doi: 10.1111/j.1469-1809.2006.00305.x

Variation of Estimates of SNP and Haplotype Diversity and Linkage Disequilibrium in Samples from the Same Population Due to Experimental and Evolutionary Sample Size

P. M. Visscher[1,2]

[1] *Queensland Institute of Medical Research, Brisbane, Australia*
[2] *Institute of Evolutionary Biology, School of Biological Sciences, University of Edinburgh, UK*

Summary

Studies of genetic polymorphisms and diversity between and within human populations are increasingly characterised by a very large number of genetic markers but using a relatively small number of individuals from which DNA samples were taken. In this report we examine the limitations of a small experimental sample size relative to a large genomic sample size, and quantify the sampling variance of a number of measures of diversity and linkage disequilibrium. The relationship between sample size and observed levels of polymorphism and haplotype diversity at the level of a gene is investigated under a neutral model of sequence evolution, using coalescent simulations. It is shown that the effect of evolutionary sampling, as manifested by differences between samples (genes) in measures of diversity estimated using very large sample sizes, is substantial, with a coefficient of variation of the number of detected polymorphic SNPs or haplotypes in the order of 15%. The effect of experimental design (sample size) is also very large, and a number of 'significant' results reported in the literature can be explained by sampling alone. The expected correlation coefficient of measures of linkage disequilibrium across samples from the same population has been quantified and found to be consistent with empirical estimates from the literature.

Keywords: haplotype diversity, linkage disequilibrium, coalescent, SNP, sample size

Introduction

Genetic polymorphism and diversity studies in human populations are increasingly characterised by a very large number of genetic markers but with a relatively small number of individuals from which DNA samples are taken. For example, the HapMap project (Altshuler *et al*. 2005) has published the analysis of one million SNPs typed in four ethnic samples, each sample varying from 44 to 90 individuals. The larger samples (90 individuals) are from 30 trios, equivalent to population data on a sample size of 60 individuals. Phase II of the HapMap project aims to score a further 4.6 million SNPs in the same individuals. The SeattleSNPs project (http://pga.gs.washington.edu/) has (re)sequenced 100s of genes in two samples, consisting of 24 individuals of European descent and 23 of African-American descent (Crawford *et al*. 2004, 2005). As a third example, the DNA Polymorphism Discovery Resource (http://locus.umdnj.edu/nigms/products/pdr.html) uses a total of 450 individuals from 5 populations and subsets (without ethnicity identifiers) of 90 individuals or less for human genetic variation studies.

As well as making comparisons across genes within a sample, studies of human genetic diversity often compare allele frequencies or haplotype frequencies between samples from two or more populations. Commonly used

Corresponding author: Dr. Peter M. Visscher, Genetic Epidemiology, Queensland Institute of Medical Research, 300 Herston Road, Herston 4006, Australia. Tel. +61 7 3362 0166, fax: +61 7 3362 0101. E-mail: peter.visscher@qimr.edu.au

summaries of diversity data include the number of segregating sites (or SNPs), the number of haplotypes, and various measures of linkage disequilibrium. For example, the HapMap Consortium reported the correlation between linkage disequilibrium in samples from 4 populations (Altshuler *et al*. 2005), as did Evans & Cardon (2005) for two samples from Caucasian populations and Willer *et al*. (2006) for a comparison of samples from Finland and HapMap. However, none of these studies quantified the sampling variance of this correlation under the hypothesis that the samples were from the same population.

However genetic diversity is measured, we argue that it is instructive to consider two sources of variation when drawing inference: evolutionary sampling and experimental sampling. *Evolutionary sampling* is caused by genetic drift; the alleles found in a population are a finite sample of the alleles found in that population some time in the past. Evolutionary sampling is governed by *effective population size* (N_e). This is determined by historical and present population size, by variation in reproductive output across individuals in the population, by population stratification and subdivision, and by natural selection (Hartl & Clark, 1997). *Experimental sampling* is caused simply by genotyping a sample of individuals instead of the whole population from which they come. The effects of evolutionary and experimental sampling are often blurred together, because widely used coalescent models describe both simultaneously. For many aspects of data analysis and hypothesis testing the distinction is unimportant, and a simultaneous treatment of both is an advantage of coalescent modelling. However, evolutionary and experimental sampling can be separated by making the sample size in a coalescent model very large. In this paper we use this device to describe the relative effects of evolutionary and experimental sampling on measures of haplotype diversity and similarity, and on measures of linkage disequilibrium. We consider situations where independent (not closely linked) genes are studied in a single sample, and where two independent samples have been taken from the same population. The question we address is how variable measures of diversity and linkage disequilibrium are, as a function of experimental and evolutionary sample size.

