European Journal of Human Genetics (2006) 14, 426–437

www.nature.com/ejhg

ARTICLE

An utter refutation of the 'Fundamental Theorem of the HapMap'

Joseph D Terwilliger*[,1,2,3,4,5] and Tero Hiekkalinna[5,6]

[1]*Department of Genetics and Development, Columbia University, New York, NY, USA;* [2]*Department of Psychiatry, Columbia University, New York, NY, USA;* [3]*Columbia Genome Center, Columbia University, New York, NY, USA;* [4]*Division of Medical Genetics, New York State Psychiatric Institute, New York, NY, USA;* [5]*Finnish Genome Center, University of Helsinki, Helsinki, Finland;* [6]*Department of Molecular Medicine, National Public Health Institute, Helsinki, Finland*

The International HapMap Project was proposed in order to quantify linkage disequilibrium (LD) relationships among human DNA polymorphisms in an assortment of populations, in order to facilitate the process of selecting a minimal set of markers that could capture most of the signal from the untyped markers in a genome-wide association study. The central dogma can be summarized by the argument that if a marker is in tight LD with a polymorphism that directly impacts disease risk, as measured by the metric r^2, then one would be able to detect an association between the marker and disease with sample size that was increased by a factor of $1/r^2$ over that needed to detect the effect of the functional variant directly. This 'fundamental theorem' holds, however, only if one assumes that the LD between loci and the etiological effect of the functional variant are independent of each other, that they are statistically independent of all other etiological factors (in exposure and action), that sampling is prospective, and that the estimates of r^2 are accurate. None of these are standard operating assumptions, however. We describe the ramifications of these implicit assumptions, and provide simple examples in which the effects of a functional variant could be unequivocally detected if it were directly genotyped, even as markers in high LD with the functional variant would never show association with disease, even in infinite sample sizes. Both theoretical and empirical refutation of the central dogma of genome-wide association studies is thus presented.

European Journal of Human Genetics (2006) **14,** 426–437. doi:10.1038/sj.ejhg.5201583; published online 15 February 2006

Keywords: linkage disequilibrium; correlation coefficients; HapMap; association studies

Introduction

Genome-wide association studies have been very successful at identifying loci involved in rare Mendelian traits in population isolates, and as such have been suggested in recent years to be potentially useful for dissection of the etiology of complex traits as well. With this aim in mind, the International HapMap project has been developing a dense genome-wide map of single-nucleotide polymorphisms (SNPs) and characterizing the linkage disequilibrium (LD) among them. The goal is ultimately to allow scientists to select subsets of the SNPs, which are in strong enough LD with the untyped SNPs to allow them to serve as useful surrogates (ie to reduce dimensionality of a genome scan by selecting a maximal set of markers showing a minimal amount of LD among themselves). This is the essence of linkage analysis as well – in which a relatively sparse marker map is used to infer the inheritance vectors in families at every genomic position with reasonable and predictable certainty. There are, however, many ways in which application of the correlations owing to LD differ

*Correspondence: Dr JD Terwilliger, 60 Haven Avenue, #15-C, New York, NY 10032, USA. Tel: +358 40 7467 277;
E-mail: joseph.terwilliger@helsinki.fi
Received 16 July 2005; revised 12 December 2005; accepted 27 December 2005; published online 15 February 2006

quantitatively and qualitatively from those due to linkage. LD and linkage are different ways of assessing essentially the same phenomenon, as LD exists only when copies of a given molecular variant shared by two individuals are clonal copies of the same ancestral alleles (identical-by-descent (IBD) in the population), and markers nearby are shared as well because of a paucity of recombination between the loci historically – that is linkage (see Terwillger[1,2] for a review of this relationship).

In genome-wide mapping studies, one does not presume to know what the etiological architecture of the trait under study is in truth. However, study design and analysis methods are predicated on sets of assumptions, because power under different approaches can only be compared under some mathematically tractable models of 'truth'. Association studies are often argued to be more powerful than linkage studies for various reasons. It is rather obvious that if you measure a functional polymorphism directly, it will never be less correlated to the trait than a marker that is both linked to and in LD with the functional site. Furthermore, it is similarly obvious that such a marker can never be less correlated to the trait than a marker that is linked to, but not in LD with the functional site. But this does not mean that linkage analysis is less powerful than association analysis, and says nothing about whether one would have more power studying families or unrelated individuals, although it is clear that having access to a well-characterized dense map of markers across the genome and understanding their LD relationships could be potentially useful. But the question of how valuable such an approach will be is a function of many parameters describing the assumed etiological models, not to mention the study design employed. In order for association studies to work, one needs the phenotype being studied to predict the genotypes of the locus to be identified to a reasonable degree – that is, to have high 'detectance'.[3,4] Furthermore, it is necessary for there to be LD between the functional variant(s) and at least one allele of one of the markers being studied. And finally, one often further assumes that the marker genotypes are statistically independent of the trait, conditional on the genotypes of the functional site in question.

The substantial debate in the literature about the prognosis of genome-wide association studies for mapping genes involved in multifactorial traits to date has focused on the first of those issues – whether or not significantly high detectance is expected for such traits, and whether or not the subset of SNPs selected for analysis will be in sufficient LD with the functional genotypes predicted by the phenotype of interest. These arguments focus on issues related to the common-variant/common-disease (CVCD) hypothesis,[1,3–15] the quality and quantity of LD[2,16–40] and the effects of different population and study design/ascertainment options.[3,8,19,38,41–50] However, the ramifications of assumptions about the independence of marker genotypes and trait phenotypes conditional on genotypes of the functional variant of primary interest have not been described in much detail, despite the centrality of this assumption. In this paper, we address the critical importance of this issue, and demonstrate numerous reasons why such conditional independence of exposures rarely exists in practice, using a combination of theoretical and empirical arguments. We thus present a further criticism of the rationale for genome-wide association studies based on structural theoretical arguments, independent of the widely known counterarguments outlined above. In fact, for purposes of this manuscript, we assume that there are common risk alleles that are detectable if they were themselves genotyped, and show that markers in high LD with the functional variants may never show evidence of association in infinite sample sizes.

Measures of LD between two SNPs

The coefficient of LD between alleles of two SNPs with alleles (*A/a*) and (*B/b*), respectively, is defined as $\delta = D_{AB} = p_{\underline{AB}} - p_A p_B$,[51,52] where $p_{\underline{AB}}$ denotes the frequency of the haplotype bearing alleles *A* and *B* at the two loci, and p_A and p_B denote the allele frequencies of alleles *A* and *B*. Since D_{AB} varies enormously as a function of the allele frequencies, it is common to quantify LD rather in terms of measures which attempt to normalize for the effects of allele frequencies. One such measure is defined as

$$D'_{AB} = \begin{cases} \frac{D_{AB}}{\min(p_A[1-p_B],[1-p_A]p_B)} & \forall\ D_{AB} > 0 \\ \frac{D_{AB}}{\max(p_A p_B,[1-p_A][1-p_B])} & \forall\ D_{AB} < 0 \end{cases}$$

This measure has been generally preferred by population geneticists, as it has predictable behavior as a function of the recombination fraction between the SNPs, the demographic history of the two polymorphisms in the population. An alternative standardized measure, the correlation coefficient, is often used by statisticians, because it has predictable behavior concerning not the evolutionary relationships among markers, but rather concerning the power to detect the correlation in a sample. This measure is defined as

$$\rho_{AB} = \frac{D_{AB}}{\sqrt{p_A[1-p_A]p_B[1-p_B]}} = [P(A|B) - P(A|b)]\sqrt{\frac{p_B[1-p_B]}{p_A[1-p_A]}}$$

