

## ARTICLE

# Pairwise linkage disequilibrium under disease models

Steven J Schrodi<sup>\*,1</sup>, Veronica E Garcia<sup>1</sup>, Charley Rowland<sup>1</sup> and Hywel B Jones<sup>2</sup>

<sup>1</sup>Statistical Genetics, Celera Diagnostics Inc., Alameda, CA, USA; <sup>2</sup>ParAllele BioScience Inc., South San Francisco, CA, USA

Many genetic studies of disease association rely heavily on linkage disequilibrium (LD) patterns between pairs of markers to detect susceptibility markers. This is true of large-scale positional mapping approaches as well as haplotype construction, selection of tagging single-nucleotide polymorphisms and population genetic analyses. Whereas the distribution of different LD measures has been investigated for randomly selected chromosomes from populations undergoing a variety of demographic effects, little is known about LD within disease-affected samples, and how various disease models influence the difference in LD between patients and the general population. As whole-genome efforts are now underway to characterize and utilize LD patterns in randomly sampled individuals, knowledge about the extent that LD differs between patients and the general population becomes crucial. Such information will allow investigators to design improved mapping experiments and better understand haplotype information arising from such experiments. In this paper, we explore two-site LD measures in the context of single gene disease models. Analytic expressions are presented for infinite populations and properties of sampling densities are reported for different disease models. Interestingly, results indicate that 'underdominant', some dominant, recessive and 'protective' disease models generate weaker LD levels in patients compared to the general population, whereas other models produce stronger LD among affected individuals. Analytic results are also presented for the ratio of LD in patients to the LD in the general population as a function of recombination fraction using a Haldane model. In addition, we explore the impact of various allele frequency combinations on LD differences.

*European Journal of Human Genetics* (2007) 15, 212–220. doi:10.1038/sj.ejhg.5201731; published online 15 November 2006

**Keywords:** mapping; linkage disequilibrium; disease models; haplotype

## Introduction

Large-scale genetic association studies generally depend at least in part on the existence of linkage disequilibrium (LD) between genetic markers and a disease locus (for a review see Clark<sup>1</sup>). LD is a concept of statistical correlation between alleles segregating at two or more loci.<sup>2</sup> Conversely, linkage equilibrium refers to the state where the alleles at a particular locus are independently distributed

with respect to the alleles at an alternative locus. There are several ways in which LD can be generated in a sample of chromosomes. Population genetics factors can produce LD through a variety of processes such as natural selection, strong genetic drift, admixture and new mutations.<sup>3,4,5</sup> Similarly, skewed sampling of chromosomes from a population, for example, the selection of disease-affected individuals, can also give rise to LD levels higher or lower than expected. Importantly, this type of skewed sampling can also produce departures from what is termed the 'fundamental theorem of the HapMap' (terminology from Terwilliger and Hiekkalinna, 2006).<sup>6</sup> The fundamental theorem states that the statistical power to detect disease association indirectly at a marker locus in LD with a

\*Correspondence: Dr SJ Schrodi, Statistical Genetics, Celera Diagnostics Inc., 1311 Harbor Bay Pkwy, Alameda, CA 94502, USA.  
Tel: +1 510 749 4270; Fax: +1 510 743 6250;  
E-mail: Steven.Schrodi@celeradiagnostics.com  
Received 13 March 2006; revised 28 September 2006; accepted 4 October 2006; published online 15 November 2006

disease-susceptibility locus is approximately the same as the power to detect disease-association directly at the susceptibility locus, if the sample size is increased by a factor of  $1/r^2$ , where  $r^2$  is the commonly used measure of pairwise LD. Put concretely, imagine two loci, one directly involved in disease and the other in LD with the first, causative locus with  $r^2 = \frac{1}{3}$  between them. If 500 cases and 500 controls are required to obtain 80% power to detect disease association at the causative locus, then the fundamental theorem states that approximately 1500 cases and 1500 controls are necessary to reach the same 80% power level at the marker locus in LD with the causative locus. As pairwise LD is correlated with disease status, some level of departure from the fundamental theorem is to be expected. This point was made explicitly by Terwilliger and Hiekkalinna<sup>6</sup> where they argued that the fundamental theorem only applies, among other conditions, if the 'LD between loci and the etiological effect of the functional variant are independent of each other.'

Recently, there has been considerable interest in creating and utilizing whole-genome haplotype maps for the purposes of disease-susceptibility mapping in humans<sup>7,8</sup> and investigation of genetic structure of populations. These maps allow one to quantify the strength of LD across the entire genome. By choosing representative markers from sets of markers that are in high LD with each other, investigators aim to reduce drastically the number of markers necessary to interrogate adequately the genome.<sup>9,10</sup> Additionally, the pattern of disease association decay with decreasing pairwise LD can be used to identify regions that are more likely to carry predisposing chromosomal segments. As LD maps are typically constructed from randomly sampled individuals, understanding the effect that different disease models have on modifying the level of LD in patients is important: such information can be used to (i) better select tagging markers for large-scale studies and (ii) construct statistical tests to better understand if specific regions are disease-predisposing. To these ends, we wanted to investigate the impact that different disease models have on traditional measures of LD.

In this paper, we derive simple analytic results for commonly used measures of LD under general single gene disease models, defining regions of the parameter space that give rise to LD levels in disease-affected individuals either above or below the general population LD level. We then investigate LD sampling properties given general population haplotype frequencies under a neutral coalescent. The results characterize the effect of disease models on LD patterns. This work may change how current HapMap data are used to select tagging single-nucleotide polymorphisms (SNPs). For example, in some instances, it may be desirable to genotype densely a small set of affected individuals alongside a small set of control individuals, just as the HapMap project densely genotyped randomly

selected individuals. Additionally, selection of SNPs to perform fine-scale mapping, once associated markers are identified, may be informed by explicitly modeling LD patterns differently between cases and controls. Lastly, this type of information may also enable improved statistical tests for identifying regions with disequilibrium patterns that correspond to those expected under certain disease models.

