Evidence for population growth in humans is confounded by fine-scale population structure

Susan E. Ptak and Molly Przeworski

Although many studies have reported human polymorphism data, there has been no analysis of the effect of sampling design on the patterns of variability recovered. Here, we consider which factors affect a summary of the allele-frequency spectrum. The most important variable to emerge from our analysis is the number of ethnicities sampled: studies that sequence individuals from more ethnicities recover more rare alleles. These observations are consistent with fine-scale geographic differentiation as well as population growth. They suggest that the geographic sampling strategy should be considered carefully, especially when the aim is to infer the demographic history of humans.

DOI: 10.1016/S0168-9525(02)02781-6

Susan E. Ptak
Molly Przeworski*
Max Planck Institute for Evolutionary Anthropology,
Inselstrasse 22,
04103 Leipzig, Germany.
*e-mail: przewors@eva.mpg.de

Genetic variation among extant humans carries information about the evolutionary history of our species. Unlinked regions in the genome represent independent realizations of this evolutionary process and thus, with polymorphism data from enough loci, it should be possible to infer many aspects of our evolution [1–4]. Conversely, a better understanding of the evolutionary history of humans should help us to predict patterns of genetic variability, thereby aiding in the design and interpretation of genome-wide association studies [5,6]. It will also help us to interpret polymorphism data from regions of the genome that have experienced natural selection [7].

With these myriad goals in mind, researchers have collected polymorphism data from more than 400 regions of the human genome in over 40 studies. The loci have been sequenced in different laboratories and distinct strategies have been implemented regarding the number and variety of geographic sampling localities, the number of individuals considered and so forth. The patterns of variability recovered have been extremely varied. To some extent, this variability is expected: patterns of polymorphism will differ greatly from locus to locus by chance, even if they have been generated by exactly the same evolutionary process [8,9]. However this variance might also reflect differences among study designs. If there are aspects of the sampling strategy that influence patterns of variation, their identification should inform the design of future studies. It can also point to important features of the evolutionary history of human populations [10].

It is commonly quoted that 85% of human genetic diversity is found within populations [11], a finding usually interpreted as evidence that human populations are genetically very similar to one another [12]. Although this is certainly true (if only because, on average, two humans are identical at 99.9% of their DNA), this level of population structure is sufficient to have profound effects on levels of linkage disequilibrium in some contexts [5,13]. Furthermore, a high proportion of alleles seem to be specific to samples from single populations [14]. It therefore seems plausible that the geographic sampling scheme influences the allele-frequency spectrum as well as levels of allelic associations. To investigate this possibility, we tabulated a widely used summary of the allele-frequency spectrum, Tajima's D [15] (Box 1), henceforth abbreviated as TD, for all large studies of humans. The studies that we included reported an overall value of TD, including synonymous, non-synonymous and non-coding sites (see legend of Fig. 1). Our choice of TD as a statistic was motivated in part by convenience – almost all studies reported the value of TD – and partly because TD has been shown to be sensitive to demographic history [16,17].

Importance of geographic sampling design

To evaluate the effect of the number and variety of populations sampled, we categorized the studies as geography or population based. To avoid as far as possible decisions about what constitutes a population, we defined a population-based sample as one that included more than ten individuals each from at most three localities (e.g. Ref. [1]). We defined geography-based samples as those containing at most ten individuals each from at least three sampling localities (e.g. Ref. [18]). Most studies can be neatly categorized as one or the other, except for five, three of which sample various ethnicities in the USA (e.g. Ref. [14]). We group these latter three in an additional category, labeled 'USA'. Figure 1 presents the distribution of TD values for geography-, USA- and population-based surveys of polymorphism. The mean TD value is significantly lower in geography-based samples than in population-based ones ($t = 3.99, P < 0.001, \mathrm{df} = 44.9$). Geography-based surveys also have a lower mean TD value than USA-based samples ($t = 2.56, P = 0.011, \mathrm{df} = 350$). Surveys that sample few individuals from many localities appear to recover more rare alleles than those that sample many individuals from few localities.

