Annals of human genetics

doi: 10.1111/j.1469-1809.2010.00571.x

Evaluating differences in linkage disequilibrium between populations

Birgir Hrafnkelsson[1,2*], Agnar Helgason[1], Gudbjorn F. Jonsson[1], Daniel F. Gudbjartsson[1], Thorlakur Jonsson[1], Sverrir Thorvaldsson[1], Hreinn Stefansson[1], Valgerdur Steinthorsdottir[1], Nanna Vidarsdottir[1], Derek Middleton[3], Henning S. Petersen[4], Conrado Martinez[5,6], Jon Snaedal[1], Palmi V. Jonsson[1], Sigurbjorn Bjornsson[1], Jeffrey R. Gulcher[1] and Kari Stefansson[1]

[1] *deCODE genetics, Sturlugata 8, 101 Reykjavik, Iceland*
[2] *Division of Applied Mathematics, Science Institute, University of Iceland, Dunhagi 5, 107 Reykjavik, Iceland*
[3] *Northern Ireland Histocompatibility and Immunogenetics Laboratory City Hospital, Belfast BT9 7TS, UK*
[4] *Primary Health Care Clinic, 3900 Nuuk, Greenland*
[5] *Castellon Province Hospital Foundation, Ave. Doctor Clara 19, 12002 Castellon, Spain*
[6] *Institute of Biological Anthropology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK*

Summary

We propose two methods to evaluate the statistical significance of differences in linkage disequilibrium (LD) between populations, where LD is measured by the standardised parameter D'. The first method is based on bootstrapping individuals within populations in order to test LD differences for each pair of loci. Using this approach we propose a solution to the problem of testing multiple locus-pairs by means of a single test for the number of pairs that exhibit significant LD differences among populations. The second method provides the Bayesian posterior probability that one population has greater LD than the other for each locus pair. Both methods can handle genotypes with unknown phase, and are demonstrated using two data sets. For the purpose of demonstration, we apply the methods to two different sets of data from humans. First, we explore the issue of LD differences between reproductively isolated populations using a new data set of twelve Xq25 microsatellites, typed in four European populations. Second, we examine evidence for LD differences between Alzheimer cases and controls from the Icelandic population using 19 single nucleotide polymorphisms (SNPs) from a 97 kb region flanking the Apolipoprotein E (*APOE*) gene on chromosome 19.

Keywords: Alzheimer, Apolipoprotein E gene, Bayesian inference, bootstrap based inference, the D' parameter as a measure of linkage disequilibrium

Introduction

Following the publication of the human genome sequence and the rapid growth in genotyping capacity, research in human genetics is increasingly focused on the issue of linkage disequilibrium (LD) - in particular, how the correlation between alleles at different loci in a population can be exploited to pinpoint genetic variants underlying the risk of disease phenotypes. Theory suggests that the extent of LD at a given region in the genome will vary among human populations dependent on the relative influence of demographic factors such as genetic drift, natural selection and admixture (Slatkin, 1994; Terwilliger et al., 1998; Ardlie et al., 2002; Nordborg & Tavare, 2002; Wall & Pritchard, 2003). Considerable attention has been given to the question of whether small, isolated populations (so-called genetic isolates) have higher background levels of LD than large outbred and heterogeneous populations (Wright et al., 1999; Shifman & Darvasi, 2001; Heutink & Oostra, 2002).

A number of studies have examined small segments of the genome with the aim of assessing LD differences between populations with varying demographic histories. While some studies have been interpreted as indicating greater LD in small isolated populations (Laan & Pääbo, 1997; Shifman &

*Corresponding author: Birgir Hrafnkelsson, Division of Applied Mathematics, Science Institute, University of Iceland, Dunhagi 5, 107 Reykjavik, Iceland. Tel: +354-5254669; Fax: +354-5254632; E-mail: birgirhr@hi.is

Darvasi, 2001; Kaessmann et al., 2002; Katoh et al., 2002; Latini et al., 2004; Laan et al., 2005; Sawyer et al., 2005), others indicate that no such differences exist (Dunning et al., 2000; Eaves et al., 2000; Taillon-Miller et al., 2000). Although this disagreement may partly be explained by the populations, genomic regions and types of loci used in these studies, it is also attributable to the lack of a satisfactory statistical method for comparing levels of LD in different populations. Conclusions in most previous studies were based on informal comparisons of D' or r^2 values between populations or of P-values obtained from the application of the Markov Chain Monte Carlo (MCMC) adaptation of the Fisher-Exact test (Guo & Thompson, 1992). More conclusive answers to questions about LD differences between populations and the level of background LD in genetic isolates requires more formal statistical methods that take into account the sampling distribution of the statistics used to measure LD.

Important contributions to the development of such statistical tests have been made by Ayres & Balding (2001) in a Bayesian framework and Zapata et al. (2001) in a frequentist framework. In this paper, we build on these contributions to develop two statistical procedures for making LD comparisons, both of which can handle genotype data with unknown phase for loci with any number of alleles.

The first procedure is based on bootstrapping individuals within each population to obtain a P-value for the null hypothesis that the difference between the D' values of two populations for a given locus pair, $\Delta D'$, is zero, versus the alternative hypothesis that $\Delta D' > 0$. To obtain a single P-value for the comparison of D' values across multiple locus pairs, and thereby sidestep the problem of multiple testing, we propose a single test that uses the joint null distribution of $\Delta D'$ values for multiple pairs to provide the tail probability for the number with significantly different D' values. The rationale behind this procedure is similar to that of Zaykin et al. (2006), who proposed a single statistical test that simultaneously tests differences in standardised composite coefficients (Hamilton & Cole, 2004; Zaykin, 2004). Wang et al. (2007) propose a LD contrast test based on a regression type model that takes into account background LD. However, while the approaches of Zaykin et al. (2006) and Wang et al. (2007) are only designed to handle SNPs, ours can handle both SNPs and microsatellites.

The second procedure expands on the Bayesian approach proposed by Ayres & Balding (2001), where posterior distributions for D' are generated for each locus pair using an MCMC algorithm. These distributions are then compared for each pair of populations to yield posterior probabilities corresponding to one population having a higher value of D' than another population for each locus pair.

We demonstrate these two LD comparison procedures using two previously unpublished data sets. The first consists of twelve X-chromosome microsatellite loci, typed in males from four European populations. The second data set contains genotypes, with unknown phase, from a set of Alzheimer patients and controls from Iceland, typed for 19 SNPs from a 97 kb region on chromosome 19 containing the Apolipoprotein E (*APOE*) gene. Numerous studies have demonstrated a strong association between Alzheimer disease and variation in this gene (Strittmatter et al., 1993; Martin et al., 2000), and the Icelandic data are no exception. We explore the evidence that the different configuration of haplotypes in patients and controls results in significantly different patterns of LD.

Materials and Methods

The D' Parameter of Gametic Disequilibrium

The D' parameter, initially proposed by Lewontin (1964), was defined by Hedrick (1987) for two multiallelic loci as follows. Let A_k be the k-th allele of the first locus, $k = 1, \ldots, K$, and let B_l be the l-th allele of the second locus, $l = 1, \ldots, L$. Denote the frequency of gamete A_kB_l by p_{kl}, the frequency of allele A_k by $p_{k.}(p_{k.} = \sum_{l=1}^{L} p_{kl})$ and the frequency of allele B_l by $p_{.l}(p_{.l} = \sum_{k=1}^{K} p_{kl})$. Let $\pi = (p_{kl})_{k,l}$ denote the vector containing the gamete frequencies. A standardised measure for gametic disequilibrium between alleles A_k and B_l can be defined as $D'_{kl} = D_{kl}/D_{\max}$, where $D_{kl} = p_{kl} - p_{k.}p_{.l}$, and $D_{\max} = \min\{p_{k.}p_{.l}, (1 - p_{k.})(1 - p_{.l})\}$ when $D_{kl} < 0$ or $D_{\max} = \min\{p_{k.}(1 - p_{.l}), (1 - p_{k.})p_{.l}\}$ when $D_{kl} > 0$. If $p_{kl} = 0$ and/or $(1 - p_{k.} - p_{.l} + p_{kl}) = 0$ then $D'_{kl} = -1$ and if $(p_{k.} - p_{kl}) = 0$ and/or $(p_{.l} - p_{kl}) = 0$ then $D'_{kl} = 1$. A weighted measure for the disequilibrium between all the alleles at the two loci can be defined as

$$D' = \sum_{k=1}^{K} \sum_{l=1}^{L} p_{k.}p_{.l}|D'_{kl}|. \quad (1)$$

The D' parameter thus defined has a minimum of 0 and maximum equal to or very close to 1, the latter depending on allele frequencies and the number of alleles, see Zapata (2000). Robinson et al. (1991) showed that higher-order systems impose additional constraints on bounds of D_{kl} and thus on D'.

A Test for Differences in D' Based on the Bootstrap

The test procedure introduced here is based on the bootstrap approach which involves randomly sampling individuals with replacement within populations (Efron & Tibshirani, 1993). The aim is to test the null hypothesis of no difference in D' between the two populations ($\Delta D' = 0$), against the alternative hypothesis that $\Delta D' > 0$.

The D' parameter is estimated by plugging the maximum likelihood estimator (MLE) for π into the definition of D' as given in (1) when the phase of alleles is known (e.g., Zapata et al.,

2001). When phase is unknown the expectation-maximization (EM) algorithm is used to first estimate haplotype frequencies for each locus pair (Excoffier & Slatkin, 1995). The test is not designed to handle missing data but could be extended to do so. Although other phase algorithms could be used, such as the PHASE algorithm (Stephens et al., 2001), the EM algorithm has the advantage of being both simple and quick in the case of two loci.

