Robustness of the inference of human population structure: A comparison of X-chromosomal and autosomal microsatellites

Sohini Ramachandran,[1] Noah A. Rosenberg,[2] Lev A. Zhivotovsky[3] and Marcus W. Feldman[1]*

[1]Department of Biological Sciences, Stanford University, Stanford, CA 94305-5020, USA
[2]Program in Molecular and Computational Biology, 1042 W. 36th Place DRB 289, University of Southern California, Los Angeles, CA 90089, USA
[3]Vavilov Institute of General Genetics, Russian Academy of Sciences, 3 Gubkin Street, Moscow 117809, Russia
**Correspondence to:* Tel: +1 650 723 1893; E-mail: sohini@stanford.edu

Date received (in revised form): 24th October 2003

Abstract

In this paper, data on 20 X-chromosomal microsatellite polymorphisms from the HGDP-CEPH cell line panel are used to infer human population structure. Inferences from these data are compared to those obtained from autosomal microsatellites. Some of the major features of the structure seen with 377 autosomal markers are generally visible with the X-linked markers, although the latter provide less resolution. Differences between the X-chromosomal and autosomal results can be explained without requiring major differences in demographic parameters between males and females. The dependence of the partitioning on the number of individuals sampled from each region and on the number of markers used is discussed.

Keywords: *AMOVA, Bayesian inference, clustering, human evolution, population divergence, X chromosome*

Introduction

Differences in patterns of human genetic variation across genetic systems — such as autosomes, the X and Y chromosomes and the mitochondrial genome — can be attributed to two main sources: (1) differences between males and females in demographic parameters such as population size and migration rate; and (2) differences across systems in the mechanism of inheritance. Past studies have reported on differences between evolution in males and females by comparing autosomal data with the non-recombining portion of the Y chromosome (NRY) and to mitochondrial DNA (mtDNA);[1–3] however, the X chromosome has generally not been utilised in these studies.

Unlike the NRY and mtDNA, the X chromosome undergoes recombination and contains numerous independent markers. Additionally, selection, if present in the uniparental systems, will affect every locus; on the X chromosome, however, it affects only those loci that are closely linked to selected sites. Consequently, the differences in variation between autosomes and the X chromosome may be more directly ascribed to male/female demographic differences than those between autosomes and the uniparental systems.

Here the HGDP-CEPH Human Genome Diversity Cell Line Panel[4] is used to test whether individual multilocus genotypes defined by X-linked markers produce different inferences about population structure from those obtained using autosomal genotypes. Results from X-chromosomal analysis of molecular variance (AMOVA) and cluster analysis (as implemented in *structure*[5]) are compared with those found using the same techniques on 377 autosomal markers.[6] These comparisons are used to study the extent to which the differences between their findings and the results reported by Rosenberg *et al.*[6] stem from differences in the mechanism of inheritance and from the smaller amount of information available on the X chromosome. This analysis leads to conclusions about the robustness of population structure inference with respect to number of microsatellite markers and the number of individuals sampled per region.

Methods and results

Data

The 1,056 individuals (677 males, 379 females) analysed by Rosenberg *et al.*[6] and Zhivotovsky *et al.*,[7] who derive from 52 populations in seven regional groups, were typed for X-linked markers. The X-chromosomal data were compared to autosomal data from these same individuals.[6,7]

The loci studied on the X chromosome consist of 20 polymorphic microsatellites — 4 di-, 2 tri- and 14 tetranucleotide repeats — from Marshfield Screening Set #10, with 5.2 per cent missing data. Three of the markers were pseudoautosomal (tetranucleotide DXYS218, tetranucleotide DXS9900 and dinucleotide DXYS154), so that the males were not hemizygous but homo- or heterozygous at these loci.

In both the autosomal data and the X-chromosomal data, markers are sufficiently widely spaced that, within individual populations, linkage disequilibrium as estimated by homozygosity-based statistics[8] is generally not observed (results not shown). Thus, these loci can be treated as independent markers.

