GENOME-WIDE ASSOCIATION STUDIES: THEORETICAL AND PRACTICAL CONCERNS

William Y. S. Wang[*‡], *Bryan J. Barratt*[*§], *David G. Clayton*[*] *and John A. Todd*[*]

Abstract | To fully understand the allelic variation that underlies common diseases, complete genome sequencing for many individuals with and without disease is required. This is still not technically feasible. However, recently it has become possible to carry out partial surveys of the genome by genotyping large numbers of common SNPs in genome-wide association studies. Here, we outline the main factors — including models of the allelic architecture of common diseases, sample size, map density and sample-collection biases — that need to be taken into account in order to optimize the cost efficiency of identifying genuine disease-susceptibility loci.

[*]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Cambridge Institute for Medical Research, University of Cambridge, Cambridge CB2 2XY, UK.
[‡]Basic and Clinical Genomics Laboratory, School of Medical Sciences and Institute for Biomedical Research, University of Sydney, NSW 2006, Australia.
[§]Research and Development Genetics, AstraZeneca, Alderley Park, Macclesfield, Cheshire SK10 4TG, UK.
Correspondence to J.A.T.
e-mail: john.todd@cimr.cam.ac.uk
doi:10.1038/nrg1522

The development of common disease results from complex interactions between numerous environmental factors and alleles of many genes. Identifying the alleles that affect the risk of developing disease will help in understanding disease aetiology and sub-classification. Over the past 30 years, genetic studies of multifactorial human diseases have identified ~50 genes and their allelic variants that can be considered irrefutable or true positives[1,2]. However, there are probably hundreds of susceptibility loci that increase the risk for each common disease. The key question is how to harness the marked recent improvements in our knowledge of the genome sequence and its variation in populations, together with advances in genotyping technologies, to accelerate susceptibility-locus discovery at the lowest cost.

In an accompanying review in this journal, Hirschhorn and Daly[3] put a case forward for the genome-wide association approach, "in which a dense set of SNPs across the genome is genotyped to survey the most common genetic variation for a role in disease or to identify the heritable quantitative traits that are risk factors for disease". They recommend caution in applying the latest high-throughput methods for genotyping[4–8], as the cost of failure is potentially huge for studies that are designed and executed with low statistical power and inadequate quality control. Here, in the context of genome-wide association studies and of minimizing the cost per true positive, we discuss in more detail the rationale for using large sample sizes in light of the smallest allelic risks that are feasible to detect, the choice of SNPs to be genotyped, study-design efficiencies and certain aspects of the statistical analyses of such data. We are not advocating an abandonment of linkage studies of common disease[9–12]. We still cannot say whether the LINKAGE ANALYSIS approach has 'failed' in a general sense, because almost all published studies have used small sample sizes[13] (fewer than 500 AFFECTED SIB-PAIRS), so this alone cannot be used as a justification for carrying out genome-wide association studies. Genome-wide linkage analysis will remain an essential approach until technology is available that allows the association analysis of both rare and common variants at a practical cost and high throughput.

Furthermore, as described previously[14], we view genome-wide association studies not as a new approach in itself, but as a more cost-efficient way to survey common genetic variation compared with the gene-by-gene functional-candidate approach. The latter approach has been successful but, as only small numbers of genes have been studied so far and, as we discuss, sample sizes might have been too small, few true positives have been identified, despite numerous studies and enormous effort. By exploiting the non-random association of alleles at nearby loci (LINKAGE DISEQUILIBRIUM (LD)), which is an

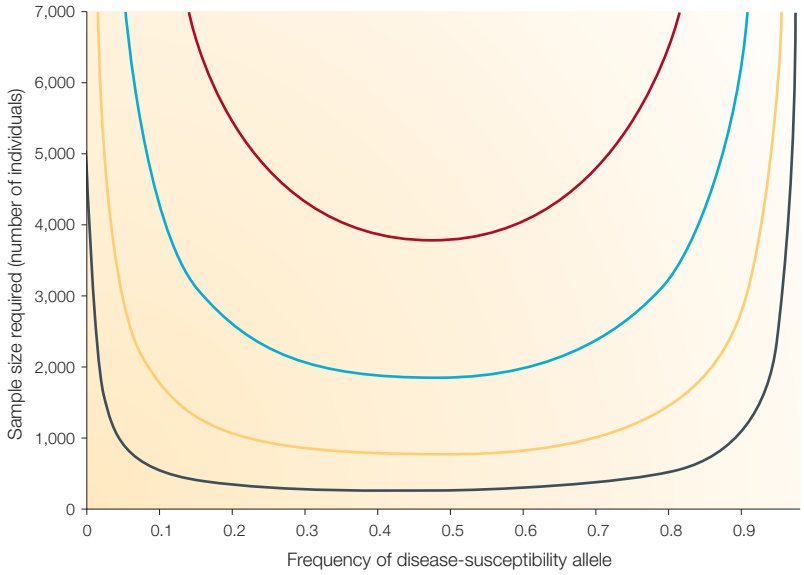

Figure 1 | **Effects of allele frequency on sample-size requirements.** The numbers of cases and controls that are required in an association study to detect disease variants with allelic odds ratios of 1.2 (red), 1.3 (blue), 1.5 (yellow) and 2 (black) are shown. Numbers shown are for a statistical power of 80% at a significance level of $P < 10^{-6}$, assuming a multiplicative model for the effects of alleles and perfect correlative linkage disequlibrium between alleles of test markers and disease variants.

important and widespread feature of the genome[5,15–18], it is now possible to survey in an association study a significant proportion of the common variation of a large number of genes that occur in regions of high LD. Cost efficiency can be gained, as it is not necessary to genotype SNPs that are in strong LD with other SNPs; this can be done by choosing a subset set of SNPs (known as tag SNPs (see Online links box)) that capture most of the allelic variation in a region[19]. The rationale and limitations of this strategy will be discussed, bearing in mind the inadequacy of tag SNPs in detecting rare susceptibility variants and, by definition, their lack of cost-saving advantage in regions of low LD, which might constitute about 20% of the human genome. As well as discussing these more practical issues, we first discuss theoretical considerations concerning two as yet unknown parameters that determine the potential statistical power of an association study — the frequency of susceptibility alleles among the population and the size of their effects on disease phenotypes.

Allelic spectra of common diseases

The allelic spectrum or architecture of a disease refers to the number of disease variants that exist, their allele frequencies and the risks that they confer[9,20,21]. Numerous sources, from both theoretical models and practical experiments, have provided insights into the allelic architecture of common diseases, demonstrating the multiplicity of loci that are involved and their range of effects. Regardless of the exact shape of the spectra, which will differ between diseases, the allele frequencies of variants that predispose to disease and the strength of their phenotypic effects indicate the potential statistical power of genetic association studies, and therefore their likelihood of success and the cost per true-positive result. Here, we first discuss the impact that these two factors are likely to have on the feasibility of genome-wide association studies, and then provide an overview of what is known so far about the allelic spectra of common diseases. It should be noted that other factors also affect statistical power — for example, confounding factors, such as population structure and geography, misclassification errors and selection biases — and some of these factors are discussed in a later section.

Implications for association studies. FIGURE 1 shows that if susceptibility alleles have MINOR ALLELE FREQUENCIES (MAFs) of less than 0.1 and their effect sizes are less than an ODDS RATIO of 1.3, then unrealistically large sample sizes of more than 10,000 cases and 10,000 controls (or 10,000 families) would be required to achieve convincing statistical support for a disease association. We cannot estimate with any accuracy what proportion of disease-susceptibility alleles will lie outside this range (that is, those with odds ratios of 1.3 or above and MAFs of >0.1) and therefore be feasible for detection in genome-wide association studies, and this limitation is discussed below. However, we suggest that studies aimed at detecting such alleles — requiring the analysis of thousands of samples, rather than hundreds of samples — will provide an overall lower cost per true-positive result compared with current candidate-gene and linkage-based approaches.

