Available online at www.sciencedirect.com

ScienceDirect

Mathematical Biosciences 208 (2007) 33–47

Mathematical Biosciences

www.elsevier.com/locate/mbs

Sampling properties of homozygosity-based statistics for linkage disequilibrium

Noah A. Rosenberg [a,b,c,*], Michael G.B. Blum [a]

[a] Department of Human Genetics, University of Michigan, 1241 East Catherine Street, Ann Arbor, MI 48109-0618, USA
[b] Bioinformatics Program, University of Michigan, 2017 Palmer Commons, 100 Washtenaw Avenue, Ann Arbor, MI 48109-2218, USA
[c] Life Sciences Institute, University of Michigan, 210 Washtenaw Avenue, Ann Arbor, MI 48109-2216, USA

Received 14 November 2005; received in revised form 24 April 2006; accepted 6 July 2006
Available online 27 July 2006

Abstract

Homozygosity-based statistics such as Ohta's identity-in-state (IIS) excess offer the potential to measure linkage disequilibrium for multiallelic loci in small samples. However, previous observations have suggested that for independent loci, in small samples these statistics might produce values that more frequently lie on one side rather than on the other side of zero. Here we investigate the sampling properties of the IIS excess. We find that for any pair of independent polymorphic loci, as sample size n approaches infinity, the sampling distribution of the IIS excess approaches a normal distribution. For large samples, the IIS excess tends towards symmetry around zero, and the probabilities of positive and of negative IIS excess both approach 1/2. Surprisingly, however, we also find that for sufficiently large n, independent loci can be chosen so that the probability of a sample having positive IIS excess is arbitrarily close to either 0 or 1. The results are applied to interpretation of data from human populations, and we conclude that before employing homozygosity-based statistics to measure LD in a particular sample, especially for loci with either very small or very large homozygosities, it is useful to verify that loci with the observed homozygosity values are not likely to produce a large bias in IIS excess in samples of the given size.

© 2006 Elsevier Inc. All rights reserved.

[*] Corresponding author. Tel.: +1 734 615 9556; fax: +1 734 615 6553.
E-mail address: rnoah@umich.edu (N.A. Rosenberg).

0025-5564/$ - see front matter © 2006 Elsevier Inc. All rights reserved.
doi:10.1016/j.mbs.2006.07.001

Keywords: Convergence; Correlation coefficient; Haplotype phase; Heterozygosity; Slutsky's theorem

1. Introduction

The measurement of linkage disequilibrium (LD) is an important tool helpful across a wide range of problems in population genetics, such as association mapping of genes that influence phenotypes [15,36], estimation of the age of alleles [17,25], inference of demographic and selective history [6,29], and evaluation of the nature of the recombination process [8,16]. Of the various statistics available for measuring LD [3,5,11], in diploid organisms, many assume that data are available in the form of haplotypes. In other words, it is assumed that the loci of interest are linked and that the assignments of alleles to individual chromosomes, or the haplotype phases of sets of loci, are known. When data consist of two-locus or multilocus genotypes of linked loci, so that haplotype phase is not known, these statistics rely on computational haplotype prediction algorithms that take the genotypes as inputs, and that then measure LD from the estimated haplotypes or haplotype frequencies. Although haplotype inference can often be performed quite accurately [9,10,26], this process may lose some information, in that many possible sets of haplotypes may be consistent with the collection of genotypes, and depending on the particular haplotype sets that are found to be most likely, haplotype estimation may introduce a bias into the LD measurement.

An alternative solution to the problem of unknown haplotype phase in diploid organisms is to make direct use of the genotypes [2,12,20,22,30,31,33,34]. For example, Ohta [12–14] observed that for a pair of loci at linkage equilibrium—the state in which all allelic combinations (with one allele at each locus) have frequency equal to the product of the frequencies of their constituent alleles—the probability of an individual being homozygous at both loci is the product of the probability of being homozygous at the first locus and the probability of being homozygous at the second locus. Thus, LD can be measured as $\Delta = (X_{12n}/n) - (X_{1n}/n)\ (X_{2n}/n)$, where X_{1n}, X_{2n}, and X_{12n} denote the numbers of homozygotes at the first locus, homozygotes at the second locus, and double homozygotes in a sample of n individuals. Ohta termed Δ the "identity excess," which we label here the "identity-in-state excess" or "IIS excess," to avoid confusion with identity-by-descent.

Similarly to the haplotype inference approach, the IIS excess method also introduces a loss of information into LD measurement, in that its calculation requires the collapsing of a larger number of distinct genotypes into a smaller number of types, namely, double heterozygotes, homozygotes at the first locus only, homozygotes at the second locus only, and double homozygotes. However, it avoids the potential bias introduced by haplotype prediction algorithms, whose success at obtaining the haplotype frequencies required in LD estimation may vary with the underlying amount of LD. Additionally, in contrast to many LD statistics, the IIS excess is straightforwardly applicable to multiallelic loci; for such loci, the fact that genotypes are collapsed into a small number of types may actually be beneficial, in that this compression can enable the computation of reasonably accurate estimates of LD without requiring an impractically large sample.

Several recent articles have further developed the connection between homozygosity and LD. Vitalis and Couvet explored the influence of various population processes on the IIS excess

[28], and then used the IIS excess to devise estimators of evolutionary parameters [27]. Sabatti and Risch [20] defined additional statistics based on the IIS excess, exploring aspects of their estimation and devising asymptotic tests of the null hypothesis of zero IIS excess. Importantly, Sabatti and Risch also noted that while IIS excess statistics are nonzero with some forms of LD, the presence of LD does not guarantee that the IIS excess will be nonzero. Yang [34] further showed that if the loci are not in Hardy–Weinberg equlibrium, a nonzero IIS excess reflects the combination of LD with other disequilibria. However, given that Hardy–Weinberg disequilibrium is not likely to be severe in most scenarios of interest, it does not substantially affect the general utility of statistics based on the IIS excess [21]. Nevertheless, interpretation of the IIS excess as a measure of LD is most straightforward when Hardy–Weinberg equilibrium holds at both loci.

Most recently, Rosenberg and Calabrese [18] showed that the IIS excess is sensitive to population structure, producing an excess of positive values in structured populations. Additionally, we observed a dependence of the IIS on sample size in a situation where pairs of loci were not associated, that is, not in linkage disequilibrium. In particular, we noticed that in smaller samples from human populations, more pairs of loci tended to produce negative values of the IIS excess than in larger samples. This effect was found to result from two observations: (i) for values of the true homozygosities of two loci between 0 and 0.5, the IIS excess tends to be negative more often than positive, with the effect stronger in small samples; (ii) in most of the human populations considered, most loci had homozygosities between 0 and 0.5.