Genetic diversity

Heterozygosity in the seven regions, computed using the unbiased estimator,[9] ranged as follows: 0.57 (America), 0.64 (Oceania), 0.67 (East Asia), 0.71 (Central/South Asia), 0.72 (Europe), 0.74 (Middle East) and 0.78 (Africa). Of the 237 total alleles in the data, 34 were confined to a single population; 29 of these 'private' alleles appeared only once in the sample. Of the 208 alleles found more than once in the sample, 7.2 per cent were exclusive to one of the seven geographical regions listed above.

AMOVA for the X chromosome

It has generally been observed that the within-population component of genetic variation, $W = 1 - F_{st}$, is the largest component of human genetic diversity.[1,6,10,11] Using *Genetic Data Analysis* (GDA)[12] and assuming Hardy-Weinberg proportions within populations, the variance of allelic indicator variables were partitioned in the same manner as was done for autosomal loci from the same individuals.[6] For the 17 non-pseudoautosomal X-chromosomal markers, the within-group variance component accounted for 87–93 per cent of variation among individuals (Table 1). Note that these values are generally smaller than the corresponding autosomal values in Table 1 of Rosenberg *et al.*[6] (Table 2).

This observation may be explained by a faster rate of genetic drift for X-chromosomal markers, by comparison to that for autosomal markers. Because populations contain fewer copies of X chromosomes than of any given autosomes, drift may proceed more rapidly for X-chromosomal markers, leading to greater X-chromosomal differentiation across populations and larger among-population and among-region variance components.

This argument can be investigated using Slatkin's[13,14] formulation of F_{st} in a set of d populations, each with constant 'effective population size' of N individuals. Consider a marker for which t_0 and t_1 are the mean coalescence times for two alleles from the same population and from different populations, respectively, and for which the mean coalescence time for two alleles chosen from any two populations is $t = t_0/d + (d-1)t_1/d$. Assuming mutation rates are small, Slatkin[13] obtained:

$$F_{st} = (t - t_0)/t \tag{1}$$

Suppose that the d populations diverged simultaneously at time Q in the past, from an ancestral population also with effective population size N, where Q is measured in the same units as t_0, t_1 and t. Noting that $t_1 = t_0 + Q$ and substituting u for $(d-1)Q/d$, (1) gives:

$$F_{st} = u/(t_0 + u) \tag{2}$$

The value of t_0, in units of generations or years, is proportional to the effective population size, and therefore differs across marker systems. Let F_{aut} and F_X denote autosomal and X-chromosomal values of F_{st}, and let T_{aut} and T_X denote autosomal and X-chromosomal values of t_0. Following a similar calculation to that of Pérez-Lezaun *et al.*,[15] to determine the relationship between F_{aut} and F_X, one can equate expressions for u obtained from autosomal and X-chromosomal versions of (2):

$$F_{aut} = \frac{T_X F_X}{T_{aut} - F_X(T_{aut} - T_X)} \tag{3}$$

$$F_X = \frac{T_{aut} F_{aut}}{T_X + F_{aut}(T_{aut} - T_X)} \tag{4}$$

The within-population components of genetic variation, $W_{aut} = 1 - F_{aut}$ and $W_X = 1 - F_X$ for autosomal and X-chromosomal loci, respectively, satisfy:

$$W_{aut} = \frac{T_{aut} W_X}{T_X + W_X(T_{aut} - T_X)} \tag{5}$$

$$W_X = \frac{T_X W_{aut}}{T_{aut} - W_{aut}(T_{aut} - T_X)} \tag{6}$$

Writing $N = N_f + N_m$, where N_f and N_m are effective population sizes of females and males, respectively, $r = N_f/N$ is the female fraction of the effective population size. Using the expressions for autosomal and X-chromosomal effective population sizes, $N_{aut} = 4r(1-r)N$ and $N_X = 9r(1-r)N/[2(2-r)]$,[16–18] together with the fact that $T_{aut}/N_{aut} = T_X/N_X$, (3–6) can be simplified.