We define $\Delta\hat{D}'_{\text{obs}}$ as the difference between the observed D' values of two populations for a given locus pair computed using the MLE, such that,

$$\Delta\hat{D}'_{\text{obs}} = \hat{D}'_{2,\text{obs}} - \hat{D}'_{1,\text{obs}},$$

with the subscripts 1 and 2 denoting the two populations. For each locus pair, we compute point estimates of $\Delta D' = D'_2 - D'_1$, for each of the B bootstrapped data sets within each population. Note that in this bootstrap algorithm it is the individuals (along with their genotypes at all the loci being examined) that are resampled. Further, for each bootstrapped data set $\Delta\hat{D}'$ is computed for all pairs of loci to capture the correlation between the $\Delta\hat{D}'$ estimators for all pairs. The point estimates of D'_1, D'_2 and $\Delta D'$ for the b-th bootstrapped data set are denoted by $\hat{D}'_{1b}$, $\hat{D}'_{2b}$ and $\Delta\hat{D}'_b$, respectively, where $\Delta\hat{D}'_b = \hat{D}'_{2b} - \hat{D}'_{1b}$, $b = 1, \ldots, B$. One problem that arises in this approach is that the distribution of $\Delta\hat{D}'_b$ values is not centred around zero in the accordance with the sampling distribution of $\Delta\hat{D}'$ under the null hypothesis, where $\Delta D' = 0$, albeit their variances are similar. A standard solution in the bootstrap literature (Efron & Tibshirani, 1993) is to make the $\Delta\hat{D}'_b$ values mean zero by subtracting their mean which is denoted by $\overline{\Delta\hat{D}'}$, and compare these values to the bias corrected $\Delta\hat{D}'_{\text{obs}}$ given by

$$\Delta\hat{D}'_{\text{obs,unb}} = \Delta\hat{D}'_{\text{obs}} - \widehat{\text{bias}}_{\Delta D}$$

where

$$\widehat{\text{bias}}_{\Delta D} = \overline{\Delta\hat{D}'} - \Delta\hat{D}'_{\text{obs}}.$$

A P-value is approximated by finding the proportion of values such that $\Delta\hat{D}'_b - \overline{\Delta\hat{D}'} > \Delta\hat{D}'_{\text{obs,unb}}$ which is the same as finding the proportion of values such that $\Delta\hat{D}'_b - \Delta\hat{D}'_{\text{obs}} > \Delta\hat{D}'_{\text{obs}}$. However, since $\Delta\hat{D}'_b$ is in the interval $[-1, 1]$, this subtraction causes trouble when $\Delta\hat{D}'_{\text{obs}} > 0.5$, in which case $\Delta\hat{D}'_b - \Delta\hat{D}'_{\text{obs}} > \Delta\hat{D}'_{\text{obs}}$ is never true. We overcome this problem by using twice the Fisher transformation (Fisher, 1915)

$$h(\Delta D') = \log(1 + \Delta D') - \log(1 - \Delta D'),$$

and subtracting $h(\Delta\hat{D}'_{\text{obs}})$ from the $h(\Delta\hat{D}'_b)$ values. In this scheme the proportion of values such that $h(\Delta\hat{D}'_b) - h(\Delta\hat{D}'_{\text{obs}}) > h(\Delta\hat{D}'_{\text{obs}})$ provides the P-value. This P-value can be presented in a computional formula as

$$\hat{P}\text{-value} = \frac{1}{B}\sum_{b=1}^{B} I\left\{h(\Delta\hat{D}'_b) > 2h(\Delta\hat{D}'_{\text{obs}})\right\}$$

where I is an indicator function such that $I(A) = 1$ if A is true and $I(A) = 0$ otherwise, and in the case of $\Delta D' = \pm 1$ then set $h(\Delta D') = \pm 10^6$.

The proposed bootstrap test evaluates the statistical significance of LD differences between two groups for a single locus pair. Thus, when multiple locus pair are tested, it is necessary to correct for the number of locus pairs. As most standard correction methods, such as Bonferroni, tend to be overly conservative, we propose a single test for all locus pairs that is performed as follows:

1. Let $M(\alpha)$ be the number of locus pairs with a bootstrap P-value less than α and denote its observed value by $M_{\text{obs}}(\alpha)$, where α is a common threshold of significance for all the tests.
2. For each locus pair find the $100(1-\alpha)\%$ percentile of the $\Delta\hat{D}'_b$ values and then count the number of locus pairs in each bootstrap data set that exceed their $100(1-\alpha)\%$ percentile. The distribution of these counts approximates the null distribution of $M(\alpha)$, denoted by $M_0(\alpha)$.
3. An overall P-value corresponding to the test for the alternative hypothesis that the two populations have significantly different D' values for at least one locus pair is given by $P(M_0(\alpha) \geq M_{\text{obs}}(\alpha))$.

The above test takes into account the correlation between the estimators of D' at all the pairs of loci.

A Bayesian Test for Differences in D'

The Bayesian approach for D' developed by Ayres & Balding (2001) yields a posterior distribution of D', given a model for the data and a prior distribution for the haplotype frequencies. As demonstrated by Ayres & Balding (2001), the posterior distribution of D' provides the basis for statistical evaluations of LD differences between populations. We have adopted this approach with some modification to the MCMC algorithm, which in our version uses a Gibbs sampler without a Metropolis–Hastings step, obviating the need to tune proposal densities and making the generation of posterior samples of haplotype frequencies automatic. Moreover, we propose a different prior that yields results that are less sensitive to the number of alleles per locus than that proposed by Ayres & Balding (2001). Here we describe the Bayesian estimation procedure when phase of alleles is known and haplotypes with missing alleles are ignored. Additional procedures to deal with unknown phase and missing data are presented in the Appendix.

Let $C_{kl,v}$ be the count of haplotypes for gamete A_kB_l in population v, $v = 1, 2$. Let $C_v = (C_{kl,v})_{k,l}$ be a $(K \times L)$ vector of these counts for population v (as before, K and L are the numbers of distinct alleles at the first locus and second locus, respectively, and the corresponding gamete frequencies are $\pi_v = (p_{kl,v})_{k,l}$). The total count of haplotypes in C_v is denoted by m_v. Assume *a priori* that

$$\pi_v \sim \text{Dir}_{KL}(\beta), \quad v = 1, 2$$

where $\mathrm{Dir}_J(\beta')$ denotes a Dirichlet distribution on a $(J-1)$ dimensional simplex with parameter vector β' (e.g., Johnson & Kotz, 1972). The model could be extended to have a prior distribution on β and then have the joint posterior distribution of β and π_v evaluated, but here it will be assumed that the vector β is determined beforehand. The selection of β is discussed in the section *Evaluation of prior distributions for gamete frequencies*. A statistical model that describes the data C_v is as follows

$$C_v|\pi_v \sim \mathrm{Mult}_{KL}(m_v, \pi_v)$$

where $\mathrm{Mult}_J(n, s)$ denotes a J dimensional multinomial distribution with n trials and a probability vector s. So, the data, C_v, given the haplotype frequencies, π_v, follow a multinomial distribution while the prior distribution of π_v is a Dirichlet distribution. The resulting posterior distribution of π_v given C_v is the following Dirichlet distribution

$$\pi_v|C_v \sim \mathrm{Dir}_{KL}(C_v + \beta),$$

thus, the posterior mean of π_v is given by

$$\mathrm{E}(\pi_v|C_v) = \frac{C_v + \beta}{m_v + \sum_{j=1}^{KL} \beta_j}.$$

The posterior distribution of D'_v is then found by drawing samples from the posterior distribution of π_v and computing D'_v with (1) for each of the sampled π_v's, $v = 1, 2$. The posterior probability that Population 2 has a greater D' value than Population 1 is given by $P(D'_2 > D'_1|C_1, C_2)$ and can be computed using samples from the posterior distributions of D'_1 and D'_2. This estimated value of $P(D'_2 > D'_1|C_1, C_2)$ is then used to evaluate the hypothesis $D'_2 > D'_1$, which might for example be accepted if $P(D'_2 > D'_1|C_1, C_2) > 0.95$.

Different Assumptions Underlying the two LD Comparison Procedures

Although both the bootstrap procedure and the Bayesian procedure described above are designed to tackle the same problem, the interpretation of the bootstrap P-values and the Bayesian posterior probability values is conceptually different. The bootstrap P-values state how likely the observed D' differences are to occur relative to their sampling distribution, when it is assumed that there is no difference in D' between the populations. In this test no assumptions are made about the sampling distribution of D'. In contrast, the Bayesian approach describes the uncertainty of the statement that one population has a greater D' value than another population based on a multinomial model, the Dirichlet prior and the observed haplotype frequencies within each population. This uncertainty is quantified with a probability measure that does not reflect sampling frequencies, but can be described as a subjective prior probability measure that is updated after seeing the data.

Two Sets of Genotype Data

Two data sets are used to illustrate the proposed LD comparison procedures. The first data set consists of twelve dinucleotide repeat microsatellite loci, see Table S1 in supporting information, spanning 1.5 Mb of X-chromosome region Xq25, and typed in males from four European populations: 114 Greenlanders, 141 Icelanders, 113 Northern Irish and 136 Spanish. In these data the phase of alleles is known because males are haploid for these loci.

The second data set consists of 19 SNPs spanning just over 97kb from the 19q13.2 region on chromosome 19, which includes the Apolipoprotein E (*APOE*) gene, see Table S2 in supporting information. The SNPs were typed in 163 Icelandic Alzheimer patients and 150 controls, none of whom were related within four meioses. All but two of the SNPs were typed with individual assays, providing diploid genotypes with unknown phase. Two of the SNPs (rs429358 and rs7412) are only 138 bases apart and were typed with an assay that provided information about their phase. In combination these two SNPs define the three widely known *APOE* allele or haplotype states ($\epsilon 2$, $\epsilon 3$ and $\epsilon 4$) that underlie the metabolically distinct isoforms of Apolipoprotein E (Fullerton et al., 2000). Due to the different genotyping method and because these two SNPs yielded only three of the four possible haplotypes (i.e., $\epsilon 2$, $\epsilon 3$ and $\epsilon 4$), they were combined into a single composite locus with 3 alleles (hereafter referred to as *APOE*) for the purposes of this analysis. The final data set therefore consisted of 18 loci, yielding a total of 153 locus pairs.

Information about the assays used for the previously described microsatellites and the SNPs can be obtained on request from the corresponding author.