A study of 6,000 cases and 6,000 controls (or 6,000 families with 2 parents and an affected offspring) would provide, under ideal conditions, approximately 0%, 3%, 43% and 94% power to detect disease susceptibility variants with an odds ratio of 1.3 and MAFs of 0.01, 0.02, 0.05 and 0.1, in corresponding order, at a significance level of $P < 10^{-6}$ (FIG. 1). Significance thresholds in the order of $P < 10^{-6}$ have been proposed for genome-wide association studies, owing to the need to allow for the very small prior probability that any given locus or region is truly associated with disease[3,14,22–24,103,104]. There is a steep decline in power for odds ratios of 1.2 or less (for example, 34% for an MAF of 0.1) (FIG. 1). Conversely, for an odds ratio of 2, even for an MAF of 0.005 there is 76% power. However, we suspect that such high odds ratios will be rare in common diseases (see below).

Undoubtedly, even the best-designed studies, aimed at a minimum MAF of 10% and an odds ratio of 1.3, will have less power than expected owing to many factors, including genotype and phenotype misclassification and confounding factors, so that even larger sample sizes might be required. It is noted, however, that in a study of 12,000 cases and controls, for example, genotyping can be performed in stages with little loss of power. This provides significant savings in genotyping costs, as most of the genotyping is performed in the first stage in a fraction (about 20–30%) of the total number of samples (see REFS 3,25 for more detail about such methods).

In the following sections we discuss theoretical models of the allelic spectra of common diseases and estimate their likely distributions.

LINKAGE ANALYSIS
Mapping genes by typing genetic markers in families to identify chromosome regions that are associated with disease or trait values within pedigrees more often than are expected by chance. Such linked regions are more likely to contain a causal genetic variant.

AFFECTED SIB-PAIR (ASP) STUDIES
Linkage studies that are based on the collection of a large number of families, consisting of affected siblings, and their parents if available. In linkage analyses, the studies rely on the principle that ASPs share half their chromosomes.

LINKAGE DISEQUILIBRIUM
The non-random association of alleles of different linked polymorphisms in a population.

MINOR ALLELE FREQUENCY (MAF). The frequency of the less common allele of a polymorphic locus. It has a value that lies between 0 and 0.5, and can vary between populations.

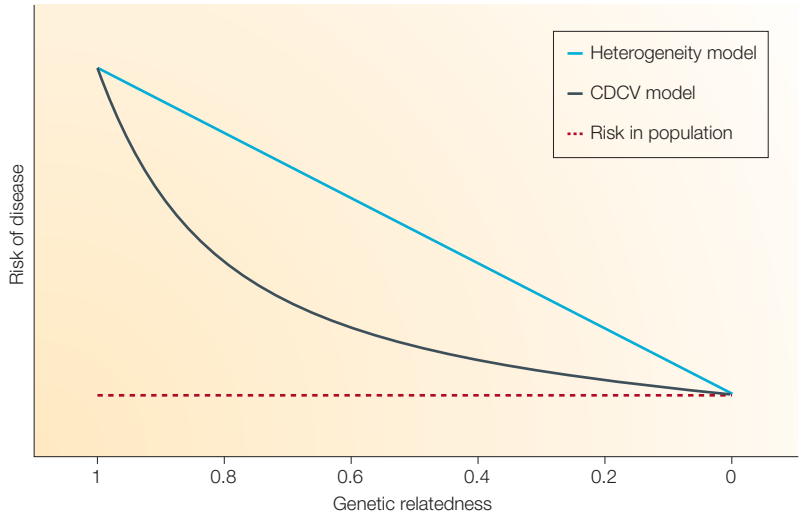

Figure 2 | **Models of the risks conferred by disease-associated variants.** The risk of disease as a function of genetic relatedness to affected individuals is shown. Two hypothetical common diseases are considered (blue and black lines), which have the same monozygotic risk (the risk of a monozygotic twin of a disease case also being affected by the disease; where genetic relatedness is 1) and the same underlying risk in the population (red dashed line). For the disease represented by the blue line, the risk of disease falls linearly with decreased genetic relatedness, consistent with disease heterogeneity, owing to the reduction in the number of shared rare alleles — the disease heterogeneity model. For the disease indicated by the black line, the fall in risk as a function of genetic relatedness is more rapid, as can occur when multiple, common, interacting alleles contribute to disease — an example of the common disease/common variant (CDCV) model.

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

QUANTITATIVE TRAIT LOCI
Genetic loci that contribute to variations in quantitative, that is continuous, phenotypes.

PURIFYING SELECTION
Evolutionary selective forces that reduce the frequency of specific polymorphisms that have phenotypic effects.

POSITIVE SELECTION
The effect of evolutionary selective forces that favour certain variants and tend to increase their allele frequencies.

Allele frequencies for susceptibility loci. Two polarized views have dominated much of the literature on the allelic frequencies of common diseases[9,21]. The common disease/common variant (CDCV) hypothesis proposes, as its name suggests, that common diseases are a result of common variants[20]. Under this model, disease susceptibility is suggested to result from the joint action of several common variants, and unrelated affected individuals share a significant proportion of disease alleles. The extreme alternative to CDCV is the classical disease heterogeneity hypothesis (or multiple rare-variant hypothesis), in which disease susceptibility is due to distinct genetic variants in different individuals and disease-susceptibility alleles have low population frequencies[26] (MAFs of less than 0.01).

The allelic spectra of most common diseases probably fall between these two extremes. The classical heterogeneity model, with multiple rare variants contributing additively and independently (in a biological sense), leads to correlations between traits in related subjects falling off linearly with the distance of the relationship between them[27] (FIG. 2). This is the result of linear reductions in sharing of disease alleles with the increasing distance of relationships. By contrast, if a common disease is largely due to the interdependent interactions of several loci with common alleles, the decline in risk with the degree of relatedness will be more rapid than a linear decline. Investigations of whether this correlation applies to different common diseases and traits have yielded different results, providing support for genetic additivity in some cancers[28] and in stature[29], and non-additivity in type 1 diabetes[30] (see Online links box).

Arguments used to support these two hypotheses have largely been based on population-genetic theories and will therefore be influenced by the underlying assumptions of these theories[20,31]. Empirical evidence indicates that both high- and low-frequency alleles contribute to common diseases[2,32–38]. For example, in a review of mapped QUANTITATIVE TRAIT LOCI (QTL), approximately 50% of the candidate causal variants had MAFs exceeding 0.05, whereas the other half had lower MAFs[9].

We suggest that it is preferable to avoid this polarization of rare versus common disease-susceptibility alleles, and instead consider the divergence of the allelic spectrum of disease variants from that of all variants (with or without phenotypic effects) in the human genome (FIG. 3). The most neutral hypothesis would be that the allelic spectrum of disease variants is the same as the general spectrum of all genetic variants[17,39,40]. Under this neutral model, although most susceptibility variants are rare (with MAFs of less than 0.01), SNPs with MAFs of greater than 0.01 would account for more than 90% of genetic differences between individuals and should contribute significantly to phenotypes[17,41]. Compared with the overall allelic spectrum, the CDCV model could be considered as a shift towards common variants and the heterogeneity model a shift towards rare variants[40] (FIG. 3). Protein-coding regions of the genome have polymorphisms with lower MAFs than the genome in general and, therefore, disease variants that cause non-synonymous changes[42,43] might contribute to a rare shift. Different evolutionary forces can result in different spectral shifts; for example, PURIFYING SELECTION might result in a rare shift[31]. By contrast, diseases that are mediated by immune responses, such as autoimmune disorders, might be caused by alleles that have been under POSITIVE SELECTION to provide resistance to infectious diseases and have therefore reached higher population frequencies[36]. Similarly, metabolic diseases such as type 2 diabetes (see Online links box), in which alleles are selected for adaptive responses to starvation or energy balance, might affect susceptibility in the modern environment — the thrifty gene hypothesis[44]. The allelic spectrum will therefore vary between different common diseases and is likely to consist of a complex mixture of allele frequencies[26,32], approximating the curved L-shaped distributions that are shown in FIG. 3 (note that the curves would be U-shaped if allele frequencies between 0 and 1.0 were represented, instead of 0 to 0.5 when considering only minor alleles).