The sample estimate of ρ_{AB} is conventionally denoted as r_{AB}. It can be shown that the test statistic for the conventional χ^2 test of independence on a 2×2 table of haplotype frequencies of the alleles of these two loci is numerically equivalent to $X^2 = Nr_{AB}^2$, where *N* denotes the sample size. Leaving aside the philosophical and scientific rationales for preferring one metric over the other, in this paper we focus on the properties of the ρ^2 estimates as

predictors of the power of an association study, owing to the following factual relationship: If SNP B has a functional relationship to phenotype C, then under simple random sampling, the χ^2 statistic relating B and C would be $X^2 = N_{BC} r^2_{BC}$, and in order to obtain the same numerical value for a χ^2 statistic relating marker A to phenotype C instead of typing marker B would require sample size

$$N_{AC} = N_{BC}\, r^2_{BC}/r^2_{AC}$$

Underestimation of sample size requirements due to upward bias in LD estimates from small samples

The so-called haplotype map (or HapMap) of SNPs spanning the genome was designed to facilitate genome-wide association analysis, based on extensions of the relationship described above relating test statistics and correlation coefficients. In a recent paper, Gabriel *et al*[53] claimed that '… the average maximal r^2 value between each additional SNP and the haplotype framework was high, ranging from 0.67 to 0.87 in the four population samples. That is, for the average untested marker, only a small increase in sample size (15–50%) would be needed for the use of a haplotype-based (as compared to direct) association study.'[53]

It is well known that these measures of LD are strongly biased in an upward direction in small samples (with the D' metric being more strongly biased than r^2), because the measures are both defined to be nonnegative.[37] While the bias in the measure r^2 is often not large in magnitude, the effects of the bias on estimates of $1/r^2$, which is claimed to be linearly related to the required sample size and thus to the power of association studies can be enormous even with samples much larger than those being used in the International HapMap project to characterize LD across the genome.[53] Figure 1 shows graphically the magnitude of the bias for a variety of sample sizes, based on simulation of 1 000 000 data sets, in which the allele frequency for the rare allele at each locus was 0.1. In this figure, the x-axis shows the true value of $1/\rho^2$ with the y-axis showing the expected value of $1/r^2$, which is theorized by Gabriel *et al*[53] and others to be a measure of the increased sample size needed when replacing a functional variant in an analysis by a tag SNP in LD with it.[54] In each graph, if the estimates were unbiased, Figure 1 would contain only the straight line $x = y$. It is apparent that $1/r^2$ is estimated to be much lower than $1/\rho^2$, and thus provides a gross underestimate of the needed sample size. But this is just the tip of the iceberg when it comes to problems in the theory, which we now examine in detail.

Sample size for genome-wide linkage analysis

The reason why genome-wide linkage analysis has been successful in reducing the dimensionality of a genome scan is that the correlations in inheritance due to linkage are strictly a function of meiotic recombination frequencies, which have highly regular behavior, as reflected by the existence of mapping functions. That is to say that in a linkage analysis, one measures cosegregation of loci in families, which is solely a function of the genetic distance between the loci. If there are three loci, X, Y, and Z, and a recombination event occurs between X and Y, and no recombination occurs between Y and Z, then the alleles of loci X and Z must be likewise recombinant in that meiosis, whether or not they are syntenic. Such simple deterministic relationships make linkage mapping mathematically tractable. It can be shown that, in general, for an ordered set of marker loci, $X—Y—Z$, $\theta_{XZ} = \theta_{XY} + \theta_{YZ} - 2c\theta_{XY}\theta_{YZ}$, where θ_{XY} is the probability of recombination between loci X and Y, and where c is the 'coefficient of coincidence', a measure of the strength of crossover 'interference', or nonindependence of recombination in adjacent intervals. For the most part, while there is evidence of weak interference in real data, most statistical analyses assume that $c = 1$ (no interference). This is a close approximation to the truth over large distances, and on small distances, where these measures are most relevant in linkage analysis, interference has little or no effect on the analysis outcomes, because c is strongly bounded with the following equation:[51]

$$\max\left(\frac{\theta_{XY} + \theta_{YZ} - 0.5}{2\theta_{XY}\theta_{YZ}}, 0\right) \le c \le 1$$

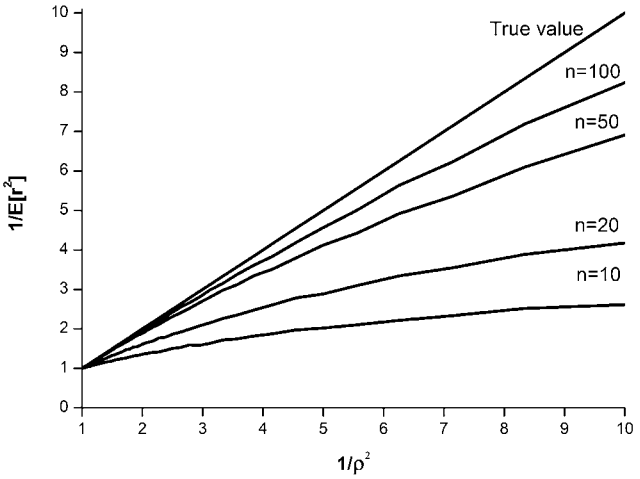

Figure 1 A total of 1 000 000 replicates were simulated of data sets of varying size, from 10 to 100 samples to demonstrate the small sample bias in estimates of the squared correlation coefficient, and its reciprocal. This figure shows the bias in $1/r^2$ as an estimator of $1/\rho^2$ for sample sizes 10, 20, 50, and 100 in order from the lowermost to uppermost curves in the figure. The straight line $x = y$ would represent an absence of bias in the estimators. Note that $1/r^2$ is used as an estimator of the multiplier for the sample size needed for equivalent power using marker *A* instead of functional variant *B* in an association study according to the 'Fundamental Theorem of the HapMap'. In this graph, the line $x = y$ would represent the theoretical predictions, and the other curves show the effects of underestimation of this term owing to small sample bias. For purposes of figure, both loci have minor allele frequency of 0.1, the best-case scenario expected for LD mapping with SNPs.

Let us assume that we have two markers, X and Y, as above, whose positions are known, and we want to compute the value of some statistic relating each of those to some disease locus, Z. Now, let us make the further assumption that the loci are actually in the order $X-Y-Z$, and that we know θ_{XY} and θ_{YZ}. Looking at meioses from a parent with phased genotype ($\underline{1_X\_1_Y\_1_Z}/\underline{2_X\_2_Y\_2_Z}$), we can express the correlation coefficient between the inheritance of the alleles at loci X and Y as

$$\begin{aligned}\rho_{XY} &= [P(1_X|1_Y) - P(1_X|2_Y)]\sqrt{\frac{P(1_X)P(2_X)}{P(1_Y)P(2_Y)}} \\ &= 1 - 2\theta_{XY}\end{aligned}$$

such that

$$\begin{aligned}\rho_{XZ} &= 1 - 2(\theta_{XY} + \theta_{YZ} - 2c\theta_{XY}\theta_{YZ}) \\ &= \rho_{XY}\rho_{YZ} + 4(c-1)\theta_{XY}\theta_{YZ}\end{aligned}$$

Note that when we assume the absence of interference ($c=1$), this implies that $\rho_{XZ} = \rho_{XY}\rho_{YZ}$. Thus, if one wanted to use marker X as a surrogate for marker Y in a linkage analysis, the sample size increase needed would be, following the theory above,

$$\begin{aligned}N_{XZ} &= N_{YZ}\frac{\rho_{YZ}^2}{\rho_{XZ}^2} \\ &= N_{YZ}\left(\frac{\rho_{YZ}}{\rho_{XY}\rho_{YZ} + 4(c-1)\theta_{XY}\theta_{YZ}}\right)^2 \leq \frac{N_{YZ}}{(1-2\theta_{XY})^2}\end{aligned}$$

Thus, the correlations among loci owing to linkage are sufficiently strong to allow for significant reduction in the number of positions across the genome at which one needs to examine chromosomal segregation in families, and thus reduction in dimensionality in genome scans.