Important developments in this area can be found in a study by Nielsen *et al*,<sup>11</sup> where the authors construct a statistical test using LD differences between cases and controls, thereby providing researchers an additional method for testing for association, aside from the more traditional haplotype-based contingency table tests of homogeneity. More recently, an extension of this work was published showing analytic and graphical methods for this LD-contrast-type test.<sup>12</sup>

### Theory

To characterize pairwise LD in preferentially selected groups of individuals, we will define a simple single gene disease model and explore two commonly used measures,  $D$  and  $r^2$ , as a function of penetrance parameters and allele frequencies. Both asymptotic and sampling results are presented. For a two-locus model, say loci  $A$  and  $B$ , in which two alleles segregate at each locus,  $D$  is defined as  $p_{11}p_{22} - p_{12}p_{21}$ , where  $p_{ij}$  is the frequency of the  $A_iB_j$  haplotype. Denote  $A_1B_1$  and  $A_2B_2$  as parental haplotypes, and the remaining two as recombinant haplotypes. See Devlin and Risch,<sup>13</sup> for a review of these and other measures of LD.  $r^2$ , the squared correlation coefficient between alleles at the two loci,<sup>14</sup> is a normalized version of  $D$  and is defined as

$$r^2 = \frac{D^2}{p_{1\bullet}(1 - p_{1\bullet})p_{\bullet 1}(1 - p_{\bullet 1})} \quad (1)$$

denoting the margins (single-locus allele frequencies)  $p_{11} + p_{12}$  and  $p_{11} + p_{21}$  by  $p_{1\bullet}$  and  $p_{\bullet 1}$ , respectively. We will treat both LD measures as being calculated in two distinct ways: as population parameters and as sampling statistics from a small set of chromosomes from affected individuals, the former of which will be called 'asymptotic results' and the latter 'sampling results'.

Suppose now that the  $A$  locus postulated above has a variant that predisposes carriers to a disease phenotype. Following this characterization, designate the  $B$  locus as the marker locus with no causal relationship to the disease phenotype. Denote the two alleles segregating at the  $A$  locus by  $A_1$  and  $A_2$ . Further define three genotypic penetrances to specify a single gene disease model,  $f_{11} = P[Dz|A_1A_1]$ ,  $f_{12} = P[Dz|A_1A_2]$  and  $f_{22} = P[Dz|A_2A_2]$ . Let us define the frequencies of the haplotypes in affected individuals as  $\tilde{p}_{ij}$ , using analogous definitions as in the general population. Assuming Hardy-Weinberg equilibrium (HWE), we arrive at the set of affected haplotype

frequencies by applying Bayes' rule:

$$\begin{aligned} \tilde{p}_{11} &= \frac{p_{11}}{K} [f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})]; & \tilde{p}_{12} &= \frac{p_{12}}{K} [f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})] \\ \tilde{p}_{21} &= \frac{p_{21}}{K} [f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})]; & \tilde{p}_{22} &= \frac{p_{22}}{K} [f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})] \end{aligned} \quad (2 - 5)$$

where the prevalence of disease is  $K = P[Dz]$ , and making use of the Hardy-Weinberg assumption,  $K = f_{11}p_{1\bullet}^2 + 2f_{12}p_{1\bullet}(1 - p_{1\bullet}) + f_{22}(1 - p_{1\bullet})^2$ . Similar equations, using different notation, can be found in Nielson et al.<sup>11</sup> The single-locus allele frequencies in affected individuals are simply

$$\tilde{p}_{1\bullet} = \frac{p_{1\bullet}}{K} [f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})] \quad (6)$$

$$\tilde{p}_{\bullet 1} = \frac{p_{11}}{K} [f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})] + \frac{p_{21}}{K} [f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})] \quad (7)$$

It is now a matter of simple algebra to calculate LD measures in patients under the general single gene model. Combining the above results allows  $\tilde{D}_\infty$ , the LD in all affected individuals, to be expressed in terms of the general population  $D_\infty$  and multiplicative factor:

$$\tilde{D}_\infty = \frac{D_\infty}{K^2} [f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})][f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})] \quad (8)$$

where the infinity subscript is shown to indicate that this is an asymptotic result applying to the population as a whole. If we set each of the penetrances equal to the same constant, it can be easily verified that  $|\tilde{D}_\infty| = |D_\infty|$  as expected. Similarly, an expression for  $\tilde{r}_\infty^2$  can be obtained

$$\tilde{r}_\infty^2 = r_\infty^2 \left( \frac{\tilde{D}_\infty}{D_\infty} \right)^2 \frac{[p_{1\bullet}p_{\bullet 1}(1 - p_{1\bullet})(1 - p_{\bullet 1})]}{[\tilde{p}_{1\bullet}\tilde{p}_{\bullet 1}(1 - \tilde{p}_{1\bullet})(1 - \tilde{p}_{\bullet 1})]} \quad (9)$$

For the sake of brevity, we write  $\tilde{r}_\infty^2$  in terms of the allele frequencies in the affected individuals in the denominator:

$$\begin{aligned} \tilde{r}_\infty^2 &= \frac{r_\infty^2}{K^4} \left( \frac{1}{\tilde{p}_{1\bullet}\tilde{p}_{\bullet 1}(1 - \tilde{p}_{1\bullet})(1 - \tilde{p}_{\bullet 1})} \right) \\ &\quad \times p_{1\bullet}p_{\bullet 1}(1 - p_{1\bullet})(1 - p_{\bullet 1}) [f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})]^2 \\ &\quad \times [f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})]^2 \end{aligned}$$

The ratio of  $\tilde{r}_\infty^2$  to  $r_\infty^2$  is therefore

$$\frac{[f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})][f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})][p_{\bullet 1}(1 - p_{\bullet 1})]}{C[f_{11}p_{11}p_{1\bullet} + f_{12}(p_{11} - 2p_{11}p_{1\bullet} + p_{1\bullet}p_{\bullet 1}) + f_{22}(1 - p_{1\bullet})(p_{\bullet 1} - p_{11})]} \quad (10)$$

where

$$C = f_{11}p_{1\bullet}(p_{11} - p_{1\bullet}) + f_{12}[p_{1\bullet}(p_{\bullet 1} + 2p_{1\bullet} - 2) - 2p_{11}p_{1\bullet} + p_{11}] + f_{22}(1 - p_{1\bullet})(p_{\bullet 1} + p_{\bullet 1} - p_{11} - 1)$$