In this article, we more fully characterize the nature of the sampling dependence of the IIS excess, producing two results that are surprising when considered together: for any values of the true homozygosities H_1 and H_2 of two loci, as $n \to \infty$, the probability of a sample of size n having positive IIS excess at the loci approaches 1/2; at the same time, however, for any $\epsilon > 0$, if n is sufficiently large, there exist values of H_1 and H_2 for the homozygosities of two loci so that the probability of a sample of size n having positive IIS excess at the two loci exceeds $1 - \epsilon$. Similarly, there also exist values for the two homozygosities so that the probability of a sample of size n having positive IIS excess is less than ϵ.

Lemma 1 and 2 identify symmetry in the distribution of the IIS excess, showing that the IIS excess for a pair of loci with homozygosities H_1 and H_2 has a close relationship to the IIS excess for three other pairs of loci: those with homozygosities $1 - H_1$ and H_2, H_1 and $1 - H_2$, and $1 - H_1$ and $1 - H_2$. Proposition 3 then demonstrates that if two loci are independent (that is, in linkage equilibrium), $\sqrt{n}$ times the IIS excess approaches a normal distribution as the sample size $n \to \infty$. Corollary 4 then shows that a standardized version of the IIS excess approaches a standard normal distribution. Also as a consequence of Proposition 3, Corollary 5 shows that as $n \to \infty$, for any true homozygosities H_1 and H_2, the probability of a sample having positive IIS excess approaches 1/2. However, Proposition 6 then demonstrates that considering all possible values of (H_1, H_2), as $n \to \infty$, the supremum of the probability of positive IIS excess approaches 1 and the infimum approaches 0. In other words, for sufficiently large n, despite the fact that with increasing n the probability of positive IIS excess approaches 1/2 for any two values of H_1 and H_2, there is some pair (H_1, H_2) for which the probability of positive IIS excess is close to 1 and another pair for which it is close to 0. Lastly, Proposition 7 gives a large-sample Poisson approximation for the distribution of the IIS excess, which is simplified in Corollary 8 to produce the approximate probability of positive IIS excess in the case that both H_1 and H_2 are near 0 or 1.

2. Results

Consider two loci that are in linkage equilibrium and that have homozygosities H_1, $H_2 \in [0,1]$. Let X_{1n}, X_{2n}, and X_{12n} denote the random numbers of homozygotes at the first locus, homozygotes at the second locus, and double homozygotes in a sample of n diploid individuals, respectively (X_{1n} and X_{2n} are independent). For these two loci, denote the IIS excess statistic by $\Delta_n(H_1, H_2)$:

$$\Delta_n(H_1, H_2) = \frac{X_{12n}}{n} - \frac{X_{1n}}{n}\frac{X_{2n}}{n}. \tag{1}$$

As no stipulations are made here regarding allele frequencies, no assumptions are made about the presence or absence of Hardy–Weinberg equilibrium at the two loci.

For a pair of loci, Sabatti and Risch [20] defined a statistic HR, whose numerator equals Δ_n. Analogously to the LD statistic r [5,11,15], the statistic HR can be viewed as the sample correlation coefficient of the indicator variable for homozygosity at the first locus (1 if homozygous, 0 if heterozygous) and the corresponding indicator variable for the second locus:

$$HR_n(H_1, H_2) = \frac{\frac{X_{12n}}{n} - \frac{X_{1n}}{n}\frac{X_{2n}}{n}}{\sqrt{\frac{X_{1n}}{n}\left(1 - \frac{X_{1n}}{n}\right)\frac{X_{2n}}{n}\left(1 - \frac{X_{2n}}{n}\right)}}. \tag{2}$$

$HR_n(H_1,H_2)$ is assumed to be zero if the denominator (and consequently, the numerator) is zero.

For loci with homozygosities H_1 and H_2, the random variables X_{1n}, X_{2n}, and X_{12n} are binomially distributed with parameters H_1, H_2, and H_1H_2, respectively. The joint probability that $X_{1n} = x_1$, $X_{2n} = x_2$, and $X_{12n} = x_{12}$ is given by [18]

$$R_{n,x_1,x_2,x_{12}}(H_1, H_2) = \binom{n}{x_1} H_1^{x_1}(1 - H_1)^{n-x_1} \binom{n}{x_2} H_2^{x_2}(1 - H_2)^{n-x_2} \frac{\binom{x_1}{x_{12}}\binom{n - x_1}{x_2 - x_{12}}}{\binom{n}{x_2}}. \tag{3}$$

It is required that $\max(0,x_1 + x_2 - n) \leqslant x_{12} \leqslant \min(x_1,x_2)$ and $0 \leqslant x_1,x_2 \leqslant n$. The formula is symmetric with respect to transposition of the two loci.

Using Eq. (3),

$$Pr[\Delta_n(H_1, H_2) < 0] = \sum_{x_1=0}^{n} \sum_{x_2=0}^{n} \sum_{x_{12}=\max(0,x_1+x_2-n)}^{\gamma_*(x_1,x_2)} R_{n,x_1,x_2,x_{12}}(H_1, H_2), \tag{4}$$

$$Pr[\Delta_n(H_1, H_2) = 0] = \sum_{x_1=0}^{n} \sum_{x_2=0}^{n} R_{n,x_1,x_2,\gamma(x_1,x_2)}(H_1, H_2), \tag{5}$$

$$Pr[\Delta_n(H_1, H_2) > 0] = \sum_{x_1=0}^{n} \sum_{x_2=0}^{n} \sum_{x_{12}=\gamma^*(x_1,x_2)}^{\min(x_1,x_2)} R_{n,x_1,x_2,x_{12}}(H_1, H_2), \tag{6}$$

where $\gamma(x_1, x_2) = x_1x_2/n$, $\gamma_* = \gamma - 1$ if γ is an integer and $\gamma_* = \lfloor\gamma\rfloor$ otherwise, $\gamma^* = \gamma + 1$ if γ is an integer and $\gamma^* = \lceil\gamma\rceil$ otherwise, and $R_{n,x_1,x_2,\gamma(x_1,x_2)}$ is set to 0 if $\gamma(x_1, x_2)$ is not an integer.

Denote $H_1^* = 1 - H_1$ and $H_2^* = 1 - H_2$. Some symmetry arguments will be useful in proving the main results.

Lemma 1.

(i) $R_{n,x_1,x_2,x_{12}}(H_1^*, H_2^*) = R_{n,n-x_1,n-x_2,n-x_1-x_2+x_{12}}(H_1, H_2)$,
(ii) $R_{n,x_1,x_2,x_{12}}(H_1^*, H_2) = R_{n,n-x_1,x_2,x_2-x_{12}}(H_1, H_2)$,
(iii) $R_{n,x_1,x_2,x_{12}}(H_1, H_2^*) = R_{n,x_1,n-x_2,x_1-x_{12}}(H_1, H_2)$.