Methods

Haplotype Diversity Measures

We simulated DNA polymorphism data, assuming the neutral coalescent under the infinite sites model of mutation (Hudson, 1985, 2002). Replicate genes were simulated assuming that $\theta = 4N_e\mu = 17$ and $\rho = 4N_e r = 5$, with μ and r the mutation and recombination rate for the whole sequence, respectively, and N_e the effective population size. The parameters θ and ρ are the scaled population mutation rate and the scaled population recombinate rate, respectively (Ewens, 2004, Hartl & Clark, 1997). (Note that θ and ρ are the standard symbols used for these quantities in the population genetics literature, but that the same symbols are often used to denote recombination fraction and population correlation in the statistical genetics literature.)

For a gene length of 16.5 kb and an effective population size of $N_e = 10{,}000$ these parameters correspond to a nucleotide mutation rate of $2.6\ 10^{-8}$ and a recombination fraction of 0.0076 per Mb (approximately 0.8 cM/Mb). These parameters were chosen because they resulted in average numbers of SNPs and haplotypes that were consistent with those reported in a recent paper that investigated haplotype diversity in two populations (Crawford *et al*. 2004).

Samples of $2n$ chromosomes were drawn and randomly subdivided into two sub-samples of size n. This simulation approach results in two finite experimental samples of the same population, because chromosomes in both samples have experienced the same evolutionary sampling process. For each of the two sub-samples we calculated the number of SNPs, number of haplotypes, and effective number of haplotypes. Analogous to the effective number of alleles, the effective number of haplotypes is defined as $1/\Sigma p_i^2$, where p_i is the sample frequency of the i-th haplotype. The closely related sample heterozygosity is defined as $1 - \Sigma p_i^2$. Following common practice these measures were calculated using only SNPs with a minor allele frequency (MAF) $\geq$ 5%, with this threshold applied to each sub-sample separately. We also calculated the number of haplotypes shared in the two sub-samples, and the number unique to each sub-sample. These comparative measures were calculated using only SNPs above a threshold

MAF $\geq$ 5% applied to the whole sample. For each sample size n that we considered, we performed 10,000 replicate simulations, which can be viewed as being samples from different independent genes. Variation between genes reflects evolutionary sample size, whereas variation between sub-samples from the same population reflects experimental sample size.

Linkage Disequilibrium

Simulations examining the effects of sampling on measures of linkage disequilibrium were identical to those described above, except that we used parameters $\theta = 100$ and $\rho = 100$. This approximately represents a 250 kb region in the genome. As before, samples of $2n$ chromosomes were randomly split into two equal sized sub-samples. We used an estimate of the population recombination rate as a measure of LD within each sub-sample, which we calculated using only SNPs above a MAF $\geq$ 0.05 threshold as before. For each sub-sample we estimated ρ for the entire segment using linear regression: ρ_d was estimated from the average r^2 between all pairs that were a distance d apart, using $E(r_d^2) = (10 + \rho_d)/(22 + 13\rho_d + \rho_d^2) + 1/n$ (Hill, 1975; McVean, 2002; Ohta & Kimura, 1969), and the estimate of ρ_d was regressed on d using weighted least squares, the weights equal to the number of pairs that were used to calculate the average r^2. In addition we estimated the correlation between pairwise r^2 across the two samples for those pairs of SNPs that were polymorphic (MAF $\geq$ 0.05) in both samples. This scenario is analogous to that in which the correlation is estimated from two samples that are essentially from the same (ancestral) population (Evans & Cardon, 2005). For each sample size we performed 1000 replicate simulations.

Results

In Appendix I we show the expectation and variance of the number of polymorphic SNPs as a function of sample size and mutation rate, in the absence of recombination. In Tables 1–3 we present results for a gene of length 16.5 kb (equal to the average length of the genes studied by Crawford *et al*., 2004). Replicate simulations can be thought of as representing a different (and not closely linked) gene, so the variation between replicates is the same as between-gene variation. In Table 4 we present the results of the experimental sampling variation of the estimation of recombination rate from population data and that of the estimation of LD.