In contrast, in the case of LD, the correlation coefficients are not multiplicative. To illustrate the lack of multiplicativity of r^2 estimates for SNP markers, we used all the pairwise r^2 estimates from the International HapMap Project,[5] release date 16 June 2005. In this analysis, we considered all triples of markers $X-Y-Z$, and in Figure 2, on the horizontal axis we give the reported estimate for r^2_{XZ}, and on the vertical axis we give the product $r^2_{XY}r^2_{YZ}$ for all triples of markers. If the correlation coefficients were multiplicative, as they are for linked marker loci in linkage analysis, the graph would be basically a straight diagonal line through the origin ($x=y$). As you can clearly see, however, this is not the case at all. Rather, there is precious little information about the correlation coefficient between X and Z, which can be gleaned from knowing the values of the correlation coefficients between markers X and Y, and that between markers Y and Z. Thus, the theory described below, which drives the HapMap project clearly does not hold in general for correlation coefficients, and in fact can be grossly misleading, most strikingly so when there is substantial LD – the very situation HapMap is designed to model.

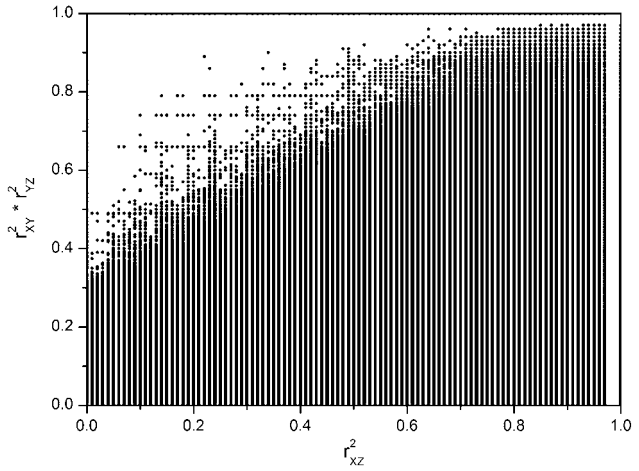

Figure 2 Non-multiplicativity of r^2 estimates: r^2 estimates reported in the 16 June 2005 release from the International HapMap Consortium, based on CEPH (CEU), CHB, JPT and YRI data sets are demonstrated to be nonmultiplicative. All triples of SNPs were considered: $X-Y-Z$, from all autosomal chromosomes and on the x-axis is the estimate of r^2_{XZ}, and on the y-axis is the product of the estimators of r^2_{XY} and r^2_{YZ}. If correlation coefficients were multiplicative, the graph should be the straight line $x=y$, but as can be clearly seen, this is not the case, and the values of r^2_{XY} and r^2_{YZ} can be seen to provide very little information about the correlation coefficient between the flanking markers X and Z, which would not be the case if the assumptions of the Fundamental Theorem of the HapMap held in general.

Fundamental theorem of the HapMap

Statements about the relationships between LD and power of association studies like those made in the Gabriel *et al*[53] paper are based on theory, which assumes a multiplicative relationship among estimated correlation coefficients for different factors[32], although it is well known that correlation coefficients are not generally multiplicative. For example, Czechs have higher alcohol consumption than Finns, and men have higher alcohol consumption then women.[55] If the correlation coefficients describing these relationships were multiplicative, then one would arrive at the false conclusion that this implies that being male was correlated with being Czech. Justification for moving forward with HapMap as a tool for genome-wide association studies has been based on extrapolations from the aforementioned theory relating χ^2 statistics to correlation coefficients. Let us define this hypothesized relationship formally as follows:

Theorem ('Fundamental Theorem of The HapMap'): If ρ_{AB} is the correlation coefficient between alleles of two SNPs, A (with alternate allele a) and B (with alternate allele b), and if sample size N_{BC} would be sufficiently large to detect a correlation between phenotype C (with alternate phenotype c) and functional allele B, then the sample size, N_{AC}, needed to detect a correlation between nonfunctional allele A and the same phenotype would be $N_{AC} = N_{BC}/\rho^2_{AB}$.

However, remember that earlier we demonstrated that in general, $N_{AC} = N_{BC}\ r_{BC}^2/r_{AC}^2$, such that the implicit assumption is that $r_{BC}^2/r_{AC}^2 = 1/\rho_{AB}^2$, or in other words, that $\rho_{AB}^2 r_{BC}^2 = r_{AC}^2$. This is analogous to the highly deterministic relationship among linked marker loci in the case of linkage analysis, $\rho_{XZ} = \rho_{XY}\rho_{YZ}$, which holds in the absence of crossover interference, as shown above. Note that this multiplicative relationship only holds because of the great regularity of the correlations generated by the recombination process in meiosis. The generalization of this relationship to LD studies (and the theory of correlation coefficients in general) is far from straightforward and requires strong additional assumptions for it to hold, not to mention large samples, as r^2 is an upwardly biased estimator of the squared correlation coefficient, as shown above.

It can be shown in general (see Appendix A) that the relationship $\rho_{AC} = \rho_{AB}\rho_{BC}$ implies the independence of A and C conditional on B, that is to say $P(A|BC) = P(A|Bc) = P(A|B)$, and so forth. In the context of association studies, the Fundamental Theorem of the HapMap implies that the frequency of SNP allele A conditional on allele B at the functional site is invariant between cases and controls, implying that the only reason allele A might differ in frequency between cases and controls would be because of differences in the frequency of allele B. For purposes of the following discussion, we ignore the effects of ploidy, to simplify the algebra and need to specify specific dominance relationships. In the context of this discussion, we refer to A and B as alleles on a haplotype drawn at random from an individual, such that $P(C|B)$ refers to the probability that the person from whom a B allele is selected at random is affected. Note that in terms of a penetrance model, $P(C|B) = \{P(C|BB)P(BB) + 0.5P(C|B\text{b})P(B\text{b})\}/P(B)$, which is exactly what would be compared in a cohort study comparing allele frequencies with a dichotomous phenotypic outcome.

Bounds on unconditional and conditional ρ_{AC} given B or b

Because all pairwise haplotype frequencies are probabilities constrained to be between 0 and 1, there are algebraic restrictions on the range of ρ_{AC} as a function of $P(A)$ and $P(C)$, as follows:

$$\max\left(-\sqrt{\frac{P(A)P(C)}{P(a)P(c)}}, -\sqrt{\frac{P(a)P(c)}{P(A)P(C)}}\right) \le \rho_{AC} \le \min\left(\sqrt{\frac{P(a)P(C)}{P(A)P(c)}}, \sqrt{\frac{P(A)P(c)}{P(a)P(C)}}\right)$$

Furthermore, if we know ρ_{AB} and ρ_{BC}, $P(A)$, $P(B)$, and $P(C)$, then we also know uniquely $P(A|B)$, $P(A|\text{b})$, $P(C|B)$, and $P(C|\text{b})$ as

$$P(A|B) = P(A) + \rho_{AB}\sqrt{\frac{P(A)P(a)P(b)}{P(B)}},$$

and

$$P(C|B) = P(C) + \rho_{BC}\sqrt{\frac{P(b)P(C)P(c)}{P(B)}},$$

and

$$P(A|b) = P(A) - \rho_{AB}\sqrt{\frac{P(A)P(a)P(B)}{P(b)}},$$

and

$$P(C|b) = P(C) - \rho_{BC}\sqrt{\frac{P(B)P(C)P(c)}{P(b)}}$$

These restrictions imply that the restrictions on the conditional values of ρ_{AC} given B and b are different, since

$$\max\left(-\sqrt{\frac{P(A|B)P(C|B)}{P(a|B)P(c|B)}}, -\sqrt{\frac{P(a|B)P(c|B)}{P(A|B)P(C|B)}}\right) \le \rho_{AC|B} \le \min\left(\sqrt{\frac{P(a|B)P(C|B)}{P(A|B)P(c|B)}}, \sqrt{\frac{P(A|B)P(c|B)}{P(a|B)P(C|B)}}\right)$$

This implies that there are additional restrictions on the unconditional values of ρ_{AC}, and thus ρ_{AB} and ρ_{BC} contain information about ρ_{AC}, even though they do not determine it uniquely.