Proof. These identities follow directly from Eq. (3). □

Lemma 2.

(i) $Pr[\Delta_n(H_1^*, H_2^*) < 0] = Pr[\Delta_n(H_1, H_2) < 0]$,
(ii) $Pr[\Delta_n(H_1^*, H_2^*) = 0] = Pr[\Delta_n(H_1, H_2) = 0]$,
(iii) $Pr[\Delta_n(H_1^*, H_2^*) > 0] = Pr[\Delta_n(H_1, H_2) > 0]$,
(iv) $Pr[\Delta_n(H_1^*, H_2) < 0] = Pr[\Delta_n(H_1, H_2) > 0]$,
(v) $Pr[\Delta_n(H_1^*, H_2) = 0] = Pr[\Delta_n(H_1, H_2) = 0]$,
(vi) $Pr[\Delta_n(H_1^*, H_2) > 0] = Pr[\Delta_n(H_1, H_2) < 0]$,
(vii) $Pr[\Delta_n(H_1, H_2^*) < 0] = Pr[\Delta_n(H_1, H_2) > 0]$,
(viii) $Pr[\Delta_n(H_1, H_2^*) = 0] = Pr[\Delta_n(H_1, H_2) = 0]$,
(ix) $Pr[\Delta_n(H_1, H_2^*) > 0] = Pr[\Delta_n(H_1, H_2) < 0]$.

Proof. We prove (i). The remaining proofs are similar, substituting Eqs. (5) or (6) for Eq. (4), and Lemma 1(ii) or 1(iii) for 1(i). Proving two of (i), (ii), and (iii) trivially implies the third statement; a similar relationship holds among (iv), (v), and (vi) and among (vii), (viii), and (ix). Applying Eq. (4) and Lemma 1(i),

$$Pr\left[\Delta_n\left(H_1^*, H_2^*\right) < 0\right] = \sum_{x_1=0}^{n}\sum_{x_2=0}^{n}\sum_{x_{12}=\max(0,x_1+x_2-n)}^{\gamma_*(x_1,x_2)} R_{n,n-x_1,n-x_2,n-x_1-x_2+x_{12}}(H_1, H_2).$$

If we substitute $y_1 = n - x_1$, $y_2 = n - x_2$, and $y_{12} = n - x_1 - x_2 + x_{12}$, and determine the limits of summation with these new variables, the sum becomes

$$\sum_{y_1=0}^{n}\sum_{y_2=0}^{n}\sum_{y_{12}=\max(0,y_1+y_2-n)}^{\gamma_*(y_1,y_2)} R_{n,y_1,y_2,y_{12}}(H_1, H_2),$$

which by Eq. (4) equals $Pr[\Delta_n(H_1, H_2) < 0]$. □

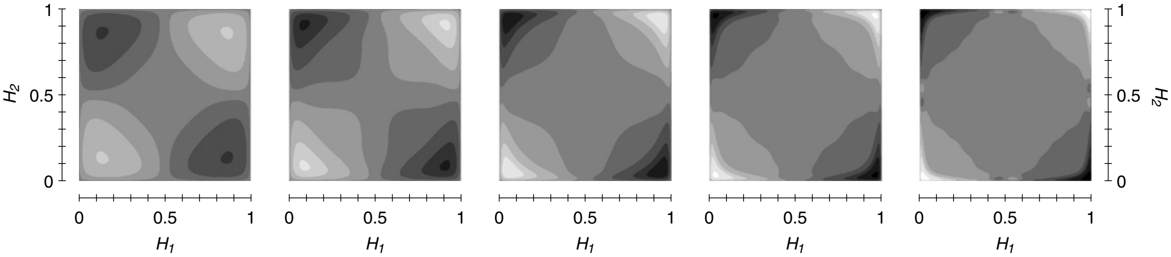

Fig. 1. Probability of positive IIS excess (plus half the probability of IIS excess equal to zero) as a function of homozygosities (H_1, H_2), computed from Eqs. (5) and (6). From left to right, the plots represent $n = 10$, 20, 40, 80, and 160. From lightest to darkest, the shades represent positive IIS excess probabilities in $[0, 0.175)$, $[0.175, 0.3)$, $[0.3, 0.375)$, $[0.375, 0.45)$, $[0.45, 0.49)$, $[0.49, 0.51]$, $(0.51, 0.55]$, $(0.55, 0.625]$, $(0.625, 0.7]$, $(0.7, 0.825]$, and $(0.825, 1]$.

The symmetry in the distribution of the IIS excess, proven in Lemma 2, is illustrated in Fig. 1. Symmetry across the line $H_2 = H_1$ is a consequence of the fact that H_1 and H_2 have interchangeable roles in Eq. (3). Symmetry across the line $H_1 + H_2 = 1$ in the probability of positive IIS excess results from Lemma 2(i)–(iii), and antisymmetry across the lines $H_1 = 1/2$ and $H_2 = 1/2$ results from Lemma 2(iv)–(ix).

We now prove that $\sqrt{n}\Delta_n(H_1, H_2)$ is asymptotically normally distributed. For Proposition 3 and its corollaries, it is useful to think of an infinite sequence of individuals, so that among the first n of the individuals, X_{1n}, X_{2n}, and X_{12n} respectively represent the numbers of homozygotes at the first locus, homozygotes at the second locus, and double homozygotes. As X_{1n}, X_{2n}, and X_{12n} can each be viewed as the sum of independent and identically distributed Bernoulli random variables, this perspective enables the application of results stated using means of partial sums of sequences of IID random variables.

It is also useful to notice that $\Delta_n(H_1, H_2)$ is the estimator based on a sample of size n of the covariance of two Bernoulli random variables. Given the asymptotic normality of many other statistics associated with Bernoulli trials and multinomial sampling [1, pp. 419–425] it is not surprising that asymptotic normality applies to $\sqrt{n}\Delta_n(H_1, H_2)$ as well. The proof employs the delta method [1,7], a commonly used procedure for obtaining asymptotic normality of sample statistics [1, p. 419].

Proposition 3. *For fixed values of H_1 and H_2 not equal to 0 or 1, as $n \to \infty$,*

$$\sqrt{n}\Delta_n(H_1, H_2) \xrightarrow{d} N(0, H_1(1 - H_1)H_2(1 - H_2)),$$

where $N(\mu, \sigma^2)$ denotes a normal random variable with mean μ and variance σ^2, and $\xrightarrow{d}$ denotes convergence in distribution.