PII: S0168-9525(02)02781-6

Box 1. Tajima's *D*

Tajima's *D* (*TD*) is the (approximately) normalized difference between π, the average nucleotide heterozygosity [a], and θ_W, a measure of diversity based on the number of segregating sites in the sample [b]. Under the standard neutral model of a randomly mating population of constant size, the expected value of *TD* is roughly 0 for all sample sizes [c]. *TD* will be negative if there are more rare alleles than expected under the standard neutral model. This excess is sometimes referred to as a skew in the allele-frequency spectrum towards rare alleles. Such a skew is expected under models of population growth [d,e], as well as if the polymorphic sites are under weak purifying selection [f]. Under these models, the expectation of *TD* is no longer independent of sample size: larger samples will tend to have a higher proportion of rare alleles and therefore more-negative *TD* values [g,h].

References

[a] Tajima, F. (1983) Evolutionary relationship of DNA sequences in finite populations. *Genetics* 105, 437–460

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

[c] Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595

[d] Tajima, F. (1989) The effect of change in population size on DNA polymorphism. *Genetics* 123, 597–601

[e] Slatkin, M. and Hudson, R.R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129, 555–562

[f] McVean, G.A. and Charlesworth, B. (2000) The effects of Hill–Robertson interference between weakly selected mutations on patterns of molecular evolution and variation. *Genetics* 155, 929–944

[g] Simonsen, K.L. *et al.* (1995) Properties of statistical tests of neutrality for DNA polymorphism data. *Genetics* 141, 413–429

[h] Gillespie, J.H. (2000) Genetic drift in an infinite population. The pseudohitchhiking model. *Genetics* 155, 909–919

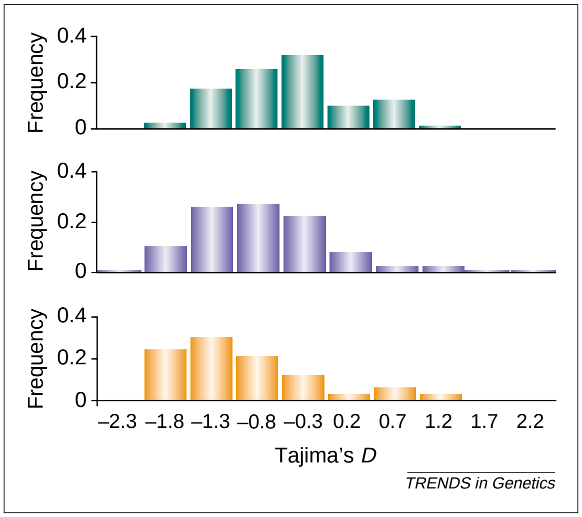


Fig. 1. The frequency distributions for Tajima's *D* (*TD*) for population-, USA- and geography-based sampling schemes (from top to bottom). The data examined include 437 loci from single-nucleotide polymorphism (SNP) detection studies of autosomal and X-linked loci (published or communicated to us by April 15th 2002) with at least five segregating sites (the data are available from http://email.eva.mpg.de/~przewors/). Not all studies rely on resequencing to identify variants, with some using DHPLC and others variant detection arrays. However, all studies call each base pair in each individual, rather than typing a prescreened set of SNPs. Requiring a minimum number of segregating sites is meant to ensure that every dataset is informative. Non-recombining loci (including mitochondrial DNA and the Y chromosome) are excluded, because they are especially susceptible to the effects of natural selection. In the 11 cases in which the same region was considered by different researchers, or the promoter and coding region of the same gene were sequenced, we choose studies based on the following criteria: (1) we kept the sampling scheme that we could most easily classify; (2) if there was no difference in that respect, we kept the sample with the most segregating sites (if the number of segregating sites was not published for one of the versions, we discarded it). All the parametric analyses use a log-transformed value of *TD* [$TD \rightarrow \ln(TD{+}3)$]. The mean values of *TD* are –0.69 for population-based sampling [standard error (SE) = 0.07, sample size (n) = 82], –0.99 for USA-based sampling (SE = 0.04, n = 319) and –1.25 for geography-based sampling (SE = 0.13, n = 33). The mean *TD* value is significantly lower in geography-based samples than in population-based ones (t = 3.99, P < 0.001, df = 44.9). Similarly, the mean *TD* value is significantly lower in geography-based samples than in USA-based ones (t = 2.56, P = 0.011, df = 350). (We report results for unequal or equal variances depending on the outcome of an *F* test for equality of variances at the 10% level; no result depends on this decision.) For both the geography- and the population-based surveys, the transformed data do not depart significantly from normality (based on a Kolmogorov–Smirnov test, P = 0.200 for both). The distribution of transformed *TD* values is only approximately normally distributed within the USA-based sampling scheme. However, nonparametric tests yield similar results to parametric analyses (results not shown). Samples for the APOE locus and data from SeattleSNP (http://pga.mbt.washington.edu/) include African-Americans and European-Americans as their two 'populations'. If we reclassify these as USA rather than population based, both comparisons are still significant at the 5% level (results not shown).