By definition,

$$\rho_{AC} = \frac{1}{\sqrt{P(A)P(a)P(C)P(c)}}\left\{P(A|B)P(C|B)P(B)\left[1 + \rho_{AC|B}\sqrt{\frac{P(a|B)P(c|B)}{P(A|B)P(C|B)}}\right] + P(A|b)P(C|b)P(b)\left[1 + \rho_{AC|b}\sqrt{\frac{P(a|b)P(c|b)}{P(A|b)P(C|b)}}\right] - P(A)P(C)\right\}$$

and this quantity has its lower bound when

$$\rho_{AC|B} = \max\left(-\sqrt{\frac{P(A|B)P(C|B)}{P(a|B)P(c|B)}}, -\sqrt{\frac{P(a|B)P(c|B)}{P(A|B)P(C|B)}}\right),$$

and

$$\rho_{AC|b} = \max\left(-\sqrt{\frac{P(A|b)P(C|b)}{P(a|b)P(c|b)}}, -\sqrt{\frac{P(a|b)P(c|b)}{P(A|b)P(C|b)}}\right)$$

and its upper bound when

$$\rho_{AC|B} = \min\left(\sqrt{\frac{P(A|B)P(c|B)}{P(a|B)P(C|B)}}, \sqrt{\frac{P(a|B)P(C|B)}{P(A|B)P(c|B)}}\right),$$

and

$$\rho_{AC|b} = \min\left(\sqrt{\frac{P(A|b)P(c|b)}{P(a|b)P(C|b)}}, \sqrt{\frac{P(a|b)P(C|b)}{P(A|b)P(c|b)}}\right)$$

To see what these constraints mean about the information about tertiary correlations contained within sets of pairwise correlation coefficients, consider the following set

of assumptions about marker and disease loci, which are best-case scenarios for association studies: $P(A)=0.1$ (tag SNP with rare allele frequency of 10%), $P(B)=0.1$ (functional variant with rare allele frequency of 10%), $P(C)=0.025$ (disease prevalence of 2.5%), $\rho_{AB}=0.9$ ($r^2=0.81$ between the rare alleles of the functional variant and the tag SNP). Translating the constraints described above into parameters that are more comprehensible to the genetic epidemiologist concerned about practical ramifications, we graph the bounds on r^2_{AC} in Figure 3a, as a function of the relative risk of allele B, which is defined as

$$RR_B = \frac{P(C|B)}{P(C|b)} = \frac{P(C) + \rho_{BC}\sqrt{\frac{(1-P(B)P(C)(1-P(C))}{P(B)}}}{P(C) - \rho_{BC}\sqrt{\frac{P(B)P(C)(1-P(C))}{1-P(B)}}}$$

The horizontal bar on top is the upper bound on r^2_{AC} imposed by $P(A)$ and $P(C)$ alone, the upper curve is the upper bound as a function of RR_B ranging from 0.1 to 10, and the lower curve is the predicted value of r^2_{AC} from Gabriel *et al*.[53] Note that throughout this range, the lower bound on r^2_{AC} is 0, that is, it is possible to have no correlation whatsoever between the phenotype and the tag SNP, even with correlation coefficient of 0.9 between the SNP and the functional polymorphism!

Since it is r^2_{AC} that is directly related to power, and since it was indicated above that $N_{AC}\, r^2_{AC}=\chi^2$ for a given data set, then if we wanted to obtain the same significance from an association study using SNP A instead of functional variant B, to satisfy the relationship $N_{AC}r^2_{AC}=N_{BC}r^2_{BC}$, the relative sample size needed using the surrogate tag SNP would be $N=N_{AC}/N_{BC}=r^2_{BC}/r^2_{AC}$. Figure 3b graphs the upper bound on this relative sample size over the range of RR_B extending from 1 to 100. The upper value is well beyond the plausible range for variants being sought in complex trait association studies, with RR_B between 1.5 and 3 being the range most people claim to be interested in. Note that for these fairly reasonable assumptions about the frequencies of the variants, and the high correlation coefficient of 0.9 between the SNP and the functional variant, the sample size has an upper bound of infinity over much of the range considered. Note that the thin horizontal line at $N=1.23$ is the predicted increase in sample size needed claimed by Gabriel *et al*[53] in their naïve application of the theory of correlation coefficients.

For the reader interested in exploring the effects of altered values of the various parameters, this can be done using the Excel spreadsheet found at http://linkage.cpmc.columbia.edu/excel/rsquaredAC.xls.

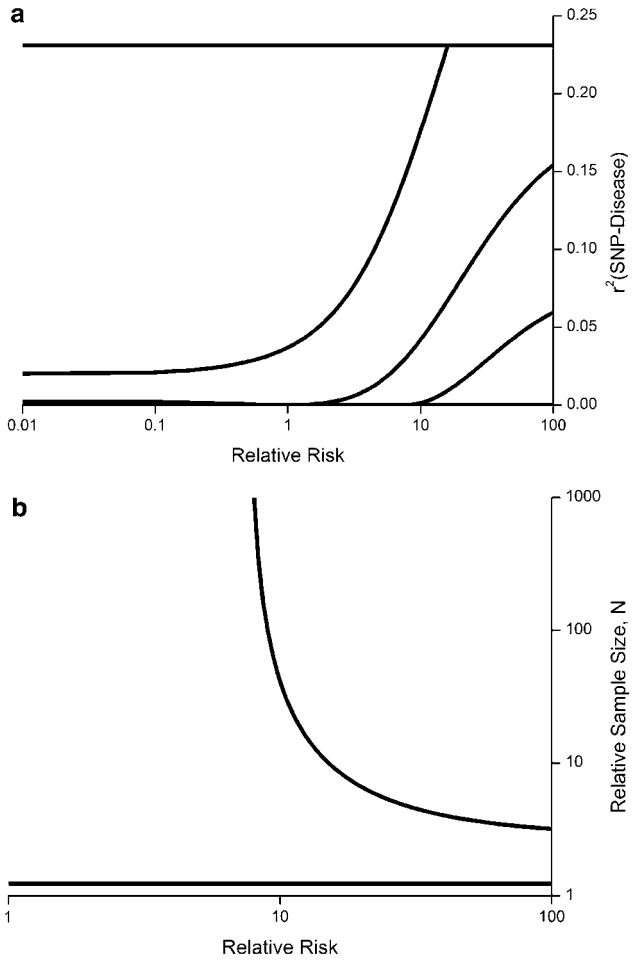

Figure 3 Bounds on r^2_{AC} are shown graphically for simple random sampling. In this case, the allele frequencies for minor alleles of both functional locus B and tag SNP A are set to 0.1 and the correlation coefficient between them, $\rho_{AB}=0.9$. (**a**) Upper and lower bounds on r^2_{AC} as a function of the relative risk of the functional variant B on phenotype C, with the curve in the middle representing the theoretical prediction under multiplicativity of correlation coefficients. (**b**) The same bounds but with the y-axis representing ρ^2_{BC}/ρ^2_{AC}, which is the increase in sample size actually needed when typing SNP A instead of functional site B. Note that in (**b**), the theoretical prediction is that this ratio should be 1.2 for all values of the relative risk.