Table 1. Analysis of molecular variance (AMOVA) for 17 (non-pseudoautosomal) X-chromosomal markers. Ninety-five percent confidence intervals (in parentheses) were calculated using 1,000 bootstraps across loci. The World-B97 sample[6,19] consists of 14 populations that were chosen in order to approximate the sample of Barbujani *et al.*[10]

Sample	Number of regions	Number of populations	Variance components (%)					
			Within populations		Among populations within regions		Among regions	
World	1	52	91.1	(89.0, 92.9)	8.9	(7.1, 11.0)		
World	5	52	89.3	(86.5, 91.8)	4.8	(4.4, 5.3)	5.8	(3.6, 8.5)
World	7	52	90.4	(88.0, 92.5)	4.6	(4.2, 5.1)	4.9	(3.0, 7.3)
World-B97	5	14	85.4	(81.4, 88.7)	7.1	(5.9, 8.2)	7.5	(4.0, 11.8)
Africa	1	6	93.1	(91.2, 94.8)	6.9	(5.2, 8.8)		
Eurasia	1	21	96.2	(95.4, 96.8)	3.8	(3.2, 4.6)		
Eurasia	3	21	95.9	(95.1, 96.7)	3.2	(2.7, 3.8)	0.9	(0.4, 1.4)
Europe	1	8	97.2	(96.3, 98.0)	2.8	(2.0, 3.7)		
Middle East	1	4	97.7	(96.9, 98.4)	2.3	(1.6, 3.1)		
Central/ South Asia	1	9	95.8	(95.0, 96.5)	4.2	(3.5, 5.0)		
East Asia	1	18	95.3	(94.4, 96.1)	4.7	(3.9, 5.6)		
Oceania	1	2	91.3	(87.9, 94.6)	8.7	(5.4, 12.1)		
America	1	5	86.9	(84.7, 88.9)	13.1	(11.1, 15.3)		

Restricting attention to (6), leads to:

$$W_X = \frac{9W_{aut}}{8(2-r) - W_{aut}(7-8r)} \quad (7)$$

In terms of the relative rate of drift in females compared with males, denoted as $z = [1/(2N_f)]/[1/(2N_m)] = N_m/N_f$, (7) gives:

$$W_X = \frac{9(z+1)W_{aut}}{8(2z+1) - W_{aut}(7z-1)} \quad (8)$$

Note that there are two special cases of interest (Table 2). At $r = 1/2\,(z = 1)$, drift proceeds at the same rate in males and females, so that $W_X = 3W_{aut}/(4 - W_{aut})$. At $r = 7/8 = 0.875\,(z = 1/7 \approx 0.143)$, the slow speed of drift in females compared with males reduces the drift rate of X chromosomes exactly enough to counteract the increase in X-chromosomal drift rate that results from their smaller number in the population. In other words, $W_X = W_{aut}$. The fact that the hypothesis $W_X = W_{aut}$ (Table 2) at $P = 0.05$ for 11 of the 13 groupings of data in Table 2 can be rejected means that the hypothesis $z = 1/7$ can also be rejected.

For each of the 13 datasets, the values of r, the female fraction of effective population size, were varied from 0 to 1. At each choice of r, the transformation in (7) was applied and P-values for the two-sided Wilcoxon test between the list of 377 transformed autosomal within-population variance components and the within-population variance components observed at the 17 non-pseudoautosomal X-linked markers were obtained (Figure 1a–c).

At $r = 0.5$, when drift proceeds at the same rate in males and females, significant P-values ($P < 0.05$) were found for Africa, Eurasia (treated both as one region and three regions), Europe, Central/South Asia and East Asia. Therefore, for the remaining seven of the 13 samples, the differences in autosomal and X-chromosomal F_{st} values can be explained by assuming that $N_m = N_f$ and by using the smaller effective population size of X chromosomes alone. Because Rosenberg *et al.*[19] found that repeat size affected divergence, Wilcoxon tests were also performed between transformations of the 274 autosomal tetranucleotide repeats and the 14 X-linked tetranucleotides and similar results were obtained, whether or not the two pseudoautosomal tetranucleotides were included in analysis (not shown).

Table 2. Results for two-sided Wilcoxon tests with $r = 0.875$ and $r = 0.5$. For 11 of the 13 datasets, the observed within-population variance components on the autosomes are significantly different ($P < 0.05$) from those observed using X-linked markers. For seven of the 13 groupings, the observed differences can be explained by accounting for the smaller effective population size of X chromosomes compared with autosomes ($r = 0.5$). For six regions where a value of r is not given in the rightmost column, no value of r produces a high P-value.