Table 1 Mean (and SD) of polymorphism and haplotype variation as a function of the sampled number of chromosomes (n)*

n	No. SNPs	No. SNPs for MAF $\geq$ 5%	No. haplotypes for MAF $\geq$ 5%	Eff. no. haplotypes for MAF $\geq$ 5%
47	75.4 (18.0)	49.1 (16.7)	15.0 (2.6)	10.1 (2.5)
75	82.9 (18.1)	51.1 (16.7)	17.4 (2.9)	11.0 (2.6)
100	88.0 (18.6)	51.9 (16.9)	18.7 (3.2)	11.3 (2.8)
150	94.8 (18.6)	49.9 (16.7)	19.5 (3.4)	11.0 (2.7)
200	99.8 (18.8)	50.9 (16.5)	20.7 (3.6)	11.2 (2.7)
1000	127.7 (19.6)	50.5 (16.9)	25.5 (4.6)	11.2 (2.8)

*From 10,000 simulations.

Table 2 Differences between two sub-samples of size n, both sampled from the same population. The MAF $\geq$ 5% threshold was applied to each sample separately*

	SD of difference between the sub-samples			
n	No. SNPs for MAF $\geq$ 5%	No. haplotypes for MAF $\geq$ 5%	Eff. no. haplotypes for MAF $\geq$ 5%	Correlation between no. haplotypes
47	7.6	2.6	2.4	0.51
75	6.7	2.7	2.3	0.59
100	6.2	2.7	2.1	0.64
150	5.5	2.6	1.8	0.71
200	5.1	2.6	1.7	0.75
1000	3.3	2.3	1.0	0.87

*From 10,000 simulations.

Table 3 Haplotype diversity (and SD) in two sub-samples of size n that were obtained from drawing a sample of $2n$ chromosomes, when MAF $\geq$ 5% for the combined sample

$2n$	Total no. haplotypes	No. shared haplotypes	No. unique haplotypes to sub-sample	Prop. common haplotypes in sub-samples	Prop. chromosomes in shared haplotypes
94	18.1 (3.1)	12.3 (2.1)	2.9 (1.7)	0.81 (0.09)	0.903 (0.066)
150	19.4 (3.5)	14.2 (2.4)	2.6 (1.6)	0.85 (0.08)	0.947 (0.038)
200	20.7 (3.6)	15.7 (2.6)	2.5 (1.6)	0.87 (0.08)	0.962 (0.029)
300	22.0 (3.9)	17.2 (2.9)	2.4 (1.6)	0.88 (0.07)	0.976 (0.018)
400	22.8 (4.1)	18.1 (3.2)	2.4 (1.5)	0.89 (0.07)	0.983 (0.013)
2000	27.8 (4.9)	23.4 (4.1)	2.2 (1.5)	0.92 (0.05)	0.997 (0.003)

*From 10,000 simulations.

Table 4 Linkage disequilibrium differences between two sub-samples of size n from the same population that were obtained from drawing a sample of $2n$ chromosomes, when MAF $\geq$ 5% for each sample

n	SD of difference in estimate of ρ	Correlation between r^2 across samples
50	24.2	0.88
75	20.3	0.92
100	19.1	0.94
200	14.6	0.97

*From 1000 simulations.

Table 1 shows results concerning variation in diversity across genes, within a single sample of n chromosomes. The standard deviations (SD) are relatively large, and reflect both evolutionary and experimental sampling. For any sample size between $n = 47$ and $n = 1000$, the number of SNPs with a minor allele frequency (MAF) of 5% or more has a mean of approximately 51 and an SD of approximately 17, so the coefficient of variation (CV) is about one-third. Thus, the typical range of the number of sampled SNPs across genes of the same size (16.5 kb) is very large (approximately 51 ± 2 SD, so ~ 17 to ~ 85), corresponding to five-fold differences in the density of detected SNPs. The differences would be even larger under a more realistic model that allowed variation in rates of recombination and mutation, and effects of natural selection. These results are not too surprising; it is well known in the theoretical population genetics literature that the variance in the total number of segregating sites (SNPs) in the whole population is at least as large as the mean, regardless of the recombination rate (Ewens, 2004; Watterson, 1975). When there is no recombination, the CV of the number of SNPs with MAF $\geq$ 0.05 can be calculated using previous results (Fu, 1995). Figure 1 shows the CV of the number of SNPs as a function of the mutation rate (θ) for the gene, and the experimental sample size n, for MAF of $\geq$ 0.00 and $\geq$ 0.05. When SNPs are selected with MAF $\geq$ 0.05 then the CV is approximately 40–50% and fairly independent of sample size and mutation rate (θ). Hence, relative to the average number of common SNPs identified per gene in a sample of chromosomes, there will be a large amount of variation between genes, even under the simple model employed here. When there is complete ascertainment of SNPs (MAF $\geq$ 0.00) then even with $n = 1000$ chromosomes the CV is approximately 20%, so that large differences in SNP density across genes will be observed.