Proof. Because X_{1n} is binomially distributed, X_{1n}/n has mean H_1 and variance $H_1(1 - H_1)/n$. Similarly X_{2n}/n has mean H_2 and variance $H_2(1 - H_2)/n$, and X_{12n}/n has mean H_1H_2 and variance $H_1H_2(1 - H_1H_2)/n$. The covariance of X_{1n}/n and X_{12n}/n can be calculated using the independence of X_{1n} and X_{2n}:

$$\begin{aligned}\mathrm{Cov}\left[\frac{X_{1n}}{n},\frac{X_{12n}}{n}\right] &= \mathbb{E}\left[\frac{X_{1n}}{n}\frac{X_{12n}}{n}\right]-(H_1)(H_1H_2)\\ &= -H_1^2H_2+\sum_{x_1=0}^{n}\sum_{x_2=0}^{n}\frac{x_{1n}}{n}\mathbb{E}\left[\frac{X_{12n}}{n}\middle|X_{1n}=x_1,X_{2n}=x_2\right]Pr[X_{1n}=x_1]Pr[X_{2n}=x_2]\\ &= -H_1^2H_2+\sum_{x_1=0}^{n}\sum_{x_2=0}^{n}\frac{x_{1n}^2}{n^2}\frac{x_{2n}}{n}Pr[X_{1n}=x_1]Pr[X_{2n}=x_2]\\ &= -H_1^2H_2+\frac{\mathrm{Var}[X_{1n}]+\mathbb{E}[X_{1n}]^2}{n^2}\frac{\mathbb{E}[X_{2n}]}{n}\\ &= H_1H_2(1-H_1)/n.\end{aligned} \tag{7}$$

Similarly $\mathrm{Cov}[X_{2n}/n, X_{12n}/n] = H_1H_2(1 - H_2)/n$. Because X_{1n}, X_{2n}, and X_{12n} can each be written as the sum of IID Bernoulli random variables, we can use a multivariate central limit theorem stated for sums of this type [4, p. 26],

$$\sqrt{n}\begin{pmatrix}X_{1n}/n-H_1\\ X_{2n}/n-H_2\\ X_{12n}/n-H_1H_2\end{pmatrix}\xrightarrow{d}N\left(\begin{pmatrix}0\\0\\0\end{pmatrix},\begin{pmatrix}H_1(1-H_1) & 0 & H_1H_2(1-H_1)\\ 0 & H_2(1-H_2) & H_1H_2(1-H_2)\\ H_1H_2(1-H_1) & H_1H_2(1-H_2) & H_1H_2(1-H_1H_2)\end{pmatrix}\right), \tag{8}$$

where $N(\boldsymbol{\mu},\boldsymbol{\Sigma})$ denotes a multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance–covariance matrix $\boldsymbol{\Sigma}$. Define $g:\mathbb{R}^3\to\mathbb{R}$ by $g(a,b,c)=c-ab$. Using the delta method [7, p. 148] and denoting the variance–covariance matrix in Eq. (8) by $\boldsymbol{V}$,

$$\sqrt{n}[g(X_{1n}/n, X_{2n}/n, X_{12n}/n)-g(H_1, H_2, H_1H_2)]\xrightarrow{d}N(0, \boldsymbol{ZVZ}^{\mathrm{T}}),$$

where

$$\boldsymbol{Z}=\left(\frac{\partial g}{\partial a},\frac{\partial g}{\partial b},\frac{\partial g}{\partial c}\right)\Bigg|_{(a,b,c)=(H_1,H_2,H_1H_2)}=(-H_2, -H_1, 1).$$

Simplifying, we obtain that $\sqrt{n}[X_{12n}/n-(X_{1n}/n)(X_{2n}/n)]\xrightarrow{d}N(0, H_1(1-H_1)H_2(1-H_2))$. □

Corollary 4. *For fixed values of* H_1 *and* H_2 *not equal to* 0 *or* 1, *as* $n\to\infty$,

$$\sqrt{n}HR_n(H_1, H_2)\xrightarrow{d}N(0, 1).$$

Proof. From Proposition 3,

$$\sqrt{n}\frac{\frac{X_{12n}}{n}-\frac{X_{1n}}{n}\frac{X_{2n}}{n}}{\sqrt{H_1(1-H_1)H_2(1-H_2)}}\xrightarrow{d}N(0, 1).$$

Repeated application of Slutsky's theorem [24, p. 19], using the fact that as means of the first n terms in sequences of IID Bernoulli random variables, $X_{1n}/(nH_1)$, $X_{2n}/(nH_2)$, $(n - X_{1n})/[n(1 - H_1)]$, and $(n - X_{2n})/[n(1 - H_2)]$ all converge in probability to 1 [24, p. 9], yields

$$\sqrt{n}\frac{\frac{X_{12n}}{n} - \frac{X_{1n}}{n}\frac{X_{2n}}{n}}{\sqrt{\frac{X_{1n}}{n}\left(1 - \frac{X_{1n}}{n}\right)\frac{X_{2n}}{n}\left(1 - \frac{X_{2n}}{n}\right)}} \xrightarrow{d} N(0, 1). \quad \square \tag{9}$$

Corollary 5. *For a real number c and fixed values of* H_1 *and* H_2 *not equal to* 0 *or* 1,

$$\lim_{n\to\infty} Pr[\sqrt{n}\Delta_n(H_1, H_2) \leqslant -c] = \lim_{n\to\infty} Pr[\sqrt{n}\Delta_n(H_1, H_2) \geqslant c].$$

Consequently,

$$\lim_{n\to\infty} Pr[\Delta_n(H_1, H_2) \leqslant 0] = \lim_{n\to\infty} Pr[\Delta_n(H_1, H_2) \geqslant 0] = 1/2.$$

Proof. The first statement is a direct consequence of Proposition 3 together with the definition of convergence in distribution. The second statement follows by choosing $c = 0$, and noting that for any sample size n, $Pr[\sqrt{n}\Delta_n(H_1, H_2) \leqslant 0] = Pr[\Delta_n(H_1, H_2) \leqslant 0]$, and $Pr[\sqrt{n}\Delta_n(H_1, H_2) \geqslant 0] = Pr[\Delta_n(H_1, H_2) \geqslant 0]$. □

Proposition 6.

(i) $\lim_{n\to\infty}\sup_{H_1,H_2\in[0,1]} Pr[\Delta_n(H_1, H_2) < 0] = 1$
(ii) $\lim_{n\to\infty}\inf_{H_1,H_2\in[0,1]} Pr[\Delta_n(H_1, H_2) < 0] = 0$.