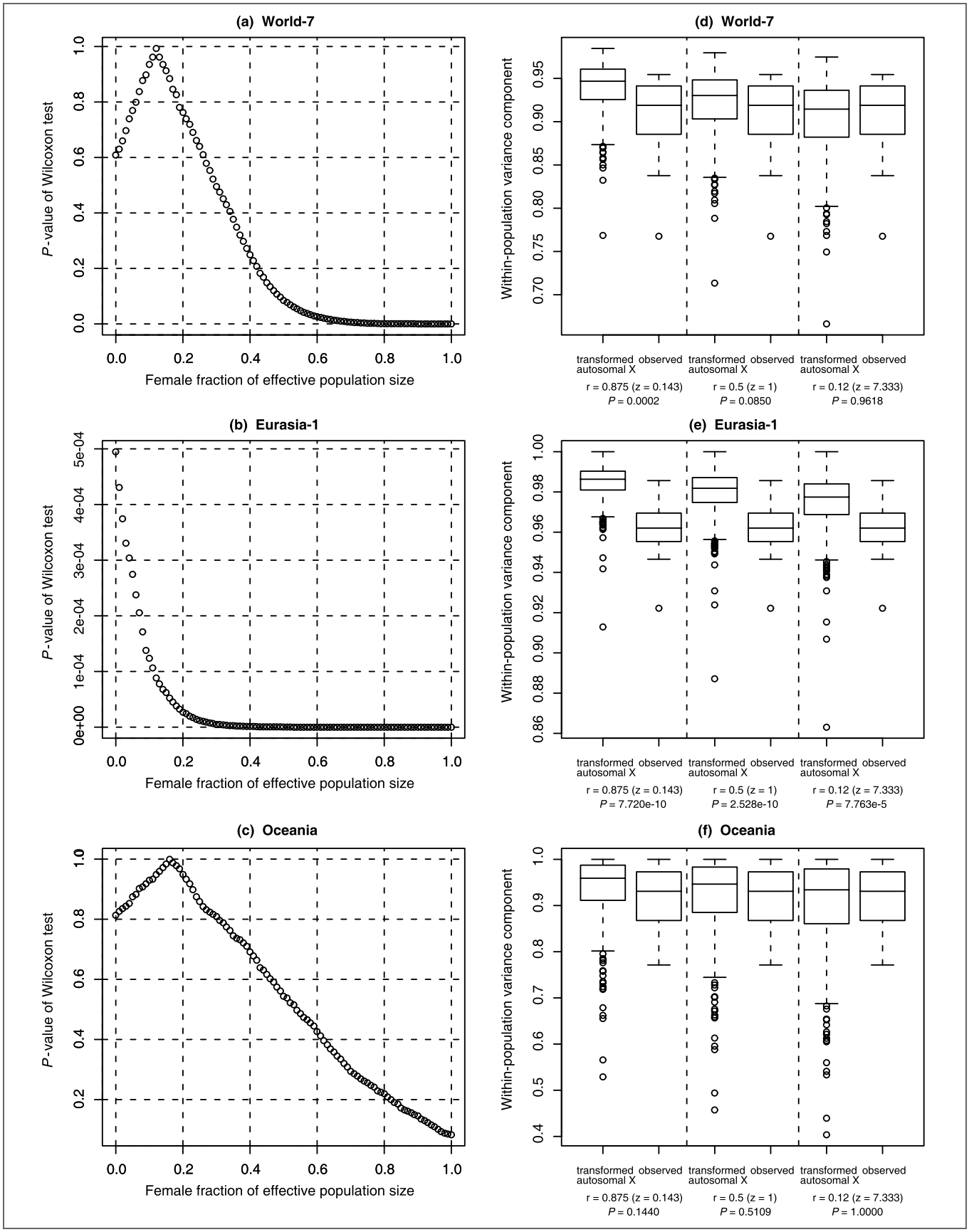
Sample	Number of regions	Number of populations	P-value for two-sided Wilcoxon test with $H_0: W_X = W_{aut}$ ($r = 0.875$, $z = 0.143$)	P-value for two-sided Wilcoxon test with $H_0: W_X = 3W_{aut}/(4 - W_{aut})$ ($r = 0.5$, $z = 1$)	Value of r that produces highest P-value
World	1	52	5.43×10^{-5}	0.05	0.08
World	5	52	1.01×10^{-3}	0.16	0.18
World	7	52	2.42×10^{-4}	0.09	0.12
World-B97	5	14	9.04×10^{-3}	0.49	0.35
Africa	1	6	4.44×10^{-6}	7.69×10^{-4}	
Eurasia	1	21	7.72×10^{-10}	2.53×10^{-7}	
Eurasia	3	21	3.63×10^{-9}	1.22×10^{-6}	
Europe	1	8	1.15×10^{-6}	5.49×10^{-5}	
Middle East	1	4	6.08×10^{-3}	0.16	
Central/South Asia	1	9	5.47×10^{-10}	3.60×10^{-8}	
East Asia	1	18	2.43×10^{-10}	1.19×10^{-8}	
Oceania	1	2	0.16	0.54	0.16
America	1	5	0.26	0.50	0.66

