genetic effects alone are unlikely to be sufficient explanation. It must be noted that there is very little evidence from GWASs for either G×E or G×G interactions[108], but such effects could be mild at the level of individual associations and could be below the power of detection.

Very few common variants for disease have been functionally validated. A technical argument against all common variant models is that association alone is insufficient evidence of function: correlation is not causation. Very few of the thousands of significant GWAS associations have been shown using molecular genetics, biochemistry or biophysics to be the actual risk variant[38,109]. In this context, prudence suggests an open mind in each individual case as to whether the variant, another common variant in LD or a series of less common variants of large effect that are synthetically associated may be responsible. The case of protection against anaemia in chronic hepatitis C — where the initial GWAS clearly captured two less common causal coding variants at the inosine triphosphate pyrophosphatase (*ITPA*) locus — is a good example[110].

What accounts for the missing heritability? Finally, there is the argument that there truly is missing heritability in GWASs. It is rarely appreciated that GWASs do not actually measure heritability (that is, the ratio of genetic to total phenotypic variance in a population), but rather they just measure the genetic variance. Missing heritability is inferred with respect to heritability estimates that generally derive from family studies, where, ironically, direct estimates of the genetic contribution are lacking. Consequently, it is difficult to estimate actually how much of the variance GWAS-captured variants should explain in outbred populations. However, in the case of height, where heritability is as high as 80%, over 50% of the phenotype can be attributed to common variants using the mixed linear modelling approach[71], and after adjustment for allele frequency skews and incomplete LD, essentially the entire genetic contribution has been ascribed to genotypic variation. Yet the same methods applied to BMI do not capture much more than half of the expected genotypic variance, and it is not clear whether they are as efficient for schizophrenia, arthritis or intelligence[75,76]. In these cases, it can be argued that the infinitesimal model

Cryptic variation
Genetic variation with effects that are only seen under perturbed conditions, such as in the presence a particular mutation or environmental exposure.

Transgressive segregation
The appearance of traits in the offspring that are more extreme than those observed in either parent.

Beavis effect
Also called the 'winner's curse', this is the observation that the effect sizes estimated in a discovery sample tend to be overestimates of the true effect sizes, as they typically receive the benefit of sampling variance in the same direction as the true effect in order to exceed strict genome-wide significance levels.