Proof that Proposition 6(i) implies Proposition 6(ii). Choose $\epsilon > 0$. Using Proposition 6(i), for any sufficiently large n, there exist H_{1n}, H_{2n} with $Pr[\Delta_n(H_{1n}, H_{2n}) < 0] > 1 - \epsilon$. It then follows from Lemma 2(ix) that $Pr[\Delta_n(H_{1n}, H^*_{2n}) > 0] > 1 - \epsilon$, from which $Pr[\Delta_n(H_{1n}, H^*_{2n}) < 0] < \epsilon$. □

This argument verifies that symmetry can be used to obtain Proposition 6(ii) from 6(i), whose proof we will defer until after Proposition 7. One way of thinking about why Proposition 6(i) holds is that for large n, we can find H_1 and H_2 that are large enough so that it is likely that homozygotes at each of the two loci will separately be observed, but small enough so that double homozygotes are unlikely to be observed. For such loci X_{12n}/n is very likely to be 0, but $(X_{1n}/n)(X_{2n}/n)$ is very likely to be positive, so that $\Delta_n(H_1, H_2)$ is likely to be negative.

The sequence of plots in Fig. 1 shows how the probability of positive IIS excess approaches 1/2, as proven in Corollary 5. The center region of the space of possible values of (H_1, H_2), in which this probability is close to 1/2, expands as n increases, forcing the regions with positive IIS excess probability far from 1/2 into the corners of the space. The locations where the probability is not close to 1/2 tend towards having positive IIS excess probability farther and farther from 1/2, as is demonstrated by Proposition 6. This trend is also illustrated in Fig. 2, which shows a cross section of the plots in Fig. 1 along the line $H_2 = H_1$. It can also be seen in Fig. 2 that for a given pair (H_1, H_2), there may be some sample size for which the positive IIS excess probability is far from 1/2, but as the sample size increases, the probability tends back towards 1/2.

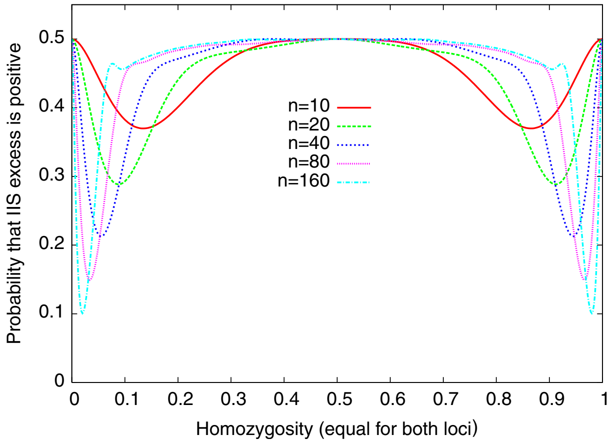

Fig. 2. Probability of positive IIS excess (plus half the probability of IIS excess equal to zero) as a function of homozygosity, computed from Eqs. (5) and (6), for $H_1 = H_2$ and $n = 10, 20, 40, 80$, and 160.

This tendency of the positive IIS excess probability to move away from and then back towards 1/2 as sample size increases is more directly illustrated in Fig. 3. For homozygosities close to 0.5, the probability remains near 1/2 for small sample sizes; the approach to the large-n limit of 1/2, however, is not monotonic. For more extreme homozygosities, sample size must be considerably larger before the probability moves back towards 1/2. At $H_1 = H_2 = 0.04$, it only reaches its minimum of ≈ 0.158 at $n = 82$ before climbing back towards its large-n limit of 1/2. At $H_1 = H_2 = 0.01$, the probability continues to decline as n reaches the rightmost edge of the plot. Notice that regardless of the values of H_1 and H_2, for $n = 2$, the probabilities of positive and negative IIS excess are equal, so that the probability of positive IIS excess plus half the probability of zero IIS excess is always

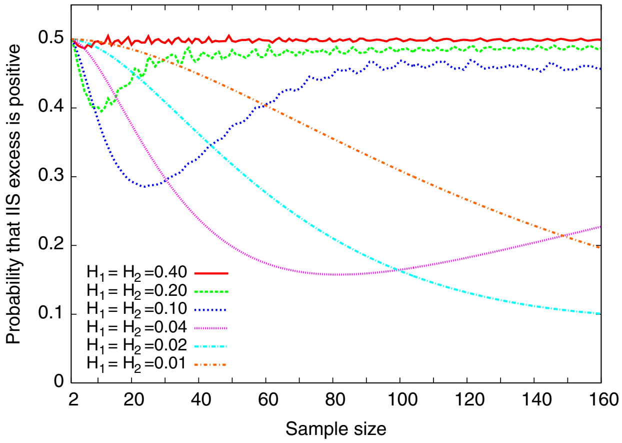

Fig. 3. Probability of positive IIS excess (plus half the probability of IIS excess equal to zero) as a function of sample size, computed from Eqs. (5) and (6), for $H_1 = H_2 = 0.40, 0.20, 0.10, 0.04, 0.02$, and 0.01.

1/2. This can be seen by observing that for a pair of individuals, IIS excess is positive if and only if one individual is a double homozygote and the other is a double heterozygote, and negative if and only if one individual is homozygous only at one of the loci, and the other is homozygous only at the other locus. These events have the same probability, $2H_1(1-H_1)H_2(1-H_2)$.

We now give a Poisson approximation to the distribution of the IIS excess, which applies if both H_1 and H_2 are small. To obtain this approximation, in Proposition 7, we consider sequences with decreasing homozygosities as n increases, rather than sequences in which H_1 and H_2 are constant across values of n. Proposition 7 is used to prove Proposition 6i, as well as Corollary 8. Corollary 8 is illustrated in Fig. 4, which shows the probability of positive IIS excess for increasing n and decreasing H_{1n} and H_{2n}, subject to $H_{1n}=\lambda_1/\sqrt{n}$, $H_{2n}=\lambda_2/\sqrt{n}$.

Proposition 7. *Suppose $H_{1n}=\lambda_1/\sqrt{n}$ and $H_{2n}=\lambda_2/\sqrt{n}$ for positive constants λ_1 and λ_2. Then*

$$n\Delta_n(H_{1n},H_{2n})+\lambda_1\lambda_2 \xrightarrow{d} Poisson(\lambda_1\lambda_2). \tag{10}$$

Proof. Proving the proposition amounts to showing that

$$X_{12n}-\left(\frac{X_{1n}}{\sqrt{n}}\frac{X_{2n}}{\sqrt{n}}-\lambda_1\lambda_2\right)\xrightarrow{d}\text{Poisson}(\lambda_1\lambda_2).$$

Because $\lim_{n\to\infty}H_{1n}=0$ and $\mathbb{E}[(X_{1n}/\sqrt{n}-\lambda_1)^2]=\text{Var}[X_{1n}/\sqrt{n}]$, $\lim_{n\to\infty}\mathbb{E}[(X_{1n}/\sqrt{n}-\lambda_1)^2]=\lim_{n\to\infty}H_{1n}(1-H_{1n})=0$. Thus $X_{1n}/\sqrt{n}$ converges in second mean to λ_1, and consequently [24, p. 10], it converges in probability to λ_1. Similarly, $X_{2n}/\sqrt{n}\xrightarrow{p}\lambda_2$.