These asymptotic LD measures can be examined under specific disease models by positing relationships between the three penetrances. Evaluation of four classic models, dominant, recessive, additive and multiplicative, will shed

some light on how phenotype-based sampling modifies levels of LD. First, consider the pure dominant model where  $f_{11} = f_{12}$  and  $f_{22} = 0$  (this and other models are considered 'pure' models when one or more of the penetrances of genotypes not carrying a predisposing allele is 0 - that is, the prevalence is zero in the absence of the predisposing allele). Under this dominant model,

$$\tilde{D}_\infty = \frac{D_\infty}{p_{1\bullet}(2 - p_{1\bullet})^2} \quad (11)$$

Hence, when the probability of the predisposing allele,  $p_{1\bullet}$ , is less than  $(3 - \sqrt{5})/2$  (approximately 0.381966),  $|\tilde{D}_\infty| > |D_\infty|$ ; otherwise, the LD in patients is less than the general population value, ignoring trivial solutions. Generalizing this dominant model by considering  $f_{22} \geq 0$ ,

$$\tilde{D}_\infty = \frac{D_\infty \gamma [p_{1\bullet}(\gamma - 1) + 1]}{[-p_{1\bullet}^2(\gamma - 1) + 2p_{1\bullet}(\gamma - 1) + 1]^2} \quad (12)$$

For this and subsequent results, the notation is changed to genotype relative risk (represented as  $\gamma$ ), such that  $(f_{11}/f_{12}) = (f_{12}/f_{22}) = \gamma$  for the above model.  $\tilde{D}_\infty$  in patients under the pure recessive mode of inheritance, where  $f_{12} = f_{22} = 0$ , is zero regardless of allele frequency. The reason for this is that all patients must have the  $A_1A_1$  genotype and therefore the only two possible haplotypes,  $A_1B_1$  and  $A_1B_2$ , necessarily yielding  $\tilde{D}_\infty = 0$ . The general recessive model,  $f_{22} = f_{12}$ ,  $(f_{11}/f_{22}) = \gamma$ , has richer dynamics:

$$\tilde{D}_\infty = \frac{D_\infty(\gamma p_{1\bullet} - p_{1\bullet} + 1)}{(\gamma p_{1\bullet}^2 - p_{1\bullet}^2 + 1)^2} \quad (13)$$

Analysis of equation (13) shows that for high-frequency predisposing alleles,  $|\tilde{D}_\infty| < |D_\infty|$ ; otherwise low-frequency predisposing alleles with a recessive mode of inheritance produce higher LD levels in patients (see Table 1). Considering the two intermediate models, general additive  $(f_{11}/f_{22}) = 2\gamma - 1, (f_{12}/f_{22}) = \gamma$  and multiplicative  $(f_{11}/f_{22}) = \gamma^2, (f_{12}/f_{22}) = \gamma$  models, give

$$\tilde{D}_\infty = \frac{D_\infty(\gamma p_{1\bullet} - p_{1\bullet} + 1)(\gamma p_{1\bullet} - p_{1\bullet} + \gamma)}{(2\gamma p_{1\bullet} - 2p_{1\bullet} + 1)^2} \quad (14)$$

and

$$\tilde{D}_\infty = \frac{D_\infty \gamma}{(\gamma p_{1\bullet} - p_{1\bullet} + 1)^2} \quad (15)$$

respectively. Under the additive model,  $|\tilde{D}_\infty| < |D_\infty|$  for allele frequencies,  $p_{1\bullet}$ , above  $(\gamma - 3 + \sqrt{\gamma^2 + 6\gamma - 3})/6 \times (\gamma - 1)$ . Similarly, for the multiplicative model,  $|\tilde{D}_\infty| < |D_\infty|$  when  $p_{1\bullet} > (1 - \sqrt{\gamma})/(1 - \gamma)$ . Additionally, modes of inheritance where the penetrance of the heterozygote is smaller than either of the homozygotes, or an 'underdominant' model, the inequality  $|\tilde{D}_\infty| < |D_\infty|$  always holds. Table 1 shows a summary of results under various inheritance models with analytic results of LD isoclines.

As the allele frequencies  $p_{1\bullet}$  and  $p_{\bullet 1}$  are invariant to the effects of recombination, the ratio of  $\tilde{D}_\infty$  to  $D_\infty$  does not vary with recombination fraction. However, this is not the

**Table 1** LD relationships under various disease models

Model	Penetrance relationships	$\tilde{D}_\infty/D_\infty$	Allele frequency
Pure dominant	$f_{11} = f_{12}$ $f_{22} = 0$	$\frac{1}{p(2-p)^2}$	$p > \frac{3-\sqrt{5}}{2}$
Pure recessive	$f_{12} = f_{22} = 0$	0	All $p$
General recessive	$f_{11} > f_{12} = f_{22}$ ( $f_{11}/f_{22} = \gamma$ )	$\frac{(\gamma-1)p+1}{[(\gamma-1)p^2+1]^2}$	Expr I
Additive	$(f_{12}/f_{22}) = \gamma$ $(f_{11}/f_{22}) = 2\gamma-1$	$\frac{[(\gamma-1)p+1][(\gamma-1)p+\gamma]}{[2(\gamma-1)p+1]^2}$	$p > \frac{\gamma-3+\sqrt{\gamma^2+6\gamma-3}}{6(\gamma-1)}$
Mult	$(f_{12}/f_{22}) = \gamma_2$ $(f_{11}/f_{22}) = \gamma^2$	$\frac{\gamma}{[(\gamma-1)p+1]^2}$	$p > \frac{1-\sqrt{\gamma}}{1-\gamma}$
Symmetric underdominant	$f_{12} < f_{11}$ $f_{12} < f_{22}$ $f_{11} = f_{22} = \gamma(f_{12})$	$\frac{[(\gamma-1)p-\gamma][(1-\gamma)p-1]}{[2(\gamma-1)p^2+2(1-\gamma)p+\gamma]^2}$	All $p$
Symmetric overdominant	$f_{12} > f_{11}$ $f_{12} > f_{22}$ $(f_{11}) = \gamma(f_{22}) = (f_{12})$	$\frac{[(\gamma-1)p+1][(1-\gamma)p+\gamma]}{[2(\gamma-1)p^2+2(1-\gamma)p-1]^2}$	Never

Expl:  $\frac{\sqrt[3]{2(G^2/3)+4}\sqrt[3]{3(\gamma-1)}}{\sqrt[3]{36(G^{1/3})^{\gamma-1}}}$  where  $G = \lfloor \sqrt{3(27\gamma+5)(\gamma-1)} + 9\gamma - 9 \rfloor (\gamma-1)$