The results in Table 2 represent the situation where the same gene is sequenced in two sub-samples drawn from the same population. (This is the situation under the null hypothesis in an association study, for example.) The expected value of each of the diversity measures is the same for each sub-sample, because they are from the same population, and therefore we only present the standard deviation of the difference. We see that there is considerable variation in the number of haplotypes found in each sub-sample, and that the level of variation is fairly constant (~ 2.6) with sample size. This occurs because, although the frequency of common SNPs and haplotypes are more precisely estimated with a larger sample size, more relatively rare SNPs (say, 0.05 < MAF < 0.10) are detected in each sub-sample, and haplotypes that include a number of these relatively rare SNPs tend not be shared between the sub-samples. The correlation across genes in the number of haplotypes between the two sub-samples is relatively low, unless the sample size is very large. If two

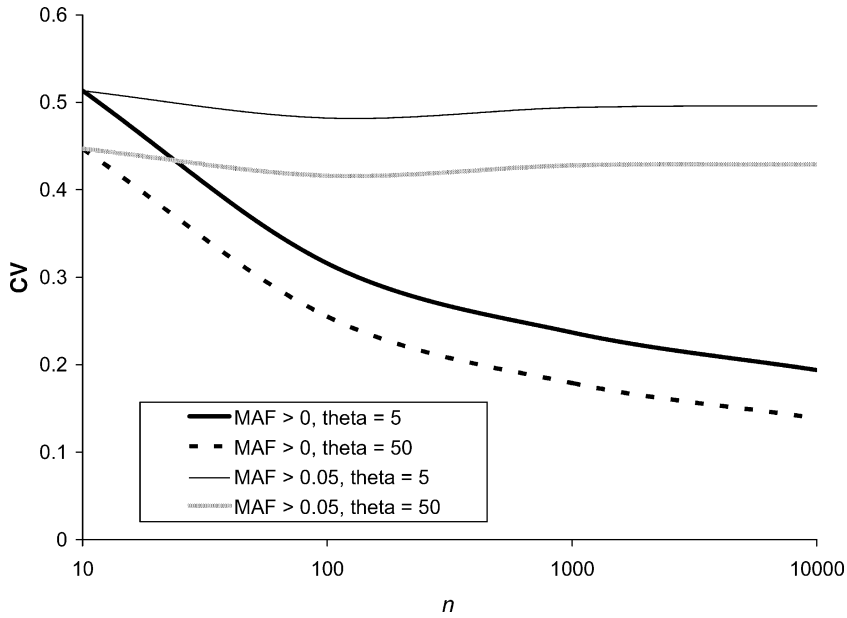

Figure 1 Coefficient of variation (CV) of the number of segregating sites (SNPs) as a function of theta (θ) $= 4N_e\mu$ and experimental sample size (n), for MAF $\geq$ 0.00 and MAF $\geq$ 0.05.

small samples from different populations were studied in this way, observing a correlation coefficient that is significantly smaller than one therefore does not provide any evidence that the two populations are genetically distinct.

Table 3 shows measures of haplotype diversity and differences between the sub-samples, calculated using SNPs above a threshold MAF $\geq$ 0.05 in the combined sample of size $2n$. If the two sub-samples are from the same population then selecting common SNPs in the combined sample is unlikely to result in the complete absence of the polymorphisms in a sub-sample. Hence most SNPs and haplotypes are represented in both samples. The proportion of haplotypes in a sub-sample that is shared with the other sample ranges from 0.81 to 0.92. The proportion of chromosomes in each sub-sample that are in shared haplotypes is very large, ranging from 0.903 ($2n = 94$) to 0.997 ($2n = 2000$).