Results

Evaluation of Prior Distributions for Gamete Frequencies

The prior distribution of the haplotype frequencies, π, in the Bayesian estimation method needs to be carefully chosen, as it can have considerable effect on the posterior distribution of D' and therefore conclusions about LD differences between populations. In most cases, researchers will have no prior empirical information about haplotype frequencies (for a discussion of informative priors, see Lockwood et al., 2001). One of the suggested priors for π by Ayres & Balding (2001) is a Dirichlet prior such that the average of the elements of the prior's parameter vector, β, is equal to 1 (equivalent to having one haplotype on average in each cell *a priori*). If all the elements of β are equal to 1 then the Dirichlet prior is a uniform distribution on the $(KL-1)$ dimensional simplex. Here, this uniform prior will be referred to as the Ayres & Balding prior. Another candidate is the Jeffreys (1961) prior, which in the case of the multinomial distribution is a

Dirichlet distribution with a uniform parameter vector where all elements are equal to 0.5 (equivalent to having 0.5 haplotypes in each cell *a priori*). Jeffreys prior is a non-informative prior for haplotype frequencies, but does not necessarily result in a non-informative prior for the D' parameter, which is a transformation of the haplotype frequencies (see Bernardo & Ramón 1998 on non-informative priors for transformed parameters).

The two aforementioned priors have the disadvantage of not taking into account the varying size of (KL), the maximum possible number of haplotypes, for pairs of loci with more than two alleles. This is particularly important for microsatellites, which typically have numerous rare alleles, with the result that many of the possible haplotypes for pairs of microsatellites are not observed. In such cases, the priors described above will tend to overestimate the frequencies of rare or non-existent haplotypes. To address this problem, we propose a vague, but proper prior that reflects the maximum possible number of haplotypes for each locus pair, based on a Dirichlet distribution with a parameter vector β such that $\beta = (KL)^{-1}\Upsilon$, where Υ is a vector of ones. In this scheme, the impact of the prior on the frequency of each haplotype is proportional to the number of possible haplotypes at the locus pair. This will be referred to as the $(KL)^{-1}$ prior.

We also consider the improper prior $\beta = 0 \times \Upsilon$, i.e., each element of β is equal to zero, referred to as the zero prior. One feature of this prior is that in the case of SNPs where one of the haplotypes is not observed then $D' = 1$ with a posterior probability equal to one. In general, if one or more haplotypes are not observed then the frequency, p_{kl}, of each of these unobserved haplotypes will be zero with a posterior probability equal to one. However, it is still possible to sample from the posterior distribution of the p_{kl}'s of the observed haplotypes and compute D'.

In the scheme described in the subsection *A Bayesian test for differences in* D' the marginal prior and posterior distributions for the proportion of the j-th gamete frequency, π_j, become beta distributions with parameters $(\eta_1, \eta_2) = (\beta_j, \sum_{k\neq j}\beta_k)$ and $(\eta_3, \eta_4) = (\beta_j + C_j, \sum_{k\neq j}(\beta_k + C_k))$, respectively, based on properties of the Dirichlet distribution. In case of $\beta = 0 \times \Upsilon$ then $(\eta_1, \eta_2) = (0, 0)$ (an improper prior), if $\beta = (KL)^{-1} \times \Upsilon$ then $(\eta_1, \eta_2) = ((KL)^{-1}, (KL-1)/(KL))$, Jeffreys prior with $\beta = 0.5 \times \Upsilon$ gives $(\eta_1, \eta_2) = (0.5, 0.5(KL-1))$ and the Ayres & Balding prior, $\beta = \Upsilon$, results in $(\eta_1, \eta_2) = (1, KL - 1)$. Both the Jeffreys prior and the Ayres & Balding prior will have a strong influence on the posterior distribution of π_j if m is small and KL is relatively large, e.g., the posterior means are $(1 + C_j)(m + KL)^{-1}$ and $(0.5 + C_j)(m + 0.5KL)^{-1}$, respectively, while the $(KL)^{-1}$ prior yields the posterior mean $\{(KL)^{-1} + C_j\}\,(m+1)^{-1}$.

We evaluated the effect of the four different Dirichlet priors for the haplotype frequencies on the posterior mean of D' through simulation. The priors have parameter vectors $\beta = 0 \times \Upsilon$ (our suggestion), $\beta = (KL)^{-1}\Upsilon$ (our suggestion), $\beta = 0.5\Upsilon$ (Jeffreys) and $\beta = \Upsilon$ (Ayres & Balding). Samples of sizes 20, 35, 50, 75, 100, 150, 200, 350, 500, 750, 1000, 1500, 2000, 3500, 5000, and 10000 were randomly drawn (2000 times for each sample size) from predefined tables of haplotype frequencies. The D' parameter was estimated using the four priors and the Bayesian posterior mean as a point estimator. For comparison, the MLE of D' was also calculated. A comparison of the sampling distributions of the four Bayesian estimators, reveals the impact each prior has on the estimation of D'. A comparison with the sampling distribution of the MLE provides reference to an estimator that does not depend on a prior and is expected to have good large sample properties.

The four haplotype frequency tables were as follows. First is a 2×2 table with haplotype frequencies $p_{11} = 0.005$, $p_{12} = 0.33$, $p_{21} = 0.33$, $p_{22} = 0.335$, and a D' value of 0.955, designed to represent the scenario of strong LD between a pair of SNPs. The other three tables were based on a single 8×8 table corresponding to a pair of microsatellite loci based on real data from two adjacent X-chromosome microsatellites typed in 4096 Icelandic males, see Table S3 in supporting information. This table is arbitrarily chosen, but is useful to demonstrate the effect of the four priors on D' calculations for large tables where some of the possible haplotypes are not sampled. We first examine this 8×8 table ($D' = 0.302$), then a 4×4 table ($D' = 0.542$) obtained by further combining pairs of consecutive alleles for both loci according to the allele order of the 8×8 table, and finally a 2×2 table ($D' = 0.567$) obtained by further combining pairs of consecutive alleles from the 4×4 table.

Figure 1 shows the variance of the five D' estimators for the four tables and the different sample sizes, and demonstrates that as the size (KL) of the table increases, so does the impact of the Jeffreys and Ayres & Balding priors on the variance of the posterior mean estimator. In contrast, under our $(KL)^{-1}$ prior, the posterior mean estimator has variance closer to that of the MLE and is less affected by (KL). In other words, the $(KL)^{-1}$ prior is less informative about D' and thus more suitable for general inference about D' than the other two priors. The zero prior has variance even closer to that of the MLE and is hardly affected by (KL). Thus, the improper prior has good properties for estimating D'.

In case of these four tables, the MLE and the four posterior mean estimators for D' are biased. The bias depends on the value of the D' parameter and the sample size, see Figure 1. The $(KL)^{-1}$ prior tends to bias the posterior mean estimator a little downwards when $D' > 0.6$ while the zero prior gives results that are almost unbiased when $D' > 0.6$. Both the $(KL)^{-1}$ prior and the zero prior result in a positive bias as D' decreases below 0.6, particularly for small sample sizes. The

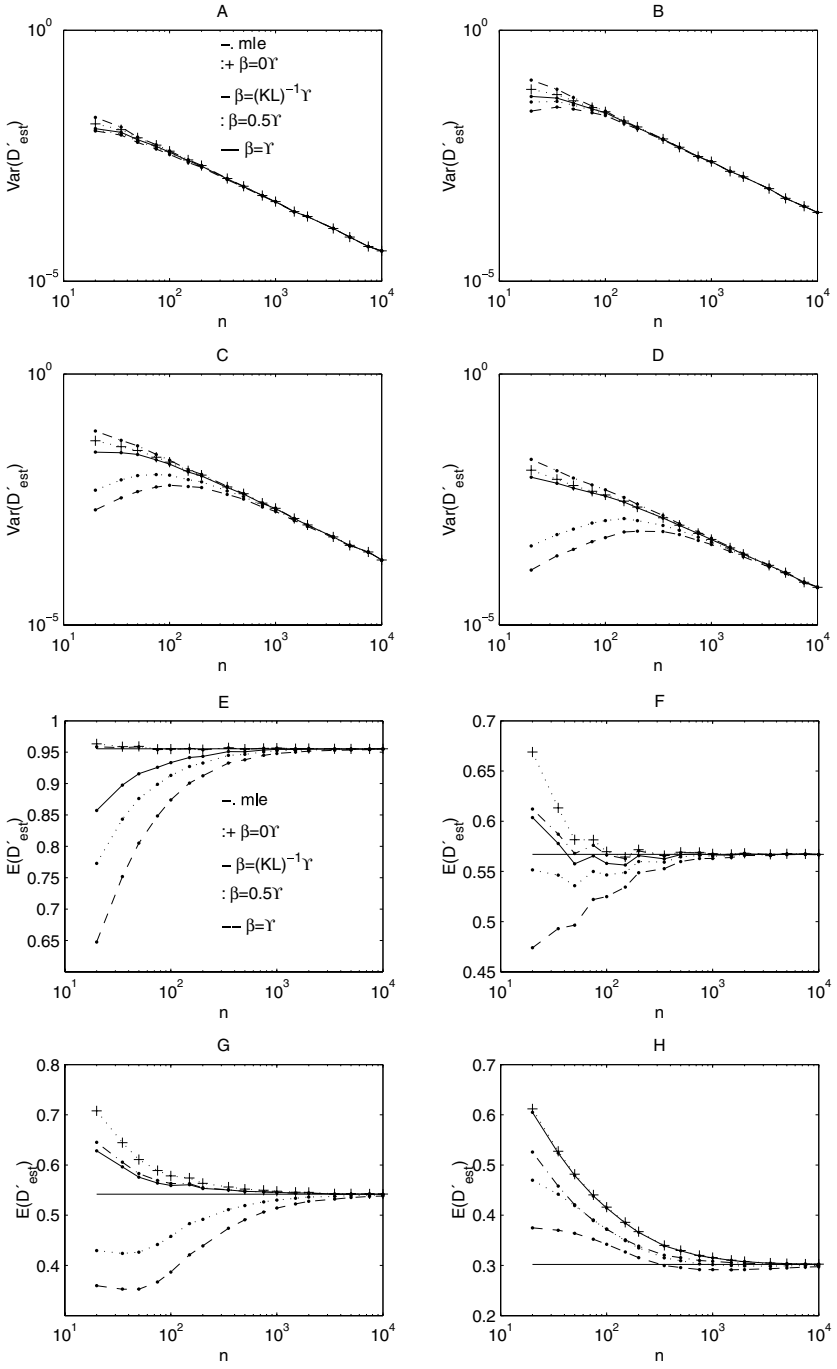

Figure 1 Comparsion of the variance and the expected value of the MLE (dotted dashed line) and the posterior mean estimators when Dirichlet priors with parameter vectors $\beta = 0 \times \Upsilon$ (dotted line with pluses), $\beta = (KL)^{-1}\Upsilon$ (solid line), $\beta = 0.5\Upsilon$ (dotted line) and $\beta = \Upsilon$ (dashed line) are used, for the 2×2 table with $D' = 0.955$ (A variance, E expected value), the 2×2 table based on the 8×8 table, $D' = 0.567$ (B variance, F expected value), the 4×4 table based on the 8×8 table, $D' = 0.542$ (C variance, G expected value), and the 8×8 table, $D' = 0.302$ (D variance, H expected value). A log-log scale is used to obtain a straight line relationship between the variance of the D' estimators and the sample size.