Set of predictor variables

To consider the effect of the number of sampling localities together with four other – potentially confounding – explanatory variables, we ran a multiple linear regression. To do so, we used an alternative classification of geographic sampling schemes: we counted the number of different 'ethnicities' in the sample (*nethn*), where ethnicity is defined by the label assigned to the individuals by the authors of the original paper. We also recorded the number of chromosomes sampled (*nchrom*). The human population has obviously increased in size over time, and we would expect larger samples to have more-negative *TD* values under simple growth models (Box 1). Our third variable was the proportion of the sample that is African or African-American (*prAfr*). Several researchers have remarked that the *TD* values tend to be lower in Sub-Saharan African or African-American samples than in non-African samples [1,19] (see also http://pga.mbt.washington.edu/). We also noted whether a locus is in a genic region or not (*genic*), because genic regions might evolve under higher levels of constraint.

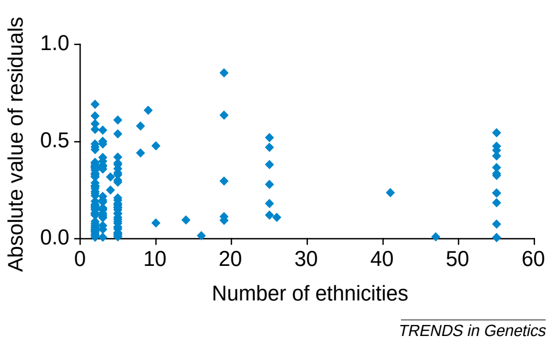
Finally, we considered whether the locus was X-linked (*sex*). Under simple growth models, samples from X-linked loci are expected to have more-negative *TD* values than do autosomal loci. Indeed, as long as there is little sexual selection acting on males, the effective population size of X chromosomes will be smaller than that of the autosomes [20]. The genealogical history of X-linked loci will therefore be shorter. As a result of this shorter history, samples from X-linked loci will tend to reflect recent changes (e.g. in the population size) more than autosomal ones, because a larger proportion of their history will take place during these changes.

Best predictor variable: *nethn*

To identify which variables among the five (*nethn, nchrom, prAfr*, *genic* and *sex*) were the best predictors

Box 2. Validity of the assumptions of a linear regression model of *nethn* on *D*, excluding data from Stephens, *et al.* [a].

Linear regression assumes a linear relationship between *nethn* and the recoded *TD* values, as well as constant variances across *nethn* values. A visual inspection of the Fig. I suggests that the first assumption is reasonable. In addition, the scatter along the *y* axis appears to be approximately constant. To confirm this impression, we used several tests that consider whether there is a monotonic increase or decrease in variance as a function of *nethn*: the modified Levene test ($t = 0.33$, $P = 0.746$, df = 158); the Breusch–Pagan test ($X^2 = 1.09$, $P = 0.296$, df = 1); and a rank correlation of absolute residuals and *nethn* (Kendall's $\tau = 0.012$, $P = 0.838$, df = 160). Furthermore, a Kolmogorov–Smirnov test and a visual inspection confirm that these residuals show no statistical deviation from normality ($KS = 0.063$, $P = 0.200$, $n = 160$), as assumed by the analysis.



