Available online at www.sciencedirect.com
SCIENCE DIRECT®
Theoretical Population Biology 66 (2004) 381–391
Theoretical Population Biology
www.elsevier.com/locate/ytpbi

Polyploid and multilocus extensions of the Wahlund inequality

Noah A. Rosenberg*, Peter P. Calabrese

Program in Molecular and Computational Biology, Department of Biological Sciences, University of Southern California, 1042 West 36th Place—DRB 289, Los Angeles, CA 90089, USA

Received 9 April 2003

Abstract

Wahlund's inequality informally states that if a structured and an unstructured population have the same allele frequencies at a locus, the structured population contains more homozygotes. We show that this inequality holds generally for ploidy level P, that is, the structured population has more *P-polyhomozygotes*. Further, for M randomly chosen loci ($M \geqslant 2$), the structured population is also expected to contain more *M-multihomozygotes* than an unstructured population with the same single-locus homozygosities. The extended inequalities suggest multilocus identity coefficients analogous to F_{ST}. Using microsatellite genotypes from human populations, we demonstrate that the multilocus Wahlund inequality can explain a positive bias in "identity-in-state excess".

Keywords: Homozygosity; Identity coefficients; Linkage disequilibrium; Polyploidy; Population subdivision; Wahlund effect

The fundamental principle in the study of genetically structured populations is the inequality of Wahlund (1928). Informally, for any locus, this inequality states that given structured and unstructured populations with the same allele frequencies, the structured population includes more homozygotes (Crow and Kimura, 1970, p. 54). An extreme case illustrates the result: if all of the subgroups within the structured population are fixed for different alleles, the structured population is fully homozygous, whereas an unstructured Hardy–Weinberg population that contains all of these alleles necessarily contains heterozygotes. In the form of the F-quantities of Wright (1951), much effort has been devoted to measurement of this excess homozygosity in theoretical population structure models, and to its estimation from genetic data (Excoffier, 2001; Rousset, 2002; Weir and Hill, 2002).

We show here that the Wahlund inequality extends to ploidy levels P larger than 2. For $P \geqslant 2$, at a given locus, the *P-polyhomozygosity* is the fraction of individuals in a population who have P copies of the same allele. Indeed, the proportion of P-polyhomozygotes is larger in a structured population than in the corresponding unstructured population.

More generally, the Wahlund inequality has a multilocus analogue for an arbitrary number of loci. Consider a set of M loci that are *genotypically unassociated* within subpopulations (loci for which single-locus genotypes combine randomly into multilocus genotypes), and structured and unstructured populations whose corresponding single-locus homozygosities are equal (for all M loci). Informally, the multilocus Wahlund inequality states that the expected proportion of *M-multihomozygotes* (homozygotes at all M loci) is larger in the structured than in the unstructured population. The expectation is taken over random sets of M genotypically unassociated loci. Corresponding inequalities apply for any ploidy $P \geqslant 2$. We use *M, P-homozygosity* to describe the proportion of the population that for each of M loci, has P copies of the same allele. *Homozygosity* is equivalent to 1, 2-*homozygosity*, and 2, 2-homozygosity is abbreviated *double homozygosity*.

*Corresponding author. Fax: +213-740-2437.
E-mail address: noahr@usc.edu (N.A. Rosenberg).

0040-5809/$ - see front matter © 2004 Elsevier Inc. All rights reserved.
doi:10.1016/j.tpb.2004.07.001

Previous treatments of Wahlund's results (and extensions to two loci) have quantified differences between unstructured and structured populations whose allele frequencies are equal (Sinnock, 1975), and have explored their dynamics (Feldman and Christiansen, 1975; Christiansen, 1988). Here we focus on the sign of these differences. After introducing notation in Section 1, in Sections 2 and 3 we prove the polyploid and multilocus extensions of Wahlund's inequality as consequences of Hölder's inequality. We then discuss implications for identity coefficients and genotypic associations in Section 4, providing examples of the two-locus Wahlund inequality using genotypes from human populations.

1. Notation

Suppose that locus m has N_m alleles, and that allele j at locus m has frequency q_{imj} in (infinitely large) population i. As usual, $q_{imj} \in [0,1]$ for all i, m, and j, and $\sum_{j=1}^{N_m} q_{imj} = 1$ for all i and m. For ploidy level P, the P-polyhomozygosity of locus m in population i is $H_{im(P)} \in [0,1]$. For a set of M loci, the M,P-homozygosity of population i is $H_{i(P)} \in [0,1]$. Henceforth we leave off the subscript P and use only H_{im} and H_i, as the ploidy should be clear from the context. Population i and locus m satisfy *Hardy–Weinberg proportions* if alleles at locus m combine randomly into sets of P alleles. In particular, under Hardy–Weinberg proportions,

$$H_{im} = \sum_{j=1}^{N_m} q_{imj}^P. \tag{1}$$

Loci $1,2,\ldots,M$ are *genotypically unassociated* in population i if single-locus genotypes combine randomly into multilocus genotypes. In particular, genotypically unassociated loci satisfy

$$H_i = \prod_{m=1}^{M} H_{im}. \tag{2}$$

Consider a set of (unstructured) populations $i = 1,2,\ldots,I$ in which loci $m = 1,2,\ldots,M$ are genotypically unassociated. Loci need not satisfy Hardy–Weinberg proportions in any of these populations. Let S be a structured population such that for $i = 1,2,\ldots,I$, the proportion of individuals in S drawn from population i is $f_i \in (0,1]$, with $\sum_{i=1}^{I} f_i = 1$. Then

$$H_{Sm} = \sum_{i=1}^{I} f_i H_{im}, \tag{3}$$

$$H_S = \sum_{i=1}^{I} f_i H_i. \tag{4}$$

Eq. (3) holds for all m.

Consider also an *unstructured* population V in which each locus has the same P-polyhomozygosity as in the structured population S. Loci in V need not satisfy Hardy–Weinberg proportions, and allele frequencies at a locus in V need not equal the corresponding frequencies in S. The frequencies in V are only restricted by the requirement that they be compatible with the P-polyhomozygosity of the locus.

Assume that loci $m = 1,2,\ldots,M$ are genotypically unassociated in population V. In contrast to the structured S, for which the M,P-homozygosity is a weighted average of the M,P-homozygosities of the component subpopulations, in the unstructured V, the M,P-homozygosity is the product across loci of average locus P-polyhomozygosities:

$$H_{Vm} = \sum_{i=1}^{I} f_i H_{im} = H_{Sm}, \tag{5}$$

$$H_V = \prod_{m=1}^{M} H_{Vm}. \tag{6}$$

Eq. (5) holds for all m.

Finally, consider a second unstructured population T that for each locus has the same *allele frequencies* as the structured population S. Assume that loci $1,2,...,M$ are genotypically unassociated in T. Also assume that each locus satisfies Hardy–Weinberg proportions in T. Then

$$q_{Tmj} = \sum_{i=1}^{I} f_i q_{imj}, \tag{7}$$

$$H_{Tm} = \sum_{j=1}^{N_m} q_{Tmj}^P, \tag{8}$$

$$H_T = \prod_{m=1}^{M} H_{Tm}. \tag{9}$$

Eq. (7) holds for all j and m and (8) holds for all m.

The motivation for introducing two distinct unstructured populations is as follows. Wahlund inequalities for one locus relate P-polyhomozygosities of population S and population T, whose allele frequencies correspond to those in S. In contrast, Wahlund inequalities for multiple loci relate M,P-homozygosities of population S and population V, whose P-polyhomozygosities correspond to those in S.