Ascertainment bias

Of course, in real-world epidemiological studies, one does not use simple random sampling of the sort for which these mathematical models were derived to fit. In practice, one would ascertain individuals from the population conditional on the trait (outcome C in our nomenclature) because this systematically increases the power by enriching for the rare outcome variable. Mathematically this has the effect of increasing the magnitude of ρ_{BC} as follows: since under simple random sampling (r.s.),

$$\rho_{BC}(\text{r.s.}) = [P(B|C) - P(B|c)]\sqrt{\frac{P(C)P(c)}{P(B)P(b)}}$$

then if one were to sample from the population conditional on outcome such that the sample had proportion p_C of cases and $(1-p_C)$ of controls, the value of ρ_{BC} under case

control sampling (c.c.) would be

$$\rho_{BC}(\text{c.c.}) = [P(B|C) - P(B|c)]\sqrt{\frac{p_C(1-p_C)}{p_B(1-p_B)}}$$

where $p_B = P(B|C)p_c + P(B|c)(1-p_c)$ such that the correlation coefficient is increased by a factor of

$$\frac{\rho_{BC}(\text{c.c.})}{\rho_{BC}(\text{r.s.})} = \sqrt{\frac{p_C[1-p_C]P(B)[1-P(B)]}{P(C)[1-P(C)]p_B[1-p_B]}}$$

Thus, the sample size needed for an equivalent expected χ^2 statistic under case–control sampling with proportion of cases sample set at p_C would be

$$N_{BC}(\text{c.c.}) = N_{BC}(r.s.)\frac{P(C)[1-P(C)]p_B[1-p_B]}{p_C[1-p_C]P(B)[1-P(B)]}$$

A definition of invariant LD among SNPs under case–control sampling in the spirit of the underlying assumptions of the 'Fundamental Theorem of the HapMap' would be that $P(A|B)$ and $P(A|b)$ remain invariant with phenotype. Even if this were true, ρ_{AB} must be different, since

$$\rho_{AB} = [P(A|B) - P(A|b)]\sqrt{\frac{p_B[1-p_B]}{p_A[1-p_A]}}$$

While the first part of this equation might be invariant in random sampling or case-control sampling, the second term cannot, since both p_B and p_A would vary due to ascertainment bias if B were functional.

Ascertainment bias influences the value of N_{AC}, the sample size requirement under case control sampling as follows:

$$\begin{aligned} N_{AC}(\text{c.c.}) &= N_{BC}(\text{c.c.})\frac{\rho^2_{BC}(\text{c.c.})}{\rho^2_{AC}(\text{c.c.})} \\ &= N_{BC}(\text{r.s.})\frac{\rho^2_{BC}(\text{c.c.})}{\rho^2_{AC}(\text{c.c.})}\frac{P(C)[1-P(C)]p_B[1-p_B]}{p_C[1-p_C]P(B)[1-P(B)]} \end{aligned}$$

where N_{AC}(c.c.) refers to sample size needed under case control sampling, while N_{AC}(r.s.) refers to the analogous quantity under simple random sampling. While we demonstrated above that the sample size requirement under case–control sampling is reduced whenever $P(C) < 0.5$ in the population, it is not necessarily true that the sample size requirement is decreased when typing SNP A as a surrogate for the functional variant, even when ρ_{AB} is high in the population. Figure 4a shows the upper and lower bounds on ρ^2_{AC} over the same range given for random sampling in Figure 3a, while Figure 4b shows the same bounds for $1/\rho^2_{AC}$ as a function of the relative risk, as in Figure 3b. Note that while ρ^2_{AC} may obtain much higher values under case-control sampling than simple random sampling, the range includes 0 for even high relative risks for B. This means that the sample size requirement (equivalent to that in Figure 3b for random sampling) must include infinity as an upper bound over the entire range. Thus, case–control sampling, theoretically, can lead to even less power than simple random sampling under some models, even when ρ_{AB} in the population is as high as 0.9, as it was in the example. It is important to note as well that ρ_{AB} in a case–control design cannot be uniquely determined as a function of the population ρ_{AB}, the relative risk of disease given functional variant B, and the frequencies of A, B, and C, further complicating predictions about power in that context. Nevertheless, the fact remains that case–control sampling can potentially reduce power over random sampling, when using a tag SNP, A, as a surrogate for some functional variant, B, in an association study with disease C!

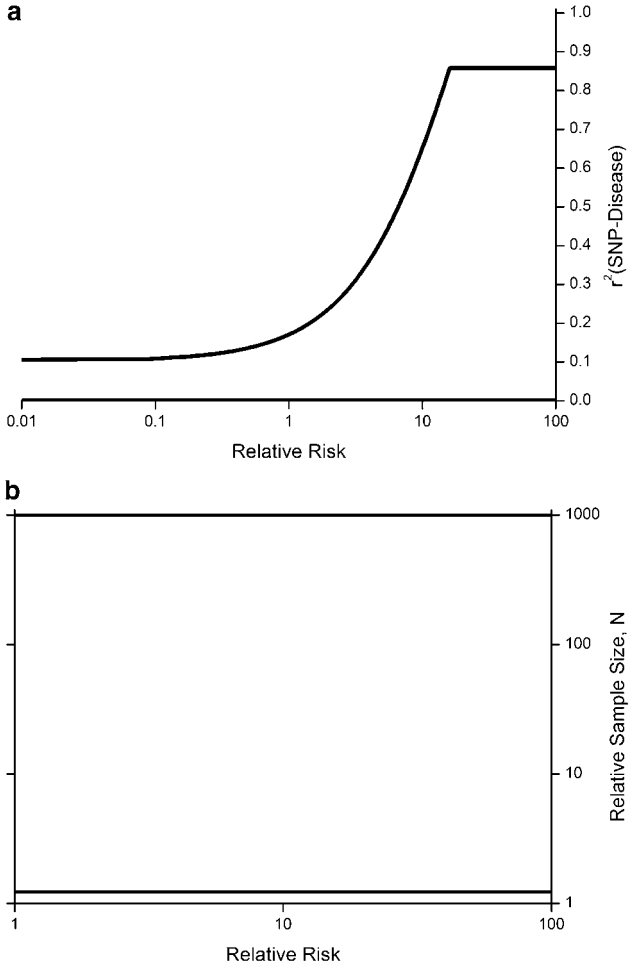

Figure 4 Bounds on r^2_{AC} are shown graphically for case-control sampling. (**a**) is the equivalent graph to Figure 3a for case-control sampling, and (**b**) is the analog of Figure 3b for case-control sampling. The assumed models are the same as in Figure 3.

