REVIEWS

THE COMPLEX INTERPLAY AMONG FACTORS THAT INFLUENCE ALLELIC ASSOCIATION

Krina T. Zondervan and Lon R. Cardon

Small effect sizes, common-disease/common-variant versus rare variant influences, biased single nucleotide polymorphism ascertainment and low linkage disequilibrium have recently been discussed as impediments to association studies. Such a focus on the individual factors that highlight their maximum potential effect (whether positive or deleterious) is often optimistic as, in practice, they do not operate in isolation. Instead, they work jointly to generate the disease gene architecture and to determine the ability of a study to discover it. Here, we consider how the effect size of the susceptibility locus, the frequency of the disease allele(s), the frequency of the marker allele(s) that are correlated with the disease allele(s) and the extent of linkage disequilibrium together influence genetic association studies.

Since the initial excitement surrounding the sequencing of the human genome, there has been a continuing debate on how this wealth of information can be used to investigate the genetic contribution to common complex diseases — at present, the principal health burden in the Western world[1]. COMPLEX TRAITS arise as a result of the interplay between many genetic variants and environmental exposures. Tracing patterns of gene segregation in families — previously used successfully for single-gene Mendelian disorders — has provided little success in the identification of genes that underlie complex traits. Population-based association methods were initially heralded as more promising[2], yet many investigators have since remarked on their relative lack of success so far[3]. This has resulted in a more tempered, if not pessimistic, view of the many challenges that are faced in mapping disease genes for complex traits[4–6].

Several factors, such as genetic and phenotypic complexity, environmental influences, sub-optimal sampling and data overinterpretation, have been cited as contributors to the lack of success in detecting complex trait loci[3,5]. Although these and other factors are almost certainly to blame, as a first approximation it is useful to consider the complex trait system in terms of four framework parameters: the EFFECT SIZE of a disease locus;

the frequency of the disease allele(s); the frequency of the marker allele(s); and the extent of LINKAGE DISEQUILIBRIUM (LD) between the marker and disease locus. These four parameters are the result of the more subtle hallmarks of MULTIFACTORIAL DISEASES, including interactions among disease loci that are related to effect size or the fact that the disease allele frequencies might reflect a range of mutations at the same locus. This set of parameters provides a convenient summary of the basic aim of the association-study design: to correlate genotypes and disease phenotypes that are obtained from a sample of individuals. In this review, we discuss how we can optimize our chances of finding complex disease associations by examining the interplay of the factors that influence the size of an observed association and, therefore, our ability to find associations. We focus here on genetic variants that are amenable to detection in population-based association studies; that is, we exclude those that are so rare (<0.01–0.001 in frequency) that only family-based studies would realistically provide sufficient numbers of cases to explore the association.

Here, we are concerned exclusively with 'indirect' association studies, namely, those in which there is good reason to believe that the actual susceptibility or aetiological allele is not genotyped but might be located near

an otherwise anonymous marker that is genotyped[7]. For common complex diseases, strong hypotheses about the role of specific variants are generally not available, so the indirect association approach has become widespread. This approach encompasses positional cloning applications, as well as studies of candidate regions and genes (but not candidate mutations). It also underpins what are commonly known as HAPLOTYPE TAGGING methods that aim to capture the haplotype structure in a candidate region[8]. This contrasts with hypothesis-driven, 'direct' association studies in which the susceptibility or causal alleles are themselves evaluated. In this latter case, marker allele frequency and LD parameters would be irrelevant, so the dominant factors would simply be the effect size and disease allele frequencies.

The CASE–CONTROL STUDY is the most commonly used population-based study design for allelic association. It is well suited to the investigation of many genetic and environmental risk factors because of the effective randomization of genotype groups[9]. Accordingly, we discuss the joint effects of the four parameters in the case–control context. We first review the EPIDEMIOLOGICAL principles of case–control study design. We then briefly discuss the concept of LD (as it relates to the analysis of association studies) and current understanding of its variation across the genome. We then discuss the underappreciated importance of the marker allele frequency relative to the frequency of the disease variant in influencing the probability of finding the association. Finally, we describe how all of these features can be considered together if, instead of treating POWER, sample size, study design, and so on, in terms of the effect size of an unidentified disease locus, we consider them in terms of the effect size of the marker loci — some properties of which we do know, or at least can estimate. Assuming that the disease of interest is dichotomous (that is, the disease is either present or absent), we describe this cumulative effect as 'apparent effect size', or operationally as the 'marker odds ratio'. We show how to predict the apparent effect size for a marker on the basis of the four parameters, and we highlight the relationships between disease and marker loci using five established examples of associations between complex diseases and genetic polymorphisms. We conclude by showing how different values of these parameters affect the power of finding associations in large case–control studies.

The case–control study and its effect size

For many decades, classical epidemiology has focused on the role of environmental factors in disease processes. Various study designs have been developed to obtain accurate and unbiased estimates of disease risk that is related to environmental factors. The gold standard in the field has been the PROSPECTIVE COHORT STUDY, because of its ability to obtain unbiased risk estimates[10]. The case–control study was developed as an alternative and less costly design that would allow the exploration of a greater number of factors in relatively rare conditions. In a case–control study, the objective is to compare exposure to risk factors (environmental or genetic) between affected individuals and unaffected controls,

who have been selected from the same population as the cases, to find associations between risk factors and disease[11]. Appropriate selection of controls is an epidemiological principle that is crucial to the validity of the case–control study[12]. Here, we assume that case–control studies are well-designed in the sense of being relatively homogeneous (BOX 1), and that remaining confounding factors are addressed as much as possible by careful matching of cases and controls for the relevant variables or by using methods of GENOMIC CONTROL[13].

The standard measure of effect in the case–control study is the odds ratio (OR), defined as the odds of exposure among cases divided by the odds of exposure among controls. The OR provides a good approximation of the relative risk of the risk factor in question (that is, the ratio of risk of disease in people with the risk factor to that in people without) if sampling for the case–control study mimics that of a prospective cohort study. This means that newly diagnosed (incident) cases and controls are sampled consecutively and prospectively from the same source population, that controls are sampled with replacement (they can be sampled more than once over time) and that, theoretically, cases are eligible to be sampled as controls until the onset of disease[14]. If instead, incident or prevalent cases and controls are sampled cross-sectionally (that is, at a specific point in time after a period of accruement), the OR only provides a good approximation of relative risk if the disease under study is relatively rare (prevalence $<\sim10\%$)[15]. An OR of 1 means that exposed people are at no increased risk compared with the unexposed, an OR of >1 implies an increased risk (or positive association) and $0 < OR < 1$ implies a protective effect (or negative association). Various statistical tests can be used to determine the probability that the OR differs significantly from unity[16], in which case the result is interpreted as significant genetic association.