2. Wahlund inequalities: one locus

We will need the following Hölder inequality (Beckenbach and Bellman, 1961, p. 68).

Theorem 1 (*Hölder inequality*). *For a real number $z > 1$, given two sequences of I nonnegative real numbers*

$a_1, a_2, \ldots, a_I$ *and* $b_1, b_2, \ldots, b_I$,

$$\left(\sum_{i=1}^{I} a_i^z\right)^{1/z}\left(\sum_{i=1}^{I} b_i^{z^*}\right)^{1/z^*} \geqslant \sum_{i=1}^{I} a_i b_i, \tag{10}$$

where $z^* = z/(z-1)$. *Equality holds in* (10) *if and only if there exists a constant* c *such that for all* i, $a_i^z = cb_i^{z^*}$.

The informal Wahlund inequality is that a structured population has higher homozygosity than predicted by its allele frequencies under the assumption that loci satisfy Hardy–Weinberg proportions within subpopulations, with equality if and only if all subpopulations have the same allele frequencies. The following theorem gives the formal statement.

Theorem 2 (*Diploid Wahlund inequality*). *For* $M = 1$, $P = 2$ *and any locus* m, *if the* I *component subpopulations of* S *each satisfy Hardy–Weinberg proportions at locus* m, *then*

$$H_{Sm} \geqslant H_{Tm}, \tag{11}$$

with equality if and only if for all j *and any* i_1, i_2, $q_{i_1mj} = q_{i_2mj}$.

Proof. Consider a particular allele j. Applying Theorem 1 with $z = 2$, $a_i = q_{imj}\sqrt{f_i}$ and $b_i = \sqrt{f_i}$,

$$\sum_{i=1}^{I} f_i q_{imj}^2 \geqslant \left(\sum_{i=1}^{I} f_i q_{imj}\right)^2, \tag{12}$$

with equality if and only if $q_{1mj} = q_{2mj} = \cdots = q_{Imj}$. Summing (12) across alleles and applying (1), (3), (7), and (8), both (11) and the equality condition follow. □

Note that this proof verifies the stronger statement that for *each* allele, the proportion of homozygotes for that allele is greater in the structured population than in the unstructured population, with equality if and only if the allele has the same frequency in all subpopulations.

Theorem 2 also extends to P-polyhomozygotes, that is, a structured population contains at least as many P-polyhomozygotes as predicted by its allele frequencies under the assumption of Hardy–Weinberg proportions within subpopulations. The proof follows that of Theorem 2, using $z = P$, $a_i = q_{imj} f_i^{1/P}$, $b_i = f_i^{(P-1)/P}$.

Theorem 3 (*Poly-Wahlund inequality*). *For* $M = 1$, $P \geqslant 2$ *and any locus* m, *if the* I *component subpopulations of* S *each satisfy Hardy–Weinberg proportions at locus* m, *then*

$$H_{Sm} \geqslant H_{Tm}, \tag{13}$$

with equality if and only if for all j *and any* i_1, i_2, $q_{i_1mj} = q_{i_2mj}$.

Similarly to Theorem 2, Theorem 3 shows that for each allele, the proportion of P-polyhomozygotes for that allele is at least as large in the structured as in the unstructured population, with equality if and only if the allele has equal frequency in all subpopulations.

3. Wahlund inequalities: M loci

The extension of Theorem 2 to multiple loci is less straightforward than its extension to polyploidy. Theorem 2 guarantees that for *any* diploid locus that satisfies Hardy–Weinberg proportions within subpopulations, the structured population has higher homozygosity than the corresponding unstructured population; Theorem 3 produces a similar conclusion for polyploid loci. However, as demonstrated below, sets of M loci exist for which a structured population has lower M, P-homozygosity than the corresponding unstructured population (in other words, H_S can be smaller than H_V). This remains true even if the restriction of Hardy–Weinberg proportions within subpopulations is imposed on the unstructured population (H_S can be smaller than H_T).

Consider a two-locus, two-subpopulation, diploid case with $f_1 = f_2 = 0.5$. Suppose $H_{11} = H_{22} = 1$ and $H_{12} = H_{21} = 0.66$. Then the double homozygosity H_S is $(0.5)(1)(0.66) + (0.5)(0.66)(1) = 0.66$, whereas H_V is $[(0.5)(1) + (0.5)(0.66)]^2 = 0.6889$. To produce a scenario with $H_S < H_T$, we can further suppose that both loci satisfy Hardy–Weinberg proportions in both subpopulations, that both loci have three alleles, and that $(1, 0, 0)$, $(0.8, 0.1, 0.1)$, $(0.8, 0.1, 0.1)$, and $(1, 0, 0)$ represent the allele frequency vectors of population 1 and locus 1, population 1 and locus 2, population 2 and locus 1, and population 2 and locus 2, respectively. We then obtain $H_T = [0.9^2 + 2(0.05)^2]^2 = 0.664225$.

By appending $M - 2$ monomorphic loci to form a set of M loci, this counterexample can be extended to arbitrary $M \geqslant 2$ and $P \geqslant 2$. Setting $H_{11} = H_{22} = 1$ and $H_{12} = H_{21} = 0.8^P + 2(0.1^P)$, we have $H_V = [(1 + 0.8^P + 2(0.1^P))/2]^2$ and $H_S = 0.8^P + 2(0.1^P)$. If we assume that all loci satisfy Hardy–Weinberg proportions in both subpopulations and apply the allele frequencies from the previous paragraph, $H_T = [0.9^P + 2(0.05)^P]^2$. It is then straightforward to prove that for any $P \geqslant 2$, $H_V > H_T > H_S$ (for example, at $P = 6$, $H_V \approx 0.40$, $H_T \approx 0.28$, $H_S \approx 0.26$).

Thus, the direct analogue of the Wahlund inequality, claiming for *any* set of M loci greater M, P-homozygosity in a structured population than in the corresponding unstructured population, does not hold. We will see however that a multilocus extension of the Wahlund inequality does exist, in that the *expectation* of M, P-homozygosity taken over sets of M loci that are genotypically unassociated in unstructured populations is at least as large in the structured S as in the unstructured V (and in T, if Hardy–Weinberg proportions are satisfied in the subpopulations of S).

Suppose that for population i, H_{i1}, H_{i2}, ..., H_{iM} are independently and identically distributed with mean μ_i and variance σ_i^2. Denote the mean P-polyhomozygosity of a randomly chosen locus m, in populations S and V, by μ. That is,

$$\mu = \mathbb{E}[H_{Sm}] = \mathbb{E}[H_{Vm}] = \sum_{i=1}^{I} f_i \mathbb{E}[H_{im}] = \sum_{i=1}^{I} f_i \mu_i. \quad (14)$$

The expectations and variances of H_S and H_V over sets of M loci that are genotypically unassociated in unstructured populations are (Appendix A)

$$\mathbb{E}[H_S] = \sum_{i=1}^{I} f_i \mu_i^M = \mu_S, \quad (15)$$

$$\mathbb{E}[H_V] = \left(\sum_{i=1}^{I} f_i \mu_i \right)^M = \mu^M = \mu_V, \quad (16)$$

$$\mathrm{Var}[H_S] = \sum_{i=1}^{I} f_i^2 [(\mu_i^2 + \sigma_i^2)^M - (\mu_i^2)^M], \quad (17)$$

$$\mathrm{Var}[H_V] = \left(\mu^2 + \sum_{i=1}^{I} f_i^2 \sigma_i^2 \right)^M - (\mu^2)^M, \quad (18)$$

$$\mathrm{Cov}[H_S, H_V] = \sum_{i=1}^{I} f_i [(\mu_i \mu + f_i \sigma_i^2)^M - (\mu_i \mu)^M]. \quad (19)$$

Theorem 4 (*Multilocus Wahlund inequality*). *For* $M \geqslant 2$ *and* $P \geqslant 2$,

$$\mathbb{E}[H_S] \geqslant \mathbb{E}[H_V], \quad (20)$$

with equality if and only if $\mu_1 = \mu_2 = \cdots = \mu_I$.