MLE estimator behaves similarly to these two estimators. In contrast, the Jeffreys prior and Ayres & Balding prior tend to bias downwards when $D' > 0.4$, particularly for small sample sizes, and bias slightly upwards when $D' < 0.4$. In all cases the bias is relatively small for sample sizes greater than 1500. Although these results may not be general for all tables of haplotype counts, they do indicate a considerable bias in D' estimation for sample sizes smaller than 1500.

The bias in the estimation of D' is caused by a complex interaction of several different factors. When the Dirichlet prior is of the form $\beta = \kappa\Upsilon$ (which is the case here with $\kappa = 0, (KL)^{-1}, 0.5, 1$) it follows that the larger the value of κ and the greater the number of alleles at the two loci, the more concentrated the prior density of D' becomes around a value which is less than 0.5 and depends on the number of alleles at the two loci and κ, see Ayres & Balding (2001). On the other hand there is an inherited positive bias in the MLE for D' which increases as D' becomes smaller and decreases as the sample size becomes larger. Note that the posterior density for D' is proportional to the product of the likelihood and the prior density. So, the likelihood part of the posterior density for D' has a maximum that is usually above the true value and the prior part has a maximum that is between zero and 0.5. How much the posterior density is shaped by these two parts depends on the sample size, such that the larger the sample size, the greater the influence of the likelihood part.

The Effect of Unequal Sample Size on the Bootstrap Test and the Bayesian Test

In order to evaluate the impact of the bias in D' estimation on our proposed bootstrap and Bayesian LD comparison tests, we explored the nature of this bias for situations where the null hypothesis $D'_1 = D'_2$ was true for a range of different sample sizes and values of D'. The probability of incorrectly accepting the hypothesis that $D'_2 > D'_1$ for the bootstrap test and the Bayesian test was evaluated with a simulation study (see Tables S4 and S5 respectively in supporting information). Within each of the seven cases that were simulated, the underlying haplotype frequencies and thus the D' values, were the same. The sample sizes in the two populations were 50, 100, 400, 1000, 1500 and 2000, and equal and unequal sample sizes were used. Four of the haplotype frequency tables used in this analysis are the same as in the simulation study described in the subsection *Evaluation of prior distributions for gamete frequencies*. In addition, three new tables were generated, using the marginal allele frequencies from the previously described 8×8, 4×4 and 2×2 tables, but assuming independence between loci and hence $D' = 0$.

Our findings indicate that the D' estimation bias does not produce an excess of false-positive results in our LD comparison tests when sample sizes from the groups being compared are equal or when samples sizes are unequal but both are greater than 1500. In the special case $D'_1 = D'_2 = 0$, this is not true in general regardless of sample sizes and whether the sample sizes are equal or not. Problems emerge in all cases where sample sizes are unequal and smaller than 1500. Thus, in the case of the Bayesian test, as D' approaches zero, there is a tendency to return an excess of false-positive results, indicating greater D' for the population with the smaller sample size. The effect of this bias increases as the number of alleles per locus increases, but is negligible in the case of loci with only two alleles (such as SNPs). As D' approaches one under the Bayesian test, the bias tends to be negative, with the Bayesian test having a tendency to return an excess of false-positive results indicating greater D' for the population with the larger sample size. In the case of the bootstrap test, when sample sizes are unequal and smaller than 1500, there is a tendency for false-positive results, indicating greater D' for the population with the smaller sample size, but this tendency seems to be to a less extent than the Bayesian test. We note that the bootstrap test is conservative in case of SNPs with D' close to one.

Overall, both the Bayesian and bootstrap tests are very sensitive to unequal sample sizes when D' approaches zero in both populations and as the number of alleles per locus increases. This is due to the difficulty of estimating D' when its true value is close to zero, in which case the bias of the estimator for D' is positive and directly related to the variance of the D' estimator.

The Power of the Bootstrap Test

The power to correctly determine difference in D' between two populations was evaluated for 2×2 tables under three different scenarios for equal sample sizes of 25, 50, 100, 400 and 1000. The first involves Population 1 with frequencies $p_{11} = p_{12} = p_{21} = p_{22} = 0.25$ resulting in $D'_1 = 0$ while the frequencies for Population 2 are $p_{11} = p_{22} = 0.25 - q$, $p_{12} = p_{21} = 0.25 + q$, resulting in $D'_2 = 4q$, $q \in (0, 0.25]$. The power of the test was computed for values of q in $(0, 0.25]$ (Fig. 2, panel A). In the second scenario Population 1 has frequencies $p_{11} = p_{22} = 0.125$, $p_{12} = p_{21} = 0.375$, and $D'_1 = 0.5$, and Population 2 has frequencies $p_{11} = p_{22} = 0.125 - q$, $p_{12} = p_{21} = 0.375 + q$, and $D'_2 = 0.5 + 4q$, $q \in (0, 0.125]$. The power was computed for values of q in $(0, 0.125]$, (Fig. 2, panel B). The last scenario is such that Population 2 has fixed frequencies $p_{11} = p_{22} = 0.125$, $p_{12} = p_{21} = 0.375$, which yields $D'_2 = 0.5$ while Population 1 has frequencies $p_{11} = p_{22} = 0.125 + q$, $p_{12} = p_{21} = 0.375 - q$, and $D'_2 = 0.5 - 4q$, $q \in (0, 0.125]$. The power was computed for the same sample sizes as for the previous scenarios (Fig. 2, panel C).

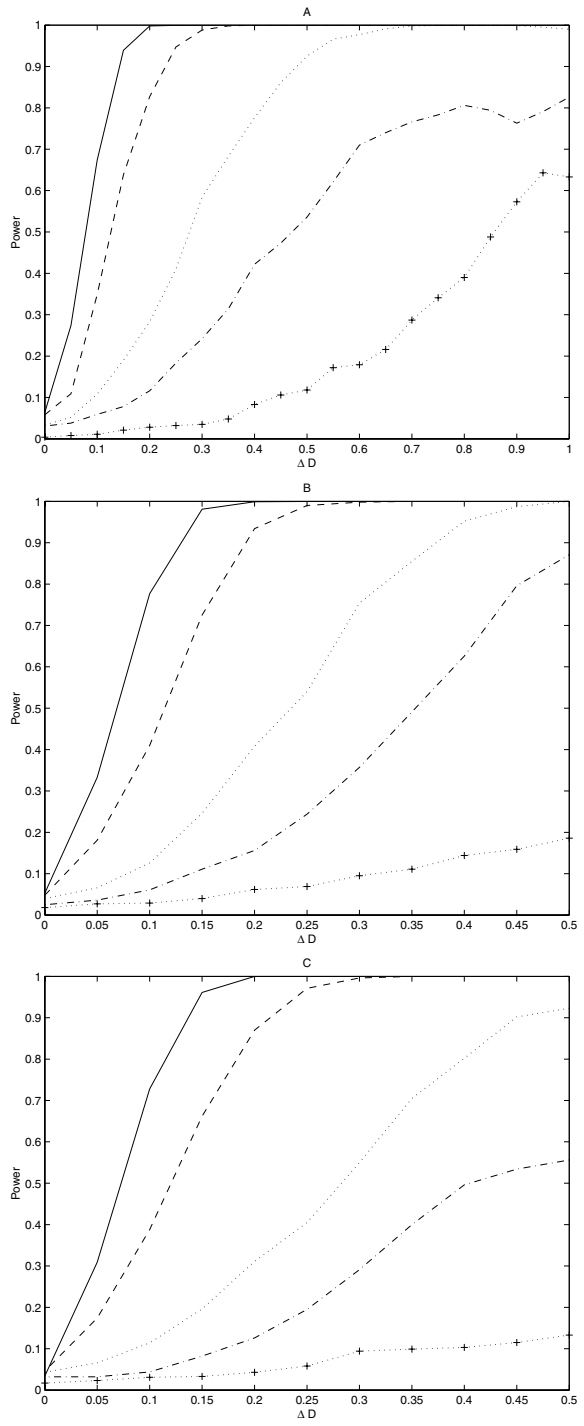

Figure 2 Bootstrap power curve comparison for three cases of 2×2 tables. $\Delta D'$ takes values between zero (H_0 true) and a positive value such that either one of or both of D'_1 and D'_2 are at the end of the parameter space. The sample sizes from the two populations are equal with $n = 25$ (dotted line with pluses), $n = 50$ (dashdot), $n = 100$ (dotted), $n = 400$ (dashed) and $n = 1000$ (solid).

As expected the power increases with sample size. From Figure 2 it can be seen that large sample sizes are needed to detect a D' difference less than 0.15. To consistently detect $\Delta D' = 0.15$ (power > 0.95) a sample size of 1000 is needed. A sample size of 100 results in power around 0.5 when detecting 0.25 difference in D'. Comparison of the second and the third scenarios reveals that it is slightly harder to detect differences between smaller values of D' than between larger values of D'.

LD Differences for 12 Xq25 Microsatellites in four European Populations

We applied the bootstrap and Bayesian LD comparison methods to the set of X chromosome microsatellites typed in Greenlanders, Icelanders, Northern Irish and Spanish, to illustrate the use of these methods for evaluating differences in LD among reproductively isolated populations. Table 1 shows the overall differences in LD between each pair of populations in terms of bootstrap single-test P-values using a threshold of $\alpha = 0.05$. Our results clearly indicate that the Greenland sample exhibits significantly larger D' values across the Xq25 region than any of the other three populations while the differences between the Icelanders, Northern Irish and Spanish are not statistically significant although there is a suggestion of greater LD in Icelanders when compared to Northern Irish.