TRENDS in Genetics

Fig. I. Plot of the absolute value of the unexplained variation in the transformed Tajima's *D* (*TD*) values in a regression model with the number of ethnicities (*nethn*).

The regression estimates are sensitive to outliers. We identified influential outliers using a range of approaches (studentized residuals, leverage statistics, Cook's values, standardized dfBeta and standardized dffit). Three loci (*CYP1A1*, *FOXP2* and tubulin M40) were identified as outliers by at least one of these methods. Removing these loci and rerunning the regression did not substantially alter the results ($F = 42.15$, $P < 0.001$, $n = 156$). The analysis further assumes that the loci are evolving independently, so that the covariance in *TD* values caused by shared genealogical history is 0. This might not be the case if loci are closely linked, as for many loci sequenced by SeattleSNPs (http://pga.mbt.washington.edu/), or if there are high levels of population differentiation [b]. It is unclear what effect this covariance might have. If we exclude the loci from SeattleSNPs, only *nethn* is identified as a predictor variable in multiple regression analyses (results not shown).

References

a Stephens, J.C. *et al.* (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293, 489–493

b Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69, 1–14

of *TD* values, we performed a forward stepwise regression (*P*-to-enter = 0.05, *P*-to-remove = 0.10) and a backwards regression (*P*-to-remove = 0.05). Both procedures identify only *nethn* as explaining a significant proportion of the variance in transformed *TD* values (see legend of Fig. 1 for details of transformation). This is also true if the stepwise and backwards regressions are implemented with less-stringent criteria (*P*-to-enter = 0.10 and *P*-to-remove = 0.15, and *P*-to-remove = 0.10, respectively).

Linear-regression analyses make several assumptions (Box 2) that are unlikely to be met exactly in this context. We therefore verified that our analyses were valid by running several standard diagnostics on a regression model of *TD* values on *nethn*. These tests revealed that the variance of the error terms was not constant across values of *nethn*. We addressed this problem in two ways. First, we performed a weighted regression using a linear regression of the absolute value of the residuals on *nethn* to estimate the weights [21] (see legend of Fig. 2). None of the conclusions changed and the diagnostics suggested that the assumptions were then met (results not shown). We also reran the stepwise and backwards regression analyses excluding the data from Ref. [14]. Without these data, there is no evidence of a departure from the assumptions of linear regression in a regression model of *TD* values on *nethn* (Box 2). Again, both stepwise and backwards regressions identify only *nethn* as a predictor variable but, with less stringent criteria, both stepwise and backwards regressions select *prAfr* and *nethn*. In the following analyses, we exclude the data from Ref. [14].

A regression of transformed *TD* values on *nethn* alone explains an appreciable proportion of the variance ($R^2 = 0.22$). *TD* values decrease sharply with increasing number of ethnicities ($t = -6.76$, $P < 0.001$, df = 158). Thus, using the number of ethnicities to classify geographic sampling schemes, we again find that more-diverse samples contain a higher proportion of rare alleles.

Contrary to our expectation, we do not detect a relationship between *nchrom* and *TD* values. To gain a sense of how unexpected this result is under simple growth models, we simulated a model with random mating and constant population size, followed by a period of exponential growth to the present (c.f. [19]). We explored a range of parameters, constrained to yield roughly the same mean *TD* value as is observed (–0.81, excluding data from Ref. [14]). For a given set of parameters, we generated a *TD* value for each locus using coalescent simulations with a fixed number of segregating sites [22], (conservatively) assuming no recombination within each locus. We then tabulated the proportion of 500 runs in which the regression coefficient of *TD* on *nchrom* was significantly different from 0 at the 5% level. Over the range of parameters explored, this proportion was always greater than 60%. In addition, in a linear regression of *TD* on *nchrom*, over 90% of simulated runs had a lower significance level than is observed