The expectations are taken over sets of M loci that are genotypically unassociated in unstructured populations.

Proof. Using (15) and (16), we wish to prove

$$\sum_{i=1}^{I} f_i \mu_i^M \geqslant \left(\sum_{i=1}^{I} f_i \mu_i \right)^M. \quad (21)$$

The truth of (21) follows from Theorem 1, using $z = M$, $a_i = \mu_i f_i^{1/M}$, and $b_i = f_i^{(M-1)/M}$. Equality holds in (21) if and only if for all i, $(\mu_i)^M$ equals a constant c. □

Theorem 4 states that the expected M, P-homozygosity is larger in a structured population than is predicted from its single-locus P-polyhomozygosities under the assumption of no genotypic association within subpopulations. Regarding individual loci, the theorem requires only that their P-polyhomozygosities be specified and it imposes no requirements on allele frequencies in the component subpopulations, other than that they be compatible with the P-polyhomozygosities. In particular, the loci need not satisfy Hardy–Weinberg proportions within subpopulations in order for the expected proportion of M, P-homozygotes to be larger in population S than in population V.

If the loci do all satisfy Hardy–Weinberg proportions in all component populations of S, however, we can also relate H_S and H_T. Applying Theorem 3 to each locus and multiplying across loci, $\prod_{m=1}^{M} H_{Sm} \geqslant \prod_{m=1}^{M} H_{Tm}$. Using (5), (6), and (9), $H_V \geqslant H_T$. Taking expectations over sets of M loci that are genotypically unassociated in unstructured populations and using Theorem 4, $\mathbb{E}[H_S] \geqslant \mathbb{E}[H_V] \geqslant \mathbb{E}[H_T]$. Explicit expansion of $\mathbb{E}[H_T]$ is difficult unless assumptions about allele frequency distributions are made; nevertheless, it is still true that a structured population is expected to contain more M, P-homozygotes than predicted from its allele frequencies under Hardy–Weinberg proportions and absence of genotypic association within component populations.

4. Applications

4.1. Identity coefficients

In the diploid case of a single locus, an important quantity for the structured population S is the identity coefficient F_{ST} (Wright, 1951; Excoffier, 2001; Rousset, 2002; Weir and Hill, 2002; Balding, 2003). For a random locus, assuming Hardy–Weinberg proportions in the diploid populations $1, 2, \ldots, I$, one formulation of F_{ST} is as a measurement of the excess homozygosity in the structured population S compared to the unstructured population T whose allele frequencies equal those in S:

$$F_{ST} = \frac{H_S - H_T}{1 - H_T}. \quad (22)$$

Theorem 2 guarantees that $F_{ST} \geqslant 0$, and the fact that homozygosities are in [0,1] guarantees $F_{ST} \in [0, 1]$. Because homozygosity is the probability that two alleles at a locus are identical in state, F_{ST} gives a normalized measure of the excess identity in a structured population compared to a corresponding unstructured population.

Theorem 3 suggests that the inequality $0 \leqslant F_{ST} \leqslant 1$ applies for any ploidy level. Thus, it is sensible to discuss higher order identity coefficients defined by the same eq. (22). More precisely, F_{ST} can be labeled $F_{ST(2)}$ as an excess in the probability that *two* randomly chosen alleles are identical in state. The coefficient $F_{ST(P)}$, which, like $F_{ST(2)}$, is in [0,1], then refers to the excess probability that P randomly chosen alleles are identical in state. Note that this quantity is sensible even if the organism under consideration does not have ploidy P. Under the infinitely many alleles mutation model, in which each mutation produces a novel allele, $F_{ST(P)}$ also equals the excess probability that P randomly chosen alleles are identical by descent.

Theorem 4 also enables a multilocus analogue of F_{ST}:

$$F_{SV(M,P)} = \frac{\mathbb{E}[H_S] - \mathbb{E}[H_V]}{1 - \mathbb{E}[H_V]}. \quad (23)$$

$F_{SV(M,P)}$ is the excess probability of M, P-homozygosity in the structured population S compared to the unstructured population V: it is the excess probability that at each of M randomly chosen genotypically unassociated loci, P randomly chosen alleles are identical in state.

For each i, suppose n_i individuals are sampled from population i and that $\hat{\mu}_i$ estimates μ_i. If $n = \sum_{i=1}^{I} n_i$, using (15) and (16), $F_{SV(M,P)}$ can be estimated by

$$\hat{F}_{SV(M,P)} = \frac{\sum_{i=1}^{I}(n_i/n)\hat{\mu}_i^M - [\sum_{i=1}^{I}(n_i/n)\hat{\mu}_i]^M}{1 - [\sum_{i=1}^{I}(n_i/n)\hat{\mu}_i]^M}. \quad (24)$$

4.2. Homozygosity and genotypic association

It is often of interest to measure differences between frequencies of multilocus diploid genotypes and the products of the frequencies of their constituent diploid genotypes. These differences, or *genotypic associations*, are related to *gametic association*, the difference between the frequency of a multilocus *haplotype* and the product of the frequencies of its constituent *alleles*. *Linkage disequilibrium* refers to gametic association for two loci. For closely linked loci, cotransmission of alleles at neighboring loci from common ancestral haplotypes causes gametic association (Nordborg and Tavaré, 2002, for example). When pairs of haplotypes are joined to form multilocus diploid genotypes, gametic association gives rise to genotypic association.

Analogously to the occurrence of gametic association in structured populations (Nei and Li, 1973; Ohta, 1982), genotypic association also occurs in structured populations, even if component subpopulations have no genotypic association. Because multilocus genotype frequencies vary across the subpopulations, individual genotypes at one or more loci provide information about which subpopulation they belong to, and are thus informative about the genotypes of the individual at the other loci.

For the case of $M = 2$, Ohta (1980) suggested that if M loci are genotypically associated, then the proportion of M-multihomozygotes will differ from the product of the constituent single-locus homozygosities. Thus, for a population i and a set of M loci, the difference Δ_M between M-multihomozygosity and the product of M single-locus homozygosities—the "identity excess"—is a measure of genotypic association (an especially convenient one, if haplotype phase is not known):

$$\Delta_M = H_i - \prod_{i=1}^{M} H_{im}. \quad (25)$$

By (2), if population i is unstructured and if the M loci are genotypically unassociated, then Δ_M, termed here the *identity-in-state excess* or *IIS excess*, equals zero. For genotypically unassociated loci in a structured population S, however, as shown by the multilocus Wahlund inequality (Theorem 4) for any $M \geqslant 2$, both $\mathbb{E}[\Delta_M]$ and genotypic association coefficients that equal Δ_M divided by positive quantities (Ohta, 1980; Vitalis and Couvet, 2001; Sabatti and Risch, 2002) have positive expectation.