Table 1 Application of the bootstrap method to the data on the European populations. Columns one and two specifiy which population is referred to as Population 1 and Population 2 when testing whether Population 2 and has greater D' values than Population 1. The third column contains the count of locus pairs out of a total of 66 that exceed the threshold $\alpha = 0.05$ in the bootstrap tests for each pair of populations while the fourth column gives the corresponding overall P-value.

Pop. 1	Pop. 2	Counts	P-value
Iceland	Greenland	27	$<10^{-4}$
N. Ireland	Greenland	31	$<10^{-4}$
Spain	Greenland	32	$<10^{-4}$
Greenland	Iceland	6	0.1921
N. Ireland	Iceland	9	0.0697
Spain	Iceland	5	0.2616
Greenland	N. Ireland	1	0.8316
Iceland	N. Ireland	5	0.2697
Spain	N. Ireland	3	0.5120
Greenland	Spain	2	0.6382
Iceland	Spain	2	0.6673
N. Ireland	Spain	2	0.6719

SNPs Flanking the *APOE* Gene in Alzheimer Patients and Controls

Figure 3 shows the posterior mean of D', i.e., the Bayesian estimate of D', for each of the 153 locus pairs in the *APOE* data set for Alzheimer patients (upper left matrix) and controls (lower right matrix), based on 10000 samples from the posterior distribution of D' for each locus pair using the Dirichlet prior with $\beta = (KL)^{-1}\Upsilon$. Here, both the Bayesian procedure and the bootstrap are applied to the *APOE* data set with the phase of alleles estimated beforehand using the EM algorithm (Excoffier & Slatkin 1995) and assumed known thereafter, and individuals with missing genotypes were omitted. Most of the locus pairs fall within a 26kb fragment containing the *APOE* gene. A visual inspection of Figure 3 reveals this fragment to contain three LD blocks in the controls, which are connected to each other by somewhat weaker LD. The Alzheimer patients exhibit a similar pattern of LD, but appear to have stronger LD between the first and third blocks, such that the whole 26kb fragment seems more like a single LD block than in the case of the controls. The final four SNPs, spanning about 30kb, reveal a fourth small LD block present in both controls and patients.

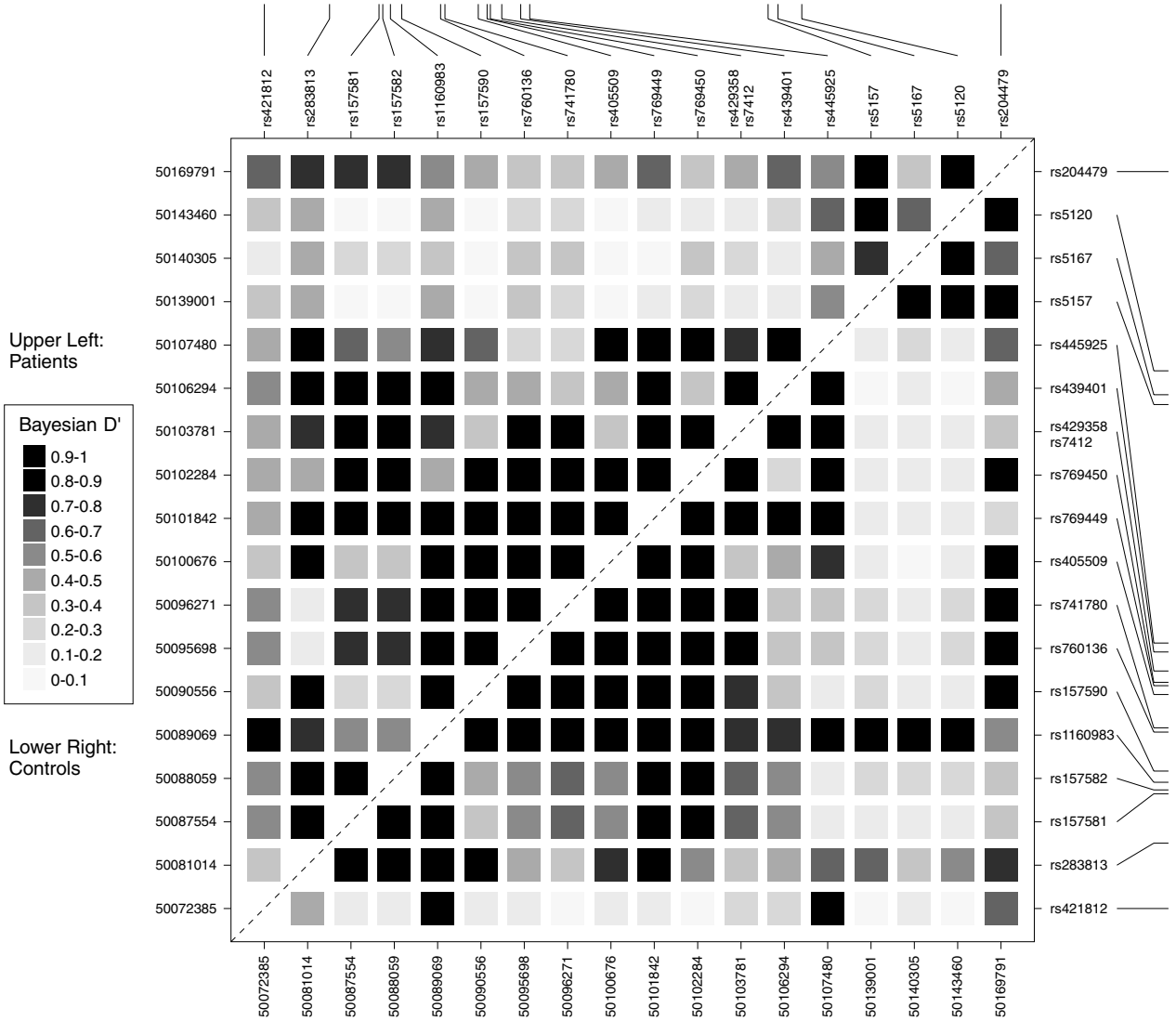

Figure 3 The posterior mean of D' for Alzheimer patients and controls, estimated for the *APOE* SNP data set, where the rectangles in the upper left corner represent values for the patients while the rectangles in the lower left corner represent values for the controls. Darker shades of grey indicate strong LD (D' close to 1) while lighter shades of gray indicate little or no LD (D' close to 0). The x-axis and the y-axis give the names and the location of the loci in each pair, on the side is a transformed physical scale.

Figure 4 shows the results obtained from applying the two LD comparison procedures to the *APOE* data set. The upper left half of Figure 4 shows, for each locus pair, the Bayesian posterior probability that the patients have greater values of D' than the controls, denoted by $P_{\text{Bayes}}(D'_{\text{pat}} > D'_{\text{ctrl}})$, derived from a comparison of the posterior distributions of D' for these two groups. A total of 19 out of the 153 locus pairs exhibited $P_{\text{Bayes}}(D'_{\text{pat}} > D'_{\text{ctrl}}) \geq 0.95$, thereof two where this probability was $\geq$0.999. In contrast, only one locus pair yielded $P_{\text{Bayes}}(D'_{\text{pat}} > D'_{\text{ctrl}}) \leq 0.05$, where the posterior probability was 0.014. Among the locus pairs with the highest posterior probability are a cluster of eight that describe the relationship between the first and third LD blocks in the 26kb region. Interestingly, the two locus pairs that show the most significant excess of D' in patients both include the composite *APOE* locus that is thought to contribute to disease risk.

The lower right half of Figure 4 shows, for each locus pair, one minus the *P*-value obtained from testing whether patients have greater D' values than controls with the bootstrap procedure, hereafter denoted as $P_{\text{boot}}(D'_{\text{pat}} > D'_{\text{ctrl}})$. These *P*-values were obtained using 10000 bootstrap samples of individuals within the groups. A total of 14 locus pairs exhibited $P_{\text{boot}}(D'_{\text{pat}} > D'_{\text{ctrl}}) \geq 0.95$ (thereof one where this probability was $\geq$0.999), whereas 17 locus pairs yielded