For the genome as a whole, it has been predicted that of the expected 10 to 15 million SNPs with MAFs of greater than 0.01 (REFS 41,45), approximately half have MAFs of greater than 0.1, and the other half have MAFs that are between 0.01 and 0.1 Given that the number of disease variants conferring mild to moderate risks might be large (as explained in the next section), then unless shifts in allelic spectra are severe — which seems unlikely, given the multiplicity of genetic and environmental effects in common disease — there are likely to be hundreds of common and rare variants contributing to the familial clustering of each common human disease.

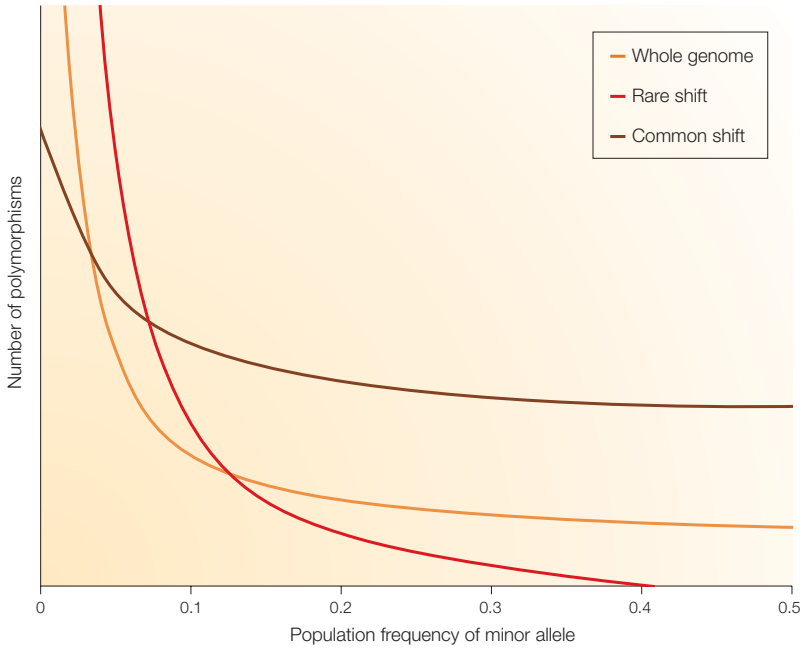

Figure 3 | **Possible allelic spectra of human diseases.** Three possible distributions of disease-associated variants with different population frequencies are shown. The orange line shows an allelic spectrum that is similar to that of the genome as a whole (that is, similar to that for all known variants, whether disease-associated or not). The red line shows a 'rare shift', with more rare disease-susceptibility variants and fewer common ones, leading to greater disease heterogeneity. The brown line shows a 'common shift', in which disease alleles tend to have high population frequencies. Modified, with permission, from REF. 40 © (2004) Elsevier Science.

As an example, using the hypothetical spectra in FIG. 3, consider a complex disease in which there are 20 disease-susceptibility variants contributing to that disease under the neutral model, in which MAFs of these variants are greater than 0.1 and their odds ratios are high enough for them to be identified in genome-wide association studies. In this case, a rare shift might result in ~10 variants with MAFs of greater than 0.1, and a common shift might result in ~40 variants. The implication for genome-wide association analyses is that experiments that are based on the existence of common variants are likely to yield a significant number of positive results unless there have been extreme shifts in the allelic spectra.

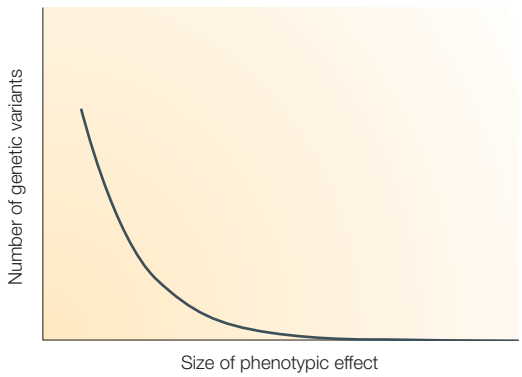

Figure 4 | **Putative distribution of phenotypic-effect sizes among disease-susceptibility variants.** A probable distribution of genetic variants that is based on an exponential distribution is the existence of a small number of variants with large effects and a large number of variants with small effects.

GENETIC DRIFT
Changes in allele frequencies in a population from one generation to another as the result of chance events in mating, meiosis and number of offspring.

SIBLING RELATIVE-RECURRENCE RISK
The risk of developing disease in a sibling of an affected individual relative to that of an individual in the general population. Commonly used as an indication of the heritability of a disease.

HAPLOTYPE
A set of alleles that is present on a single chromosome.

Risks associated with disease-susceptibility variants. The second main question concerning allelic architecture is the distribution of genetic risks conferred by individual variants. Although it is not possible to predict an accurate distribution of allelic effects for any given common disease, several lines of evidence point to potential underlying distributions. For example, such evidence has come from using mutagenesis, selection and linkage approaches in studies of QTLs in *Drosophila melanogaster*, crops and livestock and studies of rodent models of human disease. These studies have indicated that the distribution of phenotypic-effect sizes of genetic variants is consistent with the existence of few genetic loci with large effects and numerous loci with small effects[9,46–54]. The resulting curved, L-shaped distributions have been modelled by using either exponential or γ-distributions (see the graph in FIG. 4, which has a different shape and origin from the curves in FIG. 3). These results are consistent with current evolutionary theories in which, by factoring GENETIC DRIFT and mutational effects into classical models of adaptation[55], the expected distribution of QTL effects is exponential[56]. The potential for a large number of variants with small individual contributions to human phenotypes is further supported by recent findings that allelic variation frequently affects gene expression and exon splicing[57–60] — which is likely to have smaller effects than polymorphisms that affect the coding sequence — and that loci with alleles that affect the regulation of gene expression can be detected by linkage analyses[61,62].

Most irrefutable disease-susceptibility variants that have been identified so far — mainly from functional-candidate association studies — have allelic odds ratios that are in the order of 1.1–1.5 (REFS 1,2) and contribute little to familial recurrence risks[11,22,63]. For example, assuming a multiplicative model for the effects of alleles and interactions between loci, a disease-susceptibility allele with a frequency of 0.1 that confers a 1.5-fold increase in risk would be responsible for a SIBLING RELATIVE RECURRENCE RISK (λ_s) of less than 1.02, which if the overall λ_s was 5, would equate to a 1.2% contribution. It is not unreasonable to expect that QTLs would contribute effects of similar sizes to a quantitative trait. We do not know, however, if this is a representative range of effect sizes in common diseases, as only a small fraction of the genome has been evaluated in well-designed association studies (see, for example, the T1DBase database in the Online links box for genes studied in type 1 diabetes). Nevertheless, we believe that it would be unwise to undertake genome-wide association studies that do not have sufficient power to detect disease and quantitative-trait effects of this magnitude.

