Allele Frequencies and the r^2 Measure of Linkage Disequilibrium: Impact on Design and Interpretation of Association Studies

Naomi R. Wray

Department of Medical Sciences, University of Edinburgh, United Kingdom

The design and interpretation of genetic association studies depends on the relationship between the genotyped variants and the underlying functional variant, often parameterized as the squared correlation or r^2 measure of linkage disequilibrium between two loci. While it has long been recognized that placing a constraint on the r^2 between two loci also places a constraint on the difference in frequencies between the coupled alleles, this constraint has not been quantified. Here, quantification of this severe constraint is presented. For example, for $r^2 \geq .8$, the maximum difference in allele frequency is ± .06 which occurs when one locus has allele frequency .5. For $r^2 \geq .8$ and allele frequency at one locus of .1, the maximum difference in allele frequency at the second locus is only ± .02. The impact on the design and interpretation of association studies is discussed.

Association studies test for a relationship between genetic variants and disease status and are an important tool in the search for genes involved in complex diseases (Risch & Merikangas, 1996). Often there are no strong hypotheses about the functional role of a specific genotyped variant. Under these circumstances any association detected must be interpreted as being caused either by the variant or something correlated with it. The observed effect size of the association can be interpreted in terms of the effect size at the causal variant, the frequency of the causal variant, the frequency of the genotyped variant and the extent of linkage disequilibrium (LD) between the casual and genotyped loci. These parameters are, of course, unknown but understanding the relationship between them and the boundaries placed upon them helps with the interpretation of the observed association. The same parameters are also crucial to the design of association studies. The effective size of an association study (N') for a genotyped variant in LD with the causal variant is related to the size of an association study (N) needed when the causal variant itself is genotyped by a factor of $1/r^2$ (Pritchard & Przeworski, 2001; Risch & Teng, 1998) where r^2 is the squared correlation coefficient measure of LD between the two loci. Therefore, in the design of association studies it is common practice to choose markers that are representative of the LD landscape of the region under study, such that all excluded markers are in high LD with one or a combination of chosen markers. In this way, if the causal locus is not genotyped then the association with disease may still be detected by the correlated association with the genotyped marker loci. Many methods of SNP selection have been proposed. The simplest method, motivated by the power relationship described above, excludes markers that make r^2 greater than a defined threshold with any selected marker (e.g., Carlson et al., 2004).

For all of these reasons, the interplay between the allele frequencies of genotyped and causal variants and the r^2 between them underpins any association study. The dependence of r^2 on the allele frequencies at the two loci has long been recognized (Devlin & Risch, 1995; Hedrick, 1987; Hill & Robertson, 1968; Risch, 2000; Risch & Teng, 1998) and has resulted in recommendations that SNPs should be selected on the basis of their frequencies to increase the probability of detecting an association with a nearby causative locus (Garner & Slatkin, 2003; Ohashi & Tokunaga, 2001). Muller-Myshok and Abel (1997) were the first to caution on the power of association studies when the difference in allele frequencies of genotyped and causal variant is high and when LD between them is low. Zondervan and Cardon (2004) provide a thorough exploration of the relationship between causal and correlated variants and their impact on association studies as they 'discuss the underappreciated importance of the marker allele frequency relative to the frequency of the disease variant in influencing the probability of finding the association'. However, these discussions have all fallen short of quantifying the restrictions posed on allele frequencies by the r^2 relationship between them. In this study, we quantify the maximum difference in allele frequencies between two loci as constrained by their r^2 and use these results to discuss the conse-

Received 23 January, 2005; accepted 1 February, 2005.

Address for correspondence: Naomi Wray, Department of Medical Sciences, University of Edinburgh, Edinburgh EH4 2XU, UK. E-mail: naomi.wray@ed.ac.uk

quences for the design and interpretation of association studies.

Methods

Linkage Disequilibrium

Central to our discussion is the concept of linkage disequilibrium which describes the relationship between alleles at two (or more) loci. Many statistics have been proposed to explain this relationship (reviewed, e.g., by Morton et al., 2001; Nordborg & Tavare, 2002). Two of the most commonly used measures are $|D'|$ and r^2. If there are two loci, each with two alleles, and the frequencies for allele 1 at each locus are p_A and p_B respectively, and the frequency of both 1 alleles together is p_{AB} (see Table 1), then the covariance between the loci is $D = p_{AB} - p_A p_B$, where $p_A p_B$ is the expected value of p_{AB} in the absence of allelic association (or coupling). When D is positive, p_{AB} has maximum value equal to the smaller of p_A or p_B and therefore the maximum value of D is the smaller of $p_A(1 - p_B)$ and $p_B(1 - p_A)$; when it is negative, its maximum value is the smaller of $p_A p_B$ and $(1 - p_A)(1 - p_B)$. The sign of D reflects the chance ordering of the alleles at each locus, but can be important in the comparison of LD between the same loci genotyped in different populations (e.g., cases and controls) when alleles have been ordered in the same way. The LD measure r^2 is the squared correlation, where r scales D by the standard deviations of the allele frequencies at two loci, $r^2 = D^2/\{p_A p_B(1 - p_A)(1 - p_B)\}$. In contrast, D' scales D by its maximum value given the allele frequencies:

$$D' = D/\min\{p_A(1 - p_B), p_B(1 - p_A)\} \text{ if } D > 0$$

$$D' = D/\min\{p_A p_B, (1 - p_A)(1 - p_B)\} \text{ if } D < 0$$

Whenever one pair of allele combinations is absent, $|D'| = 1$ and LD is described as 'complete' because the allelic association is as high as possible given the allele frequency at each locus. For example, if $p_A = .6$, $p_B = .1$ *and* $p_{AB} = .1$ (hence $p_{aB} = 0$), $|D'| = 1$, but $r^2 = .07$; this situation may represent, for example, a young SNP that first occurred on the background of the common allele at locus A. In contrast, 'perfect LD' is when only two of the four haplotypes are observed and can only occur when allele frequencies at the two loci are the same, in this case $r^2 = |D'| = 1$. $|D'|$ has range 0 – 1 regardless of allele frequency (although with small sample size, |D'| is often estimated to be 1 when minor allele frequency is low), whereas the maximum value for r^2 is the smaller of $p_A(1 - p_B)/(1 - p_A)\,p_B$ and its inverse. Studies which describe the observed LD landscape often quote both |D'| and r^2 which allows an at-a-glance judgment of LD together with the difference in allele frequencies of the coupled alleles. For example, high $|D'|$ and high r^2 = tendency to the presence of only two haplotypes, small difference in allele frequency of coupled alleles; high $|D'|$ and low r^2 = tendency to the presence of only three haplotypes, different allele frequencies of the coupled alleles; low |D'|, low r^2 = tendency toward random coupling of alleles and presence of all four haplotypes.

Table 1

Notation for Haplotype and Allele Frequencies

Locus A	Locus B Allele 1	Allele 2	Allele frequency
Allele 1	p_{AB}	p_{Ab}	p_A
Allele 2	p_{aB}	p_{ab}	$p_a = 1 - p_A$
Allele frequency	p_B	$p_b = 1 - p_B$	1

Quantification of Relationship Between Allele Frequencies and r²

Whilst a general judgment can be made about difference in allele frequencies from the joint knowledge of the r^2 and $|D'|$ LD measures, there has been no formal framework to quantify the relationship between r^2 and allele frequencies. The boundaries on allele frequency given the r^2 between two loci can be derived as follows. If alleles A and B are the coupled alleles at two different loci and if $p_B = p_A + v$, with $v \geq 0$, then the maximum value for p_{AB} is p_A, so that the maximum value for D is $p_A - p_A(p_A + v)$. Under these conditions and if r^2 exceeds some threshold t, then r^2 can be written as

$$r^2 = \frac{[p_{AB} - p_A p_B]^2}{p_A(1-p_A)p_B(1-p_B)} = \frac{[p_A - p_A(p_A+v)]^2}{p_A(1-p_A)(p_A+v)(1-p_A-v)}$$

$$= \frac{p_A(1-p_A-v)}{(1-p_A)(p_A+v)} \geq t$$

Rearrangement of this equation shows that if the allele frequency at locus 1 is p_A then the maximum allele frequency at the second locus, given that $r^2 \geq t$,

Table 2

Limits on Allele Frequency at Locus 2 (p_B) Given Allele Frequency at Locus 1 (p_A) and LD Measure > t for Measures of LD Whose Range is Dependent on the Allele Frequencies at the Two Loci

Measure of LD	Symbol	Estimate	v_{min}	v_{max}	Lower limit of p_B	Upper limit of p_B
Squared correlation	r^2	$D^2/p_A(1-p_A)p_B(1-p_B)$	$p_A(1-p_A)(1-t)/(1-p_A(1-t))$	$p_A(1-p_A)(1-t)/(p_A(1-t)+t)$.	$t\,p_A/(1-p_A(1-t))$	$p_A/(p_A(1-t)+t)$
Regression	b	$D/p_B(1-p_B)$	$p_A(1-t)$	$p_A(1-t)/t$	$t\,p_A$	$1/t$
Frequency difference	f or d	$D/p_A(1-p_A)$	$(1-p_A)(1-t)$	$(1-p_A)(1-t)$	$p_A-(1-p_A)(1-t)$	$1-t(1-p_A)$

Note: b and f listed in Table 2 of Morton et al. (2001); d used by Kruglyak (1999)

is $p_B = p_A + v_{max}$, where $v_{max} = (1 - p_A)(1 - t)/\{p_A(1 - t) + t\}$. The minimum allele frequency at the second locus, $p_B = p_A - v_{min}$ can be derived in a similar way, $v_{min} = (1 - p_A)(1 - t)/\{1 - p_A(1 - t)\}$. Similar limits can be derived for other measures of LD which are dependent on allele frequency and are included here for completeness (Table 2).