The values of r on the interval [0,1] that produced the largest P-value (see Figure 1a–c) are also reported in Table 2. America was the only sample where the value of r corresponding to the maximal P-value was greater than 0.5 ($r = 0.66$ resulted in $P = 1.00$ for this case).

Of special interest is the fact that five of the six samples with significant P-values ($P < 0.05$) at $r = 0.5$ recorded significant P-values as r varied over the whole range [0,1]. For example, as r varies from 0 to 1 in Figure 1b, the P-value resulting from the autosomal transformation for Eurasia (treated as one region) ranges from 4.95×10^{-4} ($r = 0$) to 1.97×10^{-10} ($r = 1$). The single exception to this pattern was Africa, where the P-value decreased monotonically as r increased and $P < 0.05$ for $r \geq 0.06$.

For these six groupings of the dataset (Africa, Eurasia both as one and as three regions, Europe, Central/South Asia and East Asia), the divergence model with constant effective population size is likely to provide a poorer approximation, as it does not account for population growth or migration[7] (Figure 1d–f).

Figure 1. (a)–(c): P-values of Wilcoxon tests plotted against the female fraction of effective population size, $r = N_f/(N_f + N_m)$, as r varies on the interval [0,1] for (a) the worldwide dataset treated as seven regions (World-7), (b) Eurasia treated as one region (Eurasia-1) and (c) Oceania. Each point represents the P-value of the Wilcoxon test between the X-chromosomal and autosomal within-population variance components, with autosomal values being transformed according to (7) for a particular value of r. (d)–(f): Boxplots of autosomal within-population variance components transformed using equation (7), and observed non-pseudoautosomal X-chromosomal within-population variance components for different values of r. $r = 0.875$ ($z \approx 0.143$) corresponds to the untransformed autosomal values; $r = 0.5$ occurs when drift proceeds at the same rate in males and females; $r = 0.12$ provides the maximum P-value across values of r for the two-sided Wilcoxon test with the World-7 dataset. Note that, as r changes, the two plots across genetic systems coincide for the World-7 and Oceania datasets, but remain quite distinct for the Eurasia-1 dataset (as reflected by the P-values of the specific comparisons shown)


(a) World-7
P-value of Wilcoxon test
Female fraction of effective population size
(b) Eurasia-1
P-value of Wilcoxon test
Female fraction of effective population size
(c) Oceania
P-value of Wilcoxon test
Female fraction of effective population size
(d) World-7
Within-population variance component
transformed autosomal X
observed autosomal X
r = 0.875 (z = 0.143)
P = 0.0002
r = 0.5 (z = 1)
P = 0.0850
r = 0.12 (z = 7.333)
P = 0.9618
(e) Eurasia-1
Within-population variance component
r = 0.875 (z = 0.143)
P = 7.720e-10
r = 0.5 (z = 1)
P = 2.528e-10
r = 0.12 (z = 7.333)
P = 7.763e-5
(f) Oceania
Within-population variance component
r = 0.875 (z = 0.143)
P = 0.1440
r = 0.5 (z = 1)
P = 0.5109
r = 0.12 (z = 7.333)
P = 1.0000