Measures of SNP and haplotype diversity within and between genes are correlated with measures of linkage disequilibrium (LD). Table 4 shows the variation in measures of LD across sub-samples. The SD of the difference in the estimate of the scaled recombination rate (ρ) from the two sub-samples from the same population is large (about 20% of the true parameter value). The SD of the estimate of ρ from a regression analysis within each sub-sample was 20.9, 21.2, 20.4 and 21.0 for sample sizes of 50, 75, 100 and 200 chromosomes, respectively (results not shown in Table 4). The correlation between the estimates of r^2 across the sub-samples was relative large (>0.88) for all sample sizes, presumably because pairs of SNPs were only included in the calculation if they were both segregating in each sub-sample and if their minor allele frequency was at least 0.05. For a commonly used measure of pairwise LD, r^2, the total sampling variance can be explicitly decomposed into evolutionary sampling variance and experimental sampling variance as follows,

$$E(r^2) = \mathrm{var}(r) = 1/(2 + \rho) + 1/n$$

(following Weir & Hill, 1980). This expression suggests that experimental sample size can only be neglected when $(2 + \rho)/n$ is small. When $\rho = 5$ and $n = 47$, this ratio is 0.15. Hence, for two SNPs at a distance of 16.5 kb, the total sampling variance of their correlation is approximately 0.16 (SD = 0.41) of which 85% is because of evolutionary sampling (i.e. what would be measured if n was very large) and 15% because of experimental sampling.

Discussion

Coalescent theory tells us that (under the standard neutral model) there will always be a large sampling variance in polymorphism and haplotype diversity measures. This is because the number of mutations that generate the variation is mostly determined by the evolutionary times between the oldest few events in the genealogy, which do not depend on sample size. Here we have quantified sampling variance in diversity measures for a realistic set of parameters, assuming a standard neutral model of molecular evolution. Sampling variance arises from two sampling processes, evolutionary sampling and experimental sampling.

Evolutionary sampling means that, even if two independent genes have identical mutation and intragenic recombination rates, *whole population samples* will often have different diversities because those two genes will often have very different histories. Because we cannot control evolutionary sampling, statistical tests based on population genetic models are needed before we can infer that genes differ in the rate of mutation or recombination. Experimental sampling is caused by taking a finite sample of chromosomes from the population of all chromosomes.

Experimental sampling can be controlled. The relatively low correlation between the number of haplotypes in sub-samples ranging in size from $\sim$50 to 200 chromosomes implies that many unique haplotypes will be detected in sub-samples, even if they are drawn from the same source population. This is important because it is the null hypothesis in a case and control or association study. For example, even for samples of 1000 case and 1000 control chromosomes, when SNPs are selected in each sample for MAF $\geq$ 0.05 there will be a number of rare SNPs and haplotypes present in one sample but not the other: for this sample size the SD for the number of SNPs and haplotypes is 3.0 and 2.6, respectively, and the correlation between the number of haplotypes is 0.84 (Table 3). When the MAF is restricted in the combined sample of 2000 chromosomes, the number of unique haplotypes in the case sample is 2.1 and about 8% of all haplotypes in the case sample are not present in the control sample (Table 3). However, the proportion of chromosomes in the case sample that are in shared haplotypes is very large (99.7%, Table 3).

The effect of evolutionary sampling is, for the range of parameters considered, much larger than that of experimental sampling. This implies that the strategy of the HapMap and other projects to focus on multiple samples from different populations with relatively small experimental sample sizes is appropriate. A larger sample from a single population would reduce the effect of experimental sample size, but at the expense of being able to draw inference between populations. SNPs genotyped in the HapMap project are mostly common (say, MAF $\geq$ 0.01), and a subset of them are subsequently selected for candidate gene or genome-wide association studies in large cohorts of cases and controls. The focus on common SNPs and large sample sizes will reduce the amount of chance variation in the number of segregating polymorphisms between the case and control samples. Although we have not considered the specific HapMap ascertainment procedure in the simulations, the results from Tables 1 and 2 for MAF $\geq$ 0.05 indicate that for $n = 1000$ (the size of a powerful case-control study), the standard deviation of the difference in the number of segregating common SNPs is about 3, relative to an average of 50 under complete ascertainment (Table 1). Hence, common variation is well captured in powerful case-control studies.