To illustrate this consequence of the multilocus Wahlund inequality, we assemble structured populations from the generally unstructured human populations in the data set of Rosenberg et al. (2002). The data set includes genotypes at 377 autosomal microsatellite loci for 1056 individuals from 52 populations. Here we use 375 of these loci, excluding D11S4463 and D20S201 because of uncertainty about their positions in the genome.

For the multilocus Wahlund inequality to apply, loci must be genotypically unassociated within component subpopulations. Genotypic associations in an unstructured population, if present, are most likely to occur for closely linked loci, producing negative correlation coefficients between IIS excess statistics and genetic distance (Hedrick and Thomson, 1986; Vitalis and Couvet, 2001). Of 70,125 possible pairs of loci, 3395 include two loci that lie on the same chromosome. In most of the populations, however, the genotypic association statistic HR^2 (Sabatti and Risch, 2002), estimated for these 3395 pairs, showed little correlation with genetic distance (Table 1). Karitiana and Surui had the most strongly negative correlations, -0.065 ($p = 2 \times 10^{-4}$) and -0.055 ($p = 0.002$), respectively. Considering only the 228 pairs of loci separated by distances of 10 cM or less, the most negative correlations are farther from zero: -0.208 in Oroqen and -0.197 in Karitiana. However, due to a smaller number of data points, the p-values are larger: 0.005 and 0.003, respectively.

These computations demonstrate that little genotypic association is present within the individual populations. Nevertheless, to ensure that any positive IIS excess observed from data can be attributed to population structure rather than to linkage, we restrict our attention to the 66,730 pairs in which the loci lie on different chromosomes.

We can construct structured populations by combining individuals from several populations in proportion to sample sizes. Denote the sample size from population i by n_i and the total sample size for a collection of populations by n. Then the contribution of population i to the structured population is $f_i = n_i/n$. Let h_{im} be the observed proportion of homozygotes in population i for a locus m, and let $h_{i(m_1,m_2)}$ be the observed proportion of double homozygotes for two loci, m_1 and m_2. Let

$\delta_{i(m_1,m_2)}$ be the IIS excess for loci m_1 and m_2 in population i, estimated by inserting $h_{i(m_1,m_2)}$, h_{im_1}, and h_{im_2} into (25) (that is, $\delta_{i(m_1,m_2)} = h_{i(m_1,m_2)} - h_{im_1}h_{im_2}$). Finally, let $\chi(x) = 1$ if $x > 0$, $1/2$ if $x = 0$, and 0 if $x < 0$. We can then compute the following quantities:

1. $\hat{\mu}_i = \bar{h}_{im}$—the mean over loci of the observed proportion of homozygotes in population i.
2. $\hat{\sigma}_i^2 = \widehat{\mathrm{Var}}[h_{im}]$—the variance over loci of the observed proportion of homozygotes in population i.
3. $\hat{v}_i = \bar{h}_{i(m_1,m_2)}$—the mean over pairs of loci of the observed proportion of double homozygotes in population i.
4. $\widehat{\mathrm{Var}}[h_{i(m_1,m_2)}]$—the variance over pairs of loci of the observed proportion of double homozygotes in population i.
5. $\bar{\delta}_{i(m_1,m_2)}$—the mean over pairs of loci of the observed IIS excess in population i.
6. $\widehat{\mathrm{Var}}[\delta_{i(m_1,m_2)}]$—the variance over pairs of loci of the observed IIS excess in population i.
7. $\bar{\chi}(\delta_{i(m_1,m_2)})$—the fraction of pairs of loci with positive IIS excess (plus half the fraction with IIS excess of zero).
8. $\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$—the fraction of bootstrap resamples from the collection of values of $\delta_{i(m_1,m_2)}$ for which the mean of the resampled quantities is positive (plus half the fraction for which this mean equals zero).

For an unstructured population i, given $\hat{\mu}_i$ and assuming that loci are genotypically unassociated in the population, the predicted mean across pairs of loci of the proportion of double homozygotes is $\hat{\mu}_i^2$, and the predicted mean across pairs of the IIS excess, $\bar{\delta}_{i(m_1,m_2)}$, is 0. Although the distribution of $\delta_{i(m_1,m_2)}$ need not be symmetric, $\bar{\chi}(\delta_{i(m_1,m_2)})$ is predicted to be near $1/2$—slightly less than $1/2$ in most populations (Appendix B). This prediction, together with the prediction $\bar{\delta}_{i(m_1,m_2)} = 0$, suggests that $\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$ should not be near 1. Observations for the individual populations generally match these predictions (Table 2), with most populations having $|\bar{\delta}_{i(m_1,m_2)}| \lesssim 10^{-3}$, $\bar{\chi}(\delta_{i(m_1,m_2)}) \approx 1/2$, and $\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)}) < 0.95$ (but see Appendix B).

For a structured population S comprised of I unstructured populations, however, conditional on the values of $\hat{\mu}_i$ and f_i, the predicted mean proportion of double homozygotes is (assuming that loci are genotypically unassociated in each component population)

$$\hat{v}_i^* = \sum_{i=1}^{I} f_i \hat{\mu}_i^2. \tag{26}$$

Using (15) and (16) with $M = P = 2$, the mean predicted IIS excess, which by Theorem 4 is positive, equals

$$\bar{\delta}^*_{S(m_1,m_2)} = \sum_{i=1}^{I} f_i \hat{\mu}_i^2 - \left(\sum_{i=1}^{I} f_i \hat{\mu}_i \right)^2 \tag{27}$$

Table 1
Correlation of genotypic association with genetic distance

Population	Same chromosome		⩽ 10 cM	
	r	p-value	r	p-value
Bantu (Kenya)	−.019	.332	−.052	.489
Mandenka	.012	.494	.015	.820
Yoruba	−.018	.311	.066	.326
San	−.014	.506	.061	.453
Mbuti Pygmy	.012	.492	−.034	.625
Biaka Pygmy	−.005	.760	−.043	.519
Orcadian	.022	.224	−.014	.838
Adygei	.001	.956	−.144	.033
Russian	.019	.284	−.005	.942
Basque	−.025	.144	−.005	.942
French	.010	.549	.012	.858
Italian	.022	.215	−.041	.558
Sardinian	.009	.597	.012	.856
Tuscan	−.016	.400	−.095	.215
Mozabite	−.007	.687	.084	.209
Bedouin	.005	.787	−.038	.570
Druze	−.024	.156	.007	.919
Palestinian	−.003	.863	−.053	.424
Balochi	.018	.290	.017	.807
Brahui	−.007	.689	.002	.975
Makrani	−.023	.185	−.005	.938
Sindhi	−.014	.405	−.076	.258
Pathan	−.002	.893	−.011	.866
Burusho	−.001	.944	−.039	.554
Hazara	−.001	.965	.015	.824
Uygur	.047	.014	−.042	.571
Kalash	−.003	.861	.038	.569
Han	.015	.390	−.028	.670
Han (N. China)	−.032	.077	−.042	.550
Dai	.011	.561	−.033	.653
Daur	−.023	.216	.053	.463
Hezhen	−.025	.185	−.014	.855
Lahu	.010	.589	.061	.386
Miao	.018	.338	−.046	.522
Oroqen	−.002	.912	−.208	.005
She	.055	.003	−.095	.183
Tujia	.022	.230	.041	.554
Tu	.004	.835	−.033	.650
Xibo	.014	.469	−.017	.832
Yi	.016	.380	−.012	.869
Mongola	−.006	.753	.016	.829
Naxi	−.005	.786	.046	.526
Cambodian	.005	.768	.028	.692
Japanese	.010	.561	−.031	.640
Yakut	−.020	.252	.126	.059
Melanesian	.000	.977	.041	.543
Papuan	.007	.674	−.085	.203
Karitiana	−.065	.0002	−.197	.003
Surui	−.055	.002	−.069	.318
Colombian	−.002	.908	.013	.851
Maya	−.003	.865	.029	.659
Pima	−.026	.133	−.010	.884