Summary of LD relationships between affecteds and the general population across a variety of models. The ratio of  $\tilde{D}_\infty$  to  $D_\infty$  is presented across several disease models along with the general population allele frequency at which the LD ratio is greater than 1. The solution to the general dominant model is complicated, and hence not presented.

case with  $r^2$ . The ratio  $\tilde{r}_\infty^2/r_\infty^2$  does change, often dramatically, with increasing recombination. This is due to the inability of the allele frequency at the marker locus within affecteds to be expressed solely in terms of penetrances and population allele frequencies in lieu of haplotype frequencies. Essentially, our immediate goal here is to evaluate the ratio  $R_t = (\tilde{r}_{\infty,t}^2/r_{\infty,t}^2)$  for infinite populations following different numbers of generations. Subscripts are shown to indicate that the quantities are now a function of accumulated recombination following the passage of  $t$  generations.

$$R_t = b[\tilde{p}_{\bullet,1,t}(1 - \tilde{p}_{\bullet,1,t})]^{-1} \quad (16)$$

where the factor  $b$  is not a function of recombination rate, and can be shown to be

$$b = \frac{p_{\bullet,1}(1 - p_{\bullet,1})[f_{11}p_{1\bullet} + f_{12}(1 - p_{1\bullet})][f_{12}p_{1\bullet} + f_{22}(1 - p_{1\bullet})]}{K^2} \quad (17)$$

To characterize the effect that recombination has on the marker locus, we first use the standard recursion-based derivation for the haplotype frequencies following  $t$  generations,

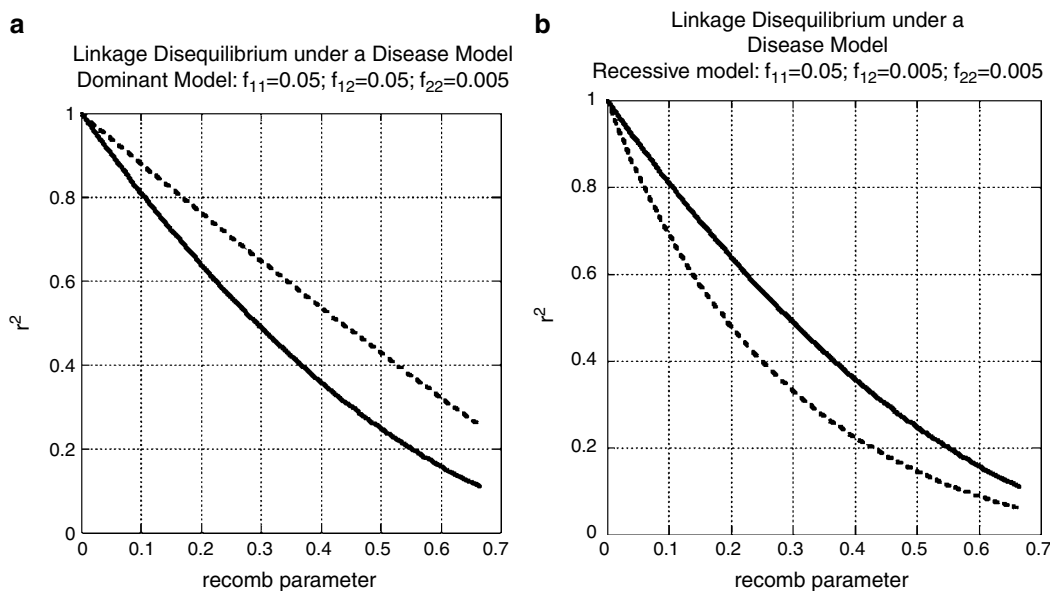
$$p_{ij,t} = (1-r)^t(p_{ij,0} - p_{i\bullet}p_{\bullet j}) + p_{i\bullet}p_{\bullet j} \quad (18)$$

Hence, substitution of the right-hand side of equation (18) into equation (7) yields

$$\begin{aligned} \tilde{p}_{\bullet,1,t} = \frac{1}{K} \{ & [D_{t=0}(1-r)^t + p_{1\bullet}p_{\bullet 1}][f_{11}p_{1\bullet} + f_{12}(1-p_{1\bullet})] \\ & - [D_{t=0}(1-r)^t + (1-p_{1\bullet})p_{\bullet 1}][f_{12}p_{1\bullet} \\ & + f_{22}(1-p_{1\bullet})] \} \end{aligned} \quad (19)$$

which, in turn, is used to complete the derivation of  $R_t$  in terms of recombination fractions, generations, and the initial state of the system.

Figures 1a and b show the decay in LD between a causative locus  $A$  and marker locus  $B$  as recombination increases between the two loci. LD levels within affected individuals and within the general population were



**Figure 1**  $r^2$  values for the general population are shown in a solid line, whereas  $r^2$  values for affected individuals are shown in the dashed line. (a) Decay of LD from cumulative recombination events with and without a disease model. Under this dominant model, the LD among affected individuals is always equal to or higher than the LD level in the general population. Switching to a recessive model (b) displays the opposite pattern with  $R_t < 1$  for any nontrivial level of recombination. The relative risk of predisposing to nonpredisposing genotypes under both models is 10.

calculated under three different disease models. In all models, initially only the parental haplotypes are present. These initial parental haplotype frequencies are 50 and 50% ( $p_{11} = 0.50$ ;  $p_{22} = 0.50$ ) across all models. The number of recombination events increases linearly in time. The first model investigated is of a disease with a dominant mode of inheritance against a background of sporadic disease. The results under this model are presented in Figure 1a. Figure 1a shows the common situation where  $R_t > 1$ .  $R_t$  is above unity regardless of the number of recombinant haplotypes, and the inceptive decay of LD is more rapid in the general population. Figure 1b shows the decay of LD situation under a recessive model. LD decay patterns under an underdominant model were studied next. The departure in LD in the affecteds from the general population under the underdominant model is more extreme than a recessive model, with certain recombination rates yielding  $R_t$  values below  $\frac{1}{2}$  (results not shown). In both models, LD levels in the disease population are lower than that in the general population across recombination fractions. It should be pointed out that empirically the patterns of LD across organisms are ubiquitously complex, and not fully determined by the effects of recombination. Although the results presented in this paper are approximate (we only model the effect of recombination on LD), we believe that it might be of interest.