The random variables X_{1n} and X_{2n} are independent. Therefore, $X_{12n}\sim$ Binomial(n, $\lambda_1\lambda_2/n$), and it follows that $X_{12n}\xrightarrow{d}X$, where X has a Poisson($\lambda_1\lambda_2$) distribution [23, p. 199]. Slutsky's theorem [24, p. 19] applied to $X_{1n}/\sqrt{n}$ and $X_{2n}/\sqrt{n}$ gives $X_{1n}X_{2n}/n\xrightarrow{d}\lambda_1\lambda_2$, or equivalently [24, p. 19], $X_{1n}X_{2n}/n\xrightarrow{p}\lambda_1\lambda_2$. Applying Slutsky's theorem again gives $X_{12n}-X_{1n}X_{2n}/n\xrightarrow{d}X-\lambda_1\lambda_2$. □

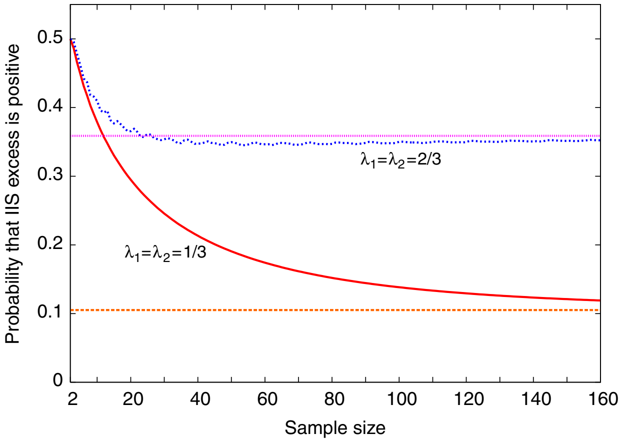

Fig. 4. Probability of positive IIS excess (plus half the probability of IIS excess equal to zero) as a function of sample size, computed from Eqs. (5) and (6), for $\lambda_1=\lambda_2=1/3$, 2/3, and $H_{1n}=\lambda_1/\sqrt{n}$, $H_{2n}=\lambda_2/\sqrt{n}$. The horizontal lines correspond to the large-n limits of $1-\ e^{-1/9}$ for $\lambda_1=\lambda_2=1/3$ and $1-\ e^{-4/9}$ for $\lambda_1=\lambda_2=2/3$, as given by Corollary 8.

Proof of proposition 6(i). Suppose $H_{1n} = \lambda_1/\sqrt{n}$ and $H_{2n} = \lambda_2/\sqrt{n}$ for positive constants λ_1 and λ_2 with $\lambda_1\lambda_2 < 2$. Then

$$Pr[\Delta_n(H_{1n}, H_{2n}) \leqslant -\lambda_1\lambda_2/(2n)] = Pr[n\Delta_n(H_{1n}, H_{2n}) + \lambda_1\lambda_2 \leqslant \lambda_1\lambda_2/2].$$

Because $0 < \lambda_1\lambda_2/2 < 1$, the value $\lambda_1\lambda_2/2$ is a point of continuity of the cumulative distribution function of a random variable X with Poisson($\lambda_1\lambda_2$) distribution. Therefore, using the definition of convergence in distribution together with Proposition 7,

$$\lim_{n\to\infty} Pr[\Delta_n(H_{1n}, H_{2n}) \leqslant -\lambda_1\lambda_2/(2n)] = Pr[X \leqslant \lambda_1\lambda_2/2] = \mathrm{e}^{-\lambda_1\lambda_2},$$

the last equality following from the Poisson distribution and from the fact that $Pr[X \leqslant \lambda_1\lambda_2/2] = Pr[X = 0]$. For any n,

$$\sup_{H_1,H_2\in[0,\,1]} Pr[\Delta_n(H_1, H_2) < 0] \geqslant Pr[\Delta_n(H_{1n}, H_{2n}) < 0] \geqslant Pr[\Delta_n(H_{1n}, H_{2n}) \leqslant -\lambda_1\lambda_2/(2n)].$$

Taking the limit as $n \to \infty$, it follows that:

$$1 \geqslant \lim_{n\to\infty} \sup_{H_1,H_2\in[0,\,1]} Pr[\Delta_n(H_1, H_2) < 0] \geqslant \lim_{n\to\infty} Pr[\Delta_n(H_{1n}, H_{2n}) \leqslant -\lambda_1\lambda_2/(2n)] = \mathrm{e}^{-\lambda_1\lambda_2}.$$

As this inequality is true for any $\lambda_1,\lambda_2 > 0$ with $\lambda_1\lambda_2 < 2$, $\lambda_1\lambda_2$ can be made arbitrarily close to 0, so that

$$\lim_{n\to\infty} \sup_{H_{1n},H_{2n}\in[0,\,1]} Pr[\Delta_n(H_{1n},H_{2n}) < 0] = 1. \quad \square$$

Corollary 8. *Suppose $H_{1n} = \lambda_1/\sqrt{n}$ and $H_{2n} = \lambda_2/\sqrt{n}$ for positive constants λ_1 and λ_2 with $\lambda_1\lambda_2 < 1$. Then*

(i) $\lim_{n\to\infty} Pr[\Delta_n(H_{1n},H_{2n}) > 0] = 1 - \mathrm{e}^{-\lambda_1\lambda_2}$
(ii) $\lim_{n\to\infty} Pr[\Delta_n(H^*_{1n},H_{2n}) > 0] = \mathrm{e}^{-\lambda_1\lambda_2}$
(iii) $\lim_{n\to\infty} Pr[\Delta_n(H_{1n},H^*_{2n}) > 0] = \mathrm{e}^{-\lambda_1\lambda_2}$
(iv) $\lim_{n\to\infty} Pr[\Delta_n(H^*_{1n},H^*_{2n}) > 0] = 1 - \mathrm{e}^{-\lambda_1\lambda_2}$