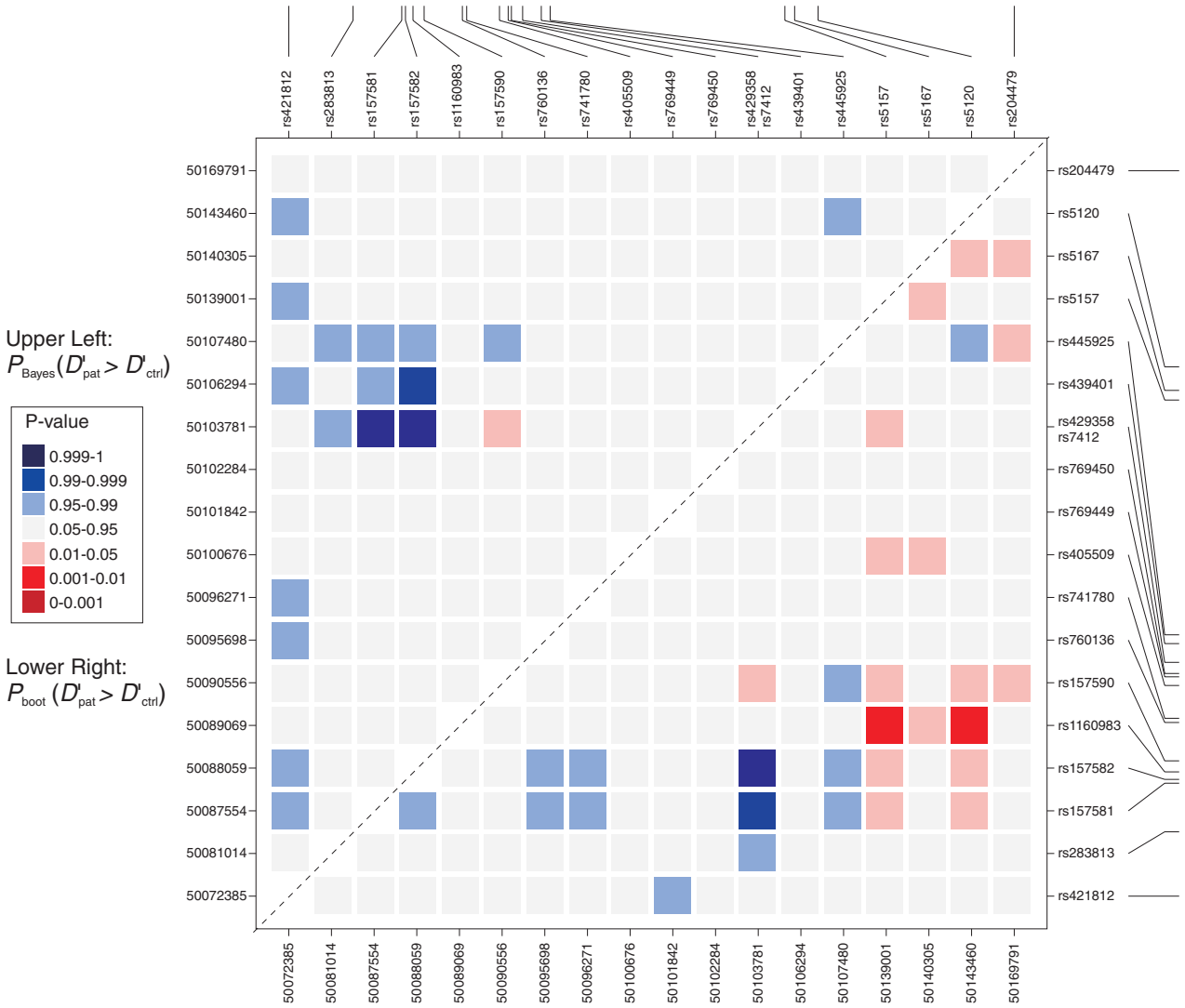

Figure 4 Comparison of D' between Alzheimer patients and controls, where the rectangles in the upper left corner represent $P_{\text{Bayes}}(D'_{\text{pat}} > D'_{\text{ctrl}})$ for the *APOE* SNP data set while the rectangles in the lower right corner represent $P_{\text{boot}}(D'_{\text{pat}} > D'_{\text{ctrl}})$ for the same data set. High values of $P(D'_{\text{pat}} > D'_{\text{ctrl}})$ and $P(D'_{\text{ctrl}} > D'_{\text{pat}})$ are represented by darker shades of blue and red, respectively. The *x*-axis and the *y*-axis give the names and the location of the loci in each pair, on the side is a transformed physical scale.

$P_{\text{boot}}(D'_{\text{pat}} > D'_{\text{ctrl}}) \leq 0.05$ (thereof none with probability ≤ 0.001). Around half of the locus pairs identified as having significantly greater D' values in Alzheimer patients using the Bayesian procedure are also identified as such by the bootstrap procedure. Furthermore, the one locus pair identified by the Bayesian procedure as having a greater D' value in controls is also identified as such by the bootstrap.

As the application of both the Bayesian and bootstrap LD comparison procedures involve 153 individual tests of differences in D', we would expect by chance on average 7.65 locus pairs to exhibit $P_{\text{boot}}(D'_{\text{pat}} > D'_{\text{ctrl}}) \geq 0.95$. To account for the problem of multiple testing we applied the single test for the number of locus pairs where D' is significantly greater (based on the bootstrap) in Alzheimer patients than controls in this data set. In the observed data there were $M_{\text{obs,pat>ctrl}}(0.05) = 14$ locus pairs with $P_{\text{boot}}(D'_{\text{pat}} > D'_{\text{ctrl}}) \geq 0.95$. The probability of obtaining $M_{0,\text{pat>ctrl}}(0.05) \geq 14$ is 0.1265. So, based on the threshold $\alpha = 0.05$ and an overall significance level 0.05 it cannot be concluded that LD for one or more locus pairs in the *APOE* region is significantly greater in Alzheimer patients than in the random controls. The probability of obtaining $M_{0,\text{pat>ctrl}}(0.001) \geq 1$ is 0.0831, which is greater than an overall significance level of 0.05 and thus not strong enough evidence to conclude that LD for one locus pair in the *APOE* region is significantly greater in Alzheimer patients than in the random controls when taking into account that multiple tests are conducted. The converse test for the number of locus pairs where D' is significantly greater (based on the bootstrap) in controls than Alzheimer patients yields a P-value of 0.0598 (i.e., the probability that $M_{0,\text{ctrl>pat}}(0.05) \geq 17$). Further, no pair exceeds the threshold $\alpha = 0.001$ when testing whether controls yield higher D' values than patients. These results do not indicate that LD for one or more locus pairs in the *APOE* region is significantly greater in controls than in Alzheimer patients when a correction for multiple tests is applied.

Association tests for each of the eighteen loci comparing Alzheimer patients and controls revealed that there are significant differences in allele frequencies after Bonferroni correction for multiple tests. Six loci out of eighteen were significant at the $0.05/18 = 0.0028$ level, and five loci out of eighteen were significant at the $0.01/18 = 0.00055$ level, see Table S2 in supporting information. Thus, in the case of the *APOE* data there are significant differences in allele frequencies between Alzheimer patients and controls, while the difference in D' is not significant.

Discussion

We have described two methods for comparing the strength of LD between reproductively isolated populations or subgroups of a single population. The former method involves bootstrapping individuals within groups to test the null hypothesis that the strength of LD is identical. The bootstrap procedure additionally allows for a single test of LD differences between groups for a large set of locus pairs – based on the number of locus pairs exhibiting significant LD differences compared with the null distribution derived from multiple bootstrap data sets. The latter method uses a Bayesian procedure to test for each locus pair whether D' is statistically different among groups. Based on a simulation study, we found that hypothesis testing based on the two methods is reasonably accurate if the sample sizes of the two populations are similar. However, if the sample sizes are different, the two methods can perform poorly especially when the number of alleles per locus is large and the sample sizes are less than 1000. The Bayesian method appears to be more sensitive than the bootstrap method to unequal sample sizes. Our suggestion is to use equal (or close to equal) sample sizes or to use sample sizes greater than 1500 when conducting hypothesis tests for D' differences. If the sample sizes are less than 1500 and there is a substantial difference between the two sample sizes, we suggest that the larger sample is reduced with random selection down to a size equal to the size of the smaller sample. Our results indicate that the bootstrap approach is more robust (prior specification not needed) and reliable than the Bayesian approach for the evaluation of LD differences between population samples.

Various features of the methodology presented here could be developed further. Thus, for example, although calculation time for both methods is relatively short for moderately large data sets, computational time could be further shortened by implementing tests based on large sample theory, along the lines suggested by Zapata et al. (2001). However, a problem with large sample approaches is that when the number of gametes is small, as is the case for pairs of SNPs, and moderate sample sizes are used, the normal approximation to the sampling distribution of D' can be poor. The Bayesian approach overcomes this problem, but the underlying MCMC algorithm requires a large number of iterations for accurate assessments of the posterior distribution of D'. By approximating the posterior densities of D', the computional time could potentially be decreased without loss of accuracy.

We have applied the proposed LD comparison methods in two case studies. The first was based on a small set of microsatellites from Xq25 typed in males from four European populations. Here we observed significantly larger D' values for the Greenlanders when compared to the other three populations. Also, there is tentative evidence for Icelanders having greater LD than Northern Irish and Spanish but more data are needed to confirm that. Broadly speaking, our results support the interpretation that differences in the demographic history of these populations are likely to be responsible for real differences in the patterns of LD observed. Our approach allowed definitive statements to be made about the nature of the

observed differences for each pair of populations since it directly compares each pair between populations with respect to LD.

The second case study was based on a set of 19 SNPs from a 97kb region on chromosome 19 that contains the *APOE* gene and was typed in Alzheimer patients and controls. The association of variation in this region to Alzheimer disease is well established. The common assumption is that the association is to the *APOE* SNPs (rs429358 and rs7412) themselves, however, it cannot be ruled out that other variants in strong LD with the *APOE* SNPs may be the causal factor (Martin et al., 2000). This means that the disease phenotype has been observed to be associated with variants located in and around the *APOE* gene. However, as far as we know, differences between patients and controls in the pattern and strength of LD between loci have not been explicitly explored in relation to Alzheimer disease. In cases where patients and controls differ significantly in the frequency of alleles at one or more loci and the loci are located within LD blocks, then patients and controls will also differ in the frequency of haplotypes constructed from loci in that region. Bearing in mind that the pattern of LD in a group of individuals is simply a statistical abstraction of its haplotype configuration, it then follows that significant differences in the frequency of haplotypes can entail significant differences in the strength and pattern of LD. Conversely, genomic regions exhibiting no association to a disease phenotype are unlikely to reveal significant LD differences between patients and controls. According to this reasoning, LD differences between patients and controls may be a by-product of a signal of association of the disease phenotype to the genomic region being examined. However, when corrected for multiple tests, there is no evidence for a statistical difference in LD between patients and controls.

Our knowledge of genome-wide patterns of LD in different human populations is currently being revolutionized by the high resolution SNP data sets that are being constructed as part of the HapMap project (International HapMap consortium, 2003) and by Perlegen (Hinds et al., 2005). One important goal in the analysis of these data is to understand the nature and magnitude of LD differences between the populations examined in these projects. The methods we propose could be used for this purpose.

References

Ardlie, K. G., Kruglyak, L. & Seielstad, M. (2002) Patterns of linkage disequilibrium in the human genome. *Nat Rev Genet* **3**, 299–309.

Ayres, K. L. & Balding, D. J. (2001) Measuring gametic disequilibrium from multilocus data. *Genetics* **157**, 413–423.

Bernardo J. M. & Ramón, J. M. (1998) An introduction to Bayesian reference analysis: Inference on the ratio of multinomial parameters. *The Statistician* **47**, 101–135.

Dunning, A. M., Durocher, F., Healey, C. S., Teare, M. D., McBride, S. E., Carlomagno, F., Xu, C. F., Dawson, E., Rhodes, S., Ueda, S., Lai, E., Luben, R. N., Van Rensburg, E. J., Mannermaa, A., Kataja, V., Rennart, G., Dunham, I., Purvis, I., Easton, D. & Ponder, B. A. J. (2000) The extent of linkage disequilibrium in four populations with distinct demographic histories. *Am J Hum Genet* **67**, 1544–1554.