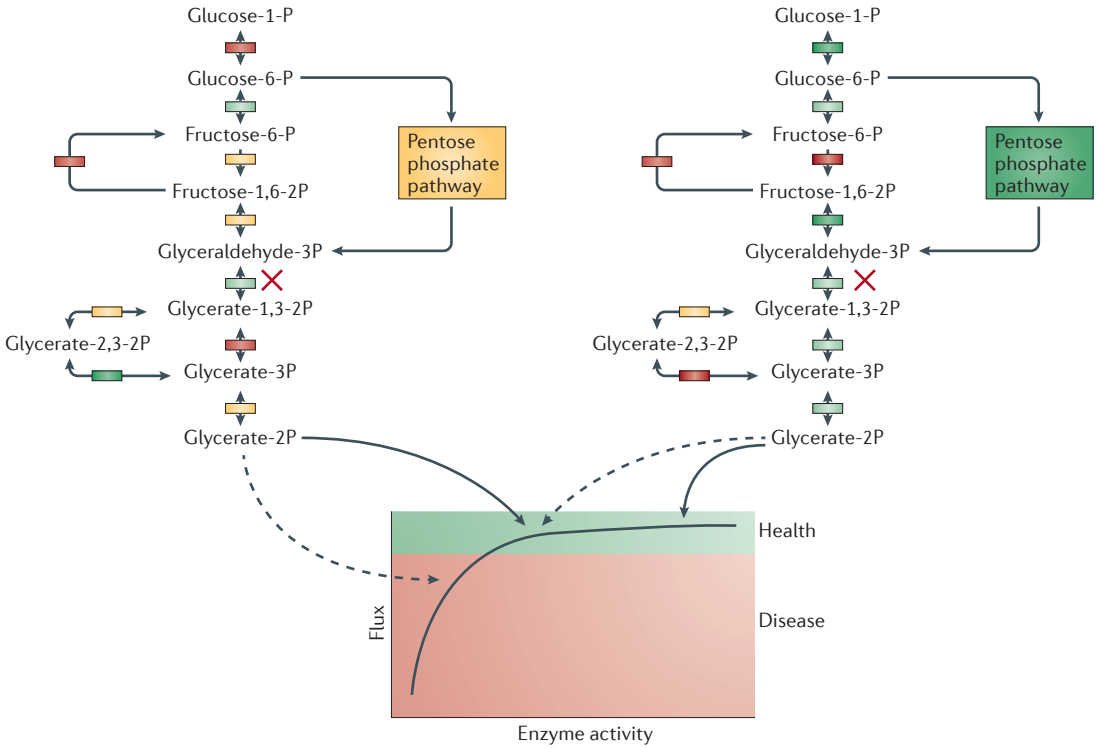

Figure 4 | **Joint effects of rare and common variants.** A straightforward reconciliation of the effects of rare and common variants supposes that pervasive common variation influences the expression and activity of genes in pathways, establishing the background liability to disease that is then further modified by rare variants with larger effects. In this hypothetical example of central metabolism, standing variation results in some individuals having lower flux than others (left versus right; coloured boxes imply enzyme activity differences from low activity (red shading) to high activity (green shading)), but according to standard biochemical theory[130], systems evolve such that most variation is accommodated within the healthy range. The impact of a rare variant that knocks out one copy of the enzyme indicated by the cross is conditional on this liability, pushing the individual on the left beyond the disease threshold, whereas the individual on the right can accommodate the mutation, given higher activity elsewhere in glycolysis.

does not capture the full range of genetic variance and that there is a true missing heritability problem that will need to be addressed with rare alleles and broad sense heritability components that were introduced in BOX 2.

Conclusions

The true debate over the source of genetic variation for disease is not one of 'is it caused by rare or common variants?' or even of 'how much does each class contribute?' but rather 'how do they work together?'[111]. An interesting thought experiment is to ask whether a population of individuals who are only polymorphic for common variants differ in their common disease risk or, conversely, whether rare variants are sufficient to explain the observed distribution of risk in the absence of common variants. The result of both experiments is likely to be negative, but the challenge is to devise strategies to test the hypotheses.

Empirically, there is ample support for both class of effect. As of June 2011, the US National Human Genome Resource Institute (NHGRI) GWAS catalogue lists 1,449 genome-wide significant associations with 237 traits and diseases spread across all chromosome except the Y chromosome[99]. Some of these may be due to LD with rare variants, but parsimony suggests that most are common variant effects. MutDB[112] contains a much larger number of rare coding variants that are either associated with disease or that are predicted to be damaging. As more individuals are sequenced, many of these appear in healthy controls, but it must be recognized that even large effect alleles are subject to background modification and incomplete penetrance. Initial resequencing of genes that have been identified by GWASs[113] has produced mixed results: only 1 of 63 genes in an initial screen for inflammatory bowel disease found evidence for rare variant effects (and these were protective[114], but see also REF. 115), whereas a significant excess of rare coding variants was found in four genes for hypertriglyceridemia[116]. Considerable resolution of the burden of deleterious rare variants will no doubt emerge in the next few years as whole-exome and whole-genome sequencing ramps up[117,118]. Interpretation will be complicated by the natural tendency to under-report negative results, by difficult statistical issues[119] and by the problem of how to define regulatory effects.

The typical resolution of the observation that disease is categorical (that is, people are either cases or controls) but genetic contributions are complex is that disease is generally a threshold-dependent response that is superimposed over a continuous liability[120].