Proof. By Lemma 2, (ii), (iii), and (iv) follow from (i). If $X \sim$ Poisson($\lambda_1\lambda_2$), then

$$\begin{aligned}\lim_{n\to\infty} Pr[\Delta_n(H_{1n}, H_{2n}) \leqslant 0] &= \lim_{n\to\infty} Pr[n\Delta_n(H_{1n}, H_{2n}) \leqslant 0]\\ &= \lim_{n\to\infty} Pr[n\Delta_n(H_{1n}, H_{2n}) + \lambda_1\lambda_2 \leqslant \lambda_1\lambda_2]\\ &= Pr[X \leqslant \lambda_1\lambda_2],\end{aligned}$$

where the last equality follows from the convergence in distribution in Proposition 7 and from the fact that because $\lambda_1\lambda_2$ is strictly between 0 and 1, it is a point of continuity of the cumulative distribution function of X. Because $\lambda_1\lambda_2 < 1$, $Pr[X \leqslant \lambda_1\lambda_2] = Pr[X = 0] = \mathrm{e}^{-\lambda_1\lambda_2}$. $\square$

3. Application to data

Investigation of the sampling properties of the IIS excess is useful in interpreting observations based on this statistic. We now illustrate how our theoretical results help to understand estimates of the IIS excess for microsatellite genotypes in human populations.

We have previously studied the patterns of genetic variation of 377 autosomal microsatellite loci in a collection of 1056 individuals from 52 human populations [18,19,35]. Using this data set, for each population, Rosenberg and Calabrese [18] computed the IIS excess for each of 66,730 (unordered) pairs of unlinked loci. It was observed that in most of the populations, the number of pairs of unlinked loci that had positive IIS excess was smaller than the number that had negative IIS excess [18].

For four of the 52 populations—French, Kalash, Surui, and Yoruba—Fig. 5 plots pairs of estimated homozygosities for all of the 66,730 pairs of loci (with both possible orderings). As the estimate of homozygosity for a given locus is obtained from the fraction of sampled individuals homozygous at the locus (excluding individuals with missing data), the points that could potentially be occupied by ordered pairs of homozygosities form a discrete set symmetric around the line $H_2 = H_1$. The fractions of pairs with positive IIS excess (including half the pairs with IIS excess equal to 0) are 0.478, 0.487, 0.501, and 0.462 for French, Kalash, Surui, and Yoruba, respectively [18]. In French, Kalash, and Yoruba, most pairs of (estimated) homozygosities are in parts of the unit square where the probability of positive IIS excess for a pair of independent loci with those homozygosities is less than 1/2 (red dots in Fig. 5). Thus, in these populations, it is not surprising that fewer than half the pairs would have positive IIS excess. In Surui, however, pairs of homozygosities are distributed over the entire unit square, and it is therefore unsurprising that in this population, the probability of positive IIS excess is very close to 1/2. Thus, by superimposing ordered pairs of homozygosities obtained from data onto plots of the theoretically computed

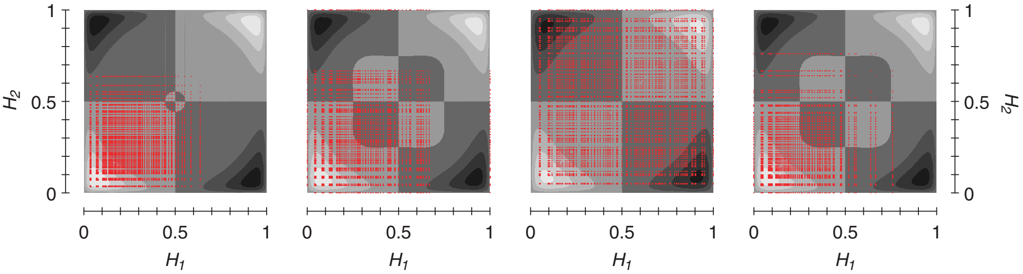

Fig. 5. Probability of positive IIS excess (plus half the probability of IIS excess equal to zero) as a function of homozygosities (H_1, H_2), computed from Eqs. (5) and (6), with estimated homozygosities from 66,730 pairs of human microsatellite markers superimposed (red dots). As a result of the finite number of possible estimates that can be obtained in small samples, multiple pairs of markers can potentially produce identical estimates; thus, many marker pairs may be represented by the same red dot. From left to right, the plots are based on data from four populations studied by Rosenberg and Calabrese [18]: French ($n = 29$), Kalash ($n = 25$), Surui ($n = 21$), and Yoruba ($n = 25$). For each population, the probability of positive IIS excess is obtained from an exact computation using the same sample size as the sample size of the population. From lightest to darkest, the shades represent positive IIS excess probabilities in [0.175, 0.3), [0.3, 0.375), [0.375, 0.45), [0.45, 50), {0.50}, (0.50, 0.55], (0.55, 0.625], (0.625, 0.7], and (0.7, 0.825].

probability of positive IIS excess, it is possible to understand why in Surui the probability of positive IIS excess is closer to 1/2 than in the other three populations.

4. Discussion

We have shown in this article that for any pair of loci in linkage equilibrium, a transformed version of the IIS excess statistic has the desirable property of converging to a normal distribution as sample size approaches infinity. This enables the development of an asymptotic test for linkage disequilibrium: under the null hypothesis of linkage equilibrium, the statistic

$$\sqrt{n}HR_n(H_1, H_2) = \sqrt{n}\frac{\frac{X_{12n}}{n} - \frac{X_{1n}}{n}\frac{X_{2n}}{n}}{\sqrt{\frac{X_{1n}}{n}\left(1 - \frac{X_{1n}}{n}\right)\frac{X_{2n}}{n}\left(1 - \frac{X_{2n}}{n}\right)}}$$

has the standard normal distribution. Note that the null hypothesis that leads to the standard normal distribution is a hypothesis of linkage equilibrium only, and not of Hardy–Weinberg equilibrium at either or both loci. As was mentioned previously, however, if Hardy–Weinberg disequilibrium is present, a nonzero value of $\sqrt{n}HR_n(H_1, H_2)$ may be a consequence both of associations of pairs of alleles (that is, linkage disequilibrium) as well as of other disequilibria [34].

The statistic $\sqrt{n}HR_n(H_1, H_2)$ complements other IIS-based statistics identified by Sabatti and Risch [20] as having standard distributions under the null hypothesis of linkage equilibrium. Indeed, one consequence of the asymptotic normality of $\sqrt{n}HR_n(H_1, H_2)$ is that $nHR_n(H_1,H_2)^2$ has an asymptotic χ_1^2 distribution, a fact that is equivalent to the asymptotic χ_1^2 distribution identified by Sabatti and Risch [20, p. 1714] for the test statistic of independence for a certain contingency table. Note also that homozygosity-based LD tests such as those using $\sqrt{n}HR_n(H_1, H_2)$ and $nHR_n(H_1,H_2)^2$ parallel corresponding tests based on LD statistics that apply only when haplotype phase is known [30–32].