It is important to recognize that any relative risk, including the OR, tells us nothing about the absolute risk associated with the risk factor — that is, the risk among all who are exposed in a population (the prevalence among exposed individuals, or penetrance). A genetic variant can therefore confer a high relative risk for a disease, but if the baseline risk of disease among unexposed individuals is small, the absolute risk associated with exposure in the general population will be small (see REF. 17 for examples). It follows that for multifactorial diseases — which are relatively common in the general population and, by definition, are influenced by multiple INCOMPLETELY PENETRANT variants — the risk-factors that could confer high ORs would be rare, so that the absolute increase in prevalence associated with the risk factor remains small. If a more common risk factor conferred a high OR, it would have to be the principal contributor to the overall prevalence of disease. Large gene effects of moderate to high frequency have not been detected in most complex traits, despite hundreds of genome scans and thousands of association studies[18]. However, apolipoprotein E(*4) (APOE(*4)) and late-onset Alzheimer disease[19], with

Box 1 | **The influence of epidemiological sampling and disease-gene mutation models**

In this review, we assume the use of well-designed case–control studies in which the cases and controls are relatively homogeneous (in terms of genetic ancestry and environmental exposures). The random sampling of cases and controls from the same source population (the main principle of a well-designed case–control study) aims to yield such homogeneity. Such sampling identifies a spectrum of allele frequencies and environmental exposures present in controls, which reflects the distribution of the underlying population, and of the allele frequency and environmental exposure patterns present in cases, which reflects that of all cases present in that population. Random ascertainment of both cases and controls from the same population provides the opportunity to conduct large-scale investigations without greatly increasing heterogeneity (both genetic and environmental) in the study sample. For example, a case–control study for which a representative sample of 500 cases and controls is randomly drawn from a general-population register (and matched on important confounders) could be extended to one of 1,000 cases and controls sampled from the same register without substantial risk of increased heterogeneity in the causes that underlie the condition (genetic or environmental).

However, random sampling of cases from the general population is only feasible for conditions for which the occurrence is automatically registered (for example, a cancer incidence register). Many (mostly benign) conditions do not lend themselves to such sampling because they are either not centrally registered or — more importantly — they might remain undiagnosed in some individuals. Therefore, diagnosed cases represent a selected subset of all cases that are present in the population, and both cases and controls generally need to be sampled from the location or medical practice at which the cases were diagnosed (which will effectively represent the new source population). In multi-centre studies that involve many such practices, there might be increased genetic and environmental heterogeneity compared with a single-centre study if individuals who attend these practices are not drawn from the same underlying population. In this situation, the increase in power that is obtained from a large multi-centre study compared with a single-centre study is not as great as might be expected.

Large-scale population-based sampling, either through random selection of cases and matched controls or through recruitment from medical practices, should generally aid the detection of common alleles with low effect sizes that underlie the disease of interest. However, it should be emphasized that population-based studies are not well-suited to the detection of very rare alleles (<0.001 in frequency) that underlie disease as they are unlikely to contribute greatly to overall risk in the population, whatever their relative risk (and, therefore, very few cases will be attributable to these alleles). In such cases, enriched ascertainment schemes could be considered to increase the relative contribution of such alleles to disease among the cases that are sampled from a particular population, thereby increasing power. An example is the selection of cases on the basis of a family history of the condition. Such selection would increase the relative frequency — and therefore the measured effect size — of rare alleles that underlie the condition among cases (at the expense of more common alleles), resulting in an increased power of detection[53]. Enrichment of cases on family history might, however, also increase the number of cases with disease that is attributable to multiple rare alleles with very low frequency, which could provide detection problems if these have arisen on different haplotypes (see BOX 2).

an allele frequency of 0.15 and an allelic OR of 3.3, is a notable exception. There are also few such examples in environmental epidemiology, a key exception being the association between smoking and lung cancer — smokers have a 20–30-fold increase in risk of lung cancer relative to never-smokers over a lifetime, and 90% of lung cancer cases could be prevented if smokers stopped before middle-age[20]. So, allelic influences on reasonably common, complex diseases will usually confer only modest ORs, although ORs can be high for genotype effects. This restricted range of effects does not apply to rare alleles or risk factors, which can confer either small or large ORs.

In the optimal marker selection model, in which the marker variant is effectively a good proxy for the disease variant (see below and BOX 2), studies of 500–1,000 cases and controls should be generally sufficient to detect the effects of an OR of >2–3. Smaller effect sizes, of ~1.2–1.3, should be detectable with studies of ~5,000 cases and controls (for reviews, see REFS 1,4). Effect sizes that are substantially below this level are difficult to reliably detect with realistic sample sizes. A key feature to note is that these are the most optimistic sampling requirements — variation in LD and marker/disease

allele relationships both work to increase the required numbers, sometimes substantially. The simple question of 'how large a sample size is needed to detect an effect of size X?', although ubiquitous in study designs, power calculations and grant applications, cannot be answered without considering the other factors.

Linkage disequilibrium

Indirect association studies of genetic variants use the principle that genetic variants (markers) that are in proximity to a disease-causing variant on a particular chromosome will be more often co-inherited with the disease-causing variant than expected under independent assortment. This lack of independence among different polymorphisms is termed linkage disequilibrium (LD), and arises because the variants share a joint population ancestry. On average, markers that are in close physical proximity are more highly correlated than those that are spaced far apart. It is, however, well-known that local variation in LD overwhelms this mean[21] — genetic markers that are immediately adjacent to each other on a chromosome might be statistically independent, whereas those that are hundreds of thousands, or sometimes even millions, of base pairs apart might be highly correlated.
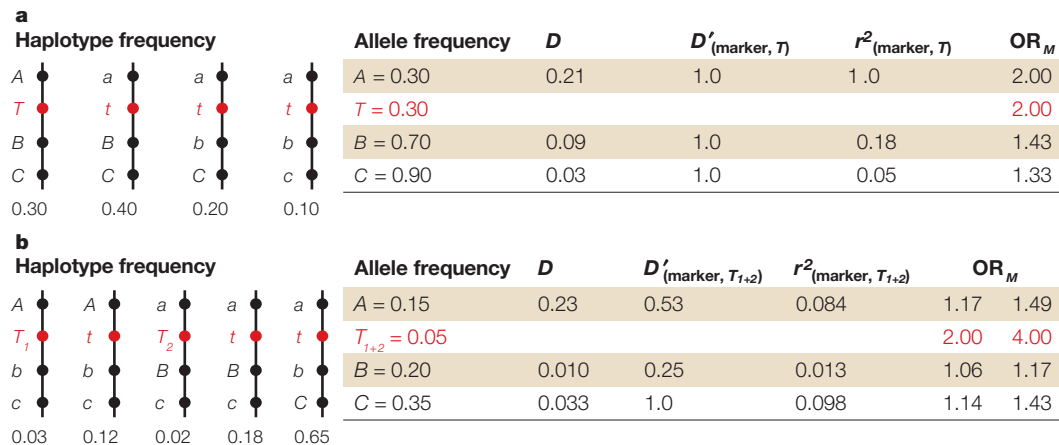
Box 2 | The relationship between LD, marker–disease allele frequency and apparent effect size

a

Haplotype frequency

A	a	a	a
T	t	t	t
B	B	b	b
C	C	C	c
0.30	0.40	0.20	0.10

Allele frequency	D	$D'_{(marker, T)}$	$r^2_{(marker, T)}$	OR_M
$A = 0.30$	0.21	1.0	1.0	2.00
$T = 0.30$				2.00
$B = 0.70$	0.09	1.0	0.18	1.43
$C = 0.90$	0.03	1.0	0.05	1.33

b

Haplotype frequency

A	A	a	a	a
T_1	t	T_2	t	t
b	b	B	B	b
c	c	c	c	C
0.03	0.12	0.02	0.18	0.65

Allele frequency	D	$D'_{(marker, T_{1+2})}$	$r^2_{(marker, T_{1+2})}$	OR_M	
$A = 0.15$	0.23	0.53	0.084	1.17	1.49
$T_{1+2} = 0.05$				2.00	4.00
$B = 0.20$	0.010	0.25	0.013	1.06	1.17
$C = 0.35$	0.033	1.0	0.098	1.14	1.43

Consider the simple situation of a diallelic polymorphism in which the disease or trait allele, *T*, gives rise to a twofold relative risk of a particular complex disease compared with allele *t*. The population frequency *T* is 0.3 (so the frequency of the other allele, *t*, is 1 – 0.3 = 0.7). Cases and controls are collected for a study of the disease, but the disease polymorphism is not genotyped (we do not know of its existence). Instead, three surrounding single nucleotide polymorphisms (*A/a, B/b* and *C/c*), which have different allele frequencies and are in varying degrees of linkage disequilibrium (LD) with the disease locus, are genotyped. Panel A shows a hypothetical haplotype set that represents different situations of LD between marker and disease loci. In a case–control study, the haplotypes that carry the *T* allele will be over-sampled in cases relative to population frequencies (in controls).