Throughout we have assumed that haplotypes are observed without error. In practice haplotypes are usually estimated, with some error, from multi-locus genotype data. Hence, when considering the total variance of haplotype-based statistics, there is an additional source of error in addition to these considered in this study.

Our simulations assumed all genes have a constant mutation rate and constant recombination rate. In reality, of course, genes differ in physical length and this will inflate variation in SNP and haplotype diversity. For example, the 100 genes from the Crawford *et al*. (2004) study have mean length 16.5 kb with SD 9.9 kb, showing considerable variation. In that study gene length was highly (and significantly) positively correlated with the number of SNPs and number of haplotypes. Correlations between gene length and the number of SNPs and haplotypes ranged from 0.67 to 0.87 in the two population samples (results not shown in Tables 1–3). These findings are consistent with a simple (e.g., neutral) model of sequence variation with a constant per site mutation rate, which would predict that

the expected number of polymorphisms is proportional to gene length (Ewens, 2004; Watterson, 1975). This expectation also holds when there is recombination (although the variance is reduced). When not corrected for gene length, the correlation between the number of haplotypes for the 100 genes in the two samples was 0.79 (Crawford *et al*. 2004). Interestingly, after a correction for gene length this correlation was 0.58, not significantly different from what would be expected if neutral DNA sequences were sampled from a single homogeneous (panmictic) population (see Table 2).

A different approach to quantify the effect of experimental sampling is to partition the observed between-gene variation in the number of SNPs (or haplotypes) into an evolutionary and experimental variance component. If V is the sampling variance for any diversity measure, then $V = V_{\mathrm{EVO}} + V_{\mathrm{EXP}}$, where V_{EVO} and V_{EXP} are caused by evolutionary and experimental sampling respectively, which are a function of N and n, respectively. For large experimental sample size V_{EXP} approaches zero but V_{EVO} does not, even if the entire population is sampled. For the number of SNPs with MAF $\geq$ 0.05, the total between-gene variance V is approximately $16.7^2 = 279$ for $n = 47$ and $16.9^2 = 286$ for $n = 1000$ (from Table 1). When a sample is split into two sub-samples, the variance between the sub-samples is $2V_{\mathrm{EXP}}$, and reflects experimental sampling only. This variance between sub-samples is approximately $7.6^2 = 58$ for $n = 47$ and $3.3^2 = 10$ for $n = 1000$ (from Table 2). Hence, the proportion of between-gene variation in the number of SNPs with MAF $\geq$ 0.05 that is attributable to evolutionary sampling is 0.90 for $n = 47$ and 0.98 for $n = 1000$.

Evans & Cardon (2005) compared LD patterns across populations empirically, and suggested that the lack of concordance between different groups may be due both to real differences and as a consequence of the LD measures themselves. They reported correlations between r^2 estimates from samples from different populations in the range of 0.73 to 0.95. The lowest correlation was from a comparison of r^2 estimates between a sample of 97 unrelated African-Americans and a sample of 42 Asian individuals. The highest correlation was from samples of 96 unrelated U.K. individuals and 46 individuals of European ancestry. Similarly, large correlations were reported from a comparison of a HapMap and Finnish sample (Willer *et al*. 2006). In this study we have quantified the correlation of LD measures across samples from the same population (Table 4), and show that the expected correlation between r^2 estimates from different samples is approximately 0.9 if the samples are from the same population and if each sample size is less than 100 chromosomes. The reported correlation coefficient of 0.95 between r^2 from two European populations by Evans & Cardon (2005) is consistent with our simulation results (Table 4), and suggests that the deviation from 1.0 is solely due to experimental sampling.

In conclusion, even with complete ascertainment of all genetic variation in a gene the sampling variation in the number of polymorphisms and haplotypes, and linkage disequilibrium remains large, and inference with respect to differences between samples of chromosomes should take this into account by formally testing the null hypothesis that the samples are from the same population. For simple statistics, such as the observed number of polymorphisms or haplotypes in a particular gene, standard statistical tests can be used. To test LD between samples, a parameter that captures the relationship between SNP association and distance can be estimated (as performed in this study) and tested across samples. Alternatively, parameter estimates can be compared to the results from (coalescent) simulations.