For locus pairs on the same chromosome and those separated by at most 10 cM, r denotes the correlation coefficient between estimates of HR^2 and sex-averaged genetic distance (Weber and Broman, 2001). For two loci, 1 and 2, HR^2 (Sabatti and Risch, 2002) was estimated as $(h_{i(1,2)} - h_{i1}h_{i2})^2/[h_{i1}(1-h_{i1})h_{i2}(1-h_{i2})]$, where h_{i1}, h_{i2}, and $h_{i(1,2)}$, respectively, denote the count estimates of homozygosity at loci 1 and 2 and the count estimate of double homozygosity. The estimate was set to 1 if differing amounts of missing data for two loci in a pair led to a value above 1. Pairs with a value of zero for the denominator of the estimate were omitted from consideration. The p-values do not account for multiple comparisons. Populations are grouped by region (Rosenberg et al., 2002).

Table 2
Homozygosity and double homozygosity statistics for individual populations

Population	Sample size	$\hat{\mu}_i$	$\hat{\sigma}_i^2$	$\hat{\nu}_i$	$\widehat{\mathrm{Var}}[h_{i(m_1,m_2)}]$ ($\times 10^{-3}$)	$\bar{\delta}_{i(m_1,m_2)}$ ($\times 10^{-4}$)	$\widehat{\mathrm{Var}}[\delta_{i(m_1,m_2)}]$ ($\times 10^{-3}$)	$\bar{\chi}(\delta_{i(m_1,m_2)})$	$\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$
Bantu (Kenya)	12	.228	.024	.052	5.47	−1.04	2.39	.433	.293
Mandenka	24	.231	.012	.053	2.77	−1.21	1.34	.459	.201
Yoruba	25	.223	.013	.050	2.67	1.28	1.24	.462	.786
San	7	.239	.032	.057	8.77	−0.32	4.19	.428	.519
Mbuti Pygmy	15	.232	.019	.054	4.35	−1.72	2.03	.427	.165
Biaka Pygmy	36	.229	.012	.052	2.39	0.34	0.98	.478	.640
Orcadian	16	.252	.019	.064	4.86	1.68	2.09	.458	.801
Adygei	17	.251	.015	.063	4.21	0.12	2.08	.454	.491
Russian	25	.250	.014	.062	3.42	1.51	1.51	.480	.824
Basque	24	.265	.013	.070	3.71	0.06	1.75	.481	.563
French	29	.252	.012	.064	2.89	1.18	1.25	.478	.839
Italian	14	.260	.021	.067	6.11	−1.80	2.89	.443	.173
Sardinian	28	.255	.011	.065	3.01	−0.94	1.39	.477	.292
Tuscan	8	.242	.029	.058	9.38	−0.41	5.33	.412	.371
Mozabite	30	.246	.010	.061	2.55	3.39	1.24	.480	.991
Bedouin	49	.272	.009	.075	2.37	8.40	0.86	.501	1
Druze	48	.277	.009	.077	2.40	7.37	0.95	.502	1
Palestinian	51	.258	.009	.067	2.06	4.15	0.77	.491	.999
Balochi	25	.289	.014	.084	4.53	7.05	1.91	.493	1
Brahui	25	.271	.013	.074	3.83	2.60	1.69	.485	.953
Makrani	25	.276	.012	.078	3.57	15.50	1.63	.500	1
Sindhi	25	.270	.013	.074	3.64	16.10	1.63	.505	1
Pathan	25	.280	.013	.081	3.92	24.87	1.74	.519	1
Burusho	25	.254	.013	.065	3.28	1.67	1.38	.475	.884
Hazara	25	.261	.012	.068	3.31	1.95	1.51	.488	.895
Uygur	10	.254	.026	.065	7.55	5.65	3.56	.442	.999
Kalash	25	.286	.018	.082	4.90	0.82	1.62	.487	.731
Han	35	.284	.016	.081	3.91	−2.57	1.20	.478	.033
Han (N. China)	10	.280	.032	.078	9.70	−1.48	3.71	.431	.311
Dai	10	.280	.031	.078	9.69	2.90	3.89	.454	.882
Daur	10	.274	.029	.075	9.51	1.60	4.35	.441	.737
Hezhen	10	.288	.034	.083	11.57	1.26	4.59	.449	.686
Lahu	10	.299	.032	.089	10.47	−1.25	3.93	.454	.304
Miao	10	.290	.030	.084	9.69	−3.85	3.73	.450	.043
Oroqen	10	.282	.032	.079	9.93	−2.04	3.99	.448	.199
She	9	.294	.033	.087	10.94	2.51	4.23	.452	.817
Tujia	10	.297	.032	.088	10.98	−1.52	4.43	.450	.273
Tu	10	.279	.032	.078	9.60	2.39	3.57	.446	.828
Xibo	9	.262	.033	.069	9.18	3.80	3.63	.448	.933
Yi	10	.276	.030	.076	9.37	2.56	3.98	.443	.854
Mongola	10	.274	.030	.075	8.94	3.56	3.73	.448	.938
Naxi	10	.284	.032	.080	9.67	1.74	3.57	.443	.703
Cambodian	11	.261	.027	.068	7.64	0.52	3.22	.435	.613
Japanese	32	.289	.017	.083	4.43	−0.40	1.35	.487	.432
Yakut	25	.282	.015	.080	4.28	1.04	1.62	.494	.677
Melanesian	22	.321	.034	.103	10.49	1.00	2.55	.471	.654
Papuan	17	.326	.025	.106	8.77	−1.24	2.75	.480	.263
Karitiana	24	.405	.043	.164	17.60	9.38	2.09	.508	1
Surui	21	.464	.057	.215	30.36	−1.16	2.68	.501	.268
Colombian	13	.383	.043	.147	17.96	5.15	3.71	.500	.981
Maya	25	.313	.019	.099	5.82	5.85	1.82	.498	1
Pima	25	.367	.033	.135	11.84	2.75	1.99	.496	.931

Calculations are based on 375 loci and 66,730 pairs of loci (66,676 in Tuscan and 66,728 in Yi, after excluding pairs for which every individual was missing genotypes at one or both loci of the pair). $\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$ was obtained using 1000 bootstrap resamples.

Table 3
Homozygosity and double homozygosity statistics for example structured populations

(Structured) population	Sample size	$\hat{\mu}_i$	$\hat{\sigma}_i^2$	$\hat{v}_i$	$\widehat{\mathrm{Var}}[h_{i(m_1,m_2)}]$ $(\times 10^{-3})$	$\bar{\delta}_{i(m_1,m_2)}$ $(\times 10^{-4})$	$\widehat{\mathrm{Var}}[\delta_{i(m_1,m_2)}]$ $(\times 10^{-3})$	$\bar{\chi}(\delta_{i(m_1,m_2)})$	$\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$
World	1056	.279	.005	.080	0.81	22.41	0.05	.620	1
				.080		19.54			
Africa + Oceania	158	.252	.007	.065	1.16	14.63	0.27	.523	1
				.065		16.37			
America	108	.384	.018	.150	6.49	27.06	0.60	.541	1
				.150		25.73			
"Structured"	100	.297	.009	.096	2.66	75.37	0.63	.602	1
				.097		78.65			

Calculations are based on 375 loci and 66,730 pairs of loci. $\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$ was obtained using 1000 bootstrap resamples. For $\hat{v}_i$ and $\bar{\delta}_{i(m_1,m_2)}$, the predicted values based on (26) and (27) are given below the values observed in the data. From top to bottom, the four examples shown are structured populations comprised of (1) the entire data of 52 populations, (2) the individuals from Africa and Oceania with each region treated as a subpopulation, (3) the five populations from the Americas, and (4) four populations from separate continents (see Fig. 1).