The trait allele *T* is present only on one haplotype, *ABC*, which has a frequency of 0.3 in the population. Marker allele *A* is also present only on *ABC,* and therefore also has a frequency of 0.3. Allele *B* occurs on *ABC,* as well as *aBC,* with a total frequency of 0.7, whereas *C* occurs on *ABC, aBC* and *abC* with a total frequency of 0.9. Allele *A* is in complete LD with *T* and because their frequencies match, the apparent odds ratio (OR) of *A* is the same as if the effect of *T* was measured directly (OR = 2.0). Alleles *B* and *C* are also in complete LD with *T* in terms of the *D'* measure (see main text) (because *T* only occurs in combination with *B* and *C*), but not all *B* or *C* alleles are co-inherited with *T* (so the *r²* (see equation 1) values are less than 1.0).

$$r^2 = \frac{D^2}{f(A_1)\ f(A_2)\ f(B_1)\ f(B_2)} \qquad (1)$$

f = frequency

The reduced effect size, OR_M, at *B* (1.43) and *C* (1.33) is calculated from the ratio of marker to disease allele frequency using the expressions in BOX 3.

Note that the allele frequencies of marker *B/b* are identical to those for marker *A/a*, as well as for the trait locus — one allele has a frequency of 0.3 and the other has a frequency of 0.7 in all cases. In addition, the *D'* values between *B* and *T* and between *A* and *T* are both 1.0. Detecting association to either *A* or *B* would seem to conform to the optimal situation of matched allele frequencies and complete LD. However, this example shows the crucial nature of the HAPLOTYPE PHASE: although the frequency patterns of the two markers match, the *B* allele, which resides on the disease haplotype, is not the one that matches the disease allele frequency. Instead, the other allele, *b*, is the one that matches and it never occurs on a haplotype with the disease allele. For equal statistical power, it would take a sample size 5.5 times greater to detect a disease association with *B* than with *A* ($1/r^2 = 1/.18 = 5.55$; see BOX 3). This example shows that markers with equal MINOR ALLELE FREQUENCY and high *D'* are not sufficient to ensure high power; they must match in phase as well. This cannot be predicted from any measurable characteristics of the markers at hand, and is not guided by the theoretical viewpoint of the investigator (common-disease/common-variant or many rare variants).

Panel A depicts a relatively simple situation, in which only one (common) polymorphism underlies the trait of interest, and this polymorphism is unique to a single haplotype (and therefore in complete LD with the markers constituting this haplotype). Panel B shows a more complex situation, in which two relatively rare polymorphisms in the same gene (which could be either separate loci or different alleles constituting a MICROSATELLITE) — T_1 with a frequency of 0.03 and T_2 with a frequency of 0.02 — increase susceptibility to a trait. In addition, T_1 and T_2 each reside on different haplotypes and are associated with opposite *A/a* and *B/b* marker alleles. Assuming equal effect sizes for the T_1 and T_2 alleles, we can consider them together as T_1 and T_2 (with an allele frequency of 0.05), which is in reduced LD with both the *A/a* and *B/b* markers (but in complete LD with *C/c*). For a moderate effect size of OR = 2.0, marker allelic ORs for *A* and *B* are greatly reduced to 1.17 and 1.06, respectively. For a large effect size (OR = 4.0), the allelic OR for *A* is reasonably detectable at 1.49, but the OR for *B* remains low at 1.17. The situation becomes more unpredictable when more rare alleles underlie the trait, and when effect sizes of these trait alleles are not equal[43]. In the latter case, epidemiological sampling strategies that are used in the case–control study will determine the mix of trait alleles that carry haplotypes among cases, and, therefore, the effect sizes that are found when individual markers are analysed (see BOX 1).

For this reason, average amounts of LD between pairs of markers for fixed physical distances are largely uninformative — for LD to inform disease studies, it must be considered locally across the entire genome[22–24].

Characterization of patterns of LD across the human genome is at present an area of highly active research[25], led by a large international initiative to evaluate LD patterns in many populations — the haplotype map or 'HapMap' project[26,27] (see online links box). Recent reviews of these data and their use in association mapping are available[28]. The amounts of LD between any two markers is influenced by the classical forces of recombination, natural selection, mutation, GENETIC DRIFT, ancestral population demographics and mating patterns. These features are of great interest for various reasons[25], but for the present purposes of direct impact on association-study outcomes, we focus on the magnitude of LD and not on how it arose or how it was maintained.

Various statistical measures can be used to summarize LD between two markers[29,30], but in practice only two, termed D' and r^2 (see BOX 2), are widely used. Both measures are built on the basic pairwise-disequilibrium coefficient, D, which is the difference between the probability of observing two marker alleles on the same haplotype and observing them independently in the population: $D = f(A_1B_1) - f(A_1)f(B_1)$[31], where A and B refer to two genetic markers, each having alleles labelled '1' and '2' and f = their frequency. D', or $|D/D_{max}|$, tends to be favoured in medical genetics, possibly owing to its simple scale [0–1] and interpretation: a value of 0.0 implies independence, whereas 1.0 means that all copies of the rarer allele occur exclusively with one of the two possible alleles at the other marker. The r^2 measure has a strong theoretical grounding in population genetics and more desirable statistical properties[32], and therefore tends to be used more for theoretical modelling. An r^2 value of 0.0 also implies independence, but $r^2 = 1.0$ has a more strict interpretation than that of D': $r^2 = 1$ only when the marker loci have identical allele frequencies and every occurrence of an allele at each of the markers perfectly predicts the allele at the other locus. By contrast, D' can reach a value of 1.0 when the allele frequencies vary widely, as it reflects the correlation only since the most recent mutation occurred[21].

There is some confusion about what amounts of LD are required for association studies, which in part stems from the statistical properties of the different LD measures, which in turn mostly relate to allele frequency and sample size dependencies. Studies of average D' levels have used values of $D' = 0.5$ or 'D' half-length' (the midpoint between minimal and maximal LD)[33,34] to describe the extent of LD along chromosome segments. Other studies have used r^2 or close derivatives of it, making use of the property that, for a given power and significance threshold, the required increase in sample size to allow for reduced LD is inversely proportional to r^2 (REFS 35,36). If, for example, in the perfect LD model ($r^2 = 1.0$) an investigator calculated that 1,000 cases and controls would be needed for an association study, then in the imperfect although more realistic situation of, say, $r^2 = 0.2$, he/she would need 1,000/0.2 =

5,000 cases and controls to achieve the same power. Kruglyak[35] suggested a value of $\sim r^2 = 0.10$ for 'useful LD', which would require 10 times the sample size of the best outcome.

These helpful deductions for power calculations are made possible by the fact that r^2 encompasses the effects of both D and all marker/disease allele frequencies. As described below and in BOX 3, this summary can be broken down further by considering the marker and disease allele frequencies separately from the LD. This breakdown is important for practical applications, as we can directly estimate one of the parameters in question — the marker allele frequency — and as we can form hypotheses about another parameter — the disease effect size. Given this information, it should be possible to increase our capacity to design realistic studies with sufficient power to detect genuine effects. In the absence of such information, we are faced with another ill-formed, incomplete, yet often-raised, question: 'how much LD is needed to detect complex disease genes?'