This interpretation actually provides a straightforward framework for integrating rare and common variant effects (FIG. 4). Liability is likely to be established by the additive and infinitesimal contributions of hundreds of polymorphisms, each regulating a series of biochemical traits that impinge on the phenotype: for example, metabolite abundance, gene expression and hormone levels. Disease arises either in individuals at the extremes of the liability scale who move beyond the threshold or in individuals close to the threshold who are pushed into adversity through environmental and behavioural agents or because they carry several rare variants. In this model, a rare variant can either increase or decrease function and whether or not it associates with disease will be conditional on the background liability. The notion that even rare variants have variable penetrance thus blurs the distinction between infinitesimal and major-effect models of disease susceptibility.

1. Lander, E. S. The new genomics: global views of biology. *Science* **274**, 536–539 (1996).
2. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
3. 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).
4. 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).
5. Maher, B. Personal genomes: the case of the missing heritability. *Nature* **456**, 18–21 (2008).
6. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
This paper provides a compendium of arguments, which were assembled by participants in a US National Institutes of Health (NIH) workshop, relating to the possible sources of missing heritability.
7. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930).
8. Visscher, P. M., Hill, W. G. & Wray, N. Heritability in the genomics era — errors and misconceptions. *Nature Rev. Genet.* **9**, 255–266 (2008).
This is an accessible modern introduction to the concept of heritability.
9. Cirulli, E. T. & Goldstein, D. B. Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nature Rev. Genet.* **11**, 415–425 (2010).
10. Feldman, M. W. The heritability hang-up. *Science* **190**, 1163–1168 (1975).
11. Eichler, E. E. *et al.* Missing heritability and strategies for finding the underlying causes of complex disease. *Nature Rev. Genet.* **11**, 446–450 (2010).
12. Park, J. H. *et al.* Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nature Genet.* **42**, 570–575 (2010).
This paper discusses how the true number of associations and their effect sizes can be inferred from observed GWAS results.
13. Speliotes, E. K. *et al.* Association analyses of 249,796 individuals reveal 18 new loci associated with body mass index. *Nature Genet.* **42**, 937–948 (2010).
The largest GWAS meta-analysis to date shows that hundreds of complex variants influence continuous traits.
14. Lango-Allen, H. *et al.* Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
15. Gibson, G. Hints of hidden heritability in GWAS. *Nature Genet.* **42**, 558–560 (2010).
16. Steinberg, M. H. & Adewoye, A. H. Modifier genes and sickle cell anemia. *Curr. Opin. Hematol.* **13**, 131–136 (2006).
17. Bodmer, W. & Bonilla, C. Common and rare variants in multifactorial susceptibility to common diseases. *Nature Genet.* **40**, 695–701 (2008).
18. McClellan, J. M., Susser, E. & King, M.-C. Schizophrenia: a common disease caused by multiple rare alleles. *Br. J. Psychiatry* **190**, 194–199 (2007).
19. Wray, N. R., Purcell, S. M. & Visscher, P. M. Synthetic associations created by rare variants do not explain most GWAS results. *PLoS Biol.* **9**, e1000579 (2011).