Effects of allelic heterogeneity

Let us now examine a very simple situation in which the implications of the 'Fundamental Theorem of the HapMap'

Table 1 Effects of allelic heterogeneity in a simple case on the multiplicativity assumption

	Detectance		*Locus B*	*Detectance*		*Locus C*	*Detectance*	
Locus A	*Case*	*Control*	*Genotype*	*Case*	*Control*	*Genotype*	*Case*	*Control*
Genotype								
$+_A/+_A$	0.417	0.574	$+_B/+_B$	0.417	0.574	$1_C/1_C$	0.25	0.25
$+_A/D_A$	0.5	0.365	$+_B/D_B$	0.5	0.365	$1_C/2_C$	0.5	0.5
D_A/D_A	0.083	0.061	D_B/D_B	0.083	0.061	$2_C/2_C$	0.25	0.25
Allele								
$+_A$	0.667	0.757	$+_B$	0.667	0.757	1_C	0.5	0.5
D_A	0.333	0.243	D_B	0.333	0.243	2_C	0.5	0.5
Odds ratio	1.55		Odds ratio	1.55		Odds ratio	1.00	

For the model outlined in the text, the detectance distributions for functional variants A and B are shown, along with that for tag SNP C. Note that both A and B would be detectable with sufficient sample size, were they genotyped, but in an infinite sample, SNP C would never show any evidence of association, as it has an odds ratio of 1, despite being in LD with both loci B and C.

would be totally misleading. Let us consider three SNPs in a haplotype block, such that two of the three, A and B are functional (with risk alleles D_A and D_B respectively and normal (wild type) alleles $+_A$ and $+_B$ respectively), and have equivalent effect on the trait, in a dominant manner, such that $P(\text{Affected}|D_A/x) = P(\text{Affected}|D_B/x) = f$, $P(\text{Affected}|+_A/+_A, +_B/+_B) = 0$, and C (a SNP with 2 alleles 1_C and 2_C) has no phenotypic effect (that is to say that presence of a disease allele at either locus A or locus B on one or both chromosomes gives an individual probability f of being affected, and if neither disease allele is present, the individual is healthy with probability 1). If we assume that all the pairwise D' values are 1, meaning that there has been no recurrent mutation or recombination historically within this block, there would be four haplotypes with nonzero frequency, for example, $H_1 = P(\underline{+_A +_B\ 1_C})$; $H_2 = P(\underline{+_A +_B\ 2_C})$; $H_3 = P(\underline{D_A +_B\ 1_C})$; $H_4 = P(\underline{+_A\ D_B\ 2_C})$. If we were to set all four haplotype frequencies to be equal $H_1 = H_2 = H_3 = H_4 = 0.25$, for example, then $\rho^2_{AC} = 0.333$; $\rho^2_{BC} = 0.333$, such that the 'Fundamental theorem of the HapMap' would predict that if marker C were used as a surrogate for A in a case-control association test, the sample size needed $N_C = N_A/0.333 = 3N_A$.

However, the true detectance distribution for this model is shown in Table 1. There would be power to detect the relationship between either functional variant and the disease, with an odds ratio of 1.55 for the risk allele at each of the disease loci, but the odds ratio is 1 with the SNP that had r^2 of 0.333 with each of the disease loci. The 'Fundamental Theorem of the HapMap' would have predicted that a sample size three times larger than needed to detect either functional variant would be sufficient to detect the association with the SNP C, but this is clearly untrue. This simple example is admittedly extreme, since both alleles are assumed to be very common, in accordance with the 'common disease/common variants' hypothesis, widely touted by the same scientists that are promoting HapMap[8,21,56–60] Nonetheless, this example clearly shows that even with tight haplotype blocks, and common disease alleles, it is possible that functional variants can be detected if they are genotyped in a sample, and yet there might be absolutely no difference between cases and controls whatsoever for other common markers within the same haplotype block.

If one allows for more substantial allelic heterogeneity, as is typically seen in most loci that have been studied in sufficient detail, this effect will be exacerbated, because the less frequent the individual variants are, the greater the likelihood that they originated on a variety of haplotypes (according to their population frequencies), so if they fall within the same 'haplotype block' they will likely be in opposite phase with any given SNP which is being genotyped. The greater the number of variants, the greater the similarity in the detectance distributions for the haplotypes in cases and controls, if one fails to genotype the functional variants themselves! Furthermore, since it appears likely that in general there is an inverse relationship between effect size and allele frequency, this would further homogenize the distributions of haplotype frequencies for common tag SNPs between cases and controls, making it trivial to construct examples for which there is substantial power to detect the functional variants themselves, if genotyped, in a case-control study, while there would be no power in an infinite sample for tag SNPs, even with very high r^2 LD of 0.8. In populations with LD extending over longer distances, the problem becomes more acute, as there are many more loci in LD with any putative marker, any of which might themselves be functional, so while fewer markers would be needed to do a genome-wide association if one chose markers based on the 'Fundamental Theorem', there would be many more sites with correlated exposure frequencies that might be potentially functional, increasing the potential magnitude of this problem. To this end, one might think twice before

deciding to use fewer markers in association studies in isolates than elsewhere. An Excel spreadsheet is available from the authors in which 3 locus haplotypes can be input under general penetrance and haplotype frequency models to examine these detectance distributions and compare them to the predictions of the Fundamental Theorem of the HapMap, can be found at http://linkage.cpmc.columbia.edu/excel/r-squared.xls.

Software, SIMQTL, for analysis of more complex models under more sophisticated ascertainment schemes is also available from the authors. It should be kept in mind that while such a multiplicity of risk alleles substantially decreases the power of association tests, it generally tends to increase the power in linkage studies. The identities of specific alleles are not examined in linkage analysis, only the sharing of any alleles (whatever their molecular configuration) IBD among relatives in a pedigree, so that whenever functional variants are linked to one another, the power of a linkage study will increase substantially, even when those loci are as far apart as several Mb!

Discussion

The proponents of association-based mapping strategies argue that since A has no functional effect on C, any correlation between A and C must be because A is correlated with B and B is correlated with C, justifying the assumption that $P(A|BC) = P(A|Bc) = P(A|B)$. Simple algebraic manipulations show that this condition is equivalent to saying that $P(C|AB) = P(C|aB)$. At first glance, this seems to be a reasonable assumption, namely that if A is nonfunctional, then the probability of any given phenotypic outcome in general is independent of A.[53,54] However, there are other ways in which having haplotype AB can influence the risk of a phenotypic outcome differently from haplotype aB. We argue that it is the exception, not the rule, for such conditional independence to hold in genetic studies of complex traits, and that the assumptions of Gabriel *et al*[53] rarely hold in practice. Certainly, blanket statements about the relationships between r^2 and sample size requirements for association studies are not factual, since typically $N_{AC} > N_{BC}/\rho^2_{AB}$. In fact, equality of these terms only holds in the best-case scenarios, analogous to linkage analysis. Conditional independence in genetics is rarely an appropriate assumption (as people have been learning the hard way in attempts to look at linkage analysis with massively dense sets of SNP markers[4,37,48,61–66]), and this is the primary reason why geneticists avoid using classical statistical techniques in favor of complex likelihood-based models when making inferences. It is imperative to remember that statistical independence is a very different thing from causal independence, and is a very strong assumption, which can have enormous consequences.

Above, we have provided a simple example where conditional independence does not hold owing to allelic heterogeneity. While allelic heterogeneity is one potential reason for deviation from the theory, it is certainly not the only way. Ethnic heterogeneity, in which the frequencies of both phenotype and SNP marker alleles vary can create not only false positives, as is by now well-appreciated, but can just as easily create 'false negatives' – that is to say tag SNPs may show no correlation at all with the phenotype, even when an easily detectable functional variant has r^2 of 0.8 with the tag SNP, for similar mathematical reasons. Likewise, environmental risk factors can have similar sorts of confounding that can easily cancel out the effects of a functional variant, when typing a tag SNP instead. And here we are only considering cases in which the functional variant does have power to be detected if it were genotyped and measured. The fact is that correlation coefficients are almost never multiplicative in practice, and in studies involving genetic risk factors for disease, we have known for decades that conditional independence of exposures never holds. In fact, this is the entire basis for the development of the complex likelihood methods we have relied upon for the past decades in understanding the genetic basis of simple diseases.