Disease and marker allele frequencies

The common-disease/common-variant (CDCV) hypothesis holds that complex traits have underlying genetic variants that occur with a relatively high frequency (>0.01–0.1), that have undergone little or no selection in earlier populations and that are likely to date back to >100,000 years ago[2,37]. Arguments that support this hypothesis include the rapid expansion of the human population from a small founder pool; the expectation that the allelic diversity for neutral or selectively equivalent alleles in a small founder pool is low (that is, these alleles will be common) and the expectation that a disease-risk allele that was common in the founder population takes a long time to be diluted out by new alleles that are generated during population growth[38].

Supporters of the rare variant model[3,39] have raised three main objections to the CDCV hypothesis. First, they argue that the fact that environmental factors have an important role in complex traits means that individual genetic variants have low attributable risks (in other words, the probability of carrying a particular genotype, given that the individual has the trait, is low). They argue that these diseases are common because of highly prevalent environmental influences, not because of common disease alleles in the population. Second, they consider recent model predictions of complex diseases in which neutral (common) susceptibility alleles contribute little to the genetic variance that underlies disease because they 'tend to be lost or close to fixation', whereas rare alleles that are under weak selection might constitute most genetic variance[40]. Furthermore, they argue that data for the late-onset Mendelian disorders, in which causal genes should have failed to influence fitness, show broad allelic diversity and therefore contrast the CDCV hypothesis[38].

The differences in assumptions and expectations have led to a heated debate about whether complex traits are most often caused by common variants or by multiple rare variants[36,41,42]. An extra dimension to these discussions relates to the ascertainment method of a

GENETIC DRIFT
The random fluctuation in population allele frequencies as genes are transmitted from one generation to the next.

Box 3 | **How four parameters determine the measured effect size: the marker odds ratio**

In the context of family studies, Risch and Teng[54] described how the effect size of interest — the odds ratio (OR) of a particular marker allele — depends on four parameters: the OR of the true disease-causing allele, the extent of linkage disequilibrium (LD) between marker and disease allele, the marker allele frequency and the disease allele frequency. In a population context, Ackerman *et al.*[55] showed that for a trait locus with alleles T/t and a marker locus with alleles M/m, and with haplotype frequencies written as f_{XY}, the OR of an association with the indirect allele M can be calculated from haplotype frequencies (see equation 2).

$$OR_M = \frac{(1-r)(OR_T \, f_{TM} + f_{Tm})}{r(OR_T \, f_{tM} + f_{tm})} \qquad (2)$$

OR_T = the trait or disease allelic OR; r = the population (control) frequency of M

As the haplotype frequencies are encompassed in the disequilibrium coefficient, $D = f_{XY} - f_X f_Y$, OR_M can be rewritten (see equation 3).

$$OR_M = 1 + \frac{D(OR_T - 1)}{r[s + (ps - D)(OR_T - 1)]} \qquad (3)$$

Alternatively, OR_M can be written as the ratio of marker to disease effects (see equation 4).

$$\frac{(OR_M - 1)}{(OR_T - 1)} = \frac{D}{r[s + (ps - D)(OR - _T1)]} \qquad (4)$$

$s = 1 - r$; p = the population frequency of the disease allele ($q = 1 - p$)

These expressions give the marker OR for situations in which a diallelic marker is either positively or negatively associated with the true disease allele, or even when the allele that is analysed at the true disease locus is not the risk-increasing one, but the one that confers a 'protective effect' ($0 < OR_T < 1$). For example, consider a disease locus of $OR_T = 2.0$, in which allele M of the marker (frequency = 0.2) is positively associated with allele T of the disease locus (frequency = 0.3) and disequilibrium $D = 0.14$ between M and T (therefore, $D' = 1.0$ and $r^2 = 0.583$), then the OR_M can be calculated as 1.78. Alternatively, the OR_M for the 'protective allele', m, (frequency $r = 0.8$; $D = -0.14$) is 0.56 (1/1.78). For ease of interpretation, we have limited the remainder of our discussion to positive associations ($D > 0$).

Complete LD
The expressions above allow direct interpretation of the effect size, marker allele frequency, disease allele frequency and LD in terms of the apparent effect size, OR_M. When LD is complete, these expressions reduce further to offer some convenient simplifications for the measured effect size. Consider the situation in which the marker and trait are in complete LD and the population frequency of the associated marker allele is at least as great as that of the disease allele ($p \leq r$ and $D' = 1$). In this case, $D = D_{max}$ and the effect size ratio reduces to the simple expression given in equation 5.

$$\frac{(OR_M - 1)}{(OR_T - 1)} = \frac{p}{r} \qquad (5)$$

That is, when LD is complete and the marker allele is more common than the disease allele, the OR at the marker differs from the true OR only by a scale factor of the allele frequency ratios. Genetically, this can be appreciated by noting that all of the relevant disease alleles reside on haplotypes that are defined by the (more common) marker allele (see BOX 2 for an example — marker *A/a*). Conversely, when the marker allele is rarer than the disease allele ($p > r$) and LD is complete, the apparent effect size still depends on the frequencies of the other alleles (see equation 6).

$$\frac{(OR_M - 1)}{(OR_T - 1)} = \frac{q}{s + (p - r)(OR_T - 1)} \qquad (6)$$

When the marker and disease allele frequencies are very similar and OR_D is relatively small, the apparent OR is again scaled by a simple ratio of allele frequencies, q/s, but this approximation breaks down quickly as the allele frequencies become discrepant and the effect size becomes larger. So, when the disease alleles are more common than the marker alleles with which they are associated, even perfect LD cannot compensate for the fact that there are further disease variants in the population that do not reside on associated marker haplotypes. More complex genetic architectures such as multiple rare variants in high LD that influence the disease would not conform to these simplifications (see BOX 2, panel B).

Matching marker–disease allele frequencies; incomplete LD
When the marker and disease allele frequencies match, the apparent effect size can be represented as shown in equation 7.

$$\frac{(OR_M - 1)}{(OR_T - 1)} = \frac{q}{s + (p - r)(OR_T - 1)} \qquad (7)$$

As noted in the main text, for traits that are influenced by common alleles, the true OR is likely to be low. So, for high values of D', $OR_M \approx 1 + D'(OR_T - 1)$. This implies that when the marker and disease allele frequencies match, the apparent effect size at the marker is close to the true OR scaled simply by the value of D' between the marker and disease locus (see FIG. 1 for examples).

Table 1 | **Examples of the disease-polymorphism associations examined**

Disease	Polymorphism	Allele frequency	Allelic OR (*D* vs. *d*)	GRR-Het (*Dd* vs. *dd*)	GRR-Hom (*DD* vs. *dd*)	References
Deep vein thrombosis	*F5* (G→A)	~0.03 (A)	3.8	4	? *DD* very rare	47
Crohn disease	*CARD15* (3 SNPs)	0.06 (composite)	4.6	3	40	48
Alzheimer disease	*APOE* (*4)	0.15	3.3	3	12	49
Bladder cancer	*GSTM1* (Null)	0.7	1.28	1	1.4	50
Type II diabetes	*PPARγ* (G→C)	0.85 (G)	1.23	1.9	2.1	51

Allelic OR, odds ratio of variant versus normal allele; *APOE*(*4), apolipoprotein E (variant*4); *CARD15*, caspace recruitment domain family, member 15, encodes NOD2; *F5*, coagulation factor V, factor V Leiden; GRR-Het, genotype relative risk (OR) for heterozygotes versus normal homozygotes; GRR-Hom, genotype relative risk (OR) for variant homozygotes versus normal homozygotes; *GSTM1*, glutathione S-transferase M1; *PPARγ*, peroxisome proliferator activated receptor γ; SNP, single nucleotide polymorphism.