(in (26) and (27), the asterisk is used to denote a predicted rather than observed value). The quantity $\bar{\chi}(\delta_{S(m_1,m_2)})$ is expected to be larger than 1/2, with $\bar{\chi}(\bar{\delta}^{boot}_{S(m_1,m_2)})$ close to 1.

As was true for the unstructured populations, observations for example structured populations also matched the predicted values (Table 3). In particular, the predicted surplus of double homozygotes was observed in the structured populations: unlike in the unstructured populations, mean IIS excess values were all positive. The structured populations generally had $\bar{\delta}_{S(m_1,m_2)}$ values larger than positive component IIS excesses by factors of 5–20, with $\bar{\chi}(\bar{\delta}^{boot}_{S(m_1,m_2)})$ values of 1. The distribution of $\delta_{S(m_1,m_2)}$ values across loci in structured populations was skewed to the right (Fig. 1), with $\bar{\chi}(\delta_{S(m_1,m_2)})$ noticeably greater than 1/2 (Table 3).

5. Conclusions

We have extended the Wahlund inequality to show that structured populations are expected to contain more M, P-homozygotes than corresponding unstructured populations. The extension enables definitions of multilocus identity coefficients analogous to F_{ST}. Moreover, the multilocus Wahlund inequality suggests that IIS excess statistics are expected to be positive in structured populations, as was observed in examples from human groups. Even in a species such as humans, in which individuals are fairly closely related, the two-locus Wahlund inequality generates a noticeable excess of double homozygotes. Positively biased IIS excess is also a property of closely linked loci (Hedrick and Thomson, 1986; Ohta, 2000; Vitalis and Couvet, 2001); thus, similarly to the multiple potential interpretations of other association measures, positive IIS excess need not be viewed as evidence of linkage when population structure might provide an alternate explanation.

Appendix A

Here we derive (15)–(19). Using the independence of the H_{im}, the expectation of a product of two or more of these random variables equals the product of the expectations. Results (17)–(19) follow from $\mathrm{Var}[H_S] = \mathbb{E}[H_S^2] - \mathbb{E}[H_S]^2$, $\mathrm{Var}[H_V] = \mathbb{E}[H_V^2] - \mathbb{E}[H_V]^2$, and $\mathrm{Cov}[H_S, H_V] = \mathbb{E}[H_S H_V] - \mathbb{E}[H_S]\mathbb{E}[H_V]$.

$$\begin{aligned}\mathbb{E}[H_S] &= \mathbb{E}\left[\sum_{i=1}^{I} f_i \prod_{m=1}^{M} H_{im}\right] \\ &= \sum_{i=1}^{I} f_i \mathbb{E}\left(\prod_{m=1}^{M} H_{im}\right) \\ &= \sum_{i=1}^{I} f_i \mu_i^M, \end{aligned} \tag{28}$$

$$\begin{aligned}\mathbb{E}[H_V] &= \mathbb{E}\left[\prod_{m=1}^{M} \sum_{i=1}^{I} f_i H_{im}\right] \\ &= \prod_{m=1}^{M} \mathbb{E}\left(\sum_{i=1}^{I} f_i H_{im}\right) \\ &= \left(\sum_{i=1}^{I} f_i \mu_i\right)^M = \mu^M, \end{aligned} \tag{29}$$

$$\begin{aligned}\mathbb{E}[H_S^2] &= \mathbb{E}\left[\left(\sum_{i=1}^{I} f_i \prod_{m=1}^{M} H_{im}\right)^2\right] \\ &= \mathbb{E}\left[\left(\sum_{i=1}^{I} f_i^2 \prod_{m=1}^{M} H_{im}^2\right)\right. \\ &\qquad \left. + \left(\sum_{i=1}^{I} \sum_{\substack{l=1 \\ l\neq i}}^{I} f_i f_l \prod_{m=1}^{M} H_{im} H_{lm}\right)\right] \end{aligned}$$

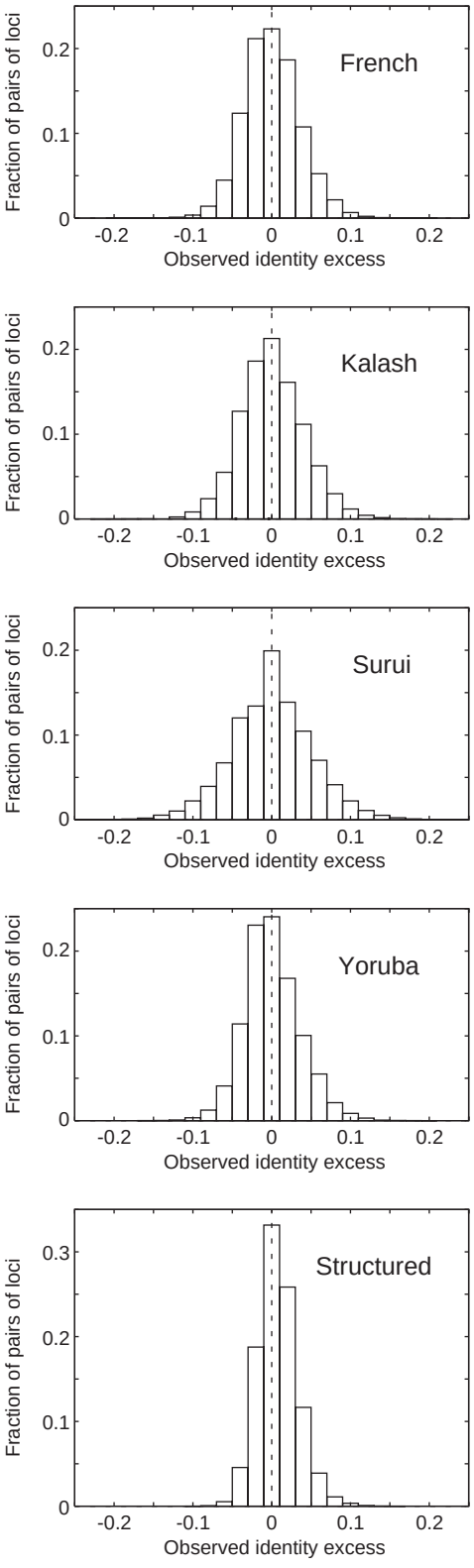

Fig. 1. Frequency distribution of the IIS excess estimate $\delta_{i(m_1,m_2)}$ across 66,730 pairs of loci. For Surui, 20 IIS excess values fell outside of $[-0.25, 0.25]$, largely as a result of a particularly small sample size at one specific locus (in general, sample size is the major determinant of the variability across pairs in IIS excess).