Eaves, I. A., Merriman, T. R., Barber, R. A., Nutland, S., Tuomilehto-Wolf, E., Tuomilehto, J., Cucca, F. & Todd, J. A. (2000) The genetically isolated populations of Finland and Sardinia may not be a panacea for linkage disequilibrium mapping of common disease genes. *Nat Genet* **25**, 320–323.

Efron B. & Tibshirani, R. J. (1993) *An Introduction to the Bootstrap*. New York: Chapman & Hall.

Excoffier, L. & Slatkin, M. (1995) Maximum-likelihood estimation of molecular haplotype frequencies in a diploid population. *Mol Biol Evol* **12**, 921–927.

Fisher, R. A. (1915) Frequency distribution of the values of the correlation coefficient in samples of an indefinitely large population. *Biometrika* **10**, 507–521.

Fullerton, S. M., Clark, A. G., Weiss, K. M., Nickerson, D. A., Taylor, S. L., Stengard, J. H., Salomaa, V., Vartiainen, E., Perola, M., Boerwinkle, E. & Sing, C. F. (2000) Apolipoprotein E variation at the sequence haplotype level: implications for the origin and maintenance of a major human polymorphism. *Am J Hum Genet* **67**, 881–900.

Guo, S. W. & Thompson, E. A. (1992) Performing the exact test of Hardy-Weinberg proportion for multiple alleles. *Biometrics* **48**, 361–372.

Hamilton, D. C. & Cole, D. E. (2004) Standardizing a composite measure of linkage disequilibrium *Ann Hum Genet* **68**, 234–239.

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

Heutink, P. & Oostra, B. A. (2002) Gene finding in genetically isolated populations. *Hum Mol Genet* **11**, 2507–2515.

Hinds, D. A., Stuve, L. L., Nilsen, G. B., Halperin, E., Eskin, E., Ballinger, D. G., Frazer, K. A. & Cox, D. R. (2005) Whole-genome patterns of common DNA variation in three human populations. *Science* **307**, 1072–1079.

International HapMap consortium (2003) The international HapMap project. *Nature* **426**, 789–796.

Jeffreys, H. (1961) *Theory of Probability, 3rd ed*. Oxford: University Press.

Johnson N. L. & Kotz, S. (1972) *Distributions in Statistics: Continuous Multivariate Distributions*. New York: Wiley.

Kaessmann, H., Zollner, S., Gustafsson, A. C., Wiebe, V., Laan, M., Lundeberg, J., Uhlen, M. & Pääbo, S. (2002) Extensive linkage disequilibrium in small human populations in Eurasia. *Am J Hum Genet* **70**, 673–685.

Katoh, T., Mano, S., Ikuta, T., Munkhbat, B., Tounai, K., Ando, H., Munkhtuvshin, N., Imanishi, T., Inoko, H. & Tamiya, G. (2002) Genetic isolates in East Asia: A study of linkage disequilibrium in the X chromosome. *Am J Hum Genet* **71**, 395–400.

Laan, M. & Pääbo, S. (1997) Demographic history and linkage disequilibrium in human populations. *Nat Genet* **17**, 435–438.

Laan, M., Wiebe, V., Khusnutdinova, E., Remm, M. & Pääbo, S. (2005) X-chromosome as a marker for population history: linkage disequilibrium and haplotype study in Eurasian populations. *Eur J Hum Genet* **13**, 452–462.

Latini, V., Sole, G., Doratiotto, S., Poddie, D., Memmi, M., Varesi, L., Vona, G., Cao, A. & Ristaldi, M. S. (2004) Genetic isolates in Corsica (France): linkage disequilibrium extension analysis on the Xq13 region. *Eur J Hum Genet* **12**, 613–619.

Lewontin, R. C. (1964) The interaction of selection and linkage. II. Optimum models. *Genetics* **50**, 757–782.
Lockwood, J. R., Roeder, K. & Devlin, B. (2001) A Bayesian hierarchical model for allele frequencies. *Genet Epidemiol* **20**, 17–33.
Martin, E. R., Lai, E. H., Gilbert, J. R., Rogala, A. R., Afshari, A. J., Riley, L., Finch, K. L., Stevens, F., Livak, K. J., Slotterbeck, B. D., Slifer, S. H., Warren, L. L., Conneally, P. M., Schmechel, D. E., Purvis, I., Pericak-Vance, M. A., Roses, A. D. & Vance, J. M. (2000) SNPing away at complex diseases: analysis of single-nucleotide polymorphisms around *APOE* in Alzheimer disease. *Am J Hum Genet* **67**, 383–394.
Nordborg, M. & Tavare, S. (2002) Linkage disequilibrium: what history has to tell us. *Trends in Genetics* **18**, 83–90.
Robinson, W. P., Asmussen, M. A. & Thomson, G. (1991) Three-locus systems impose additional constraints on pairwise disequilibria. *Genetics* **129**, 925–930.
Sawyer, S. L., Mukherjee, N., Pakstis, A. J., Feuk, L., Kidd, J. R., Brookes, A. J. & Kidd, K. K. (2005) Linkage disequilibrium patterns vary substantially among populations. *Eur J Hum Genet* **13**, 677–686.
Shifman, S. & Darvasi, A. (2001) The value of isolated populations. *Nat Genet* **28**, 309–310.
Slatkin, M. (1994) Linkage disequilibrium in growing and stable populations. *Genetics* **137**, 331–336.
Stephens, M., Smith, N. J. & Donnelly, P. (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* **68**, 978–989.
Strittmatter, W. J., Weisgraber, K. H., Huang, D. Y., Dong, L. M., Salvesen, G. S., Pericakvance, M., Schmechel, D., Saunders, A. M., Goldgaber, D. & Roses, A. D. (1993) Binding of human Apolipoprotein E to synthetic amyloid beta peptide: isoform-specific effects and implications for late-onset Alzheimer disease. *Proc Natl Acad Sci U.S.A.* **90**, 8098–8102.
Taillon-Miller, P., Bauer-Sardina, I., Saccone, N. L., Putzel, J., Laitinen, T., Cao, A., Kere, J., Pilia, G., Rice, J. P. & Kwok, P. Y. (2000) Juxtaposed regions of extensive and minimal linkage disequilibrium in human Xq25 and Xq28. *Nat Genet* **25**, 324–328.
Terwilliger, J. D., Zollner, S., Laan, M. & Pääbo, S. (1998) Mapping genes through the use of linkage disequilibrium generated by genetic drift: 'Drift mapping' in small populations with no demographic expansion. *Hum Heredity* **48**, 138–154.
Wall, J. D. & Pritchard, J. K. (2003) Haplotype blocks and linkage disequilibrium in the human genome. *Nat Rev Genet* **4**, 587–597.
Wang, T., Zhu, X. & Elston, R. C. (2007) Improving power in contrasting linkage-disequilibrium patterns between cases and controls. *Am J Hum Genet* **80**, 911–920.
Wright, A. F., Carothers, A. D. & Pirastu, M. (1999) Population choice in mapping genes for complex diseases. *Nat Genet* **23**, 397–404.
Zapata, C. (2000) The D' measure of overall gametic disequilibrium between pairs of multiallelic loci. *Evolution* **54**, 1809–1812.
Zapata, C., Carollo, C. & Rodriguez, S. (2001) Sampling variance and distribution of the D' measure of overall gametic disequilibrium between multiallelic loci. *Ann Hum Genet* **65**, 395–406.
Zaykin, D. V. (2004) Bounds and normalization of the composite linkage disequilibrium coefficient. *Genet Epidemiol* **27**, 252–257.
Zaykin, D. V., Meng, Z. & Ehm, M. G. (2006) Contrasting linkage-disequilibrium patterns between cases and controls as a novel association-mapping method. *Am J Hum Genet* **78**, 737–746.

Appendix

The Bayesian estimation of gamete frequencies for two cases will be described here. In the first case the phase of genotype data is known and haplotypes are missing at some loci for some individuals. Note that in the case introduced in *Materials and methods*, phase was known and missing data were not used. In the second case the phase of genotype data is unknown for some loci and for some individuals, and genotypes are possibly missing at some loci for some individuals.

It is assumed that the data are missing at random, that is, the data are missing independently of the true state of the genotypes, so,

$$P(\text{data missing at none/one/both of the loci} \mid \text{true state}) = P(\text{data missing at none/one/both the loci}).$$

It is also assumed that when the alleles are not missing then the true alleles are observed. Let γ denote the vector of parameters for the model that describes how the data are missing.

Bayesian Estimation of Gamete Frequencies Based on Genotype Data with Known Gametic Phase when Missing Data are Used

As in *Materials and methods*, let C_{kl} and p_{kl} be the count and frequency, respectively, of haplotypes for some population where the observed gamete is A_kB_l. Here, let $C = (C_{kl})_{k,l}$ and $\pi = (p_{kl})_{k,l}$ be $K \times L$ matrices with these counts and frequencies, respectively. Let R_k be the count of haplotypes where the allele of the second locus is missing while the observed allele of the first locus is A_k. Let $R = (R_k)_k$ be a vector of length K containing these counts. Let V_l be the count of haplotypes where the allele of the first locus is missing while the observed allele of the second locus is B_l. Let $V = (V_l)_l$ be a vector of length L containing these counts. Let Z be the count of haplotypes where the alleles of both loci are missing. The variables C, R, V and Z are the observed data.

Let X_{kl} be the true state, that is, the true count of haplotypes with gamete A_kB_l. Let $X = (X_{kl})_{k,l}$ be a $K \times L$ matrix containing these counts. Note that the matrix X is not observed directly. Let X^R and X^V be the counts in X that correspond to the counts in R and V, respectively. The relationship between X and C, X^R and X^V is

$$X = C + X^R + X^V. \tag{2}$$

The total count of haplotypes in X is denoted by m. Note that the count in Z is not in X since it will not add any information to the gamete frequencies.