### Potential impact on power

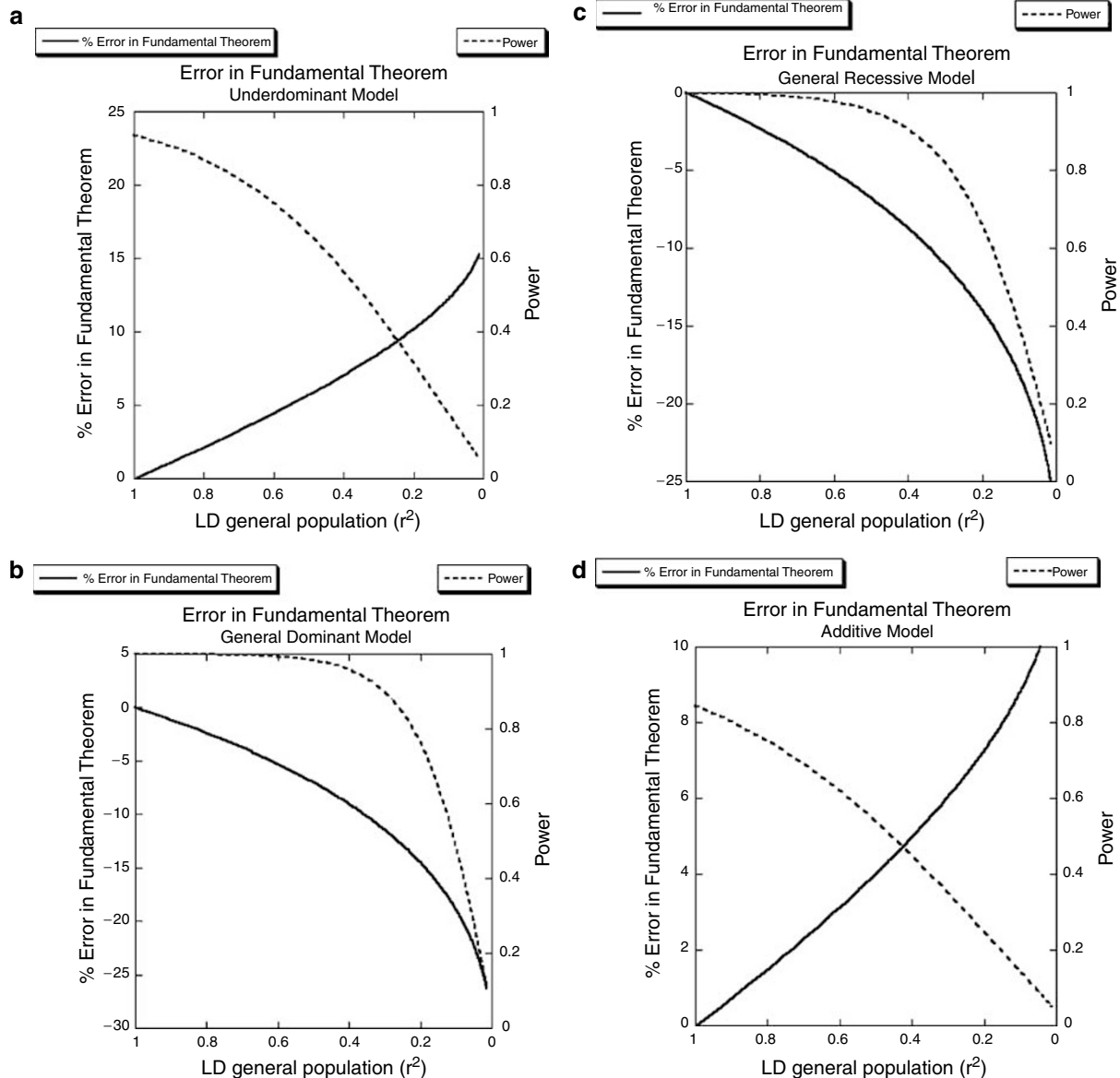
To illustrate how differences in LD patterns between cases and controls can impact disease gene mapping, consider

the following example. Suppose the general population haplotype frequencies are  $p_{11} = 0.10$ ,  $p_{12} = 0.01$ ,  $p_{21} = 0.01$  and  $p_{22} = 0.88$ . The minor allele frequency at either site is 11%, and the  $r_{\infty}^2$  statistic in the general population is 0.8061 – above many commonly used thresholds employed by many procedures to identify SNP pairs in high LD. Further consider a recessive disease model with penetrances  $f_{22} = 0.05$ ,  $f_{11} = f_{12} = 0.0001$ . Disease prevalence, assuming HWE, is 4% under these conditions. The haplotype frequencies in affected individuals are expected to be  $\hat{p}_{11} = 0.00025$ ,  $\hat{p}_{12} = 0.00003$ ,  $\hat{p}_{21} = 0.0112$ , and  $\hat{p}_{22} = 0.9885$ , yielding dramatically lower LD in the affected population:  $\hat{r}_{\infty}^2 = 0.0197$ . Trouble can arise in this situation if an investigator assumes the general population LD level before a case/control experiment. Given the high LD in the general population, one may assume that second site (locus B in the terminology used above) could be used as a tagging SNP for the first locus (the disease-predisposing locus). Assuming 250/250 case and control chromosomes used in a genome-wide association scan, power to detect disease association at the B locus is approximately 35% (taking a Bonferroni-corrected significance level of  $1 \times 10^{-7}$ ). However, if the investigator had restrained from assuming that LD patterns across cases and controls were similar, and perhaps went further to genotype densely in a small set of affected individuals, then noting that  $\hat{r}_{\infty}^2 = 0.0197$  may have persuaded this judicious researcher to genotype both the disease and marker loci (both the A and B SNPs). Had both loci been genotyped in the case/control experiment, power would more than double to 73% for a two-locus

haplotype test (using the same sample size and significance level). In both of the above power calculations, a Monte Carlo simulation running 100 000 replicates was used. Although this is an extreme example, it nonetheless demonstrates the possibility that ignoring the impact of disease models on LD can hinder mapping efforts.