$$= \sum_{i=1}^{I} f_i^2(\mu_i^2+\sigma_i^2)^M + \sum_{i=1}^{I}\sum_{\substack{l=1\\ l\neq i}}^{I} f_i f_l \mu_i^M \mu_l^M$$
$$= \sum_{i=1}^{I} f_i^2[(\mu_i^2+\sigma_i^2)^M - (\mu_i^2)^M] + \left(\sum_{i=1}^{I} f_i \mu_i^M\right)^2, \quad (30)$$

$$\mathbb{E}[H_V^2] = \mathbb{E}\left[\prod_{m=1}^{M}\left(\sum_{i=1}^{I} f_i H_{im} \sum_{l=1}^{I} f_l H_{lm}\right)\right]$$
$$= \mathbb{E}\left[\prod_{m=1}^{M}\left(\sum_{i=1}^{I} f_i^2 H_{im}^2 + \sum_{i=1}^{I}\sum_{\substack{l=1\\ l\neq i}}^{I} f_i f_l H_{im} H_{lm}\right)\right]$$
$$= \left[\sum_{i=1}^{I} f_i^2(\mu_i^2+\sigma_i^2) + \sum_{i=1}^{I}\sum_{\substack{l=1\\ l\neq i}}^{I} f_i f_l \mu_i \mu_l\right]^M$$
$$= \left[\sum_{i=1}^{I} f_i^2\sigma_i^2 + \left(\sum_{i=1}^{I} f_i\mu_i\right)^2\right]^M, \quad (31)$$

$$\mathbb{E}[H_S H_V] = \mathbb{E}\left[\left(\sum_{i=1}^{I} f_i \prod_{m=1}^{M} H_{im}\right)\left(\prod_{n=1}^{M}\sum_{l=1}^{I} f_l H_{ln}\right)\right]$$
$$= \sum_{i=1}^{I} f_i \mathbb{E}\left[\prod_{m=1}^{M} H_{im} \prod_{n=1}^{M}\left(f_i H_{in} + \sum_{\substack{l=1\\ l\neq i}}^{I} f_l H_{ln}\right)\right]$$
$$= \sum_{i=1}^{I} f_i \sum_{k=0}^{M}\binom{M}{k}\left(\mu_i \sum_{\substack{l=1\\ l\neq i}}^{I} f_l\mu_l\right)^k \times [f_i(\mu_i^2+\sigma_i^2)]^{M-k}$$
$$= \sum_{i=1}^{I} f_i(f_i\sigma_i^2 + \mu_i\mu)^M. \quad (32)$$

Appendix B

Suppose population i is unstructured and that loci 1 and 2 are genotypically unassociated in population i. Consider a sample of size n_i. If loci 1 and 2 have true homozygosities H_{i1} and H_{i2}, the probability that the sample includes l_1 homozygotes at locus 1, l_2

homozygotes at locus 2, and l_{12} double homozygotes ($0 \leqslant l_{12} \leqslant \min(l_1, l_2) \leqslant n_i$), is

$$R_{l_1 l_2 l_{12}} = \binom{n_i}{l_1} H_{i1}^{l_1} (1 - H_{i1})^{n_i - l_1} \times \binom{n_i}{l_2} H_{i2}^{l_2} (1 - H_{i2})^{n_i - l_2} \frac{\binom{l_1}{l_{12}} \binom{n_i - l_1}{l_2 - l_{12}}}{\binom{n_i}{l_2}}.$$

The ratio of binomial coefficients gives the probability that the l_1 homozygotes at locus 1 and the l_2 homozygotes at locus 2 overlap in exactly l_{12} individuals.

Because $\delta_{i(1,2)} = l_{12}/n_i - (l_1/n_i)(l_2/n_i)$, $\delta_{i(1,2)} > 0$ if $l_{12} > l_1 l_2 / n_i$. Thus, taking into account all possible sample configurations (l_1, l_2, l_{12}), given n_i, H_{i1}, and H_{i2},

$$\mathbb{E}[\chi(\delta_{i(1,2)})] = \sum_{l_1=0}^{n_i} \sum_{l_2=0}^{n_i} \left[-\frac{1}{2} R_{l_1 l_2 \gamma} + \sum_{l_{12}=\lceil \gamma \rceil}^{\min(l_1, l_2)} R_{l_1 l_2 l_{12}} \right], \quad (33)$$

where $\gamma = l_1 l_2 / n_i$. In the summation, $R_{l_1 l_2 \gamma}$ is set to zero if γ is not an integer.

For small sample sizes, with $H_{i1}, H_{i2} \in [0, 0.5]$, as was true of most locus pairs in most populations (Table 4), $\mathbb{E}[\chi(\delta_{i(1,2)})]$ was usually in $[0.4, 0.5]$ (Fig. 2). Thus, (33) predicts that in unstructured populations with the homozygosities and sample sizes typical of the Rosenberg et al. (2002) data, the proportion of pairs of genotypically unassociated loci with positive estimated IIS excess will be slightly smaller than 1/2. This prediction was generally satisfied (Table 2). In the instances when it was not satisfied, three main factors were responsible.

First, as the fraction of locus pairs with one homozygosity in $[0, 0.5]$ and the other in $(0.5, 1]$ increases, the effect of sampling is to inflate $\mathbb{E}[\chi(\delta_{i(1,2)})]$ (Fig. 2). Thus, Surui, Karitiana, and Colombian, comparatively homozygous Native American populations with relatively large numbers of such pairs (Table 4), were among the groups with the highest values of $\bar{\chi}(\delta_{i(m_1,m_2)})$.

Second, as sample size increases, the effect of sampling on $\mathbb{E}[\chi(\delta_{i(1,2)})]$ is reduced, as can be seen by comparing at different sample sizes the proportion of possible values (H_{i1}, H_{i2}) for which $\mathbb{E}[\chi(\delta_{i(1,2)})] \notin [0.49, 0.51]$ (Fig. 3). Thus, populations with larger samples tended to produce values of $\bar{\chi}(\delta_{i(m_1,m_2)})$ nearer 1/2, with the correlation coefficient between sample size and $|\bar{\chi}(\delta_{i(m_1,m_2)}) - 1/2|$ equaling -0.747 ($p < 10^{-4}$).

Finally, population structure inflates values of the IIS excess, so that structured populations are likely to have more locus pairs with $\delta_{i(m_1,m_2)} > 0$. Thus, populations of the Middle East and Central/South Asia with noticeable levels of population structure, as reflected in heterogeneous individual ancestry (Rosenberg et al., 2002,

Table 4
Fractions of locus pairs with zero, one, and two of the two loci having estimated homozygosity in (0.5, 1]

Population	0	1	2
Bantu (Kenya)	.916	.082	.002
Mandenka	.963	.037	<.001
Yoruba	.963	.037	<.001
San	.866	.130	.005
Mbuti Pygmy	.942	.057	<.001
Biaka Pygmy	.974	.026	<.001
Orcadian	.937	.062	<.001
Adygei	.953	.047	<.001
Russian	.947	.052	<.001
Basque	.942	.057	<.001
French	.958	.042	<.001
Italian	.891	.106	.003
Sardinian	.979	.021	<.001
Tuscan	.916	.082	.002
Mozabite	.979	.021	<.001
Bedouin	.963	.037	<.001
Druze	.973	.027	<.001
Palestinian	.979	.021	<.001
Balochi	.910	.088	.002
Brahui	.953	.047	<.001
Makrani	.932	.067	.001
Sindhi	.948	.052	<.001
Pathan	.927	.072	.001
Burusho	.948	.052	<.001
Hazara	.947	.052	<.001
Uygur	.927	.072	.001
Kalash	.866	.129	.005
Han	.901	.096	.002
Han (N. China)	.832	.161	.007
Dai	.861	.134	.005
Daur	.891	.106	.003
Hezhen	.813	.178	.009
Lahu	.813	.178	.010
Miao	.861	.134	.005
Oroqen	.846	.148	.006
She	.779	.208	.013
Tujia	.808	.182	.010
Tu	.831	.161	.008
Xibo	.827	.166	.008
Yi	.847	.147	.006
Mongola	.881	.115	.003
Naxi	.817	.174	.009
Cambodian	.836	.157	.007
Japanese	.875	.121	.004
Yakut	.932	.067	.001
Melanesian	.714	.262	.024
Papuan	.779	.208	.013
Karitiana	.530	.397	.073
Surui	.382	.473	.145
Colombian	.602	.348	.050
Maya	.847	.147	.006
Pima	.683	.287	.030