Association Studies

We can utilize these results to consider some aspects of design and interpretation of association studies. Zondervan and Cardon (2004; Box 3, equation 3), using the results of Ackerman et al. (2003), presented an expression for the allelic odds ratio present at a marker locus in LD with a causal variant (OR_M) in terms of the odds ratio at the causal variant (OR_T), the allele frequencies at the marker (p_M) and causal variant (p_T) loci and the disequilibrium covariance (D) between them:

$$OR_M = 1 + \frac{D(OR_T - 1)}{p_M[(1 - p_M) + (p_T(1 - p_M) - D)(OR_T - 1)]} \quad [1]$$

They considered some specific examples of well-established complex disease associations. They examined ranges of marker allele frequencies and D' and calculated the OR_M that would arise from these combinations (their Figure 1). We consider three of their examples: a) Type 2 diabetes, PPAR$_\gamma$, $p_T = .85$, $OR_T = 1.23$; c) Alzheimer's disease, APOE(*4), $p_T = .15$, $OR_T = 3.3$; d) deep vein thrombosis F5, $p_T = .03$, $OR_T = 3.8$). We include one additional example to complete a broad spectrum of causal allele frequencies: b) $p_T = .5$, $OR_T = 2.5$ and examine OR_M for the possible range of p_M given the r^2 between the disease and marker loci for a range of r^2. The method of Risch and Teng (1998) was used to determine the power of detecting the OR_M given allele frequencies and r^2. Programs used to generate the results replicated the results of Zondervan and Cardon (2004) when D' was used as the measure of LD rather than r^2.

When an association study is conducted, it is the association at the marker locus (i.e., OR_M) that is observed. Interpretation of the detected association requires us to consider the possible range of effects at the underlying causal locus (i.e., OR_T) for a range of possible p_T and r^2 between the marker and disease loci. To illustrate this and to allow comparison with examples a–d, we have considered four examples: e) $p_M = .85$, $OR_M = 1.23$; f) $p_M = .5$, $OR_M = 2.5$; g) $p_M = .15$, $OR_M = 3.3$; h) $p_M = .03$, $OR_M = 3.8$) and have estimated OR_T using a rearrangement of equation [1]:

$$OR_T = 1 + \frac{(OR_M - 1)p_M(1 - p_M)}{D - p_M(OR_M - 1)(p_T(1 - p_M) - D)}.$$

Results

The relationship between allele frequencies and r^2 is demonstrated in Figure 1a, where p_A is plotted against

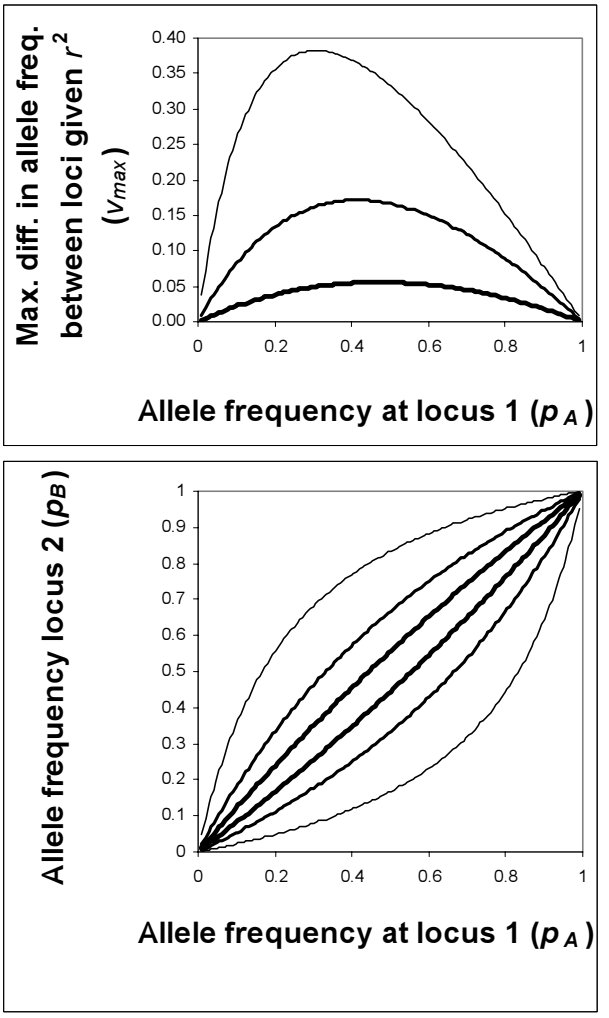


Figure 1

1a: Maximum difference (v_{max}) between allele frequency at locus 2 (p_B) given the allele frequency at locus 1 (p_A), $p_B = p_A + v_{max}$ and the LD between the two loci (r^2), for $r^2 \geq .8$ (___), .5 (__) and .2 (__).