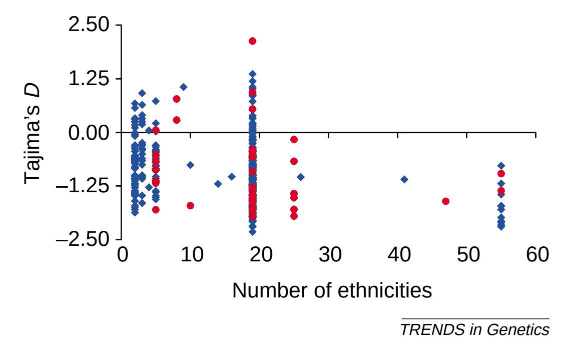


Fig. 2. A plot of Tajima's *D* (*TD*) against the number of ethnicities (*nethn*) for all 437 loci. A weighted regression of *TD* on *nethn* is highly significant ($F=41.72$, $P<0.001$, $n=437$). A non-parametric analysis leads to the same finding (Kendall's $\tau=-0.22$, $P<0.001$, $n=437$). This plot is color coded to distinguish autosomal (blue) and X-linked (red) loci. As suggested by a visual inspection of this plot and the results of the multiple regression, sex linkage does not have a detectable effect on *TD* values.

(results not shown). Thus, our results are surprising under simple growth models. However, they might be consistent with a model of population growth with population structure.

Other potential predictor variables

In a regression model with *nethn* and *prAfr*, *TD* decreases with increasing proportion of Africans and African-Americans ($t=-1.83$, $P=0.070$, df = 156), consistent with previous observations that *TD* values are lower in Sub-Saharan African populations. Examination of the residuals suggests that a linear relationship between *TD* and both independent variables is appropriate.

To confirm that genic and non-genic regions do not differ in mean *TD* values, we compared the 20 intergenic regions from Ref. [1] to the 58 genes from SeattleSNPs (http://pga.mbt.washington.edu/). These two studies have similar sampling strategies (two versus three ethnicities, respectively). No difference is found between mean *TD* values in genic and non-genic regions (as tested by *t* tests or Mann–Whitney *U* tests, $P>0.15$). This classification might be too crude to capture levels of constraints – what we classify as genic includes a large fraction of non-coding sequence, which might evolve more similarly to intergenic regions than to coding regions [23].

If the three different methods of SNP detection used in these variation surveys have different sensitivities, we might expect ascertainment method to be an additional predictor variable (see supplementary materials at http://email.eva.mpg.de/~prezewors). The mean *TD* value differs between variant detection array and sequencing as well as between denaturing high performance liquid chromatography (DHPLC) and sequencing studies ($P<0.05$; results not shown). However, the mean number of ethnicities also differs between ascertainment treatments. Using the regression estimates from a regression of *TD* on the number of ethnicities for sequencing-based studies, we can predict the mean difference in *TD* values between treatments given the mean difference in number of ethnicities. These estimates are extremely close to the observed values (results not shown). Thus, this analysis provides no clear evidence that the detection method per se has an influence; rather, the difference in mean *TD* values appears to be due to differences in the mean number of ethnicities.

Although much of the unexplained variance in *TD* values is probably due to chance variation, there are undoubtedly other predictor variables. In particular, for a subset of loci considered here, there is independent evidence for natural selection (e.g. at genes involved in resistance to malaria). In that respect, it is noteworthy that *FOXP2* is an outlier in our regression of *TD* on the number of ethnicities (Fig. 2), because patterns of polymorphism and divergence at this gene suggest the action of recent positive selection [24].

Implications

As more ethnicities are sampled, the proportion of rare alleles increases. Thus, the observation of low levels of differentiation among human populations does not imply that the geographic sampling design is unimportant. As defined, the number of ethnicities accounts for roughly one-fifth of the variance in transformed *TD* values, our summary of the frequency spectrum. Whether ethnicities as defined by the authors of the study are truly discrete populations will require much more research (e.g. [25]). It seems more likely that the number of ethnicities fortuitously turned out to be a good proxy for underlying population differentiation, perhaps because surveys that included more 'ethnicities' sampled more locations on the globe.