A hierarchical model that describes the data C, R, V and Z, the state variable X and π can be presented as follows

$$\begin{aligned} &C, R, V, Z|X, \gamma \sim \text{DataModel}(X, \gamma) \\ &\text{vec}(X)|\pi \sim \text{Mult}_{KL}(m, \text{vec}(\pi)) \\ &\text{vec}(\pi) \sim \text{Dir}_{KL}(\beta) \end{aligned} \tag{3}$$

where vec (V) denotes the vectorisation of a matrix V. The first part in (3) describes how C and the missing data are generated based on X and γ. As in *Materials and methods*, the vector β reflects the prior information on π.

The posterior distribution of π and X given the data is obtained from the distributions in (3), and is given by

$$p(\pi, X|C, R, V, Z) \propto p(C, R, V, Z|X, \gamma)p(X|\pi)p(\pi). \tag{4}$$

The following Gibbs sampler is used to generate samples from (4). The generation of X is broken into the generation of X^R and X^V, and then (2) is used to obtain X. The term "rest" denotes all the variables in the model, except for the one that is given in front.

$$p(\pi|\text{rest}) = \text{Dir}_{KL}(\text{vec}(X) + \beta)$$

$$p(X^R|\text{rest}) \propto p(X^R|\pi, R)$$

$$p(X^V|\text{rest}) \propto p(X^V|\pi, V).$$

To generate samples from $p(X^R|\pi, R)$ the following step is needed,

$$X^R_{k,1:L} \sim \text{Mult}_L\left(R_k, \frac{w_k}{\sum_{l=1}^{L} p_{kl}}\right), \qquad k = 1, \ldots, K,$$

where $X^R_{k,1:L}$ is the vector containing the elements of X^R with indices $k' = k$, $l' = 1, \ldots, L$, and $w_k = (p_{k1}, \ldots, p_{kL})^{\mathrm{T}}$, $k = 1, \ldots, K$. To generate samples from $p(X^V|\pi, V)$ the following step is needed,

$$X^V_{1:K,l} \sim \text{Mult}_K\left(V_l, \frac{q_l}{\sum_{k=1}^{K} p_{kl}}\right), \qquad l = 1, \ldots, L,$$

where $X^V_{1:K,l}$ is the vector containing the elements of X^V with indices $k' = 1, \ldots, K$, $l' = l$, and $q_l = (p_{1l}, \ldots, p_{Kl})^{\mathrm{T}}$, $l = 1, \ldots, L$.

Bayesian Estimation of Gamete Frequencies Based on Genotype Data with Unknown Gametic Phase when Missing Data are Either Used or Not

Let N_{klgs} be the count of individuals where the phase of genotypes at a given pair of loci is known, and the two gametes are A_kB_l and A_gB_s. Let $N = (N_{klgs})_{k,l,g,s}$ be a $K \times L \times K \times L$ array containing these counts. Let H_{klgs} be the count of individuals where the phase of the genotypes is unknown, and the observed alleles are A_k and A_g at the first locus ($k < g$) and B_l and B_s at the second locus ($l < s$). The count H_{klgs} corresponds to individuals that are heterozygous at both loci and the phase could not be determined from genealogical data. Let $H = (H_{klgs})_{k,l,g,s}$ be a $K \times L \times K \times L$ array containing these counts.

Let Q_{kg} be the count of individuals where both alleles of the second locus are missing while the observed alleles of the first locus are A_k and $A_g (k \leq g)$. Let $Q = (Q_{kg})_{k,g}$ be a $K \times K$ matrix containing these counts. Let U_{ls} be the count of individuals where both alleles of the first locus are missing while the observed alleles of the second locus are B_l and $B_s (l \leq s)$. Let $U = (U_{ls})_{l,s}$ be a $L \times L$ matrix containing these counts. Let W be the count of individuals where both alleles of both loci are missing. The variables N, H, Q, U and W are the observed data. When missing data are not used, then in what follows, the elements of the matrices Q and U are set to zero, and W is set to zero as well.Let Y_{klgs} be the true state, that is, the count of individuals with correctly phased genotypes where the gametes are A_kB_l and A_gB_s. Let $Y = (Y_{klgs})_{k,l,g,s}$ be a $K \times L \times K \times L$ array containing these counts. The array Y is not observed directly, however, the array N is the part of Y that is observed directly. Let M_{kl} be the true count of gamete A_kB_l. Let $M = (M_{kl})_{k,l}$ be a $K \times L$ matrix containing these counts. The relationship between M and Y is

$$M_{kl}(Y) = \sum_{i=1}^{K}\sum_{j=1}^{L}(Y_{klij} + Y_{ijkl}), \qquad k = 1, \ldots, K, \qquad l = 1, \ldots, L.$$

Let Y^H, Y^Q and Y^U be the counts of phased genotypes in Y that correspond to the counts in H, Q, and U, respectively. Thus, Y is given in terms of N, Y^H, Y^Q and Y^U by

$$Y = N + Y^H + Y^Q + Y^U. \tag{5}$$

The total count of individuals in Y is denoted by n. Note that the count in W is not in Y since it will not add any information to the gamete frequencies.

A hierarchical model that describes the data and the state variable Y can be presented as follows

$$
\begin{aligned}
&N, H, Q, U, W|Y, \gamma \sim \text{DataModel}(Y, \gamma)\\
&\text{vec}(Y)|\pi \sim \text{Mult}_{K^2L^2}(n, \text{vec}(\pi \otimes \pi))\\
&\text{vec}(\pi) \sim \text{Dir}_{KL}(\beta)
\end{aligned}
\tag{6}
$$

where $\otimes$ is the Kronecker product. The first part in (6) describes how N, H and the missing data, are generated based on Y and γ. As before, the vector β reflects the prior information on π. Note that given the state variable Y, the data does not depend on π, it just depends on the missing data model with parameter vector γ.

The posterior distribution of π and Y given the data is obtained from the distributions in (6), and is given by

$$
\begin{aligned}
&p(\pi, Y|N, H, Q, U, W)\\
&\quad \propto p(N, H, Q, U, W|Y, \gamma)p(Y|\pi)p(\pi).
\end{aligned}
\tag{7}
$$

To generate samples from (7), the following Gibbs sampler is used. The generation of Y is broken into the generation of Y^H, Y^Q and Y^U, and Y is obtained from (5), so

$$p(\pi|\text{rest}) = \text{Dir}_{KL}(\text{vec}(M(Y)) + \beta)$$

$$p(Y^H|\text{rest}) \propto p(Y^H|\pi, H)$$

$$p(Y^Q|\text{rest}) \propto p(Y^Q|\pi, Q)$$

$$p(Y^U|\text{rest}) \propto p(Y^U|\pi, U).$$

To generate samples from $p(Y^H|\pi, H)$ the following step is needed,

$$Y^H_{klgs} \sim \text{Bin}\left(H_{klgs}, \frac{p_{kl}\,p_{gs}}{p_{kl}\,p_{gs} + p_{gl}\,p_{ks}}\right),$$

$$Y^H_{glks} = H_{klgs} - Y^H_{klgs}, \quad 1 \le k < g \le K, \quad 1 \le l < s \le L.$$

To generate samples from $p(Y^Q|\pi, Q)$ the following step is needed,

$$Y^Q_{k,1:L,g,1:L} \sim \text{Mult}_{L^2}\left(Q_{kg}, \frac{\text{vec}\left(w_k w_g^{\mathrm{T}}\right)}{\sum_{l=1}^{L}\sum_{s=1}^{L} p_{kl}\,p_{gs}}\right), \quad 1 \le k \le g \le K,$$

where $Y^Q_{k,1:L,g,1:L}$ is the vector containing the elements of Y^Q with indices $k' = k$, $l' = 1, \ldots, L$, $g' = g$, and $s' = 1, \ldots, L$, and $w_t = (p_{t1}, \ldots, p_{tL})^{\mathrm{T}}$, $t = 1, \ldots, K$. To generate samples from $p(Y^U|\pi, U)$ the following step is needed,

$$Y^U_{1:K,l,1:K,s} \sim \text{Mult}_{K^2}\left(U_{ls}, \frac{\text{vec}\left(q_l q_s^{\mathrm{T}}\right)}{\sum_{k=1}^{K}\sum_{g=1}^{K} p_{kl}\,p_{gs}}\right), \quad 1 \le l \le s \le L,$$

where $Y^U_{1:K,l,1:K,s}$ is the vector containing the elements of Y^U with indices $k' = 1, \ldots, K$, $l' = l$, $g' = 1, \ldots, K$, and $s' = s$, and $q_t = (p_{1t}, \ldots, p_{Kt})^{\mathrm{T}}$, $t = 1, \ldots, L$.

Supporting Information

Additional supporting information may be found in the online version of this article:

Table S1 The names of the microsatellites of the European X chromosome Xq25 data set, their physical positions based on NCBI build 34, and their repeat motif size.

Table S2 The names of the SNPs of the *APOE* data set, with physical positions based on NCBI build 34 and *P*-values for association test for each marker comparing Alzheimer patients and controls.

Table S3 The haplotype frequencies for two adjacent X-chromosome microsatellites, each with 8 alleles.

Table S4 The probability (based on repeated sampling) of incorrectly accepting the alternative hypothesis that $D'_2 > D'_1$ when the decision to reject is based on the bootstrap method that uses transformation. The decision to reject is made if the bootstrap *P*-value is less than 0.05. This probability is computed for seven cases where D' and the allele frequencies are the same in both populations and the sample sizes are equal to 50, 100, 400, 1000, 1500 and 2000. The test has the correct significance level when the probability values in the table are close to 0.05.

Table S5 The probability (based on repeated sampling) of incorrectly accepting the alternative hypothesis that $D'_2 > D'_1$. The decision to reject is made if the posterior probability, $P(D'_2 > D'_1)$, of the Bayesian testing procedure is greater than 0.95. This probability is computed for seven cases where D' and the allele frequencies are the same in both populations and the sample sizes are equal to 50, 100, 400, 1000, 1500 and 2000. The Bayesian test procedure has the desired frequentist properties when the probability values in the table are close to 0.05.

As a service to our authors and readers, this journal provides supporting information supplied by the authors. Such materials are peer-reviewed and may be re-organised for online delivery, but are not copy-edited or typeset. Technical support issues arising from supporting information (other than missing files) should be addressed to the authors.

Received: 25 August 2009
Accepted: 7 February 2010