1b: Possible range of allele frequencies at two loci given the LD between the two loci (r^2). All possible combinations of allele frequencies are contained within the ellipses for $r^2 \geq .8$ (___), .5 (__) and .2 (__).

v_{max} for $t = r^2 = .2, .5, .8$. Figure 1b plots p_A against the minimum and maximum values for p_B for a given $r^2 = .2, .5, .8$ so that all possible combinations of allele frequencies for these r^2 are contained within the ellipses bounded by the minima and maxima.

The relationship between the allele frequency at the marker locus and the r^2 between marker and causal locus and the size of association that can be detected is examined in Figure 2 (a–d) in which OR_M is presented for the possible range of p_M given the r^2 between the disease and marker loci for a range of r^2. For each example, the maximum value of OR_M occurs when $OR_M = OR_T$ when $p_M = p_T$ and $r^2 = 1$. In Figure 2, contours of $r^2 = .2, .5, .8$ show that OR_M is always less than OR_T (as expected from definition in equation 1) and the length of the contours is shorter the higher the r^2,

reflecting the limitation on the range of p_M given the r^2 (Figure 1). From the shape of the contours, we see that for a given r^2 the observed OR_M is not highest when $p_M = p_T$; when $p_T < .5$ (examples c and d), the OR_M is highest for any given r^2 relationship when p_M is the minimum that is possible. When $p_T > .5$ (examples a and b) the OR_M is highest for any given r^2 relationship when p_M is the maximum that is possible; in this case the minor allele could be considered to be the protective allele, and the frequency of the minor allele is the minimum possible given the r^2 relationship. The shape of the contours also reflects the definition of the odds ratio: $OR_M = \{p_{MD}/(1 - p_{MD})\}/\{p_M/(1 - p_M)\}$, where p_{MD} is the allele frequency of the associated marker allele in cases, and results in the nonsymmetry of the r^2 contours in p_M about $p_M = p_T$ when $p_T = .5$. Figure 3 presents the power of detecting the OR_M shown in Figure 2; the sample size (N = number of cases = number of controls) in each example was chosen to be the minimum possible to achieve 80% power assuming a Type I error of, 5×10^{-5} when $p_M = p_T$ and $r^2 = 1$. These examples show that if a study is designed to have 80% power to detect an association between a causal variant and disease, even under favorable circumstances (high LD, $r^2 = .8$, between genotyped and causal variant, and therefore a maximum difference between the frequencies of the coupled alleles of .06), the power for detecting the association may only be 60%. For these examples, if the sample size is set to achieve 99.9% power (results not shown), then the r^2 between genotyped and causal variants must be at least .5, maximum coupled allele frequency difference of .1, to achieve 80% power for detecting an association. Figure 2 (e, f, g, h) presents OR_T for the possible range of p_T given the observed p_M and OR_M for a range of r^2 between marker and disease loci.

Discussion

We have quantified the relationship between r^2 and allele frequency and have shown that the constraints on the difference in frequency of the coupled loci are severe if a high r^2 between the loci is desired (Figure 1). For example, for $r^2 \geq .8$, the maximum difference in allele frequency is ± .06 which occurs when one locus has allele frequency .5. For $r^2 \geq .8$ and allele frequency at one locus of .1, the maximum difference in allele frequency at the second locus is only ± .02. We can utilize these results to consider some aspects of design and interpretation of association studies.

SNP Selection

In the design of association studies SNPs are selected so that the LD landscape of the region under study is adequately represented when testing for association with disease status. Figure 1a illustrates that methods that eliminate SNPs only if they make r^2 greater than some threshold t with a retained SNP are also imposing very severe restrictions on allele frequencies on eliminated SNPs. For the criterion $t = .8$ as investigated by Carlson et al. (2004), a 'bin' in which every eliminated SNP must have $r^2 \geq .8$ with a selected SNP implies that the maximum range in allele frequency occurs when the selected SNP has frequency of .5 and the eliminated SNPs have frequencies of the coupled alleles in the range .50 ± .06. For a selected SNP with minor allele frequency .10, the range in allele frequencies of the eliminated SNPs is limited to the very narrow range of .10 ± .02. Estimates of the number of SNPs required to represent the LD landscape of the whole genome (for European Americans) based on this SNP selection method are as high as 250,000 (Carlson et al., 2004), which is partly a reflection of the underlying severe restriction on difference in allele frequencies between loci as well as the imposed restriction on the LD between them. Recognition that the use of pairwise r^2 as a criterion for SNP selection results in high numbers of SNPs selected has motivated methods that utilize LD information from more than two SNPs at a time. The haplotype tagging SNP method proposed by Clayton (2002) selects SNPs ('haplotype tagging SNPs' or 'htSNPs') on the basis of the proportion of diversity of all haplotypes (in a linear regression) explained by the selected set of SNPs (i.e., coefficient of determination or haplotype r^2). This method has resulted in estimates of up to 50% fewer SNPs required to represent the whole genome compared to estimates using pairwise measures of LD (Goldstein et al., 2003). The relationship between the power of an association study and the r^2 between a genotyped SNP and a causal SNP (Pritchard & Przeworski, 2001; Risch & Teng, 1998) extends to the haplotype r^2 between a set of htSNPs and a causal variant (Goldstein et al., 2003). The properties of haplotype r^2 have been much investigated (e.g., Goldstein et al , 2003; Ke et al., 2004; Meng et al., 2003; Weale et al., 2003), but the conclusions are similar to those that have compared the use of microsatellites to SNPs in association studies, where combinations of htSNPs are analogous to the use of microsatellites: Xiong and Jin (1999) found that because microsatellites have more alleles, the probability of a microsatellite having an allele (or set of alleles) whose frequency is close to and is in coupling with the unknown causal variant is higher than for a SNP. There are usually many sets of htSNPs that all explain a similar proportion of diversity of a full set of genotyped SNPs (Johnson et al., 2001), but it is the low minor allele frequency SNPs from high |D'| blocks that are likely to be consistently eliminated when htSNPs are selected, because combinations of alleles in an htSNP haplotype (or sets of haplotypes) are likely to achieve similar frequency to the minor allele of the eliminated SNP. Therefore, if the desired level of haplotype r^2 is set too low, rare haplotypes may not be represented by the set of htSNPs because the required matching of coupled haplotype frequencies has not been achieved (Ke et al., 2004).