The median *TD* value across the 437 loci is –1.03, reflecting at most loci a skew towards rare alleles relative to the expectations for a randomly mating population of constant size. This skew in the allele-frequency spectrum has widely been interpreted as evidence for an increase in the population size beginning some 20 000–100 000 years ago [2,14,19,26–29]. Several researchers have used the frequency spectrum to estimate the onset of population growth under simple models [19,29,30]. With few exceptions [2,19], no attention has been paid to the additional assumption of random mating. The analyses reported here suggest that the skew in the frequency spectrum reflects fine-scale population differentiation as well as population growth. One consequence is that estimates of the onset of growth that assume random mating will be erroneous. In that light, it is interesting that autosomal samples from what are putatively single non-African ethnicities do not exhibit a skew in the allele-frequency spectrum towards rare alleles

[1,19], whereas a Hausa sample from Cameroon [1] shows only a slight one. Sampling one population does not circumvent the confounding effects of geographic structure; estimates of the onset of growth will be affected as long as some ancestors of the sample migrated from other populations [17]. However, the absence of a clear signal of growth in samples from single ethnicities, together with the results reported here, suggests that we currently have little unequivocal evidence for ancient population growth in humans.

Simple models of population structure such as the island model suggest that pooling samples across ethnicities should lead to more-positive, not more-negative, *TD* values [8,17]. However, if population structure was recent and if there was appreciable genetic drift associated with the founding of current ethnicities, *TD* might decrease with increasing number of ethnicities. As an illustration, we explored a model in which an ancestral population split into 50 subpopulations of the same size that experienced a brief population-size reduction and, later, a population-size increase. For the sake of simplicity, we assumed that all subpopulations had the same history and that there was no migration between them. We sampled 100/*k* chromosomes from each of *k* populations, where *k* took values between 2 and 50. Coalescent simulations of such a scenario over a range of parameters suggest that the mean *TD* can decrease sharply with increasing number of populations pooled, as is observed (results not shown).

We made no attempt to fit the data to the model; such an approach would require many more summaries of the data than *TD* alone. Our purpose was merely to verify that some models of fine-scale population structure lead to the observed relationship between *nethn* and *TD*. Furthermore, although a variant of this model might be plausible for some ethnicities, it is unlikely to apply to all, because different geographic regions probably have had distinct histories [1]. In that respect, it is noteworthy that we find weak evidence that the proportion of African-American or Sub-Saharan African chromosomes in the sample influences the allele-frequency spectrum.

Our results are consistent with the claim that many alleles are found exclusively in samples from one population [14]. If such alleles occur at appreciable frequencies within ethnicities, catalogues of single nucleotide polymorphisms identified in one or few ethnicities might be of limited utility if additional ethnicities are then to be used for genome-wide association studies. More work needs to be done to understand the scale at which allele frequencies vary across the globe, both to evaluate the feasibility of current strategies for association mapping and to better understand the history of human populations.

Acknowledgements
Thanks to Y. Gilad, R. Hudson, M. Lachman, A. Przeworski and J.D. Wall for helpful discussions, as well as to P. Andolfatto, M. Beaumont, Y. Gilad, S. Pääbo and J. Pritchard for comments on the manuscript. We thank D. Cutler, A. Di Rienzo, D. Nickerson, P. Oefner, S. Pääbo and M. Saunders for sharing unpublished data, M. Nachman, C. Ober, J. Schneider and P. Shen for providing additional information about sampling schemes, and J.D. Wall for providing a coalescent program with which to explore complex demographic histories.