20. Mackay, T. F. C. The genetic architecture of quantitative traits. *Annu. Rev. Genet.* **35**, 303–339 (2001).
21. Mackay, T. F. C. & Stone, E. A. The genetics of quantitative traits: challenges and prospects. *Nature Rev. Genet.* **10**, 565–577 (2009).
22. Feinberg, A. P. Phenotypic plasticity and the epigenetics of human disease. *Nature* **447**, 433–440 (2007).
23. Kong, A. *et al.* Parental origin of sequence variants associated with complex diseases. *Nature* **462**, 868–874 (2009).
24. Small, K. S. *et al.* Identification of an imprinted master trans regulator at the KLF14 locus related to multiple metabolic phenotypes. *Nature Genet.* **43**, 561–564 (2011).
25. Jablonka, E. & Raz, E. Transgenerational epigenetic inheritance: prevalence, mechanisms, and implications for the study of heredity and evolution. *Quart. Rev. Biol.* **84**, 131–176 (2009).
26. Bulmer, M. G. The effect of selection on genetic variability. *Am. Nat.* **105**, 201–211 (1971).
27. Barton, N. H. & Turelli, M. Evolutionary quantitative genetics: how little do we know? *Annu. Rev. Genet.* **23**, 337–370 (1989).
28. Bulmer, M. G. Maintenance of genetic variability by mutation-selection balance: a child's guide through the jungle. *Genome* **31**, 761–767 (1989).
29. Charlesworth, B. Fisher, Medawar, Hamilton and the evolution of aging. *Genetics* **156**, 927–931 (2000).
30. Lynch, M. Rate, molecular spectrum, and consequences of human mutation. *Proc. Natl Acad. Sci. USA* **107**, 961–968 (2010).
31. Hartl, D. L. & Clark, A. G. *Principles of Population Genetics* 3rd edn (Sinauer Associates, Sunderland, USA, 1998).
32. Cargill, M. *et al.* Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nature Genet.* **22**, 231–238 (1999).
33. Kryukov, G. V., Pennacchio, L. A. & Sunyaev, S. R. Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**, 727–739 (2007).
34. Zhu, Q. *et al.* A genome-wide comparison of the functional properties of rare and common genetic variants in humans. *Am. J. Hum. Genet.* **88**, 458–468 (2011).
35. Lindblad-Toh, K. *et al.* A high-resolution map of human evolutionary constraint using 29 mammals. *Nature* **478**, 476–482 (2011).
36. Wray, G. A. The evolutionary significance of *cis*-regulatory mutations. *Nature Rev. Genet.* **8**, 206–216 (2007).
37. Montgomery, S. B., Lappalainen, T., Gutierrez-Arcelus, M. & Dermitzakis, E. T. Rare and common regulatory variation in population-scale sequenced human genomes. *PLoS Genet.* **7**, e1002144 (2011).
38. Chorley, B. N. *et al.* Discovery and verification of functional single nucleotide polymorphisms in regulatory genomic regions: current and developing technologies. *Mutat. Res.* **659**, 147–157 (2008).
39. Goldstein, J. L. & Brown, M. S. The LDL receptor locus and the genetics of familial hypercholesterolemia. *Annu. Rev. Genet.* **13**, 259–289 (1979).
40. Weedon, M. N. & Frayling, T. M. Insights on pathogenesis of type 2 diabetes from MODY genetics. *Curr. Diab. Rep.* **7**, 131–138 (2007).
41. Easton, D. F. *et al.* A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the *BRCA1* and *BRCA2* breast cancer-predisposition genes. *Am. J. Hum. Genet.* **81**, 873–883 (2007).
42. Hamosh, A. *et al.* Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. *Nucl. Acids Res.* **30**, 52–55 (2002).
43. Tarpey, P. S. *et al.* A systematic, large-scale resequencing screen of X-chromosome coding exons in mental retardation. *Nature Genet.* **41**, 535–543 (2009).