Size of Effect That Can be Detected

The second aspect of association studies for which we consider the impact of the relationship between r^2 and

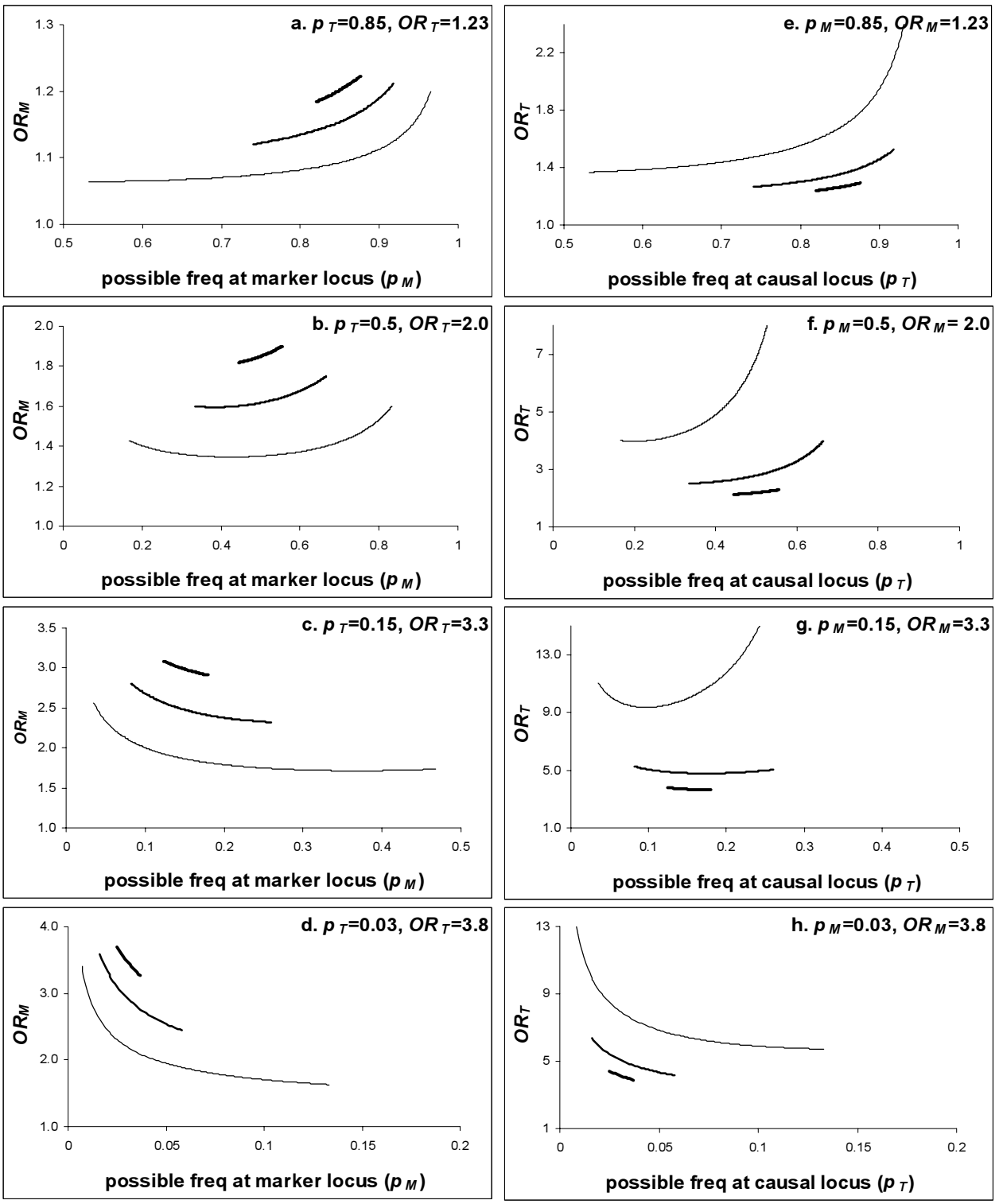


Figure 2

a, b, c, d. The allelic odds ratios (OR_M) at possible marker loci that are in LD with an underlying causal variant and the frequency of the associated allele (p_M). The causal locus has associated allele frequency p_T and allelic odds ratio OR_T. The LD contours are for r^2 = .8 (____), .5 (__) and .2 (__).

e, f, g, h. The underlying allelic odds ratios at the causal locus (OR_T) and the frequency of the associated allele at the causal locus (p_T) that could explain the observed allelic odds ratio at a genotyped locus (OR_M) given the observed associated allele frequency observed at the genotyped locus (p_M) and the LD between the causal and genotyped loci. OR_T truncated at limit of axis. The LD contours are for r^2 = .8 (____), .5 (__) and .2 (__).

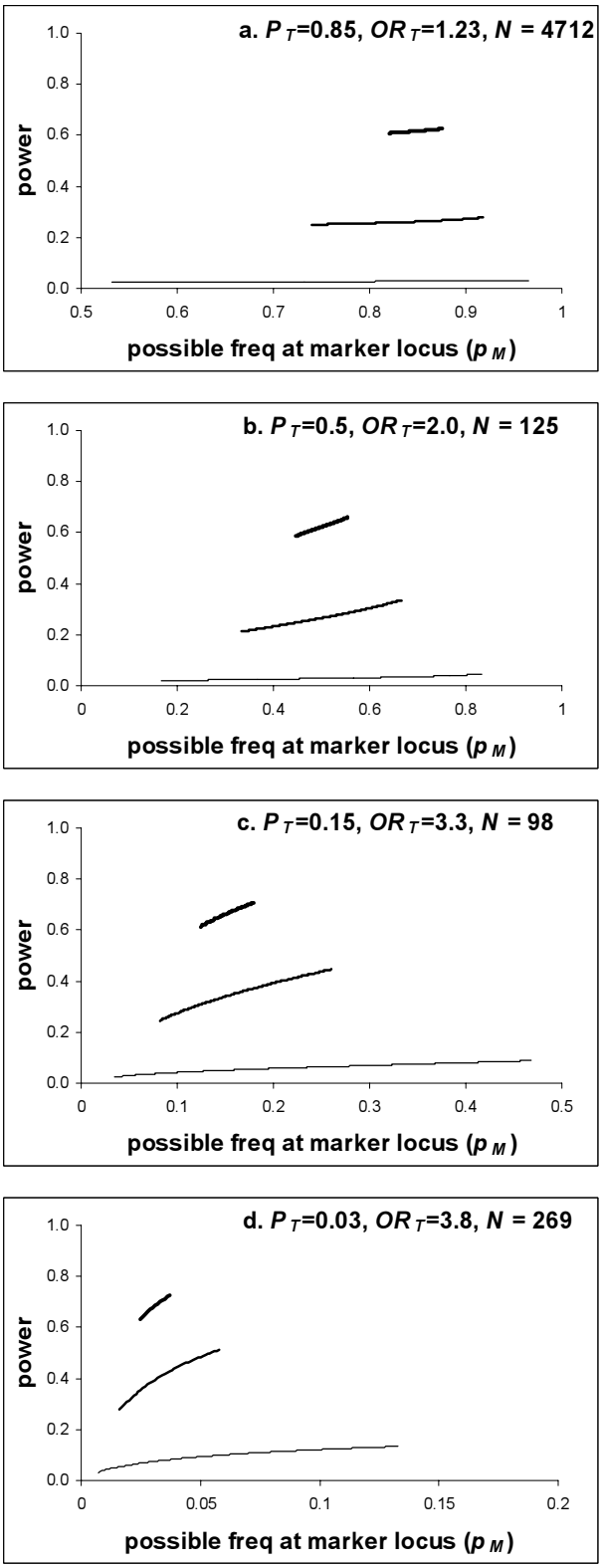


Figure 3

The power of detecting the OR_M shown in Figure 2a–d, when Type I error is 5×10^{-5} and for N cases and N controls, chosen so that 80% power is achieved for the situation when $p_M = p_T$ and r^2 = .8 (____), .5 (__) and .2 (__).

allele frequencies is the size of the association effect of a causal locus that can be detected given both the marker SNPs genotyped and the sample size of the study and thus given the power of the study. The size of association effect at the marker locus, as expressed by the odds ratio OR_M was investigated (Figure 2a–d). The results are not unsurprising: a smaller range of OR_M for higher r^2, highest OR_M when min (p_M, $1 - p_M$) is at its smallest value given p_T and r^2, nonsymmetry of the r^2 contours in p_M about $p_M = p_T$ when $p_T = .5$. These results partly reflect the relationship between the allele frequencies p_M and p_T, and the definition of the odds ratio as a ratio of ratios, $OR_M = \{p_{MD}/(1 - p_{MD})\}/\{p_M/(1 - p_M)\}$. Nonetheless, without the quantification of the relationship between allele frequencies and r^2, these results may have been underappreciated.