candidate region (by linkage analysis or function-driven methods) in determining the probability of finding loci with common versus rare disease alleles[36]. As noted previously, we only consider disease alleles that are of a detectable frequency in the population (that is, ≥ 0.001–0.01). We refer to these alleles as common if they exceed 0.1 in frequency. With respect to these alleles, we do not take a particular stand on the common versus multiple rare variant issue, other than to emphasize two crucial dimensions: those of effect size and those of marker allele frequency. In the remainder of this review, we show that, in general, problems with detection of rare variants that potentially underlie a complex trait will primarily arise if their effects sizes are small. By contrast, rare variants with moderate to large effect sizes should be more amenable to detection (provided that the genetic architecture is not complex, as would occur, for example, with multiple rare variants that have similarly large effect sizes and that are associated with opposing marker alleles; see BOX 2 and REF. 43). The second important point is that indirect association studies are built on the premise that the markers being tested are probably not those that are directly involved in the disease. Accordingly, it is the marker allele frequencies — not just those of the disease — that influence the likelihood of detecting association. For any disease allele frequency, the power of an association study is greatest when the marker and disease allele frequencies match[44–46]. For this reason, the applicability of CDCV for any particular locus is, to some extent, of secondary importance to both the effect size and the frequency difference between the measured marker and unknown disease alleles. In practice, the answer to the contentious question 'does the common-disease/common-allele hypothesis apply to my trait?' is largely dictated by the markers that are selected to be studied.

The relationships between the four parameters that affect the apparent effect size at a marker locus — the OR of the disease variant, the disease allele frequency, the marker allele frequency and the LD between marker and disease locus — are described in BOX 3. The formulae

allow direct calculation of the 'real' effects that are relevant to practical association studies and are those on which appropriate power calculations would be based. We now turn to some examples in the literature to illustrate these relationships.

A range of complex trait associations
In this section, we use five well-established examples (TABLE 1) of complex disease associations to demonstrate the relationship between the four parameters described previously and the change of the marker allelic OR in a case–control analysis. Our examples are established as plausible associations because of replication studies or meta-analyses and provide a broad range of disease allele frequencies. These include: deep vein thrombosis (DVT) and factor V Leiden[47]; Crohn disease and *NOD2/CARD15* (three single nucleotide polymorphisms (SNPs))[48]; Alzheimer disease and *APOE(*4)* (reviewed in REF. 49); bladder cancer and *GSTM1* (see REF. 50); and type II diabetes and *PPARγ* (see REF. 51). The disease allele frequency for *NOD2* (0.06) represents the combined frequency of three rare SNPs, all of which predispose the development of Crohn disease. For ease of interpretation, we considered the effect of *PPARγ* on risk of type II diabetes as a positive association of the more common G allele (frequency = 0.85), although the effect probably should be interpreted as a protective effect of the less common C allele. All of the above were treated as diallelic systems — the OR of the disease allele was calculated relative to the non-disease allele (for SNPs), or to all non-disease alleles combined, and SNPs were used as markers.

The effect sizes reported for the rare or moderately frequent disease alleles (factor V Leiden, *NOD2* and *APOE(*4)*) are substantial (TABLE 1), with allelic ORs all greater than 3, and genotype relative risks (ORs) for homozygote *T/T* versus normal type *t/t* as high as 40 for *NOD2* and Crohn disease. Effect sizes associated with the common alleles of *GSTM1* (0.7) and *PPARγ* (0.85) are much smaller, with allelic ORs not exceeding 1.3 and genotypic ORs not exceeding 2. This supports

Figure 1 | **Reduction in effect size owing to linkage disequilibrium and mismatches between marker and disease allele frequencies.** The decay of marker allelic odds ratio (OR) is shown (*y*-axis) for the five disease-association examples from TABLE 1 using markers of varying allele frequency (lines), and according to the extent of linkage disequilibrium (LD), expressed as *D′* (*x*-axis). For each disease configuration, the disease lines of matching disease–marker allele frequency are shown as black lines. Patterns for marker frequencies of 0.05, 0.10, 0.20, 0.50 and 0.80 are shown as red, blue, green, pink and pale blue lines, respectively. Allelic OR, odds ratio of variant versus normal allele; *APOE* (*4*), apolipoprotein E (variant*4); *CARD15*, caspace recruitment domain family, member 15, encodes NOD2; DAF, disease allele frequency; *F5*, coagulation factor V, factor V Leiden; *GSTM1*, glutathione S-transferase M1; *PPARγ*, peroxisome proliferator activated receptor γ.

the previously mentioned expectation that effect sizes of common polymorphisms found in complex diseases are generally small.

Influence of the four parameters on marker OR. FIGURE 1 shows the influence that marker allele frequency (MAF) and the extent of LD between marker and true disease allele in terms of *D′* have on the reduction in positive allelic effect size found for the marker. For each graph, the black line represents the situation in which MAF matches disease allele frequency (DAF). The OR at which this line crosses the *y*-axis (complete LD, with *D′* = 1) represents the true disease allelic OR; that is, the best example of complete LD and perfect match between marker and disease frequencies.

For all of the examples in FIG. 1, the black line has a slope of ~1.0 (all >0.89) indicating that when marker and disease allele frequencies match, the apparent effect size is approximately proportional to the true effect size multiplied by the *D′* value between the two loci (BOX 3). The other lines illustrate the interplay between the various factors. For the rare allele examples (FIG. 1), the dominant feature is not LD but marker allele frequencies. Even if LD is complete (*y*-axis values along the *x*-origin), at least one-third of the true effect is lost with markers of 10% minor allele frequency or greater. These types of common-allele marker are at present overly represented in public databases[25]. The frequent-allele examples, type II diabetes/*PPARγ* and bladder cancer/*GSTM1* (FIG. 1), suffer similarly from marker-related decay in effect size, but in the opposite direction: infrequent marker alleles of 20% or less lead to apparent effect sizes that are so low that they cannot be detected with even the largest samples, and even 'common allele' markers with 50% minor allele frequency result in a pronounced drop in effect size.