This was one of the first whole-exome sequencing studies that was designed to detect rare variants of large effect.
44. George, J. *et al.* Two human MYD88 variants, S34Y and R98C, interfere with MyD88–IRAK4–Myddosome assembly. *J. Biol. Chem.* **286**, 1341–1353 (2011).
45. McCarroll, S. A. & Altshuler, D. A. Copy-number variation and association studies of human disease. *Nature Genet.* **39**, S37–S42 (2007).
46. Stefansson, H. *et al.* Large recurrent microdeletions associated with schizophrenia. *Nature* **455**, 232–236 (2008).
47. Sebat, J. *et al.* Strong association of *de novo* copy number mutations with autism. *Science* **316**, 445–449 (2007).
This paper provided the first demonstration that rare copy number variants associate with psychiatric disease.
48. Cook, E. H. Jr & Scherer, S. W. Copy-number variations associated with neuropsychiatric conditions. *Nature* **455**, 919–923 (2008).
49. Davis, E. E. *et al.* TTC21B contributes both causal and modifying alleles across the ciliopathy spectrum. *Nature Genet.* **43**, 189–196 (2011).
50. Dickson, S. P., Wang, K., Krantz, I., Hakonarson, H. & Goldstein, D. B. Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**, e1000294 (2010).
This study presents the argument that common variant associations may be due to LD with rare variants.
51. Anderson, C. A., Soranzo, N. Barrett, J. C. & Zeggini, E. Synthetic associations are unlikely to account for many common disease genome-wide association signals. *PLoS Biol.* **9**, e1000580 (2011).
52. Goldstein, D. B. The importance of synthetic associations will only be resolved empirically. *PLoS Biol.* **9**, e1001008 (2011).
53. Park, J.-H. *et al.* Distribution of allele frequencies and effect sizes and their inter-relationships for common genetic susceptibility variants. *Proc. Natl Acad. Sci. USA* **108**, 18026–18031 (2011).
54. Ruderfer, D. M. *et al.* A family-based study of common polygenic variation and risk of schizophrenia. *Mol. Psychiatry* **16**, 887–888 (2011).
55. Risch, N. Linkage strategies for genetically complex traits: I. Multilocus models. *Am. J. Hum. Genet.* **46**, 222–228 (1990).
56. Slatkin, M. Genotype-specific risks as indicators of the genetic architecture of complex diseases. *Am. J. Hum. Genet.* **83**, 120–126 (2008).
57. Hemminki, K. & Bermejo, J. L. The 'common disease–common variant' hypothesis and familial risks. *PLoS ONE* **3**, e2504 (2011).
58. Slatkin, M. Exchangeable models of complex disease inheritance. *Genetics* **179**, 2253–2261 (2008).
59. Moore, J. H. & Williams, S. M. Epistasis and its implications for personal genetics. *Am. J. Hum. Genet.* **85**, 309–320 (2009).
60. Amutha, A. *et al.* Clinical profile of diabetes in the young seen between 1992 and 2009 at a specialist diabetes centre in south India. *Prim. Care Diabetes* **5**, 223–229 (2011).
61. Chan, J. C. *et al.* Diabetes in Asia: epidemiology, risk factors, and pathophysiology. *JAMA* **301**, 2129–2140 (2009).
62. Saha, S., Chant, D., Welham, J. & McGrath, J. A systematic review of the prevalence of schizophrenia. *PLoS Med.* **2**, e141 (2005).
63. Malaspina, D. *et al.* Advancing paternal age and the risk of schizophrenia. *Arch. Gen. Psychiatry* **58**, 361–367 (2001).
64. Lopez-Castroman, J. *et al.* Differences in maternal and paternal age between schizophrenia and other psychiatric disorders. *Schizophr. Res.* **116**, 184–190 (2010).