SNP choice in genome-wide association studies

To target the variation that occupies the range of MAFs of >0.1 and odds ratios of >1.3 in a statistically

Box 1 | **Applications of linkage-disequilibrium metrics**

Several metrics have been devised to measure linkage disequilibrium (LD)[96]. The two most commonly used of these are D' and r^2. Both are related to the basic unit of LD, D.

D

D measures the deviation of HAPLOTYPE frequencies from the equilibrium state[97]. LD occurs when D is significantly greater than zero. Consider two linked SNPs with alleles (A, a) and (B, b), resulting in four possible haplotypes: AB, Ab, aB and ab. D can be calculated as in equation 1, where $f(X)$ represents the frequency of the X allele or haplotype.

$$D = f(AB) - f(A)\,f(B) \qquad (1)$$

D′

D' is the absolute ratio of D compared with its maximum value, D_{max}, when $D \geq 0$, or compared with its minimal value, D_{min}, when $D < 0$ (REF. 98). $D' = 1$ denotes complete LD, and historical recombination results in the decay of D' towards zero.

r²

r^2 is the statistical coefficient of determination — a measurement of correlation between a pair of variables[99] (see equation 2).

$$r^2 = \frac{D^2}{f(A)\,f(a)\,f(B)\,f(b)} \qquad (2)$$

r^2 is of particular importance in genetic mapping as it is inversely related to the required sample size for association mapping, given a fixed genetic effect[65,66]. For example, if only one of a pair of SNPs was genotyped and r^2 between the SNPs was found to be 0.5, then to provide the same statistical power for the ungenotyped SNP compared with the case where $r^2 = 1$, the sample size would need to be doubled. When $r^2 = 1$, knowing the genotypes of alleles of one SNP is directly predictive of the genotypes of another SNP. The alternative notation R^2 is used when individual variables are predicted using the multiple regression of a constellation of other variables[67,70].

Relationship between D' and r^2

D' and r^2 can be written in terms of each other and allele frequencies. Without losing generality, the four alleles can be chosen such that $D \geq 0$ and $f(A) \geq f(B)$. So D' and D_{max} have the relations in equations 3 and 4.

$$D' = \frac{D}{D_{max}} \qquad (3)$$

$$D_{max} = f(a)\,f(B) \qquad (4)$$

and

$$r^2 = (D')^2 \times \frac{f(a)\,f(B)}{f(A)\,f(b)} \qquad (5)$$

Equation 5 shows the relationship between D', r^2 and allele frequencies. As $f(A) \geq f(B)$, r^2 has the upper bound of $(D')^2$, and reaches it only when $f(A) = f(B)$. The implication of this is that D', a commonly used measure of historical recombination, provides information on the physical extent of useful LD (in terms of association mapping and statistical power) by providing the upper limit of r^2. Dense LD maps that are based on high-frequency SNPs (MAF >0.1) can reveal regions of historical recombination. Knowing the level of D' decay in these maps directly provides the maximum potential level of useful LD in association mapping (based on r^2) for high-frequency SNPs, even if a significant proportion of common SNPs remains undiscovered. For example, if a recombination point resulted in a D' of 0.7 for SNPs on either side of it, the maximum possible r^2 for these SNPs would be 0.49, and sample sizes would need to be more than doubled to maintain the same statistical power for association mapping. It should be noted that both D' and r^2 suffer from sampling biases given a small number of individuals and for rare variants[15,68,100]. Confidence intervals for D' have been used by some investigators[15].

powerful way, we need to know all the common variants in the population that the cases and controls are taken from. Although there has been a rapid recent increase in our knowledge of human genome variation[17] — mostly in the form of SNPs — as many as 30% of common variants might remain undetected. This can be corrected by further genome resequencing for a larger set of unrelated individuals (discussed in a later section). Nevertheless, even without this resequencing, because we know that many SNPs have alleles that show strong LD with other nearby SNP alleles (the average range over which SNPs show LD is 60–200 kb in general populations[18,64]), it might be possible to carry out reasonably comprehensive genome-wide association studies that are based on known variants. We can predict that in regions of the genome with strong LD, a selection of evenly spaced SNPs, or those chosen on the basis of their LD with other SNPs (tag SNPs), can provide adequate 'coverage' of the region in an association study. In the following sections we define these terms and describe the benefits that LD provides for genome-wide association studies in the absence of both a complete SNP map and a technology that can type all known SNPs in an affordable way.

Tag SNPs. The degree of LD between alleles at two loci can be described in terms of the metric r^2 (BOX 1). r^2 is informative in association analyses because it is inversely proportional to the sample size that is required for detecting disease association given a fixed genetic risk[65,66]. For example, consider a genotyped marker SNP that is near to a susceptibility locus that is also a SNP, but which is not itself typed in an association study. If the r^2 between these two loci is 0.5, then the effective sample size for this marker, which determines the statistical power of the association study (FIG. 1), is halved (that is, the actual sample size would need to be doubled for the same statistical power). This leads to a large reduction in power from more than 90% to less than 40% for a study of 6,000 cases and 6,000 controls for an MAF of 0.1 to obtain a P value in the order of 10^{-6}. By contrast, an r^2 of 1 indicates perfect LD, and there is no loss of power when using a marker tag SNP instead of directly genotyping the disease causal varient.

Therefore, the general consensus is that an r^2 of 0.8 or greater is sufficient for tag SNP mapping to obtain a good coverage of untyped SNPs, allowing genotyping of a lower number of marker SNPs with relatively small losses in power. So, if we knew the identity of all the common SNPs in a region and there was LD between them, then by iterative pair-wise comparisons, the optimal set of tags could be chosen. Here, 'optimal' is defined as the smallest number of SNPs that needs to be genotyped to cover the other SNPs at an r^2 of 0.8 or greater[19,45,67–69]. If the LD between SNPs is strong, this could result in the need to carry out up to 70–80% less genotyping. However, if LD in a region is low, almost every SNP might have to be genotyped to ensure comprehensive coverage of the region. For example, the interferon β-1 fibroblast (*IFNB1*) gene is only 1 kb in length but requires 17 tag SNPs to cover the 19 common

SNPs present[70]. Furthermore, the well-studied common SNP *Fok*I T>C (rs10735810) in the vitamin D receptor gene is not in strong LD with any other SNP in the flanking LD blocks and needs to genotyped directly in any disease-association study: it is an obligatory member of the tag set[71].

Breaks in LD occur on average about once every 200 kb in the genome, although there is wide variability in the lengths of regions of high LD[15,18]. LD blocks can be located using the other commonly used LD metric, D', which is closely related to r^2 and provides information about the recombination breakpoints of chromosomes (BOX 1). This is one of the main outputs of the International HapMap Project (see Online links box), in which 270 DNA samples (mostly from unrelated individuals) are being genotyped for several million SNPs across the genome. The SNP map has recently been supplemented by the release of data by Perlegen Sciences, Inc. (see Online links box) for almost 1.5 million SNPs, which will be included in the HapMap project.

Regions of low LD might well be regions of intense homologous recombination and GENE CONVERSION, causing the scrambling of the association of alleles between loci — and so the reduction in LD — at a faster rate than regions with less recombination[18,72–74]. The continuing HapMap project and several recent additional studies indicate that approximately 70–80% of the genome has regions of high LD[15,16,18,75], which has significant implications for genome-wide association studies[3,17]. These patches of stronger LD that have high D' values between SNPs — and the mapping redundancy that this results in — counterbalance, to a certain extent, the current incompleteness of the SNP map (see below) and the fact that the new high-throughput technologies convert only about 50% of SNPs into robust assays[76].