One can frame this power-based argument in terms of the ‘fundamental theorem’ describing the relationship

between power to detect association indirectly at a marker locus and the  $r^2$  value between the marker and a disease-susceptibility locus. More precisely, it states that if a certain sample size is required at a disease locus to detect disease association at a given level of power, the sample size must be increased by a factor of  $1/r^2$  to obtain the same power indirectly at a marker locus. This simple relationship is described in Lai *et al*<sup>15</sup> and Pritchard and Przeworski.<sup>16</sup> This



**Figure 2** Error rate in the number of samples required to obtain a given power threshold to detect disease association as estimated by the fundamental theorem is explored for various disease models. LD is measured in terms of the  $r^2$  measure. In each case, the percentage error increases with decreasing pairwise LD. Typical error rates are in the range of  $\pm 10\%$  across much of the parameter space. Minor allele frequencies are around 10% in the general population across all models at both loci. Diploid cases/control (500/500) and 0.05 sig level assumed across all plots. (a) An underdominant model with penetrances of  $f_{11} = 0.08$ ,  $f_{12} = 0.04$ ,  $f_{22} = 0.08$ . (b) A general dominant model with penetrances of  $f_{11} = 0.08$ ,  $f_{12} = 0.08$ ,  $f_{22} = 0.04$ . (c) Results under a general recessive model with penetrances of  $f_{11} = 0.20$ ,  $f_{12} = 0.02$ ,  $f_{22} = 0.02$ . Lastly, (d) shows an additive model with penetrances of  $f_{11} = 0.05$ ,  $f_{12} = 0.10$ ,  $f_{22} = 0.15$ . Power to detect disease association at the marker locus is displayed with the error percentage.

relationship is a good rule of thumb, but in a case-control setting, deviations caused by disease models can be substantial as pointed out recently by Terwilliger and Hiekkalinna.<sup>6</sup> In most realistic instances, the underlying disease model modifies the ‘fundamental theorem’ from  $-10$  to  $+10\%$ . That is, the sample size estimated to be required to detect association at a particular power level is over- or underestimated by approximately  $10\%$ . Hence, assuming that of the several assumptions of the fundamental theorem mentioned by Terwilliger and Hiekkalinna<sup>6</sup> that the only one violated is the independence of etiology and LD patterns, our conclusions concerning the inaccuracies of the fundamental theorem are less extreme than those put forth by Terwilliger and Hiekkalinna. See Figure 2 for an evaluation of a number of these likely more realistic models. In situations where  $\tilde{r}_\infty^2/r_\infty^2 < 1$ , the fundamental theorem underestimates the sample size needed to detect disease association at a marker in LD with a disease-susceptibility site. Conversely, the fundamental theorem overestimates the sample size when  $\tilde{r}_\infty^2/r_\infty^2 > 1$ .

### Sampling properties given general population haplotype frequencies

To this point, our analysis has been restricted to properties in an infinite population. However, haplotype sampling properties are of particular interest as data invariably are in the form of a sample of chromosomes from the general population. Define  $x_{ij}$  as the number of copies of the  $A_iB_j$  haplotype in a sample of  $n$  chromosomes. Sampling of haplotypes will be based on the modified haplotype probabilities derived above

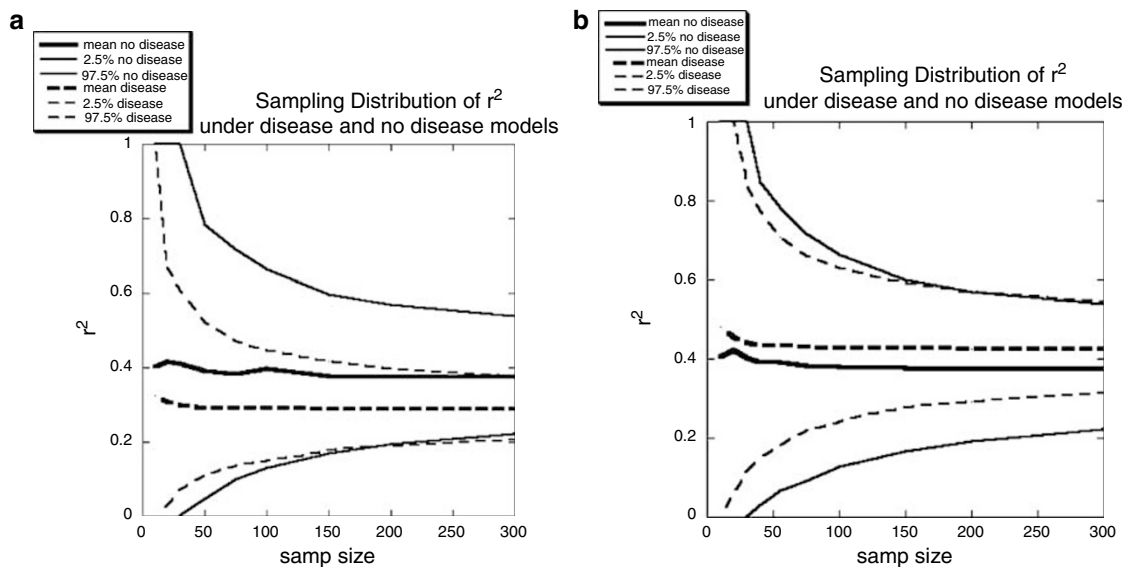
equations (2)–(5). Assuming a very large general population, the joint probability of the number of copies of each of the four haplotypes is the multinomial density

$$P[x_{11}, x_{12}, x_{21}, x_{22}] = n! \frac{\tilde{p}_{11}^{x_{11}} \tilde{p}_{12}^{x_{12}} \tilde{p}_{21}^{x_{21}} \tilde{p}_{22}^{x_{22}}}{x_{11}! x_{12}! x_{21}! x_{22}!} \quad (20)$$

As for any finite  $n$ ,  $P[p_{1\bullet} = 0] > 0$  and  $P[p_{\bullet 1} = 0] > 0$ , we can redefine the  $r^2$  statistic in either of these cases as the limit of  $r^2$  as an allele goes to fixation in the sample. It is readily shown that  $\lim_{p_{11}+p_{12} \rightarrow 0} r^2 = 0$  and  $\lim_{p_{11}+p_{21} \rightarrow 0} r^2 = 0$ . Hence, we set  $r^2 = 0$  in situations where there are no copies of one of the alleles (at either locus) in the sample.