The middle range of frequencies is clearly optimal in these examples; as the Alzheimer/APOE(*4) example shows, most of the true effect size remains when evaluated across a relatively wide range of marker allele frequencies. It is unfortunate that this particular variant is surrounded by very low LD in this region of chromosome 19, which would have rendered the allele frequency advantages largely irrelevant in a position-free gene hunt[52].
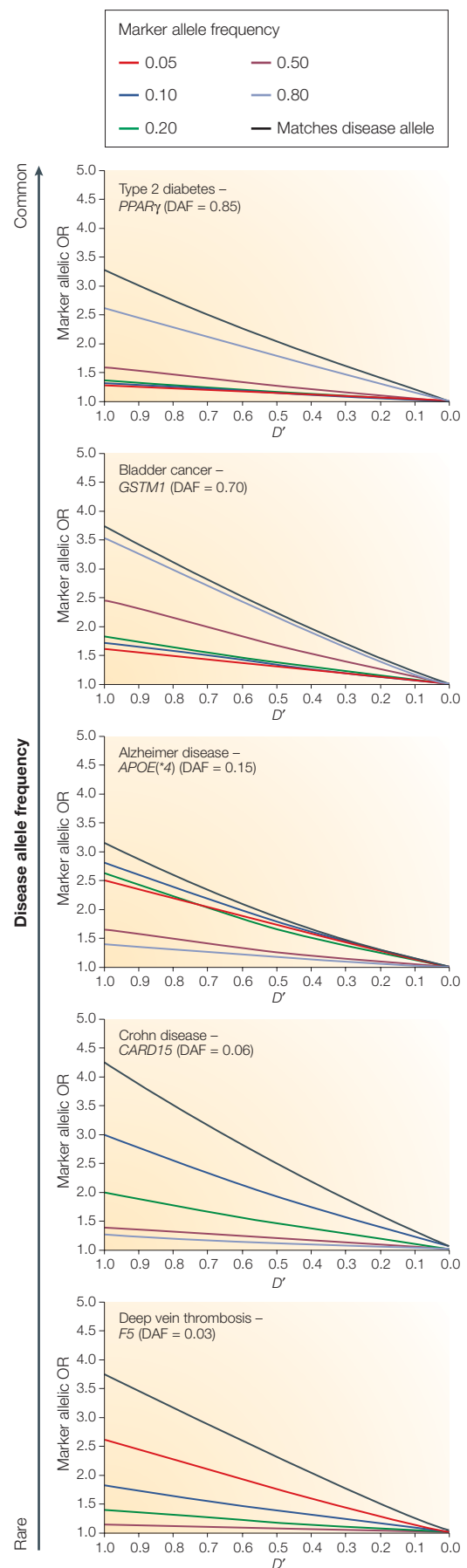
For the Alzheimer, bladder cancer and type II diabetes loci, which follow the CDCV model, incomplete LD exacerbates the information loss from a mismatch between MAF and DAF. D' values in excess of 0.8 or 0.9 are essential to retain any substantive portion of the true effect size, but even complete LD is insufficient if the allele frequencies are not close. In the bladder cancer/*GSTM1* example, having a marker of 50% allele frequency and having $D' = 0.9$ with the true disease allele would seem near optimal, but in reality, such a locus has an effective OR of less than 1.15, which is below the realistic detection threshold of most association studies.

Notably, in the rare-allele models with large effect sizes (DVT and Crohn), large marker–disease allele frequency differences coupled with LD values that are as low as $D' = 0.7$ still produce marker ORs that are similar in size to those found in the most favourable situations in the CDCV examples. On the other hand, situations of rare alleles with small effect sizes would suffer to an even greater extent than the CDCV examples from sub-optimal marker characteristics, because the combination of the very low apparent effect size with the rare allele frequency would necessitate unfeasibly large sample sizes to allow detection.
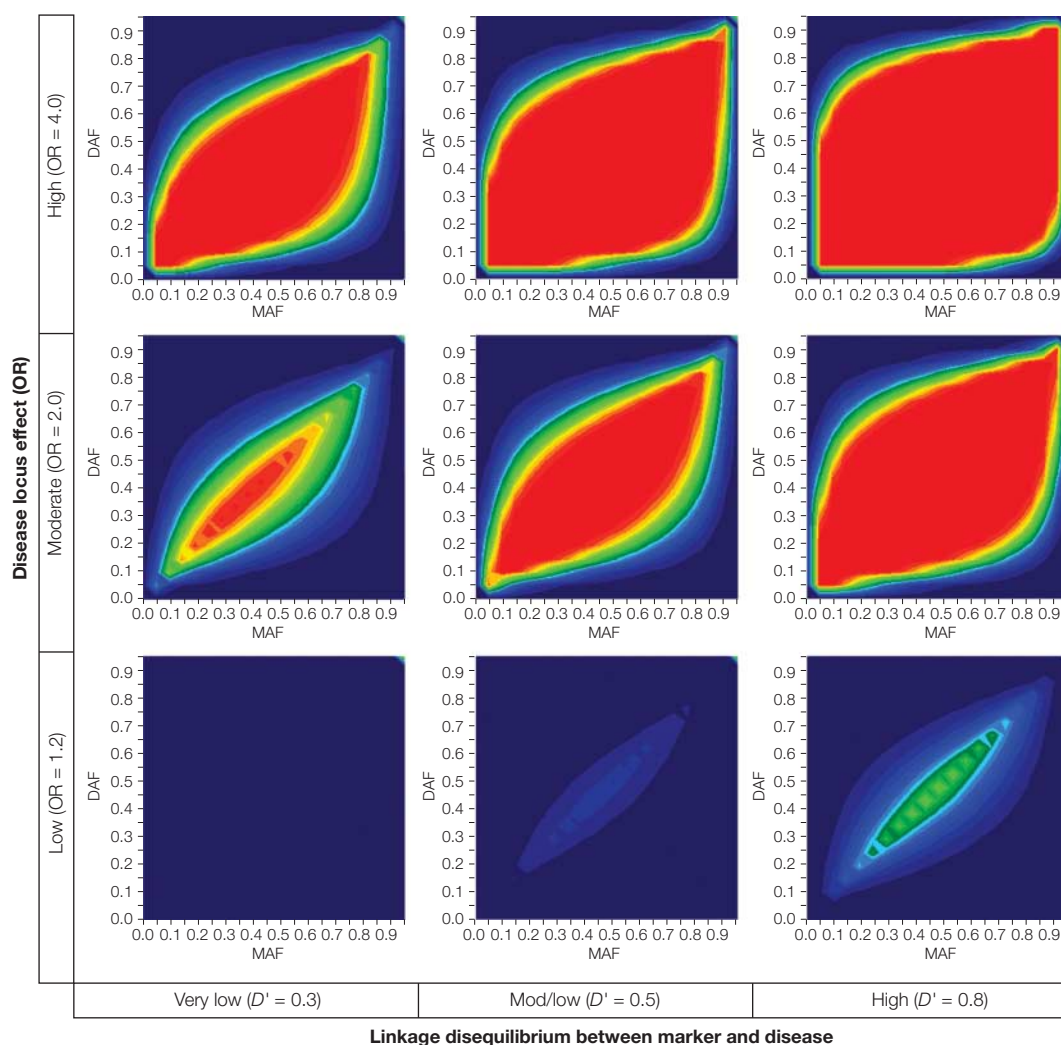

Figure 2 | **The joint influence of linkage disequilibrium, effect size and marker–disease allele frequencies on the power to detect disease association.** The power to detect apparent allelic odds ratios (ORs) is colour-coded (red indicates the highest power, navy the lowest power) for high, moderate and low disease-variant effect size (ORs of 4.0, 2.0 and 1.2, respectively) (main *y*-axis), and of high LD ($D' = 0.8$) versus moderate to low LD ($D' = 0.5$) and very low LD ($D' = 0.3$) between marker and disease variant (main *x*-axis). Each panel shows the relationship between the frequency of the marker allele (panel *x*-axis) and the disease allele (panel *y*-axis). A sample size of 5,000 case alleles and 5,000 control alleles was chosen to allow detection of a significantly increased risk ($\alpha = 0.001$), with the disease allele itself having a power of 1.0 in the situation of a common allele inferring a small effect size (OR = 1.2). This large sample size should also mitigate the bias of D' owing to sample size, which artificially inflates the baseline levels of D' in smaller samples[21]. Power was calculated using Fleiss' approximation for power of a chi-square test with continuity correction[56], which was verified as providing a good approximation to the power of an exact test even for the rare allele disease models.