Sequencing samples consisting of cases and controls in candidate regions in the genome to estimate the amount of LD measured by r^2 among the SNPs they identify, in order to select 'tag' SNPs for further study, is a strange approach, because, as we showed above, r^2 must vary between cases and controls whenever one of the markers has a functional effect. Often people look at the r^2 in cases and controls, and if it is not different, they pool the data and use that to select a marker, but as shown above, this will always bias the estimates. For that matter, if the r^2 really is invariant in cases and controls, this is evidence against alleles of either SNP being functional, which is reason to potentially not type either of them. Again, it is important to be careful about the assumptions, the theory, and their ramifications, rather than proceeding naively based on arguments like 'but that is what everyone else is doing, so it must be right', which we have all been subjected to.

We hope to make readers think twice before engaging in high risk studies without fully evaluating the potentials for confounding factors such as those described here to complicate the theoretical predictions. Human geneticists routinely rely on mathematical theory and predictions without fully understanding the assumptions driving the theory, or contemplating their implications. If nothing else, it is hoped that statistical geneticists would be more forthcoming and explicit about the theoretical ramifications of the model assumptions, as this is but one example where they fall apart. Similar difficulties and inconsistencies between theory and practice can be widely seen in such areas as studies of gene–gene and gene–environment interaction, where independence of exposures is assumed,

and deviations from independence of exposures conditional on phenotype are inferred to imply etiological interactions, when nonindependence of exposures and ascertainment bias are equally capable of explaining such phenomena without the need to invoke complex etiological interactions. Another obvious example of inconsistent theory and practice would be when linkage analyses of extremely concordant and discordant sibling pairs are performed, assuming some component of variance due to polygenic factors, and yet the null hypothesis in the linkage analysis predicts that at random genomic locations, 50% of the genome of such sibs should be IBD (when of course the polygenic factors that are individually too weak to detect, must alter this average genome-wide sharing if the analysis shows they exist…).

We hope that as gene hunting approaches increase in cost and size, that rather than becoming more cavalier about theoretical assumptions, that we be much more careful about what we believe. Technological advances are wonderful, and make it possible to do science that we could not imagine a few decades ago, but excellent technology applied to poorly designed studies (driven by assumptions the investigators themselves probably would not really believe if consciously aware of them) are not particularly wise ways to do science – it would be far better to spend more time thinking and planning before jumping in to genotyping every sample we can get our hands on, lest no one listen to us when we cry fire and there actually is one, at some point in the future.

Acknowledgements

Grant MH63749 from the National Institutes of Mental Health, along with funding from the Sigrid Juselius Foundation, the Academy of Finland, The Finnish Cultural Foundation, and the Burroughs-Wellcome Fund is gratefully acknowledged. r^2 estimates for pairs of SNPs on all Chromosomes from CEU, CHB, JPT and YRI datasets, release date June 16th 2005, from the International HapMap Consortium were used for Figure 2, and as such the consortium is acknowledged. Thanks to Harald H.H. Göring for critical comments on the manuscript.

References

1 Terwilliger JD, Weiss KM: Linkage disequilibrium mapping of complex disease: fantasy or reality? *Curr Opin Biotechnol* 1998; **9**: 578–594.
2 Terwilliger JD: On the resolution and feasiblity of genome scanning approaches. *Adv Genet* 2001; **42**: 351–391.
3 Terwilliger JD, Weiss KM: Confounding, ascertainment bias, and the blind quest for a genetic 'fountain of youth'. *Ann Med* 2003; **35**: 532–544.
4 Weiss KM, Terwilliger JD: How many diseases does it take to map a gene with SNPs? *Nat Genet* 2000; **26**: 151–157.
5 The International HapMap Consortium: The International HapMap Project. *Nature* 2003; **426**: 789–796.
6 Aquadro CF, DuMont VB, Reed FA: Genome-wide variation in the human and fruitfly: a comparison. *Curr Opin Genet Dev* 2001; **11**: 627–634.
7 Brookes AJ: Rethinking genetic strategies to study complex diseases. *Trends Mol Med* 2001; **7**: 512–516.
8 Cardon LR, Bell JI: Association study designs for complex diseases. *Nat Rev Genet* 2001; **2**: 91–99.
9 Ghosh S, Collins FS: The geneticist's approach to complex disease. *Ann Rev Med* 1996; **47**: 333–353.
10 Guyer MS, Collins FS: The Human Genome Project and the future of medicine. *Am J Dis Children* 1993; **147**: 1145–1152.
11 Kruglyak L, Nickerson DA: Variation is the spice of life. *Nat Genet* 2001; **27**: 234–236.
12 Pritchard JK: Are rare variants responsible for susceptibility to complex diseases? *Am J Hum Genet* 2001; **69**: 124–137.
13 Pritchard JK, Cox NJ: The allelic architecture of human disease genes: common disease-common variant or not? *Hum Mol Genet* 2002; **11**: 2417–2423.
14 Risch N, Merikangas K: The future of genetic studies of complex human diseases. *Science* 1996; **273**: 1516–1517.
15 Zwick ME, Cutler DJ, Chakravarti A: Patterns of genetic variation in Mendelian and complex traits. *Annu Rev Genomics Hum Genet* 2000; **1**: 387–407.
16 Ardlie KG, Kruglyak L, Seielstad M: Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* 2002; **3**: 299–309.
17 Bader JS: The relative power of SNPs and haplotype as genetic markers for association tests. *Pharmacogenomics* 2001; **2**: 11–24.
18 Barton A, Chapman P, Myerscough A *et al*: The single-nucleotide polymorphism lottery: How useful are a few common SNPs in identifying disease-associated alleles? *Genet Epidemiol* 2001; **21**: S384–S389.
19 Black WC, Baer CF, Antolin MF, DuTeau NM: Population genomics: genome-wide sampling of insect populations. *Annu Rev Entomol* 2001; **46**: 441–469.
20 Clark AG, Weiss KM, Nickerson DA *et al*: Haplotype structure and population genetic inferences from nucleotide-sequence variation in human lipoprotein lipase. *Am J Hum Genet* 1998; **63**: 595–612.
21 Collins A, Ennis S, Taillon-Miller P, Kwok PY, Morton NE: Allelic association with SNPs: metrics, populations, and the linkage disequilibrium map. *Hum Mutat* 2001; **17**: 255–262.
22 Cookson W: The extent and distribution of linkage disequilibrium: problems for SNP mappers. *J Med Genet* 2001; **38**: SP3.
23 Daly MJ, Rioux JD, Schaffner SE, Hudson TJ, Lander ES: High-resolution haplotype structure in the human genome. *Nat Genet* 2001; **29**: 229–232.
24 Dunning AM, Durocher F, Healey CS *et al*: The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* 2000; **67**: 1544–1554.
25 Goldstein DB, Weale ME: Population genomics: linkage disequilibrium holds the key. *Curr Biol* 2001; **11**: R576–R579.
26 Johnson GCL, Esposito L, Barratt BJ *et al*: Haplotype tagging for the identification of common disease genes. *Nat Genet* 2001; **29**: 233–237.
27 Jorde LB, Watkins WS, Bamshad MJ: Population genomics: a bridge from evolutionary history to genetic medicine. *Hum Mol Genet* 2001; **10**: 2199–2207.
28 Judson R, Salisbury B, Schneider J, Windemuth A, Stephens JC: How many SNPs does a genome-wide haplotype map require? *Pharmacogenomics* 2002; **3**: 379–391.
29 Maniatis N, Collins A, Xu CF *et al*: The first linkage disequilibrium (LD) maps: Delineation of hot and cold blocks by diplotype analysis. *Proc Natl Acad Sci USA* 2002; **99**: 2228–2233.
30 Morris RW, Kaplan NL: When is haplotype analysis advantageous for linkage- disequilibrium mapping? *Am J Hum Genet* 2001; **69**: 30.
31 Morton NE, Zhang W, Taillon-Miller P, Ennis S, Kwok PY, Collins A: The optimal measure of allelic association. *Proc Natl Acad Sci USA* 2001; **98**: 5217–5221.
32 Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
33 Remington DL, Thornsberry JM, Matsuoka Y *et al*: Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proc Natl Acad Sci USA* 2001; **98**: 11479–11484.