Following equation (20), multinomial-distributed haplotype counts were generated via computer simulations. The aim of these simulations was to better understand sampling properties of the  $r^2$  statistic under disease models and general population. The results of these simulations are presented in Figures 3a and b, showing the 2.5th and 97.5th quantiles as well as the mean value of  $r^2$ . A recessive model and an additive model were explored in Figures 3a and b, respectively. Not surprisingly, in both models the sample variance of  $r^2$  is smaller than  $r^2$ . This is due to the selected sampling of chromosomes for affected chromosomes, which, under many models, are more genetically homogeneous than a random sample from the general population. Hence, these sampling effects lead to reduced sampling variance.

There are many uses of the variance of different LD measures. For example, the LD contrast method of Nielsen *et al*<sup>11</sup> uses  $\text{Var}[D]$  in the test statistic. Weir<sup>17</sup> solves



**Figure 3** (a) Produced from 10 000 replicates per data point; no disease model has equal penetrances; recessive disease model parameters are  $f_{11} = 0.10$ ,  $f_{12} = f_{22} = 0.001$ ; the general population haplotype frequencies are  $p_{11} = 0.10$ ,  $p_{12} = 0.05$ ,  $p_{21} = 0.05$ ,  $p_{22} = 0.80$ . (b) Produced from 10 000 replicates per data point; no disease model has equal penetrances; additive disease model parameters are  $f_{11} = 0.09$ ,  $f_{12} = 0.05$ ,  $f_{22} = 0.01$ ; the general population haplotype frequencies are  $p_{11} = 0.10$ ,  $p_{12} = 0.05$ ,  $p_{21} = 0.05$ ,  $p_{22} = 0.80$ .

for this quantity,

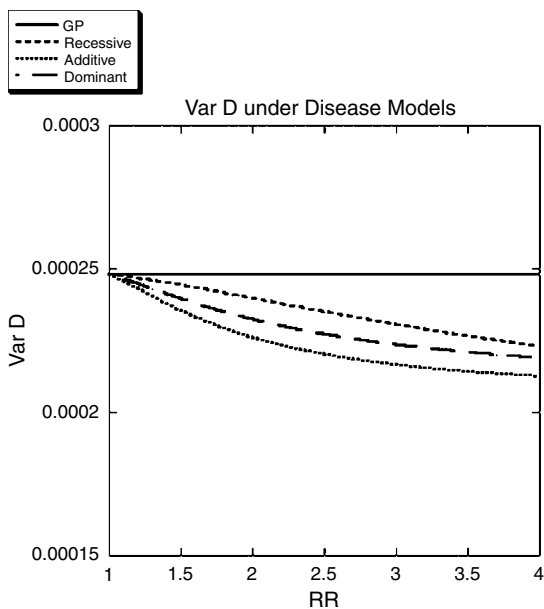
$$\text{Var}[D] = \frac{1}{n} [p_{1\bullet}(1-p_{1\bullet})p_{\bullet 1}(1-p_{\bullet 1}) + (1-2p_{1\bullet}) \times (1-2p_{\bullet 1})D - D^2] \quad (21)$$

Substitution of allele frequency and  $D$  values from disease-affected samples enables the calculation to be made for  $\text{Var}[\bar{D}]$ . Figure 4 shows results comparing the variance in  $D$  between general population samples and samples selected on the basis of disease phenotype. Hundred diploid individuals were used in the calculations. Three different disease models were evaluated: recessive, dominant and additive, and results are presented as a function of relative risk. These results show only a mild to moderate departure from the variance in the general population samples, with the variance in disease samples being approximately 10–15% lower at the more extreme relative risks.

### Sampling properties under the neutral coalescent

In the previous section, we explored the sampling properties of haplotypes preferentially or nonpreferentially sampled from the general population in accord with a disease model. In those simulations, the unselected haplotype frequencies were given. One may also be interested in the situation where those general population haplotype frequencies are randomized. A simple and flexible method to do so is to generate the general population haplotype frequencies under a Wright–Fisher model using a large-sample neutral

coalescent with recombination.<sup>18</sup> Although analytic approximations for population-based two-locus models exist, extensions to more complicated demographic models are much more straightforward under a coalescent simulation. The large-scale coalescent-generated haplotypes constitute the general population from which disease haplotypes are sampled according to penetrances. In these simulations, 5000 two-locus chromosomes were generated from which 100 chromosomes were sampled using probabilities proportional to the disease haplotype probabilities. Four different disease models were explored: dominant, recessive, underdominant and additive modes of inheritance.  $r^2$  and  $\tilde{r}^2$  were calculated for the general population and disease population samples. The mean and 0.025 and 0.975 quantiles for both correlation statistics are reported in Table 2 below for both the general population and the 100 disease haplotypes. Under most replicates, the  $A_1B_1$  haplotype is the most frequent. Table 2 summarizes this simulation study. When compared to the analytic results, these neutral coalescent results appear to corroborate the general patterns of LD with the distribution of  $\tilde{r}^2$  being shifted from  $r^2$ , with the largest reduction departures being found in recessive and underdominant models and the largest inflation departures for dominant and additive modes of inheritance. Over the models examined, the 97.5th quantile varies roughly by a factor of 4. The 95% confidence interval under the high frequency recessive model is approximately half the value in the general population (which is close to the sampling distribution value averaged across models) and the dominant model exhibits slightly greater than a twofold increase in the 95% confidence interval over the general population.



**Figure 4** A comparison between the sample variance in  $D$  under three modes of inheritance: recessive, dominant and additive, and the sample variance for general population samples. Hundred diploid individuals sampled.  $r^2 = 0.27$  in the general population. Minor allele frequency of 30% at both sites was modeled.