Implications of an incomplete SNP map. The calculations of sample sizes needed for adequate statistical power that are shown in FIG. 1 represent an ideal situation. In a genome-wide association study there will be significant losses in power and gaps in the map of all the SNPs that are covered if some SNPs are in weak LD (r^2 <0.5) with the marker SNPs that can be genotyped. Even within an LD block with a high level of D' between markers, there will often be SNPs that have low r^2 values with other SNPs within the block. Despite the intimate relationship between them (BOX 1), r^2 cannot be estimated based on D' and MAFs alone. For r^2 values to be high, the alleles of two SNPs should not only have similar frequencies, but also need to be correlated and to occur on the same ancestral haplotype. Haplotypes for a chromosome region can be visualized as a tree, with the different branches representing ancestral haplotypes, differing from each other owing to recombination, gene conversion and mutation. Minor alleles with similar frequencies that are found on different haplotypes have low r^2 values[77]. So, for a complete SNP map with contiguous coverage (with every SNP or other polymorphism having an r^2 of at least 0.8 with a SNP that can be genotyped robustly), it is necessary to obtain near-complete genome sequences from a sufficiently large number of unrelated individuals.

Towards a complete SNP map. The HapMap, including the 2005 version with the Perlegen SNPs added, originates from non-contiguous resequencing of 5–60 different chromosomes. Previously, we simulated the sampling of common SNPs (with MAFs of ≥0.1) from 73 existing near-complete SNP maps of specific genomic regions[68] (compiled by the University of Washington and Fred Hutchinson Cancer Research Center Variation Discovery Resource database (see Online links box)). This was done by near-contiguous resequencing, using a PCR-based method, for DNA samples from 47 unrelated individuals[45]. Simulated incomplete SNP maps of 1 SNP per 2.5 kb, 5 kb and 10 kb ascertained approximately 75%, 50% and 40% of the total underlying SNP variation, in corresponding order (median values)[68]. However, there were large variations in the SNP densities that were required to uncover all genetic variants for different genes, and for 7 of the 73 genes there was less than 50% SNP coverage for SNP maps of 1 SNP per 2.5 kb. To obtain 80% SNP coverage, sampling at densities of more than 1 SNP per kb was required in 10–20% of genomic regions. Similar conclusions have been reached by other studies[17,45,70,78,79].

We estimate that the coverage provided by the HapMap will reach at least 1 SNP per 5 kb by the end of 2005, and that about 300,000 tag SNPs will need to be genotyped to cover 50% of the genome at an r^2 of 0.8 or greater. However, even with an average density of 1 SNP every 2.5 kb, ~25% of SNPs will still not be captured adequately by LD and tag SNPs[68]. To tag most variants with MAFs of >0.1, as many as 500,000 more SNPs might be required to cover in a comprehensive fashion the remaining 25% of the genome that shows lower LD. Nevertheless, as few as 75,000 tag SNPs might well cover 25% of the genome (in regions of high LD), a vast increase in efficiency over a candidate gene-by-gene approach. Further cost-efficiency and power could be gained by choosing not only the regions of highest LD, but also those regions that contain the highest density of genes.

Whereas the initial selection of tags is relatively straightforward (although there are now many methods, they will all probably give similar results), the assessment of the statistical association of tags in a disease-association study requires further research. It might be that for some methods of tag selection and analysis, statistical power can also be saved by grouping tags into the LD blocks that are defined by D' patterns across the genome.

In BOX 2, we estimate that the number of human chromosomes, if resequenced in a contiguous fashion, would provide a complete map for common variants is in the order of 60. With current sequencing technology, this is a massive task, but emphasis could be placed on resequencing regions of lower LD, as detected by D' (perhaps 20% of the genome).

GENE CONVERSION
A non-reciprocal recombination process that results in an alteration of the sequence of a gene to that of its homologue during meiosis.

Box 2 | **Genome resequencing for full coverage in genome-wide association studies**

The optimal number of individuals that should be initially resequenced to provide contiguous and reliable linkage disequilibrium (LD) data for tag SNP selection is unclear. There is an obvious trade off between the laboratory effort required for resequencing and the reliability of the data: some fragments of the genome are difficult to amplify by PCR, especially G+C-rich 5′ regions of genes and common repeat-rich regions, and automated detection of SNPs is still not possible using currently available di-deoxy-sequencing chemistries.

No. of individuals in training set	R^2 <0.6 (%)	R^2 <0.7 (%)
16	5.3	9.1
20	3.0	5.7
23	1.7	3.8
32	0.6	1.7
48	0.1	0.5
96	0.0	0.0

We undertook resampling of real data to examine the reliability of the allelic R^2 method of Chapman and colleagues[67], in which the correlation for tag selection is based on multiple comparisons (R^2), rather than pairwise comparisons (r^2) (see BOX 1). For each of five regions used in previous LD simulations[68], a central contiguous genomic region containing 28 SNPs was chosen. In each trial, a smaller number or subset of individuals was randomly sampled from the complete set, and this provided information for tag SNP selection. Set sizes that were considered comprised 16, 20, 23, 32, 48 and 96 individuals, and 1,000 trials were carried out for each set size in each genomic region. For the selection of tag SNPs, an allelic R^2 cut-off of 0.8 was used[67]. The performance of the tag SNPs that were selected from the test sets was then evaluated on the basis of their ability to provide information about the whole set of SNPs. Our aim was to select tag SNPs with $R^2 \geq 0.8$ for all remaining SNPs; the average percentage of SNPs over the five genomic regions that failed to achieve R^2 of 0.6 and 0.7 in the population are shown in the table.

These results indicate that tag SNP selection requires resequencing for only a relatively small number of individuals: 32 individuals, or 64 chromosomes, might be sufficient to provide $R^2 \geq 0.8$ for more than 98% of cases, and this is consistent with previous studies for different tag SNP methods[45,68,101]. For training sets of 32 or more individuals, variation between the five genes in the level of adequate tagging was minor. However, SNPs with MAFs of <0.1 were in general more difficult to tag than common SNPs, and variations between genes can be problematic when 23 or fewer individuals are used. It is also unclear how many individuals would be required in non-European populations, although greater haplotype diversity in African populations indicates that larger sample sizes would be required[15,75].

Loss of cost efficiency by other means

In addition to the considerations described above, there are several other ways that the idealized estimated requirements for obtaining adequate statistical power that are shown in FIG. 1 might be reduced in a genome-wide association study. Some of these are discussed below.

Epistasis and subgroup analyses. Epistasis, which refers to gene–gene interactions, was originally a mechanistic and deterministic idea — William Bateson described it as one gene cancelling out the effect of another[80]. In seeking to generalize this to quantitative traits, the concept was extended to be synonymous with the statistical definition of 'interaction', which simply means non-additivity of effects that are measured on some specified scale. Although epidemiologists once believed that testing for statistical interaction would be informative for biological mechanisms, this approach has not been productive. There are two reasons for this; first, the power to detect interactions is often low; and second, even if detected, interpretation is difficult as many biological explanations could be possible[81,82].

Nevertheless, statistical interaction might be relevant to our ability to detect phenotype–genotype associations. As a possible explanation of the small effect sizes reported so far for common-disease-susceptibility loci, some authors have suggested that mathematical models describing extreme forms of epistasis are possible in which a genetic variant has no overall effect on its own, despite having strong effects within certain subgroups of the population that are defined by variants in other genes[83,84]. However, such extreme scenarios require interactions that are even stronger than those predicted by Bateson — with one gene reversing the effect of another — and this is unlikely to be a widespread phenomenon. In less extreme situations, taking into account possible statistical interactions with other genes could increase the power to detect a novel causal variant[85]. However, the need to protect against false positives that are due to SUBGROUP ANALYSES means that the power gain might be relatively modest in the presence of interactions and, of course, must be set against the inevitable loss of power when interactions are very small or absent. The prior probability that any given locus or region is truly associated with disease becomes even lower. This remains an area of some controversy, as does the closely related question of the relevance of gene–environment interactions to the discovery of genetic or environmental causes of disease[86]. In our view, consistent with that of Hirschhorn and Daly[3], the presence of epistasis places even greater pressure on the collection of well-defined, large samples, and on the necessity of replication due to the consequent increase in the subgroup analysis problem that occurs in the analysis of higher-order interactions.