Table 2 | **Approximate sample sizes needed to detect a significantly increased allelic odds ratio***

Disease allele frequency	Marker allele frequency	Allelic odds ratio of disease gene					
		3.0		2.0		1.3	
		No. cases (= no. controls)	No. cases: no. controls (= 1:4)	No. cases (= no. controls)	No. cases: no. controls (= 1:4)	No. cases (= no. controls)	No. cases: no. controls (= 1:4)
0.05	0.05	360	210:840	1110	650:2600	9500	5600:22400
	0.1	600	350:1400	2000	1200:4800	19000	11500:46000
	0.2	1170	700:2800	4150	2500:10000	40000	25000:100000
	0.3	1900	1200:4800	6800	4300:13200	70000	43000:172000
	0.5	4200	2700:10800	15000	9500:38000	160000	100000:400000
0.2	0.05	710	420:1680	1900	1090:4360	14000	8500:34000
	0.1	350	200:800	900	500:2000	6600	4400:13600
	0.2	150	85:340	360	220:880	2900	1750:7000
	0.3	210	130:520	530	360:1440	4800	3000:12000
	0.5	430	270:1080	1250	800:3200	11000	6950:27800
0.5	0.05	3150	1870:7480	6800	4000:16000	40000	25000:100000
	0.1	1500	900:3600	3200	2000:8000	19000	12000:48000
	0.2	640	390:1560	1350	850:3400	8500	5300:21200
	0.3	360	220:880	800	500:2000	5000	3100:12400
	0.5	140	90:360	320	200:800	2100	1300:5200

*Using diallelic markers with varying allele frequency and allowing linkage disequilibrium between marker and disease allele down to $D' = 0.7$, odds ratio (power ≈ 80%; $\alpha = 0.001$).

Influence of the four parameters on the power of single marker tests. For the design of new studies, it is useful to consider the relationships between apparent and real effect sizes in terms of statistical power. Although most studies are designed with consideration of the optimal properties of the disease loci, the relevant effects concern the measured marker locus. Here, we use a sample size of 5,000 case alleles and 5,000 control alleles to explore the power of a case–control study in relation to the four parameters that determine apparent effect size. This sample size was considered because it would be required to detect common disease associations with small allelic ORs (1.2–1.5), such as the type II diabetes example above, when studying the actual disease variant (assuming 80% power and a TYPE I ERROR of 0.001).

FIGURE 2 illustrates the extent to which power is affected by marker versus disease allele frequencies for models of large effect size of the disease variant (OR = 4.0), moderate effect size (OR = 2.0) and small effect size (OR = 1.2), assuming relatively high LD ($D' = 0.8$) versus moderate/low LD ($D' = 0.5$) and very low LD ($D' = 0.3$) between marker and disease loci. In variants with large effect sizes (which are typically rare), markers across the range of allele frequencies produce a high power of detection, even at moderate to low LD values. When their effect size is moderate, common disease variants (>0.1 in frequency) have a high power of detection when using common markers (>0.1 in frequency) but exact matching of DAF and MAF is not crucial, particularly for the high LD situation. Rarer variants only have a high power of detection either if LD is high (in which case, markers of up to 0.3–0.4 in frequency would be useful), or when MAF matches DAF if LD is low. In cases of low effect size, common disease variants only have any power of detection when MAF matches DAF and when LD is high. However, if LD is moderate to low, neither common nor rare variants can be detected, irrespective of MAF–DAF discrepancy. Low LD for moderate or small effects virtually eliminates any detectable signal.

The extent to which the discrepancy between marker and disease allele frequency, and LD affects the power to detect disease allele associations, and therefore, the sample size that is required, is further shown in TABLE 2. When the extent of LD and effect sizes are moderate to high ($D' = 0.7$, ORs = 2.0), associations can be generally detected using large, but feasible, sample sizes, as long as the disease–marker frequency discrepancy does not exceed ~30%. Increasing the ratio of controls to cases can help to further increase power, with the optimum in cost-benefit depending on the costs for case and control recruitment and analysis[12]. However, for disease allelic ORs of <2.0, sample sizes required are only feasible if the disease allele is common (⩾ 0.1–0.2) and if the marker allele frequency closely resembles that of the disease (or if the marker allele is in extremely high LD with the disease allele). Again, rare alleles with low effect sizes would require unfeasibly large sample sizes for reliable detection.

Interpretation and analysis considerations. The available data indicate that the few associations of rare alleles with complex diseases that have been detected were successful mostly because of their large effect sizes. Because of these large effect sizes, marker/disease-allele discrepancy and moderate LD will have had relatively little effect on the ability to detect these associations. If, on the other hand, their effects had been much smaller, optimal marker frequencies and local LD patterns would have had little beneficial influence on the probability of detection even in large-sized case–control studies (~5,000 cases/controls) as the power is so low at the outset. It is not known how many rare alleles there are that affect complex traits with small effect sizes, and we will probably never know until we find another way to uncover them. Available sample sizes and association designs are not amenable to detection of small, rare effects. Second, in contrast to rare alleles, the modest associations with common alleles are only detectable with large, yet feasible, sample sizes (several thousands

TYPE I ERROR
The probability of rejecting the null hypothesis when it is true. For genetic association studies, type I errors reflect false-positive findings of associations between allele/genotype and disease.

of cases and controls). In these cases, the marker–disease allele frequency and LD relationships become crucial, and even moderate deviations from the optimal models will prevent detection in many circumstances.

There are continuing efforts to construct LD 'maps' of the human genome[24,26], which aim to help solve the problem of unknown LD. Importantly, these maps are being constructed with generally common allele markers, thereby highlighting the issue of marker allele frequency. These markers should be useful for small effects from common alleles and even large effects from rare alleles in some circumstances. However, although there are nearly six million markers in the public domain, they are not the best markers for modest oligogenic effects from rare alleles as the power is so low at the outset, even in the presence of high LD (FIG. 2). Detection of these effects will probably require more traditional family-based linkage studies and further identification of rare variants in affected individuals.

Conclusions

One of the most important challenges in the hunt for genetic variants that underlie complex disease is how to use the wealth of information that we now have on the human genome. We expect that the effect sizes that are associated with common variants are going to be small. We have seen that single marker tests might be able to detect these effects, provided that large samples of individuals are used, that the discrepancy between marker and disease allele frequencies is as small as possible and that the LD between marker and disease allele is moderately high. The last two conditions are more likely to be met if we are able to use the pattern of genomic LD to select markers.

Although many rare alleles are likely to be associated with complex traits, it is unlikely that we will find them using association-study designs if they have small effect sizes. This would be true even if we had rare markers for our analyses. By contrast, rare disease alleles with large effect sizes are more likely to be found even when common markers are used, provided sample sizes are sufficiently large and many rare alleles of similar effect size are not associated with opposite marker alleles. To some extent, this supports the preference for using common markers to optimize the chances of finding common disease alleles with small relative risks, as well as rare ones with large relative risks. Crucial to this, however, is the realization that many thousands of cases and controls will be required for a reasonable chance to find these associations.