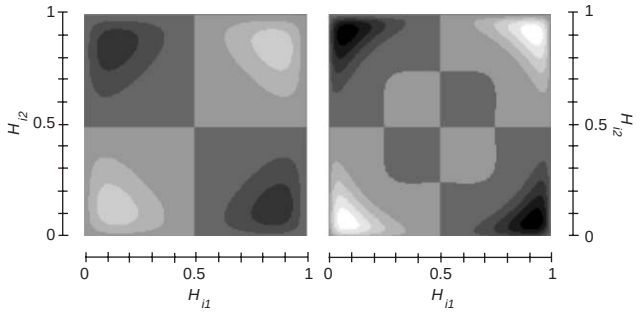

Fig. 2. $\mathbb{E}[\chi(\delta_{i(1,2)})]$, computed from (33) with $n_i = 10$ (left), $n_i = 25$ (right). From lightest to darkest, the shades represent values of $\mathbb{E}[\chi(\delta_{i(1,2)})]$ in $[0, 0.3)$, $[0.3, 0.35)$, $[0.35, 0.4)$, $[0.4, 0.45)$, $[0.45, 0.5)$, $\{0.5\}$, $(0.5, 0.55]$, $(0.55, 0.6]$, $(0.6, 0.65]$, $(0.65, 0.7]$, and $(0.7, 1]$. The shades corresponding to $[0.45, 0.5)$ and $(0.5, 0.55]$ occupy most of the area in both plots.

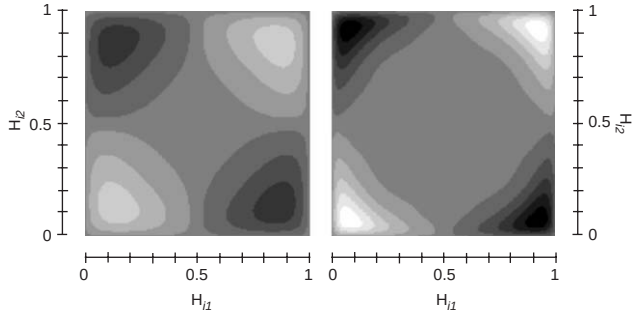

Fig. 3. $\mathbb{E}[\chi(\delta_{i(1,2)})]$, computed from (33) with $n_i = 10$ (left), $n_i = 25$ (right). This figure is based on exactly the same values of $\mathbb{E}[\chi(\delta_{i(1,2)})]$ as Fig. 2, the only difference being that the central shade, covering a substantial portion of both the left and right plots, represents $[0.49, 0.51]$ instead of $\{0.5\}$ (also, its neighboring shades represent $[0.45, 0.49)$ and $(0.51, 0.55]$ instead of $[0.45, 0.5)$ and $(0.5, 0.55]$).

Fig. 2) were among those with the highest values of $\bar{\chi}(\delta_{i(m_1,m_2)})$. In these groups, population structure also led to large values of $\bar{\delta}_{i(m_1,m_2)}$ and $\bar{\chi}(\bar{\delta}^{boot}_{i(m_1,m_2)})$. If other influences on IIS excess can be ruled out, these observations suggest the possibility of using properties of the IIS excess between unlinked loci as test statistics for the hypothesis that a population is unstructured.

Acknowledgements

We are grateful to Marc Feldman, Magnus Nordborg, Bob Sacker, and the late Ryk Ward, whose suggestions led to substantial improvements in the manuscript. This research was supported by NSF Postdoctoral Fellowships in Biological Informatics and Mathematical Sciences to N.A.R. and P.P.C., respectively, by a Burroughs Wellcome Fund Career Award in the Biomedical Sciences to N.A.R., and by Center of Excellence in Genomic Science grant 1P50HG002790-01A1 from the National Human Genome Research Institute.

References

Balding, D.J., 2003. Likelihood-based inference for genetic correlation coefficients. Theor. Pop. Biol. 63, 221–230.

Beckenbach, E., Bellman, R., 1961. An Introduction to Inequalities. Random House, New York.

Christiansen, F.B., 1988. The Wahlund effect with overlapping generations. Am. Nat. 131, 149–156.

Crow, J.F., Kimura, M., 1970. An Introduction to Population Genetics Theory. Harper & Row, New York.

Excoffier, L., 2001. Analysis of population subdivision. In: Balding, D.J., Bishop, M., Cannings, C. (Eds.), Handbook of Statistical Genetics. Wiley, Chichester, UK, pp. 271–307 (chapter 10).

Feldman, M.W., Christiansen, F.B., 1975. The effect of population subdivision on two loci without selection. Genet. Res. 24, 151–162.

Hedrick, P.W., Thomson, G., 1986. A two-locus neutrality test: applications to humans, *E. coli* and lodgepole pine. Genetics 112, 135–156.

Nei, M., Li, W.-H., 1973. Linkage disequilibrium in subdivided populations. Genetics 75, 213–219.

Nordborg, M., Tavaré, S., 2002. Linkage disequilibrium: what history has to tell us. Trends Genet. 18, 83–90.

Ohta, T., 1980. Linkage disequilibrium between amino acid sites in immunoglobulin genes and other multigene families. Genet. Res. 36, 181–197.

Ohta, T., 1982. Linkage disequilibrium due to random genetic drift in finite subdivided populations. Proc. Natl. Acad. Sci. USA 79, 1940–1944.

Ohta, T., 2000. An attempt to measure the patchwork pattern observed among alleles at major histocompatibility complex loci. J. Mol. Evol. 51, 21–25.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., Feldman, M.W., 2002. Genetic structure of human populations. Science 298, 2381–2385.

Rousset, F., 2002. Inbreeding and relatedness coefficients: what do they measure? Heredity 88, 371–380.

Sabatti, C., Risch, N., 2002. Homozygosity and linkage disequilibrium. Genetics 160, 1707–1719.

Sinnock, P., 1975. The Wahlund effect for the two-locus model. Am. Nat. 109, 565–570.

Vitalis, R., Couvet, D., 2001. Two-locus identity probabilities and identity disequilibrium in a partially selfing subdivided population. Genet. Res. 77, 67–81.

Wahlund, S., 1928. Zusammensetzung von populationen und korrelationerscheinungen vom standpunkt der vererbungslehre aus betrachtet. Hereditas 11, 65–106.

Weber, J.L., Broman, K.W., 2001. Genotyping for human whole genome scans: past, present, and future. Adv. Genet. 42, 77–96.

Weir, B.S., Hill, W.G., 2002. Estimating *F*-statistics. Annu. Rev. Genet. 36, 721–750.

Wright, S., 1951. The genetical structure of populations. Ann. Eugen. 15, 323–354.