SUBGROUP ANALYSES
In genome-wide association analyses, or any other association study, there is a very low prior probability that any given locus or region is associated with disease. If the samples or data are divided into subgroups, for example, in analysis of epistatic interactions between loci — a departure from statistical independence in the joint distributions of genotypes between the loci — then the prior probability of a true positive is even lower.

POPULATION STRATIFICATION
The presence of several population subgroups that show limited interbreeding. When such subgroups differ both in allele frequency and in disease prevalence, this can lead to erroneous results in association studies.

ADMIXTURE
The mixture of two or more genetically distinct populations.

ANCESTRY INFORMATIVE MARKERS
Genetic markers that have different frequencies between populations and can be used to readily estimate the ancestral origins of a person or population.

COHORT STUDIES
Observational studies in which defined groups of people (the cohorts) are followed over time and outcomes are compared in subsets of the cohort who were exposed to different levels of factors of interest. These studies can either be performed prospectively or retrospectively from historical records.

GENOMIC CONTROL
A statistical genetics approach that provides an adjustment of the chi-squared threshold for statistical significance in a genetic association study to help allow for population sub-structure effects.

Population substructure and other sources of error. Several problems in study design can lead to both false-positives and false-negatives in association analyses[3]. In the past decade, case-control studies and family-based association studies have emerged as the two main strategies in association analyses of common diseases[3]. Although cases and controls are generally more powerful and logistically easier to collect, they can suffer from hidden POPULATION STRATIFICATION and ADMIXTURE[87–89]; however, the magnitude of such effects remains unclear[3]. Differences in DNA quality and less-than-optimal genotyping can also lead to increased false-positive rates in both family-based and population-based studies. With the large sample sizes that we now believe to be necessary, such influences become more pronounced. Close assessment of technical difficulties — taking into account population substructure and admixture effects using ANCESTRY-INFORMATIVE MARKERS (AIMs)[90–92], replication studies in different populations, and the use of both case-control and family-based studies (with *P*-values of less than 10^{-6} obtained in at least one dataset) — will be necessary to establish irrefutable and accurately quantified evidence of genetic association. It remains to be quantified empirically, by genotyping large numbers of SNPs, how significant the effects of population substructure are. This will vary both for different populations and depending on how closely cases and controls in any particular country can be matched according to geographical subregions. AIMs for a wide variety of populations and ancestral groups should be identified, given the possibility of important effects owing to substructure even in geographically matched groups.

Selection bias. Selection bias in case-control studies arises when the case and control groups are not truly comparable. Ideally, the controls would be drawn from the same population as the cases — known as the 'study base' — and would be subject to the same selection biases. In practice, it is only possible to approach this ideal in case-control studies that are nested in COHORT STUDIES and these are rarely, if ever, sufficiently large for the purposes discussed in this review. A more realistic aim is that controls should be drawn from a population that is sufficiently similar to the study base to allow the relevant allele and haplotype frequencies to be reliably estimated.

Selection bias occurs when the control frequencies differ systematically from those in the study base. We should discriminate between selection bias that is due to causal effects, such as effects on personality leading to 'volunteer bias', which might affect a limited number of loci, and more general effects that are due to differences in population substructure between controls and the study base. These more general biases can be addressed, as described earlier, by the use of AIMs and by GENOMIC CONTROL[90,93].

A powerful protection against being misled by selection bias is the use of multiple groups. A design in which several disease groups are compared against one or two control groups allows the consistency of findings to be examined, and this strengthens the inference. An example of this is Doll and Hill's[94] classical study of the association between smoking and lung cancer, which used two different comparison groups — comprising patients suffering from two different diseases; Doll[95] cited the similarity of smoking rates in the control groups as strengthening the evidence for a causal effect on lung cancer.

Future perspectives

Despite the caveats outlined above, it seems that genome-wide association studies of the role of common variants in complex disease will be carried out in the near future. Initial studies will define more accurately the principal factors, which have been summarized above, that can reduce the power of such studies. In these studies, large sample sizes should be used, biases taken into account, multiple-testing issues addressed and replication studies carried out, therefore optimizing experimental design, statistical power and cost efficiency. Close evaluation of the yields of true susceptibility loci in relation to the cost of such rigorously designed studies will determine whether the genome-wide analyses of common SNPs is a worthwhile approach in the continuing dissection of the genetic basis of common disease.

1. Ioannidis, J. P., Trikalinos, T. A., Ntzani, E. E. & Contopoulos-Ioannidis, D. G. Genetic associations in large versus small studies: an empirical assessment. *Lancet* **361**, 567–571 (2003).
2. Lohmueller, K. E., Pearce, C. L., Pike, M., Lander, E. S. & Hirschhorn, J. N. Meta-analysis of genetic association studies supports a contribution of common variants to susceptibility to common disease. *Nature Genet.* **33**, 177–182 (2003).
3. Hirschhorn, J. N. & Daly, M. J. Genome-wide association studies for common diseases and complex traits. *Nature Rev. Genet.* **6**, 95–108 (2005).
A review of the issues that are involved in the design of large-scale association mapping, including marker selection and sources of false-positive and false-negative results.
4. Livak, K. J., Marmaro, J. & Todd, J. A. Towards fully automated genome-wide polymorphism screening. *Nature Genet.* **9**, 341–342 (1995).
5. Patil, N. *et al.* Blocks of limited haplotype diversity revealed by high-resolution scanning of human chromosome 21. *Science* **294**, 1719–1723 (2001).
6. Syvanen, A. C. Accessing genetic variation: genotyping single nucleotide polymorphisms. *Nature Rev. Genet.* **2**, 930–942 (2001).
7. Miller, R. D., Duan, S., Lovins, E. G., Kloss, E. F. & Kwok, P. Y. Efficient high-throughput resequencing of genomic DNA. *Genome Res.* **13**, 717–720 (2003).
8. Hardenbol, P. *et al.* Multiplexed genotyping with sequence-tagged molecular inversion probes. *Nature Biotechnol.* **21**, 673–678 (2003).
9. Blangero, J. Localization and identification of human quantitative trait loci: King Harvest has surely come. *Curr. Opin. Genet. Dev.* **14**, 233–240 (2004).
10. Terwilliger, J. D. & Weiss, K. M. Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann. Med.* **35**, 532–544 (2003).
11. Wang, W. Y., Cordell, H. J. & Todd, J. A. Association mapping of complex diseases in linked regions: estimation of genetic effects and feasibility of testing rare variants. *Genet. Epidemiol.* **24**, 36–43 (2003).
12. Stefansson, H., Steinthorsdottir, V., Thorgeirsson, T. E., Gulcher, J. R. & Stefansson, K. Neuregulin 1 and schizophrenia. *Ann. Med.* **36**, 62–71 (2004).
13. Altmuller, J., Palmer, L. J., Fischer, G., Scherb, H. & Wjst, M. Genomewide scans of complex human diseases: true linkage is hard to find. *Am. J. Hum. Genet.* **69**, 936–950 (2001).
This is an analyses of 101 linkage studies. It demonstrates the difficulties in achieving significant linkage, and argues for a need for larger sample sizes.
14. Neale, B. M. & Sham, P. C. The future of association studies: gene-based analysis and replication. *Am. J. Hum. Genet.* **75**, 353–362 (2004).
A review of the design of association-mapping strategies. It argues for changing the focus from SNPs to genomic regions, and outlines strategies to achieve this.
15. Gabriel, S. B. *et al.* The structure of haplotype blocks in the human genome. *Science* **296**, 2225–2229 (2002).
16. Dawson, E. *et al.* A first-generation linkage disequilibrium map of human chromosome 22. *Nature* **418**, 544–548 (2002).
17. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
This paper outlines the International HapMap Project, which is currently in progress, and will provide SNP maps, LD information and tag SNPs throughout the genome for different human populations.