Acknowledgements

This work was supported by the UK Biotechnology and Biological Research Council and the Wellcome Trust. We thank Toby Johnson for explaining coalescent theory and for many other helpful contributions to the study. We thank Bill Hill and Jay Taylor for helpful discussions and Dana Crawford and Deborah Nickerson for helpful feedback on an earlier version of the manuscript. We are grateful for the constructive comments of the referees.

References

Altshuler, D., Brooks, L. D., Chakravarti, A., Collins, F. S., Daly, M. J. & Donnelly, P. (2005) A haplotype map of the human genome. *Nature* **437**, 1299–1320.

Crawford, D. C., Akey, D. T. & Nickerson, D. A. (2005) The patterns of natural variation in human genes. *Annu Rev Genomics Hum Genet* **6**, 287–312.

Crawford, D. C., Carlson, C. S., Rieder, M. J., Carrington, D. P., Yi, Q., Smith, J. D., Eberle, M. A., Kruglyak, L.

& Nickerson, D. A. (2004) Haplotype diversity across 100 candidate genes for inflammation, lipid metabolism, and blood pressure regulation in two populations. *Am J Hum Genet* **74**, 610–622.

Evans, D. M. & Cardon, L. R. (2005) A comparison of linkage disequilibrium patterns and estimated population recombination rates across multiple populations. *Am J Hum Genet* **76**, 681–687.

Ewens, W. (2004) *Mathematical population genetics*. New York: Springer-Verlag.

Fu, Y. X. (1995) Statistical properties of segregating sites. *Theor Popul Biol* **48**, 172–197.

Hartl, D. L. & Clark, A. G. (1997) *Principles of population genetics*. Sunderland: Sinauer Associates.

Hill, W. G. (1975) Linkage disequilibrium among multiple neutral alleles produced by mutation in finite populations. *Theor Popul Biol* **8**, 117–126.

Hudson, R. R. (1985) The sampling distribution of linkage disequilibrium under an infinite allele model without selection. *Genetics* **109**, 611–631.

Hudson, R. R. (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338.

McVean, G. A. (2002) A genealogical interpretation of linkage disequilibrium. *Genetics* **162**, 987–991.

Ohta, T. & Kimura, M. (1969) Linkage disequilibrium at steady state determined by random genetic drift and recurrent mutation. *Genetics* **63**, 229–238.

Watterson, G. A. (1975) On the number of segregating sites in genetical models without recombination. *Theor Popul Biol* **7**, 256–276.

Willer, C. J., Scott, L. J., Bonnycastle, L. L., Jackson, A. U., Chines, P., Pruim, R., Bark, C. W., Tsai, Y. Y., Pugh, E. W., Doheny, K. F., Kinnunen, L., Mohlke, K. L., Valle, T. T., Bergman, R. N., Tuomilehto, J., Collins, F. S. & Boehnke, M. (2006) Tag SNP selection for Finnish individuals based on the CEPH Utah HapMap database. *Genet Epidemiol* **30**, 180–190.

Appendix I: Expected Number of Segregating Sites from a Sample of Size *n*

Under the infinite-sites neutral model of evolution, in the absence of within-gene recombination, the mean and variance of the total number of segregating sites (S) is,

$$E(S) = \theta \sum_{i=1}^{n-1} 1/i \quad \text{and} \quad \text{var}(S) = E(S) + \theta^2 \sum_{i=1}^{n-1} 1/i^2$$

(Watterson 1975; Fu 1995). It follows that the standard deviation of the number of segregating sites, scaled by the mean number of segregating sites, i.e. the coefficient of variation (CV), is

$$\sigma(S)/E(S) = CV(S) = \left[\frac{1}{\theta \sum_{i=1}^{n-1} 1/i} + \frac{\sum_{i=1}^{n-1} 1/i^2}{\left(\sum_{i=1}^{n-1} 1/i \right)^2} \right]^{0.5}$$

When only segregating sites are counted for which there are between k and l ($0 < k < l$, $l < n$) copies in the sample, the mean number of sites is,

$$E(S|k,l) = \theta \sum_{i=k}^{l} 1/i$$

(from Fu, 1995). The variance of S can be calculated using equations given by Fu (1995), using $S\,|\,k,l = \sum_{i=k}^{l} S_i$ and the variances of S_i and covariances between S_i and S_j. In the presence of recombination, the equations for the mean number of segregating sites are still valid but the variance decreases.

Received: 7 August 2005
Accepted: 12 May 2006