References

1 Frisse, L. *et al.* (2001) Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am. J. Hum. Genet.* 69, 831–843

2 Wakeley, J. *et al.* (2001) The discovery of single-nucleotide polymorphisms and inferences about human demographic history. *Am. J. Hum. Genet.* 69, 1332–1347

3 Wall, J.D. (2001) Insights from linked single nucleotide polymorphisms: what we can learn from linkage disequilibrium. *Curr. Opin. Genet. Dev.* 11, 647–651

4 Nicholson, G. *et al.* Assessing population differentiation and isolation from single nucleotide polymorphism data. *J. R. Stat. Soc.* (in press)

5 Pritchard, J.K. and Przeworski, M. (2001) Linkage disequilibrium in humans: models and data. *Am. J. Hum. Genet.* 69, 1–14

6 Goldstein, D.B. and Chikhi, L. (2002) Human migrations and population structure: what we know and why it matters. *Annu. Rev. Genomics Hum. Genet.* 3, 129–152

7 Hamblin, M.T. *et al.* (2002) Complex signatures of natural selection at the Duffy blood group locus. *Am. J. Hum. Genet.* 70, 369–383

8 Hudson, R.R. (1990) *Oxford Surveys in Evolutionary Biology* (Vol. 1), pp. 1–14, Oxford University Press

9 Slatkin, M. and Hudson, R.R. (1991) Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations. *Genetics* 129, 555–562

10 Weiss, K.M. and Clark, A.G. (2002) Linkage disequilibrium and the mapping of complex human traits. *Trends Genet.* 18, 19–24

11 Cavalli-Sforza, L.L. *et al.* (1994) *The History and Geography of Human Genes*, Princeton University Press

12 Lewontin, R.C. (1974) *The Genetic Basis of Evolutionary Change*, Columbia University Press

13 Pritchard, J.K. *et al.* (2000) Association mapping in structured populations. *Am. J. Hum. Genet.* 67, 170–181

14 Stephens, J.C. *et al.* (2001) Haplotype variation and linkage disequilibrium in 313 human genes. *Science* 293, 489–493

15 Tajima, F. (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123, 585–595

16 Tajima, F. (1989) The effect of change in population size on DNA polymorphism. *Genetics* 123, 597–601

17 Wall, J.D. (1999) Recombination and the power of statistical tests of neutrality. *Genet. Res.* 73, 65–79

18 Kaessmann, H. *et al.* (1999) DNA sequence variation in a non-coding region of low recombination on the human X chromosome. *Nat. Genet.* 22, 78–81

19 Wall, J.D. and Przeworski, M. (2000) When did the human population size start increasing? *Genetics* 155, 1865–1874

20 Caballero, A. (1995) On the effective size of populations with separate sexes, with particular reference to sex-linked genes. *Genetics* 139, 1007–1011

21 Neter, J. *et al.* (1996) *Applied Linear Statistical Models*, pp. 403–405, Irwin

22 Hudson, R.R. (1993) *Mechanisms of Molecular Evolution* (Takahata, N. and Clark, A.G., eds), pp. 23–26, Sinauer Associates

23 Chen, F.C. and Li, W.H. (2001) Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees. *Am. J. Hum. Genet.* 68, 444–456

24 Enard, W. *et al.* (2002) Molecular evolution of *FOXP2*, a gene involved in speech and language. *Nature* 418, 869–872

25 Romualdi, C. *et al.* (2002) Patterns of human diversity, within and among continents, inferred from biallelic DNA polymorphisms. *Genome Res.* 12, 602–612

26 Fay, J.C. and Wu, C.I. (1999) A human population bottleneck can account for the discordance between patterns of mitochondrial versus nuclear DNA variation. *Mol. Biol. Evol.* 16, 1003–1005

27 Harpending, H. and Rogers, A. (2000) Genetic perspectives on human origins and differentiation. *Annu. Rev. Genomics Hum. Genet.* 1, 361–385

28 Ingman, M. *et al.* (2000) Mitochondrial genome variation and the origin of modern humans. *Nature* 408, 708–713

29 Thomson, R. *et al.* (2000) Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* 97, 7360–7365

30 Rogers, A.R. and Harpending, H. (1992) Population growth makes waves in the distribution of pairwise genetic differences. *Mol. Biol. Evol.* 9, 552–569