18. McVean, G. A. *et al.* The fine-scale structure of recombination rate variation in the human genome. *Science* **304**, 581–584 (2004).
19. Johnson, G. C. *et al.* Haplotype tagging for the identification of common disease genes. *Nature Genet.* **29**, 233–237 (2001).
The authors introduce the concept of tag SNPs based on LD to minimize laboratory effort for SNP genotyping in association analyses.
20. Reich, D. E. & Lander, E. S. On the allelic spectrum of human disease. *Trends Genet.* **17**, 502–510 (2001).
21. 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).
22. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
This paper showed in explicit terms the greater power of whole-genome association studies over affected sib-pair linkage for the mapping of common diseases.
23. Dahlman, I. *et al.* Parameters for reliable results in genetic association studies in common disease. *Nature Genet.* **30**, 149–150 (2002).
24. Freimer, N. & Sabatti, C. The use of pedigree, sib-pair and association studies of common diseases for genetic mapping and epidemiology. *Nature Genet.* **36**, 1045–1051 (2004).
A clear and unbiased review of the main current genetic mapping strategies that discusses analyses using extended pedigrees, affected sib-pairs and association.
25. Lowe, C. E. *et al.* Cost-effective analysis of candidate genes using htSNPs: a staged approach. *Genes Immun.* **5**, 301–305 (2004).
26. Smith, D. J. & Lusis, A. J. The allelic structure of common disease. *Hum. Mol. Genet.* **11**, 2455–2461 (2002).
27. Fisher, R. A. Correlation between relatives on the supposition of Mendelian inheritance. *Trans. R. Soc. Edinb.* **52**, 399–433 (1918).
28. Risch, N. The genetic epidemiology of cancer: interpreting family and twin studies and their implications for molecular genetic approaches. *Cancer Epidemiol. Biomarkers Prev.* **10**, 733–741 (2001).
29. Hirschhorn, J. N. *et al.* Genomewide linkage analysis of stature in multiple populations reveals several regions with evidence of linkage to adult height. *Am. J. Hum. Genet.* **69**, 106–116 (2001).
30. Rich, S. S. Mapping genes in diabetes. Genetic epidemiological perspective. *Diabetes* **39**, 1315–1319 (1990).
31. Pritchard, J. K. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**, 124–137 (2001).
32. Todd, J. A. Human genetics. Tackling common disease. *Nature* **411**, 537–539 (2001).
33. Cohen, J. C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
34. Corder, E. H. *et al.* Gene dose of apolipoprotein E type 4 allele and the risk of Alzheimer's disease in late onset families. *Science* **261**, 921–923 (1993).
35. Bell, G. I., Horita, S. & Karam, J. H. A polymorphic locus near the human insulin gene is associated with insulin-dependent *Diabetes mellitus*. *Diabetes* **33**, 176–183 (1984).
36. Ueda, H. *et al.* Association of the T-cell regulatory gene *CTLA4* with susceptibility to autoimmune disease. *Nature* **423**, 506–511 (2003).
37. Hugot, J. P. *et al.* Association of NOD2 leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603. (2001).
38. Ogura, Y. *et al.* A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
39. Long, A. D. & Langley, C. H. The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res.* **9**, 720–731 (1999).
40. Wang, W. Y. & Pike, N. The allelic spectra of common diseases may resemble the allelic spectrum of the full genome. *Med. Hypotheses* **63**, 748–751 (2004).
41. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27**, 234–236 (2001).
Using a neutral coalescence model, this article estimates the frequency distribution of SNPs in the human genome.
42. 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).
43. Clark, A. G. Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr. Opin. Genet. Dev.* **13**, 296–302 (2003).
44. Neel, J. V. *Diabetes mellitus*: a 'thrifty' genotype rendered detrimental by 'progress'? *Am. J. Hum. Genet.* **14**, 353–362 (1962).
45. Carlson, C. S. *et al.* Additional SNPs and linkage-disequilibrium analyses are necessary for whole-genome association studies in humans. *Nature Genet.* **33**, 518–521 (2003).
46. Nezer, C. *et al.* Haplotype sharing refines the location of an imprinted quantitative trait locus with major effect on muscle mass to a 250-kb chromosome segment containing the porcine *IGF2* gene. *Genetics* **165**, 227–285 (2003).
47. Vyse, T. J. & Todd, J. A. Genetic analysis of autoimmune disease. *Cell* **85**, 311–318 (1996).
48. Robertson, A. in *Population Biology and Evolution* (ed. Lewontin, R. C.) 265–280 (Syracuse Univ. Press, New York, 1967).
49. Paterson, A. H. *et al.* Mendelian factors underlying quantitative traits in tomato: comparison across species, generations, and environments. *Genetics* **127**, 181–197 (1991).
50. Mackay, T. F., Lyman, R. F. & Jackson, M. S. Effects of *P* element insertions on quantitative traits in *Drosophila melanogaster*. *Genetics* **130**, 315–332 (1992).
51. Hayes, B. & Goddard, M. E. The distribution of the effects of genes affecting quantitative traits in livestock. *Genet. Sel. Evol.* **33**, 209–229 (2001).
52. Barton, N. H. & Keightley, P. D. Understanding quantitative genetic variation. *Nature Rev. Genet.* **3**, 11–21 (2002).
53. Wright, A., Charlesworth, B., Rudan, I., Carothers, A. & Campbell, H. A polygenic basis for late-onset disease. *Trends Genet.* **19**, 97–106 (2003).
54. Risch, N., Ghosh, S. & Todd, J. A. Statistical evaluation of multiple-locus linkage data in experimental species and its relevance to human studies: application to nonobese diabetic (*NOD*) mouse and human insulin-dependent *Diabetes mellitus* (*IDDM*). *Am. J. Hum. Genet.* **53**, 702–714 (1993).
55. Fisher, R. A. *The Genetical Theory of Natural Selection* (Oxford Univ. Press, Oxford, 1930).
56. Orr, H. A. The population genetics of adaptation: the distribution of factors fixed during adaptive evolution. *Evolution* **52**, 935–949 (1998).
57. Pagani, F. & Baralle, F. E. Genomic variants in exons and introns: identifying the splicing spoilers. *Nature Rev. Genet.* **5**, 389–396 (2004).
58. Hoogendoorn, B. *et al.* Functional analysis of human promoter polymorphisms. *Hum. Mol. Genet.* **12**, 2249–2254 (2003).
59. Lo, H. S. *et al.* Allelic variation in gene expression is common in the human genome. *Genome Res.* **13**, 1855–1862 (2003).
60. Mira, M. T. *et al.* Susceptibility to leprosy is associated with *PARK2* and *PACRG*. *Nature* **427**, 636–640 (2004).
61. Morley, M. *et al.* Genetic analysis of genome-wide variation in human gene expression. *Nature* **430**, 743–747 (2004).
62. Kleinjan, D. A. & van Heyningen, V. Long-range control of gene expression: emerging mechanisms and disruption in disease. *Am. J. Hum. Genet.* **76**, 8–32 (2005).
63. Rybicki, B. A. & Elston, R. C. The relationship between the sibling recurrence-risk ratio and genotype relative risk. *Am. J. Hum. Genet.* **66**, 593–604 (2000).
64. Jorde, L. B. Linkage disequilibrium and the search for complex disease genes. *Genome Res.* **10**, 1435–1444 (2000).
65. Sham, P. C., Cherny, S. S., Purcell, S. & Hewitt, J. K. Power of linkage versus association analysis of quantitative traits, by use of variance-components models, for sibship data. *Am. J. Hum. Genet.* **66**, 1616–1630 (2000).
66. Pritchard, J. K. & Przeworski, M. Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* **69**, 1–14 (2001).
67. 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).
This paper examines analyses of tag SNPs and suggests that it might be best to discard haplotype information and consider only the main effects of tag SNPs to avoid losing power owing to increased degrees of freedom.
68. Wang, W. Y. & Todd, J. A. The usefulness of different density SNP maps for disease association studies of common variants. *Hum. Mol. Genet.* **12**, 3145–3149 (2003).
Based on sampling simulations of published, near-complete SNP maps, this study assesses the usefulness of different density SNP maps for LD mapping.
69. Ke, X. *et al.* The impact of SNP density on fine-scale patterns of linkage disequilibrium. *Hum. Mol. Genet.* **13**, 577–588 (2004).
70. Clayton, D., Chapman, J. & Cooper, J. Use of unphased multilocus genotype data in indirect association studies. *Genet. Epidemiol.* **27**, 415–428 (2004).
71. Nejentsev, S. *et al.* Comparative high-resolution analysis of linkage disequilibrium and tag single nucleotide polymorphisms between populations in the vitamin D receptor gene. *Hum. Mol. Genet.* **13**, 1633–1639 (2004).
72. Jeffreys, A. J., Kauppi, L. & Neumann, R. Intensely punctate meiotic recombination in the class II region of the major histocompatibility complex. *Nature Genet.* **29**, 217–222 (2001).
73. Twells, R. C. *et al.* Haplotype structure, LD blocks, and uneven recombination within the *LRP5* gene. *Genome Res.* **13**, 845–855 (2003).
74. Jeffreys, A. J. & May, C. A. Intense and highly localized gene conversion activity in human meiotic crossover hot spots. *Nature Genet.* **36**, 151–156 (2004).
75. Wall, J. D. & Pritchard, J. K. Haplotype blocks and linkage disequilibrium in the human genome. *Nature Rev. Genet.* **4**, 587–597 (2003).
76. Pask, R. *et al.* Investigating the utility of combining Φ29 whole genome amplification and highly multiplexed single nucleotide polymorphism BeadArray genotyping. *BMC Biotechnol.* **4**, 15 (2004).
77. Cordell, H. J. & Clayton, D. G. Genetic association studies. *Lancet* (in the press).
78. Carlson, C. S. *et al.* Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *Am. J. Hum. Genet.* **74**, 106–120 (2004).
79. Ke, X. *et al.* Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Hum. Mol. Genet.* **13**, 2557–2565 (2004).
80. Bateson, W. *Mendel's Principles of Heredity* (Cambridge Univ. Press, Cambridge, 1909).
81. Thompson, W. D. Effect modification and the limits of biological inference from epidemiologic data. *J. Clin. Epidemiol.* **44**, 221–232 (1991).
82. Cordell, H. J. Epistasis: what it means, what it doesn't mean, and statistical methods to detect it in humans. *Hum. Mol. Genet.* **11**, 2463–2468 (2002).
83. Culverhouse, R., Suarez, B. K., Lin, J. & Reich, T. A perspective on epistasis: limits of models displaying no main effect. *Am. J. Hum. Genet.* **70**, 461–471 (2002).
84. Thornton-Wells, T. A., Moore, J. H. & Haines, J. L. Genetics, statistics and human disease: analytical retooling for complexity. *Trends. Genet.* **20**, 640–647 (2004).
85. Hoh, J. & Ott, J. Mathematical multi-locus approaches to localizing complex human trait genes. *Nature Rev. Genet.* **4**, 701–709 (2003).
86. Clayton, D. & McKeigue, P. M. Epidemiological methods for studying genes and environmental factors in complex diseases. *Lancet* **358**, 1356–1360 (2001).
87. Pato, C. N., Macciardi, F., Pato, M. T., Verga, M. & Kennedy, J. L. Review of the putative association of dopamine D2 receptor and alcoholism: a meta-analysis. *Am. J. Med. Genet.* **48**, 78–82 (1993).
88. Freedman, M. L. *et al.* Assessing the impact of population stratification on genetic association studies. *Nature Genet.* **36**, 388–393 (2004).
89. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nature Genet.* **36**, 512–517 (2004).