### Discussion

In this paper, we have explored the effect of disease models on pairwise measures of LD. Analytic work was able to delineate regions of the parameter space where  $|\bar{D}_\infty| = |D_\infty|$ . Often, the disease population exhibits higher or similar levels of LD when compared to the general population. As the affected population is selected based on the presence of an ancestral segment of DNA harboring the predisposing variant, this is the most intuitive scenario. However, all underdominant, some dominant and recessive, and protective models are capable of generating LD values substantially below those observed in the general population. This has important ramifications for disease mapping using LD-based methods. For example, a common methodology is to select a set of markers based on the observed LD in a region in the general population. The statistical power for a given sample size can then be estimated. However, if the LD in the affected population is lower than that in the general population, statistical power may be greatly overestimated. This results from the correlation between the associated allele at the marker and the predisposing allele being lower than expected based on the observations in the



**Table 2** Disease sampling from neutral coalescent-generated haplotypes

Mean $r^2$ GP	95% range $r^2$ GP	$f_{11}$	$f_{12}$	$f_{22}$	mean $r^2$ Dz	95% range $r^2$ Dz
0.01102	(1.2E-7, 0.0925)	0.05	0.005	0.005	0.00609	(0, 0.0469)
0.01012	(1.2E-7, 0.0773)	0.05	0.005	0.05	0.00660	(0, 0.0575)
0.01101	(8.0E-8, 0.0879)	0.05	0.05	0.005	0.01171	(0, 0.1021)
0.01020	(8.0E-8, 0.0871)	0.09	0.05	0.01	0.00912	(0, 0.0815)
0.01039	(1.2E-7, 0.0926)	0.05	0.05	0.05	0.01063	(0, 0.1007)
0.01053	(8.0E-8, 0.0898)	0.005	0.005	0.05	0.01255	(0, 0.1170)
0.01016	(8.0E-8, 0.0836)	0.005	0.05	0.05	0.02597	(0, 0.1909)
0.00946	(9.6E-8, 0.0735)	0.01	0.05	0.09	0.01575	(0, 0.1440)

Distribution of general population LD under a large-scale neutral coalescent and sampling results under disease models is shown. The  $r^2$  statistic is used. Replicates (10 000) were run for each disease model. Parameters of the coalescent simulations were  $4N\mu = 1.0$ ; number of chromosomes = 5000; number of segregating sites = 2;  $4Nc = 100$  across 2500 sites. GP and Dz denote 'general population' and 'disease', respectively. Average 2-site heterozygosity of 0.44 in the general population.

general population. Because of this, the allele frequency in the affected population will be less strongly influenced by the proximity of the predisposing locus and hence the difference in allele frequency between cases and controls may be less than expected under the assumption of equal LD in both populations. In this way, regions of the genome may be poorly interrogated. A possible way to alleviate these false negatives is to select a set of markers that potentially provides high levels of LD under a variety of plausible models.

Sets of tagging SNPs for mapping studies are often selected to minimize the number of SNPs assayed. That is, a single SNP may act for a number of other markers for which LD is high. As most tagging SNP programs are designed for and applied to randomly selected chromosomes from presumably unaffected individuals for use in disease association studies, incorporation of these effects may increase the efficacy of such tagging SNP procedures. This is particularly true in instances where the level of LD differs substantially between affected and unaffected individuals or where LD levels are markedly reduced in affected individuals when compared to chromosomes randomly drawn from the general population. Sets of markers based on these assumptions may be inadequate for situations where LD is greater in the affected population. For example, as we have shown, ignoring LD differences between cases and controls can produce nonoptimal power calculations.

Hopefully, this work will motivate further exploration of the impact of different LD patterns on association studies. For example, it may be possible to develop new statistical tests based both on the patterns of LD and haplotype frequency differences between case and control samples that provides high power to detect disease-predisposing regions. Any scenario where there is strong selection bias, be it positive or negative as in the case of a disease, LD may differ. Detecting such signatures of selection bias will undoubtedly add to our understanding of human disease etiology.

#### Acknowledgements

This work was supported by Celera. We thank John Sninsky, Tom White, Victoria Carlton, Ann Begovich, Xiaolan Hu and Tony Long

for key discussions and suggestions. In addition, we are grateful to the editors and two anonymous reviewers for extremely useful and discerning comments.

#### References

- Clark AG: Finding genes underlying risk of complex disease by linkage disequilibrium mapping. *Curr Opin Genet Dev* 2003; **13**: 296–302.
- Lewontin RC: The interaction of selection and linkage. I. General considerations; heterotic models. *Genetics* 1964; **49**: 49–67.
- Ohta T, Kimura M: Linkage disequilibrium between two segregating nucleotide sites under the steady flux of mutations in a finite population. *Genetics* 1971; **68**: 571–580.
- Ewens WJ: *Mathematical Population Genetics*. Berlin, Heidelberg: Springer-Verlag, 1979.
- Hudson RR: Properties of a neutral allele model with intragenic recombination. *Theor Popul Biol* 1983; **23**: 183–201.
- Terwilliger JD, Hiekkalinna T: An utter refutation of the 'Fundamental Theorem of the HapMap'. *Eur J Hum Genet* 2006; **14**: 426–437.
- Cardon LR, Abecasis GR: Using haplotype blocks to map human complex trait loci. *Trends Genet* 2003; **3**: 135–140.
- De La Vega FM, Isaac H, Collins A *et al*: The linkage disequilibrium maps of three human chromosomes across four populations reflect their demographic history and common underlying recombination pattern. *Genome Res* 2005; **15**: 454–462.
- Long AD, Langley CH: The power of association studies to detect the contribution of candidate genetic loci to variation in complex traits. *Genome Res* 1999; **9**: 720–731.
- Hu X, Schrodri SJ, Ross DA, Cargill M: Selecting tagging SNPs for association studies using power calculations from genotype data. *Hum Hered* 2004; **57**: 156–170.
- Nielsen DM, Ehm MG, Zaykin DV, Weir BS: Effect of two- and three-locus linkage disequilibrium on the power to detect marker/phenotype associations. *Genetics* 2004; **168**: 1029–1040.
- Zaykin DM, Meng Z, Ehm MG: Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am J Hum Genet* 2006; **78**: 737–746.
- Devlin B, Risch N: A comparison of linkage disequilibrium measures for fine-scale mapping. *Genomics* 1995; **29**: 311–322.
- Hill WG, Robertson AR: Linkage disequilibrium in finite populations. *Theor Appl Genet* 1968; **38**: 226–231.
- Lai C, Lyman RF, Long AD, Langley CH, Mackay TF: Naturally occurring variation in bristle number and DNA polymorphisms at the scabrous locus of *Drosophila melanogaster*. *Science* 1994; **266**: 1697–1702.
- Pritchard JK, Przeworski M: Linkage disequilibrium in humans: models and data. *Am J Hum Genet* 2001; **69**: 1–14.
- Weir BS: *Genetic Data Analysis II*. Sunderland, MA: Sinauer, 1996.
- Hudson RR: Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* 2002; **18**: 337–338.