65. Waters, K. M. *et al.* Consistent association of type 2 diabetes risk variants found in Europeans in diverse racial and ethnic groups. *PLoS Genet.* **6**, e1001078 (2010).
66. Shriner, D. *et al.* Transferability and fine-mapping of genome-wide associated loci for adult height across human populations. *PLoS ONE* **4**, e8398 (2009).
67. Sim, X. *et al.* Transferability of type 2 diabetes implicated loci in multi-ethnic cohorts from Southeast Asia. *PLoS Genet.* **7**, e1001363 (2011).
68. Waters, K. M. *et al.* Generalizability of associations from prostate cancer genome-wide association studies in multiple populations. *Cancer Epidemiol. Biomarkers Rev.* **18**, 1285–1289 (2009).
69. Gibson, G. Decanalization and the origins of complex disease. *Nature Rev. Genet.* **10**, 134–140 (2009).
70. Schork, N. J. Genome partitioning and whole genome analysis. *Adv. Genet.* **42**, 299–322 (2001).
71. Yang, J. *et al.* Common SNPs explain a large proportion of the heritability for human height. *Nature Genet.* **42**, 565–569 (2010).
This paper introduces a multivariate approach for capturing the effects of common variant associations genome-wide.
72. Goddard, M. E. & Hayes, B. J. Genomic selection. *J. Animal Breed. Genet.* **124**, 323–330 (2007).
73. Purcell, S. M. *et al.* Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
74. So, H.-C., Li, M. & Sham, P. C. Uncovering the total heritability explained by all true susceptibility variants in a genome-wide association study. *Genet. Epidemiol.* **35**, 447–456 (2011).
75. Yang, J. *et al.* Genome partitioning of genetic variation for complex traits using common SNPs. *Nature Genet.* **43**, 519–525 (2011).
76. Davies, G. *et al.* Genome-wide association studies establish that human intelligence is highly heritable and polygenic. *Mol. Psychiatry* **16**, 996–1005 (2011).
77. Falconer, D. S. *Introduction to Quantitative Genetics* Ch. 18 (Longman, New York, 1981).
78. Cannon, T. D. & Keller, M. C. Endophenotypes in the genetic analysis of mental disorders. *Annu. Rev. Clin. Psychol.* **2**, 267–290 (2006).
79. Kendler, K. S. & Meale, M. C. Endophenotype: a conceptual analysis. *Mol. Psychiatry* **15**, 789–797 (2010).
80. Kilpeläinen, T. O. *et al.* Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. *Nature Genet.* **43**, 753–760 (2011).
81. Bis, J. C. *et al.* Meta-analysis of genome-wide association studies from the CHARGE consortium identifies common variants associated with carotid intima media thickness and plaque. *Nature Genet.* **43**, 940–947 (2011).
82. Speliotes, E. K. *et al.* Genome-wide association analysis identifies variants associated with nonalcoholic fatty liver disease that have distinct effects on metabolic traits. *PLoS Genet.* **7**, e1001324 (2011).
83. Musunuru, K. *et al.* From noncoding variant to phenotype via SORT1 at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
This was an important case study showing how to go from association study to molecular function of a specific variant.
84. Wang, Y. *et al.* Whole-genome association study identifies *STK39* as a hypertension susceptibility gene. *Proc. Natl Acad. Sci. USA* **106**, 226–231 (2009).
85. Bertram, L., Lill, C. M. & Tanzi, R. E. The genetics of Alzheimer disease: back to the future. *Neuron* **68**, 270–281 (2010).
86. Cookson, W. *et al.* Mapping complex disease traits with global gene expression. *Nature Rev. Genet.* **10**, 184–194 (2009).
87. Majewski, J. & Pastinen, T. The study of eQTL variations by RNA-seq: from SNPs to phenotypes. *Trends Genet.* **27**, 72–79 (2011).
88. Lalonde, E. *et al.* RNA sequencing reveals the role of splicing polymorphisms in regulating human gene expression. *Genome Res.* **21**, 545–554 (2011).
89. Alizadeh, A. A. *et al.* Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature* **403**, 503–511 (2000).