In implementing LD tests based on the IIS excess, however, and more generally, in simply measuring the IIS excess, caution is warranted in interpreting the values of IIS-based statistics. Although for large sample sizes, the positive IIS excess probability is close to the asymptotic value of 1/2 for most homozygosity values, for extreme homozygosities, the probability of the IIS excess being positive in typical sample sizes may be quite far from the asymptotic limit. Indeed, for any sample size, there will be some extreme homozygosities for which the distribution of the IIS excess is not close to the asymptotic distribution. When using the IIS excess, properties of its exact distribution (Eq. (3)), such as those studied here, can be used to identify the potential for a bias, so that if necessary, this bias can be incorporated into interpretations of the observed values of the statistic.

Acknowledgements

We thank the reviewers for comments on an earlier draft of the manuscript. D. Ghosh provided helpful insights for the proof of Proposition 3. This research was supported by a Career Award in the Biomedical Sciences from the Burroughs Wellcome Fund to NAR and by NIH grant HL084729.

References

[1] A. Agresti, Categorial Data Analysis, Wiley, New York, 1990.

[2] R. Chakraborty, Detection of nonrandom association of alleles from the distribution of the number of heterozygous loci in a sample, Genetics 108 (1984) 719.

[3] B. Devlin, N. Risch, A comparison of linkage disequilibrium measures for fine-scale mapping, Genomics 29 (1995) 311.

[4] T.S. Ferguson, A Course in Large Sample Theory, Chapman & Hall, Boca Raton, 1996.

[5] R.R. Hudson, Linkage disequilibrium and recombination, in: D.J. Balding, M. Bishop, C. Cannings (Eds.), Handbook of Statistical Genetics, Wiley, Chichester, UK, 2001, p. 309, chapter 11.

[6] H. Innan, M. Nordborg, The extent of linkage disequilibrium and haplotype sharing around a polymorphic site, Genetics 165 (2003) 437.

[7] K. Knight, Mathematical Statistics, Chapman & Hall, Boca Raton, 2000.

[8] N. Li, M. Stephens, Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data, Genetics 165 (2003) 2213.

[9] J. Marchini, D. Cutler, N. Patterson, M. Stephens, E. Eskin, E. Halperin, S. Lin, Z.S. Qin, H.M. Munro, G.R. Abecasis, P. Donnelly, A comparison of phasing algorithms for trios and unrelated individuals, Am. J. Hum. Genet. 78 (2006) 437.

[10] T. Niu, Z.S. Qin, X. Xu, J.S. Liu, Bayesian haplotype inference for multiple linked single-nucleotide polymorphisms, Am. J. Hum. Genet. 70 (2002) 157.

[11] M. Nordborg, S. Tavaré, Linkage disequilibrium: what history has to tell us, Trends Genet. 18 (2002) 83.

[12] T. Ohta, Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families, Genet. Res. 36 (1980) 181.

[13] T. Ohta, Linkage disequilibrium due to random genetic drift in finite subdivided populations, Proc. Natl. Acad. Sci. USA 79 (1982) 1940.

[14] T. Ohta, An attempt to measure the patchwork pattern observed among alleles at major histocompatibility complex loci, J. Mol. Evol. 51 (2000) 21.

[15] J.K. Pritchard, M. Przeworski, Linkage disequilibrium in humans: models and data, Am. J. Hum. Genet. 69 (2001) 1.

[16] S.E. Ptak, K. Voelpel, M. Przeworski, Insights into recombination from patterns of linkage disequilibrium in humans, Genetics 167 (2004) 387.

[17] B. Rannala, G. Bertorelle, Using linked markers to infer the age of a mutation, Hum. Mutat. 18 (2001) 87.

[18] N.A. Rosenberg, P.P. Calabrese, Polyploid and multilocus extensions of the Wahlund inequality, Theor. Pop. Biol. 66 (2004) 381.

[19] N.A. Rosenberg, J.K. Pritchard, J.L. Weber, H.M. Cann, K.K. Kidd, L.A. Zhivotovsky, M.W. Feldman, Genetic structure of human populations, Science 298 (2002) 2381.

[20] C. Sabatti, N. Risch, Homozygosity and linkage disequilibrium, Genetics 160 (2002) 1707.

[21] C. Sabatti, N. Risch, Response to the letter "Gametic and zygotic associations" by Rong-Cai Yang, Genetics 165 (2003) 451.

[22] D.J. Schaid, Linkage disequilibrium testing when linkage phase is unknown, Genetics 166 (2004) 505.

[23] R. Sedgewick, P. Flajolet, An Introduction to the Analysis of Algorithms, Addison-Wesley, Boston, 1996.

[24] R.J. Serfling, Approximation Theorems of Mathematical Statistics, Wiley, New York, 1980.

[25] M. Slatkin, B. Rannala, Estimating allele age, Annu. Rev. Genomics Hum. Genet. 1 (2000) 225.

[26] M. Stephens, P. Scheet, Accounting for decay of linkage disequilibrium in haplotype inference and missing-data imputation, Am. J. Hum. Genet. 76 (2005) 449.

[27] R. Vitalis, D. Couvet, Estimation of effective population size and migration rate from one- and two-locus identity measures, Genetics 157 (2001) 911.

[28] R. Vitalis, D. Couvet, Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population, Genet. Res. 77 (2001) 67.

[29] J.D. Wall, P. Andolfatto, M. Przeworski, Testing models of selection and demography in Drosophila simulans, Genetics 162 (2002) 203.

[30] B.S. Weir, Inferences about linkage disequilibrium, Biometrics 35 (1979) 235.

[31] B.S. Weir, Genetic Data Analysis II, Sinauer, Sunderland, MA, 1996.

[32] B.S. Weir, C.C. Cockerham, Testing hypotheses about linkage disequilibrium with multiple alleles, Genetics 88 (1978) 633.

[33] B.S. Weir, C.C. Cockerham, Complete characterization of disequilibrium at two loci, in: M.W. Feldman (Ed.), Mathematical Evolutionary Theory, Princeton University Press, Princeton, NJ, 1989, p. 86, chapter 6.

[34] R.-C. Yang, Gametic and zygotic associations, Genetics 165 (2003) 447.

[35] L.A. Zhivotovsky, N.A. Rosenberg, M.W. Feldman, Features of evolution and expansion of modern humans, inferred from genomewide microsatellite markers, Am. J. Hum. Genet. 72 (2003) 1171.

[36] K.T. Zondervan, L.R. Cardon, The complex interplay among factors that influence allelic association, Nature Rev. Genet. 5 (2004) 89.