1. Botstein, D. & Risch, N. Discovering genotypes underlying human phenotypes: past successes for Mendelian disease, future approaches for complex disease. *Nature Genet.* **33** (Suppl), 228–237 (2003).
   Excellent overview of complex disease mapping, with anticipations for the focus of the next few years.
2. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
3. Weiss, K. M. & Terwilliger, J. D. How many diseases does it take to map a gene with SNPs? *Nature Genet.* **26**, 151–157 (2000).
4. Risch, N. J. Searching for genetic determinants in the new millennium. *Nature* **405**, 847–856 (2000).
5. Cardon, L. R. & Bell, J. I. Association study designs for complex diseases. *Nature Rev. Genet* **2**, 91–99 (2001).
6. Clark, A. G. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr. Opin. Genet. Dev.* **13**, 296–302 (2003).
7. Chakravarti, A. It's raining SNPs, hallelujah? *Nature Genet.* **19**, 216–217 (1998).
8. Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
9. Clayton, D. & McKeigue, P. M. Epidemiological methods for studying genes and environmental factors in complex disease. *Lancet* **358**, 1356–1360 (2001).
   Clear description of the advantages, disadvantages and applications of different epidemiological design methods in the study of genetic and environmental factors in complex disease aetiology. Also a useful background for the increasingly popular country-wide 'Biobank' projects.
10. Rothman, K. J. & Greenland, S. in *Modern Epidemiology* (eds Rothman, K. J. & Greenland, S.) 79–92 (Lippincott-Raven, Philadelphia, 1998).
11. Rothman, K. J. & Greenland, S. in *Modern Epidemiology* (eds Rothman, K. J. & Greenland, S.) 93–114 (Lippincott-Raven, Philadelphia, 1998).
12. Schlesselman, J. J. *Case–Control Studies. Design, Conduct, Analysis* (Oxford Univ. Press, New York, 1982).
13. Pritchard, J. K. & Donnelly, P. Case–control studies of association in structured or admixed populations. *Theor. Popul. Biol.* **60**, 227–237 (2001).
14. Greenland, S. & Rothman, K. J. in *Modern Epidemiology* (eds Rothman, K. J. & Greenland, S.) 47–64 (Lippincott-Raven, Philadelphia, 1998).
15. Kirkwood, B. R. *Cohort and Case–Control Studies. Essentials of Medical Statistics* 173–183 (Blackwell Scientific Publications, Oxford, 1988).
16. Altman, D. G. in *Practical Statistics for Medical Research* (ed. Altman, D. G.) 231–276 (Chapman and Hall, London, 1991).

17. Khoury, M. J., Beaty, T. H. & Cohen, B. H. *Fundamentals of Genetic Epidemiology* (Oxford Univ. Press, New York, 1993).
18. Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, J. Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69**, 936–950 (2001).
19. Roses, A. D. A model for suceptibility polymorphisms for complex diseases: apolipoprotein E and Alzheimer's disease. *Neurogenet.* **1**, 3–11 (1997).
20. Peto, R. *et al.* Smoking, smoking cessation, and lung cancer in the UK since 1950: combination of national statistics with two case–control studies. *BMJ* **321**, 323–329 (2000).
21. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
22. Dawson, E. *et al.* A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
23. Phillips, M. S. *et al.* Chromosome-wide distribution of haplotype blocks and the role of recombination hot spots. *Nature Genet.* **33**, 382–387 (2003).
24. Cardon, L. R. & Abecasis, G. R. Using haplotype blocks to map human complex trait loci. *Trends Genet.* **19**, 135–140 (2003).
25. Ardlie, K. G., Kruglyak, L. & Seielstad, M. Patterns of linkage disequilibrium in the human genome. *Nature Rev. Genet.* **3**, 299–309 (2002).
   Detailed review of LD in the human genome and its possible origins and applications.
26. Couzin, J. Human genome. HapMap launches with pledges of $100 million. *Science* **298**, 941–942 (2002).
27. Couzin, J. New mapping project splits the community. *Science* **296**, 1391–1392 (2002).
28. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
29. Hedrick, P. W. Gametic disequilibrium measures: proceed with caution. *Genetics* **117**, 331–341 (1987).
30. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
31. Lewontin, R. C. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
32. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
33. Reich, D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
34. Abecasis, G. R. *et al.* Extent and distribution of linkage disequilibrium in three genomic regions. *Am. J. Hum. Genet.* **68**, 191–197 (2001).

35. Kruglyak, L. Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genet.* **22**, 139–144 (1999).
36. 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).
37. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
38. Wright, A. F. & Hastie, N. D. Complex genetic diseases: controversy over the Croesus code. *Genome Biol.* **2**, comment 2007.1–2007.8 (2001).
39. Terwilliger, J. D. & Weiss, K. M. Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr. Opin. Biotechnol.* **9**, 578–594 (1998).
   A strong case against the CDCV hypothesis and details the implications of this premise for LD mapping of complex traits.
40. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
41. Smith, D. J. & Lusis, A. J. The allelic structure of common disease. *Hum. Mol. Genet.* **11**, 2455–2461 (2002).
42. Wright, A. A polygenic basis for late-onset disease. *Trends Genet.* **19**, 97–106 (2003).
43. Chapman, J. M., Cooper, J. D., Todd, J. A. & Clayton, D. G. Detecting disease associations due to linkage disequilibrium using haplotype tags: a class of tests and the determinants of statistical power. *Hum. Hered.* **56**, 18–31 (2003).
44. Thompson, E. A., Deeb, S., Walker, D. & Motulsky, A. G. The detection of linkage disequilibrium between closely linked markers: RFLPs at the AI-CIII apolipoprotein genes. *Am. J. Hum. Genet.* **42**, 113–124 (1988).
45. Muller-Myhsok, B. & Abel, L. Genetic analysis of complex diseases (comments on Risch & Merikangas). *Science* **275**, 1328–1329 (1997).
   An early description of the importance of the similarity between disease and marker allele frequency in the power of association detection.
46. Abecasis, G. R., Cookson, W. O. & Cardon, L. R. The power to detect linkage disequilibrium with quantitative traits in selected samples. *Am. J. Hum. Genet.* **68**, 1463–1474 (2001).
47. Bertina, R. M. *et al.* Mutation in blood coagulation factor V associated with resistance to activated protein C. *Nature* **369**, 64–67 (1994).
48. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).

49. Rubinsztein, D. C. & Easton, D. F. Apolipoprotein E genetic variation and Alzheimer's disease. A meta-analysis. *Dement. Geriatr. Cogn Disord.* **10**, 199–209 (1999).
50. Engel, L. S. *et al.* Pooled analysis and meta-analysis of glutathione S-transferase M1 and bladder cancer: a HuGE review. *Am. J. Epidemiol.* **156**, 95–109 (2002).
51. Altshuler, D. *et al.* The common PPARgamma Pro12Ala polymorphism is associated with decreased risk of type 2 diabetes. *Nature Genet.* **26**, 76–80 (2000).
52. Roses, A. D. Pharmacogenetics and the practice of medicine. *Nature* **405**, 857–865 (2000).
53. Antoniou, A. C. & Easton, D. F. Polygenic inheritance of breast cancer: implications for design of association studies. *Genet. Epidemiol.* **25**, 190–202 (2003).
   Recent paper that describes an example of the use of enrichment in epidemiological study design of traits with polygenic origin, and its influence on power in case–control studies.
54. Risch, N. & Teng, J. The relative power of family-based and case–control designs for linkage disequilibrium studies of complex human diseases. I. DNA pooling. *Genome Res.* **8**, 1273–1288 (1998).
   Provides the fundamental equations that describe the relationship between the marker OR detected in a case–control study and the disease OR, marker allele frequency and disease allele frequency.
55. Ackerman, H. *et al.* Haplotypic analysis of the TNF locus by association efficiency and entropy. *Genome Biol.* **4,** R24 (2003).
56. Fleiss, J. L., Tytun, A. & Ury, H. K. A simple approximation for calculating sample sizes for comparing independent proportions. *Biometrics* **36**, 343–346 (1980).

Online links

DATABASES
The following terms in this article are linked online to:
Entrez: http://www.ncbi.nih.gov/Entrez
APOE(*4) | factor V Leiden | *GSTM1* | *NOD2/CARD15* | *PPARγ*
OMIM: http://www.ncbi.nlm.nih.gov/Omim
Alzheimer disease | bladder cancer | Crohn disease | deep vein thrombosis | type II diabetes

FURTHER INFORMATION
International HapMap Project:
http://www.genome.gov/10001688
Access to this interactive links box is free online.

THE COMPLEX INTERPLAY AMONG FACTORS
THAT INFLUENCE ALLELIC ASSOCIATION

Krina T. Zondervan and Lon R. Cardon

In Figure 1, the axes of the two top panels were incorrectly labelled. The corrected version is shown below.
This correction has been made to the online enhanced text and PDF version of this review.