90. Slavov, N. & Dawson, K. A. Correlation signature of the macroscopic states of the gene regulatory network in cancer. *Proc. Natl Acad. Sci. USA* **106**, 4079–4084 (2009).
91. Voineagu, I. *et al.* Transcriptomic analysis of autistic brain reveals convergent molecular pathology. *Nature* **474**, 380–384 (2011).
92. Chaussabel, D. *et al.* A modular analysis framework for blood genomics studies: application to systemic lupus erythematosus. *Immunity* **29**, 150–164 (2008).
93. Gibson, G. & Dworkin, I. M. Uncovering cryptic genetic variation. *Nature Rev. Genet.* **5**, 681–690 (2004).
94. Aylor, D. L. *et al.* Genetic analysis of complex traits in the emerging Collaborative Cross. *Genome Res.* **21**, 1213–12122 (2011).
95. Philip, V. M. *et al.* Genetic analysis in the Collaborative Cross breeding population. *Genome Res.* **21**, 1223–1238 (2011).
96. Macdonald, S. J. & Long, A. D. Joint estimates of quantitative trait locus effect and frequency using synthetic recombinant populations of *Drosophila melanogaster*. *Genetics* **176**, 1261–1281 (2007).
97. Burke, M. K. *et al.* Genome-wide analysis of a long-term evolution experiment with *Drosophila*. *Nature* **467**, 587–590 (2010).
This paper uses an 'evolve-and-resequence' strategy to demonstrate the pervasive polygenic basis of complex traits.
98. Turner, T. L. *et al.* Population-based resequencing of experimentally evolved populations reveals the genetic basis of body size variation in *Drosophila melanogaster*. *PLoS Genet.* **7**, e1001336 (2011).
99. 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).
100. Weedon, M. N. *et al.* Genome-wide association analysis identifies 20 loci that influence adult height. *Nature Genet.* **40**, 575–583 (2008).
101. Xu, S. Theoretical basis of the Beavis effect. *Genetics* **165**, 2259–2268 (2003).
102. Zhong, R. & Prentice, R. L. Correcting "winner's curse" in odds ratios from genome-wide association findings for major complex human diseases. *Genet. Epidemiol.* **34**, 78–91 (2010).
103. Pasyukova, E. G., Vieira, C. & Mackay, T. F. C. Deficiency mapping of quantitative trait loci affecting longevity in *Drosophila melanogaster*. *Genetics* **156**, 1129–1146 (2000).
104. Fowler, J.H, Settle, J. E. & Christakis, N. A. Correlated genotypes in friendship networks. *Proc. Natl Acad. Sci. USA* **108**, 1993–1997 (2011).
105. Jelenkovic, A., Poveda, A,, Susanne, C. & Rebato, E. Contribution of genetics and environment to craniofacial anthropometric phenotypes in Belgian nuclear families. *Hum. Biol.* **80**, 637–654 (2008).
106. Rzhetsky, A., Wajngurt, D., Park, N. & Zheng, T. Probing genetic overlap among complex human phenotypes. *Proc. Natl Acad. Sci. USA* **104**, 11694–11699 (2007).
107. Ashley, E. A. *et al.* Clinical assessment incorporating a personal genome. *Lancet* **375**, 1525–1535 (2010).
This study develops a strategy that integrates whole-genome sequence and environmental exposure information to assess personal risk of disease.
108. Hamza, T. H. *et al.* Genome-wide gene-environment study identifies glutamate receptor gene *GRIN2A* as a Parkinson's disease modifier gene via interaction with coffee. *PLoS Genet.* **7**, e1002237 (2011).
109. Bauer, R. C., Stylianou, I. M., Rader, D. J. Functional validation of new pathways in lipoprotein metabolism identified by human genetics. *Curr. Opin. Lipidol.* **22**, 123–128 (2011).
110. Fellay, J. *et al. ITPA* gene variants protect against anemia in patients treated for chronic hepatitis C. *Nature* **464**, 405–408 (2010).
111. Schork, N., Murray, S. S., Frazer, K. & Topol, E. J. Common vs. rare allele hypotheses for complex disease. *Curr. Opin. Genet. Dev.* **19**, 212–219 (2009).
112. Stenson, P. D. *et al.* The Human Gene Mutation Database: 2008 update. *Genome Med.* **1**, 13 (2009).