90. Pritchard, J. K. & Rosenberg, N. A. Use of unlinked genetic markers to detect population stratification in association studies. *Am. J. Hum. Genet.* **65**, 220–228 (1999).
91. Hoggart, C. J. *et al.* Control of confounding of genetic associations in stratified populations. *Am. J. Hum. Genet.* **72**, 1492–1504 (2003).
92. Marchini, J., Cardon, L. R., Phillips, M. S. & Donnelly, P. Reply to 'Genomic control to the extreme'. *Nature Genet.* **36**, 1131 (2004).
93. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
94. Doll, R. & Hill, A. B. The mortality of doctors in relation to their smoking habits. *BMJ* **228**, 1451–1455 (1954).
95. Doll, R. *Retrospective and Prospective Studies* (ed. Witts, L. J.) (Oxford Univ. Press, London, 1959).
96. Devlin, B. & Risch, N. A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* **29**, 311–322 (1995).
97. Lewontin, R. C. & Kojima, K. The evolutionary dynamics of complex polymorphisms. *Evolution* **14**, 458–472 (1960).
98. Lewontin, R. C. The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* **49**, 49–67 (1964).
99. Hill, W. G. & Robertson, A. The effects of inbreeding at loci with heterozygote advantage. *Genetics* **60**, 615–628 (1968).
100. Weiss, K. M. & Clark, A. G. Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* **18**, 19–24 (2002).
101. Thompson, D., Stram, D., Goldgar, D. & Witte, J. S. Haplotype tagging single nucleotide polymorphisms and association studies. *Hum. Hered.* **56**, 48–55 (2003).
102. Wall, J. D. & Pritchard, J. K. Assessing the performance of the haplotype block model of linkage disequilibrium. *Am. J. Hum. Genet.* **73**, 502–515 (2003).
A review on haplotype blocks and LD in the human genome.
103. Thomas, D. C. & Clayton, D. G. Betting odds and genetic associations. *J. Natl Cancer Inst.* **96**, 421–423 (2004).
104. Wacholder, S. *et al.* Assessing the probability that a positive report is false: an approach for molecular epidemiology studies. *J. Natl Cancer Inst*. **96**, 434–442 (2004).

Acknowledgements

W.Y.S.W. received scholarships from the University of Cambridge, the University of Sydney and Gonville and Caius College, Cambridge, UK. This work was financed by the Wellcome Trust and the Juvenile Diabetes Research Foundation International.

Competing interests statement

The authors declare no competing financial interests.

Online links

DATABASES

The following terms in this article are linked online to:

OMIM: http://www.ncbi.nlm.nih.gov/Omim/
Type 1 diabetes | type 2 diabetes

FURTHER INFORMATION

David Clayton's tag SNP web site: http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP.pkg
International HapMap Project: http://www.hapmap.org
NCBI Single Nucleotide Polymorphism database web site: http://www.ncbi.nlm.nih.gov/projects/SNP
Perlegen Sciences, Inc.: http://www.els.net
TD1Base — a genetics and bioinformatics resource for type 1 diabetes researchers: http://www.t1dbase.org/cgi-bin/welcome.cgi
University of Washington and Fred Hutchinson Cancer Research Center Variation Discovery Resource database: http://pga.gs.washington.edu
Access to this interactive links box is free online.