34 Service SK, Ophoff RA, Freimer NB: The genome-wide distribution of background linkage disequilibrium in a population isolate. *Hum Mol Genet* 2001; **10**: 545–551.
35 Sham PC, Zhao JH, Curtis D: The effect of marker characteristics on the power to detect linkage disequilibrium due to single or multiple ancestral mutations. *Ann Hum Genet* 2000; **64**: 161–169.
36 Stephens JC, Schneider JA, Tanguay DA *et al*: Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 2001; **293**: 489–493.
37 Terwilliger JD, Haghighi F, Hiekkalinna TS, Goring HH: A bias-ed assessment of the use of SNPs in human complex traits. *Curr Opin Genet Dev* 2002; **12**: 726–734.
38 Varilo T, Laan M, Hovatta I, Wiebe V, Terwilliger JD, Peltonen L: Linkage disequilibrium in isolated populations: Finland and a young sub-population of Kuusamo. *Eur J Hum Genet* 2000; **8**: 604–612.
39 Varilo T, Paunio T, Parker A *et al*: The interval of linkage disequilibrium (LD) detected with multiallelic and biallelic markers in chromosomes of early and late settlement regions of Finland, 2002.
40 Weiss KM, Clark AG: Linkage disequilibrium and the mapping of complex human traits. *Trends Genet* 2002; **18**: 19–24.
41 Culverhouse R, Lin J, Liu KY, Suarez BK: Exploiting linkage disequilibrium in population isolates. *Genet Epidemiol* 2001; **21**: S429–S434.
42 Kere J: Human population genetics: Lessons from Finland. *Annu Rev Genomics Hum Genet* 2001; **2**: 103–128.
43 Peltonen L, Palotie A, Lange K: Use of population isolates for mapping complex traits. *Nat Rev Genet* 2000; **1**: 182–190.
44 Pritchard JK, Donnelly P: Case–control studies of association in structured or admixed populations. *Theor Popul Biol* 2001; **60**: 227–237.
45 Pritchard JK, Stephens M, Rosenberg NA, Donnelly P: Association mapping in structured populations. *Am J Hum Genet* 2000; **67**: 170–181.
46 Scriver CR: Human genetics: Lessons from Quebec populations. *Annu Rev Genomics Hum Genet* 2001; **2**: 69–101.
47 Terwilliger JD, Lee JH: Natural experiments in human gene mapping. In: Craford MH (ed): *Anthropol Genet*. Cambridge: Cambridge University Press, 2005.
48 Terwilliger JD, Göring HHH, Magnusson PKE, Lee JH: Study design for genetic epidemiology and gene mapping: The Korean diaspora project. *Shengming Kexue Yanjiu (Life Science Research)* 2002; **6**: 95–115.
49 Terwilliger JD, Zollner S, Laan M, Paabo S: Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'drift mapping' in small populations with no demographic expansion. *Hum Heredity* 1998; **48**: 138–154.
50 Wright AF, Carothers AD, Pirastu M: Population choice in mapping genes for complex diseases. *Nat Genet* 1999; **23**: 397–404.
51 Ott J: *Analysis of Human Genetic Linkage*. Baltimore: Johns Hopkins University Press, 1985.
52 Lewontin RC: On Measures of Gametic Disequilibrium. *Genetics* 1988; **120**: 849–852.
53 Gabriel SB, Schaffner SF, Nguyen H *et al*: The structure of haplotype blocks in the human genome. *Science* 2002; **296**: 2225–2229.
54 Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
55 Ahlstrom S, Bloomfield K, Knibbe R: Gender differences in drinking patterns in nine european countries: descriptive findings. *Subst Abus* 2001; **22**: 69–85.
56 Cargill M, Altshuler D, Ireland J *et al*: Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat Genet* 1999; **22**: 231–238.
57 Altshuler D, Clark AG: Genetics harvesting medical information from the human family tree. *Science* 2005; **307**: 1052–1053.
58 Chakravarti A: The nature and distribution of human genetic disease. *Am Naturalist* 2001; **158**: 12.
59 Chakravarti A: Single nucleotide polymorphisms to a future of genetic medicine. *Nature* 2001; **409**: 822–823.
60 Day INM, Gu DF, Ganderton RH, Spanakis E, Ye S: Epidemiology and the genetic basis of disease. *Int J Epidemiol* 2001; **30**: 661–667.
61 Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors I: complex-valued recombination fractions and complex phenotypes. *Am J Hum Genet* 2000; **66**: 1095–1106.
62 Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors II: marker-locus genotyping errors modeled with hypercomplex recombination fractions. *Am J Hum Genet* 2000; **66**: 1107–1118.
63 Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors III: marker loci and their map as nuisance parameters. *Am J Hum Genet* 2000; **66**: 1298–1309.
64 Goring HHH, Terwilliger JD: Linkage analysis in the presence of errors IV: joint pseudomarker analysis of linkage and/or linkage disequilibrium on a mixture of pedigrees and singletons when the mode of inheritance cannot be accurately specified. *Am J Hum Genet* 2000; **66**: 1310–1327.
65 Terwilliger JD: A likelihood-based extended admixture model of oligogenic inheritance in 'model-based' and 'model-free' analysis. *Eur J Hum Genet* 2000; **8**: 399–406.
66 Terwilliger JD, Goring HHH: Gene mapping in the 20th and 21st centuries: statistical methods, data analysis, and experimental design. *Hum Biol* 2000; **72**: 63–132.

Appendix

Demonstration that multiplicativity of correlation coefficients implies conditional independence.

$$\begin{aligned}\rho_{AB}\rho_{BC} &= [P(A|B) - P(A|b)]\sqrt{\frac{P(B)P(b)}{P(A)P(a)}} \\ &\quad \times [P(B|C) - P(B|c)]\sqrt{\frac{P(C)P(c)}{P(B)P(b)}} \\ &= [P(A|B) - P(A|b)] \\ &\quad \times [P(B|C) - P(B|c)]\sqrt{\frac{P(C)P(c)}{P(A)P(a)}}\end{aligned}$$

Since

$$\rho_{AC} = [P(A|C) - P(A|c)]\sqrt{\frac{P(C)P(c)}{P(A)P(a)}}$$

then $\rho_{AC} = \rho_{AB}\,\rho_{BC}$ if and only if

$$\begin{aligned}&[P(A|C) - P(A|c)] \stackrel{?}{=} \\ &\qquad [P(A|B) - P(A|b)][P(B|C) - P(B|c)]\end{aligned}$$

Expanding the right side of this equation leads to the following:

$$\begin{aligned}&[P(A|C) - P(A|c)] \stackrel{?}{=} \\ &\qquad [P(A|B)P(B|C) + P(A|b)P(b|C)] \\ &\qquad - [P(A|B)P(B|c) + P(A|b)P(b|c)] \qquad (1)\end{aligned}$$

However, if we expand the left side of the equation,

$$\begin{aligned}&[P(A|C) - P(A|c)] = \\ &\qquad [P(AB|C) + P(Ab|C)] \\ &\qquad - [P(AB|c) + P(Ab|c)]\end{aligned}$$

and according to the Chain rule from elementary probability, this implies that

$$[P(A|C) - P(A|c)] = \\ [P(A|BC)P(B|C) + P(A|bC)P(b|C)] \\ - [P(A|Bc)P(B|c) + P(A|bc)P(b|c)]$$

but this only equals the right side of Eq. (1) above, if we assume that we have conditional independence of A and C when B is true, such that, for example, $P(A|BC) = P(A|B)$.