113. 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).
114. Momozawa, Y. *et al.* Resequencing of positional candidates identifies low frequency IL23R coding variants protecting against inflammatory bowel disease. *Nature Genet.* **43**, 43–47 (2011).
115. Rivas, M. A. *et al.* Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nature Genet.* **43**, 1066–1073 (2011).
116. Johansen, C. T. *et al.* Excess of rare variants in genes identified by genome-wide association study of hypertriglyceridemia. *Nature Genet.* **42**, 684–687 (2010).
117. Ng, S. B. *et al.* Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
118. Lupski, J. R., Belmont, J. W., Boerwinkle, E. & Gibbs, R. A. Clan genomics and the complex architecture of human disease. *Cell* **147**, 32–43 (2011).
119. Bansal, V., Libiger, O., Torkamani, A. & Schork, N. J. Statistical analysis strategies for association studies involving rare variants. *Nature Rev. Genet.* **11**, 773–785 (2010).
120. Rendel, J. M. *Canalization and Gene Control* (Academic Press, New York, 1967).
121. 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).
122. Slatkin, M. Epigenetic inheritance and the missing heritability problem. *Genetics* **182**, 845–850 (2009).
123. Tong, A. H. *et al.* Systematic genetic analysis with ordered arrays of yeast deletion mutants. *Science* **294**, 2364–2368 (2001).
124. Costanzo, M. *et al.* The genetic landscape of a cell. *Science* **327**, 425–431 (2010).
125. Lucchesi, J. C. Synthetic lethality and semi-lethality among functionally related mutants of *Drosophila melanogaster*. *Genetics* **59**, 37–44 (1968).
126. Lehner, B. *et al.* Systematic mapping of genetic interactions in *Caenorhabditis elegans* identifies common modifiers of diverse signaling program. *Nature Genet.* **38**, 896–903 (2006).
127. Duggal, P., Gillanders, P. M., Holmes, T. N. & Bailey-Wilson, J. E. Establishing an adjusted p-value threshold to control the family-wide type 1 error in genome-wide association studies. *BMC Genomics* **9**, 516 (2008).
128. Li, Y., Willer, C., Sanna, S. & Abecasis, G. Genotype imputation. *Annu. Rev. Genom. Hum. Genet.* **10**, 387–406 (2009).
129. Shea, J. *et al.* Comparing strategies to fine-map the association of common SNPs at chromosome 9p21 with type 2 diabetes and myocardial infarction. *Nature Genet.* **43**, 801–805 (2011).
130. Slatkin, M. Exchangeable models of complex inherited diseases. *Genetics* **179**, 2253–2261 (2008).
131. Wray, N.R. & Goddard, M.E. Multi-locus models of genetic risk of disease. *Genome Med.* **2**, 10 (2010).
132. Kacser, H. & Burns, J. A. The control of flux. *Symp. Soc. Exp. Biol.* **27**, 65–104 (1973).
This paper presents a theoretical argument for the recessivity of naturally occurring mutations that affect metabolism.

Acknowledgements
I particularly thank F. Vannberg, D. Goldstein, P. Visscher and E. Cirulli for discussions and suggestions and the Georgia Institute of Technology and the US National Institutes of Health for funding.

Competing interests statement
The authors declare no competing financial interests.

FURTHER INFORMATION
Greg Gibson's homepage: http://www.gibsongroup.biology.gatech.edu
ENCyclopedia Of DNA Elements (ENCODE) project: http://www.genome.gov/10005107
GIANT consortium: http://www.broadinstitute.org/collaboration/giant
MutDB: http://mutdb.org
***Nature Reviews Genetics* series on Genome-wide association studies:** http://www.nature.com/nrg/series/gwas/index.html
Online Mendelian Inheritance in Man (OMIM): http://www.ncbi.nlm.nih.gov/omim
US Centers for Disease Control and Prevention (CDC): http://www.cdc.gov
US National Human Genome Research Institute (NHGRI) GWAS catalogue: http://www.genome.gov/gwastudies

ALL LINKS ARE ACTIVE IN THE ONLINE PDF