If a study is designed to have 80% power to detect an association between a causal variant and disease, even under favorable circumstances (of high LD, r^2 = .8, between genotyped and causal variant, and therefore a maximum difference between the frequencies of the coupled alleles of .06) the power for detecting the association may only be 60% (Figure 3). These results are not inconsistent with those of Risch and Teng (1998) and Pritchard and Przeworksi (2001) that a study must be of size N/r^2 to retain the same power when a marker locus is genotyped compared to size N when the causal locus is genotyped, where r^2 is the LD between marker and causal locus. In practice, the study size is often limited by the number of samples available and so Figure 3 illustrates the loss of power for a fixed sample size and reflects the relationship between allele frequencies and odds ratio. Zondervan and Cardon (2004, Figure 2) investigated the power of association studies under high, moderate and low allelic odds ratios each considered in the presence of low, moderately low and high |D'|, for all combinations of causal and marker variant. The ellipsoid shape of their graphs with highest power along the diagonal where allele frequencies are equal is a reflection of the r^2 relationship presented in Figure 1b. Where power is low when |D'| is high (their bottom right graph) is when the difference between frequencies of the causal variant and its correlated marker allele is high, and so r^2 is low.

Size of Causal Effect

The final aspect of association studies for which we consider the consequences of the boundaries on allele frequencies imposed by the r^2 between them is the interpretation of an association once it is detected. When an association study is conducted, it is the association at the marker locus (i.e., OR_M) that is observed. Interpretation of the detected association requires us to consider the possible range of effects at the underlying causal (i.e., OR_T) for a range of possible p_T and r^2 between the marker and disease loci. This was examined in Figure 2e–h) and allows us to give quantification to the usual statement made in association studies: 'the genotyped variant, or something in tight LD with it, is associated with …'. When a true

allelic association is detected, the limits on the underlying causal variant are that it is likely to be $r^2 > .5$ with the genotyped variant and allele frequency difference of .1 or less. Design of a replication study based on these results would also need to account for sampling of the cases and controls from the population (Dahlman et al., 2002).

In conclusion, we have quantified the boundaries on allele frequencies at two loci as constrained by the r^2 LD between them. The r^2 relationship between a causal variant and a marker variant has impact on three important aspects of association studies, namely, i) selection of SNPs that are representative of the LD landscape; ii) power calculations and what we can expect to detect in an association study; and iii) interpretation of an association once detected. In each of these aspects, boundaries on the allele frequencies at the two loci, given the r^2 between them, have been recognized qualitatively, but the lack of quantification may have led to underappreciation of the importance and impact of these constraints. Our results reiterate that r^2 is a complex and, perhaps, not an ideal measure of LD as shown by Hedrick (1987), nonetheless it has some useful properties and remains a commonly used measure. We have shown that the constraints on the difference in frequency of the coupled loci are severe if a high r^2 between the loci is desired. For example, for $r^2 \geq .8$, the maximum difference in allele frequency is ± .06 which occurs when one locus has allele frequency .5. For $r^2 \geq .8$ and allele frequency at one locus of .1, the maximum difference in allele frequency at the second locus is only ± .02. For studies designed to have 80% power to detect an association between a causal variant and disease, even under favorable circumstances (of high LD, $r^2 = .8$, between genotyped and causal variant, and therefore a maximum difference between the frequencies of the coupled alleles of .06), the power for detecting the association may only be 60%. Finally, when a true allelic association is detected in an association study, the limits on the underlying causal variant are that it is likely to have $r^2 > .5$ and maximum allele frequency difference of .1 with the genotyped variant.

Acknowledgments

This work was supported by grants from Organon NV and the UK Medical Research Council. I would like to thank Bill Hill, Albert Tenesa and Peter Visscher for commenting on the manuscript.

References

Ackerman, H., Usen, S., Mott, R., Richardson, A., Sisay-Joof, F., Katundu, P., Taylor, T., Ward, R., Molyneux, M., Pinder, M., & Kwiatkowski, D. P. (2003). Haplotypic analysis of the TNF locus by association efficiency and entropy. *Genome Biology, 4,* R24.

Carlson, C. S., Eberle, M. A., Rieder, M. J., Yi, Q., Kruglyak, L., & Nickerson, D. A. (2004). Selecting a maximally informative set of single-nucleotide polymorphisms for association analyses using linkage disequilibrium. *American Journal of Human Genetics, 74,* 106–120.

Clayton, D. (2002) Choosing a set of haplotype tagging SNPs from a larger set of diallelic loci. Retrieved February 3, 2005, from *http://www-gene.cimr.cam.ac.uk/clayton/software/stata/htSNP/htsnp.pdf*

Dahlman, I., Eaves, I. A., Kosoy, R., Morrison, V. A., Heward, J., Gough, S. C., Allahabadia, A., Franklyn, J. A., Tuomilehto, J., Tuomilehto-Wolf, E., Cucca, F., Guja, C., Ionescu-Tirgoviste, C., Stevens, H., Carr, P., Nutland, S., McKinney, P., Shield, J. P., Wang, W., Cordell, H. J., Walker, N., Todd, J. A., & Concannon, P. (2002). Parameters for reliable results in genetic association studies in common disease. *Nature Genetics, 30,* 149–150.

Devlin, B., & Risch, N. (1995). Comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics, 29,* 311–322.

Garner, C., & Slatkin, M. (2003). On selecting markers for association studies: Patterns of linkage disequilibrium between two and three diallelic loci. *Genetic Epidemiology, 24,* 57–67.

Goldstein, D. B., Ahmadi, K. R., Weale, M. E., & Wood, N. W. (2003). Genome scans and candidate gene approaches in the study of common diseases and variable drug responses. *Trends in Genetics, 19,* 615–622.

Hedrick, P. W. (1987). Gametic disequilibrium measures: Proceed with caution. *Genetics, 117,* 331–341.

Hill, W. G., & Robertson, A. (1968). Linkage disequilibrium in finite populations. *Theoretical and Applied Genetics, 38,* 226–231.

Johnson, G. C., Esposito, L., Barratt, B. J., Smith, A. N., Heward, J., Di Genova, G., Ueda, H., Cordell, H. J., Eaves, I. A., Dudbridge, F., Twells, R. C., Payne, F., Hughes, W., Nutland, S., Stevens, H., Carr, P., Tuomilehto-Wolf, E., Tuomilehto, J., Gough, S. C., Clayton, D. G., & Todd, J. A. (2001). Haplotype tagging for the identification of common disease genes. *Nature Genetics, 29,* 233–237.

Ke, X., Durrant, C., Morris, A., Hunt, S., Bentley, D. R., Deloukas, P., & Cardon, L. R. (2004). Efficiency and consistency of haplotype tagging of dense SNP maps in multiple samples. *Human Molecular Genetics, 13,* 2557–2565.

Kruglyak, L. (1999). Prospects for whole-genome linkage disequilibrium mapping of common disease genes. *Nature Genetics, 22,* 139–144.

Meng, Z., Zaykin, D. V., Xu, C. F., Wagner, M., & Ehm, M. G. (2003). Selection of genetic markers for association analyses, using linkage disequilibrium and haplotypes. *American Journal of Human Genetics, 73,* 115–130.

Morton, N. E., Zhang, W., Taillon-Miller, P., Ennis, S., Kwok, P. Y., & Collins, A. (2001). The optimal measure of allelic association. *Proceedings of the National Academy of Sciences USA, 98,* 5217–5221.

Muller-Myhsok, B., & Abel, L. (1997). Genetic analysis of complex diseases. *Science, 275,* 1328–1329.

Nordborg, M., & Tavare, S. (2002). Linkage disequilibrium: What history has to tell us. *Trends in Genetics, 18,* 83–90.

Ohashi, J., & Tokunaga, K. (2001). The power of genome-wide association studies of complex disease genes: Statistical limitations of indirect approaches using SNP markers. *Journal of Human Genetics, 46,* 478482.

Pritchard, J. K., & Przeworski, M. (2001). Linkage disequilibrium in humans: Models and data. *American Journal of Human Genetics, 69,* 1–14.

Risch, N. J. (2000). Searching for genetic determinants in the new millennium. *Nature, 405,* 847–856.

Risch, N., & Merikangas, K. (1996). The future of genetic studies of complex human diseases. *Science, 273,* 1516–1517.

Risch, N., & Teng, J. (1998). The relative power of family-based and case-control designs for linkage disequilibrium studies of complex human diseases I. DNA pooling. *Genome Research, 8,* 1273–1288.

Weale, M. E., Depondt, C., Macdonald, S. J., Smith, A., Lai, P. S., Shorvon, S. D., Wood, N. W., & Goldstein, D. B. (2003). Selection and evaluation of tagging SNPs in the neuronal-sodium-channel gene SCN1A: Implications for linkage-disequilibrium gene mapping. *American Journal of Human Genetics, 73,* 551–565.

Xiong, M., & Jin, L. (1999). Comparison of the power and accuracy of biallelic and microsatellite markers in population-based gene-mapping methods. *American Journal of Human Genetics, 64,* 629–640.

Zondervan, K. T., & Cardon, L. R. (2004). The complex interplay among factors that influence allelic association. *Nature Reviews: Genetics, 5,* 89–100.