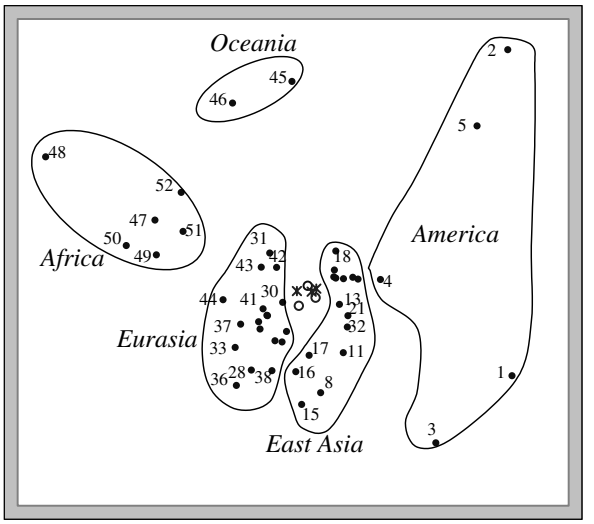


Figure 2. Principal coordinates from multidimensional scaling analysis using 17 X chromosome microsatellites (the three pseudoautosomal markers were excluded). The GDA software package[12] was used to produce a 52 × 52 matrix of pairwise F_{st} values,[9] from which two principal coordinates (*PC1* and *PC2*) were obtained by multidimensional scaling with the *SPSS 8.0.0* package. Most populations (filled circles) are indicated with numbers (see Zhivotovsky *et al.*[7]). The horizontal and vertical axes are principal coordinates *PC1* and *PC2*, respectively. Open circles and asterisks are overlapping populations of, respectively, East Asia (Dai, Cambodian, Han from North China) and Eurasia (Uygur, Hazara, Brahui). *Africa* represents sub-Saharan African populations.

Multidimensional scaling analysis

Geographic groups of populations are revealed by multidimensional scaling of pairwise F_{st} values (Figure 2): sub-Saharan Africa, (western) Eurasia (which includes Europe, the Middle East, and Central/South Asia), East Asia, Oceania and America. Three populations from Eurasia (Uygur, Hazara, Brahui) and three populations from East Asia (Dai, Cambodian, Han from North China) overlap in the plot. The American populations show much greater within-region genetic differentiation than other continental groups, with the Mayan population (labelled as 4 in Figure 2) deviating somewhat from the rest of American samples. These results agree with the analysis of the same populations using autosomal microsatellite markers.[7]

X-chromosomal population structure

The *structure*[5] program identifies subgroupings with distinctive allele frequencies and places individuals into K clusters, where K is defined beforehand by the user and can be varied across independent runs of the program. An individual's membership of a particular cluster is presented as a number between 0 and 1, with membership coefficients summing to 1 across all K clusters.

As is true of autosomal allele frequencies, X-chromosomal allele frequencies are strongly correlated across regions (Table 3). Thus, as was done for the autosomal genotypes from the same individuals,[6] the correlated allele frequencies model implemented in *structure*[5] was used with runs of the same number of iterations as those used to analyse the autosomal data.

America and Africa were the two essentially discrete regions generated at $K = 2$ for the X-chromosomal dataset (Figure 3). To compare results with Rosenberg *et al.*,[6] K was increased from 2 to 6 incrementally. At $K = 3$, Eurasian populations were somewhat identified and the Mozabites were observed to have substantial membership with Africans, as may be expected from their location in Algeria. At $K = 4$, the X-chromosomal data show noticeably different structure from the autosomal data (see Figure 1 of Rosenberg *et al.*[6]), as East Asia does not separate as a genetic cluster with good resolution. The next distinct cluster appears at $K = 6$, where the Oceanic, American and African regions are observed; Eurasia and East Asia separate less obviously, but still appear differentiated from each other.

The X chromosome polymorphisms produced similar clustering to the autosomes, but with less resolution. This raises the question of how the resolution of clusters depends on the number of markers available to study. Figure 4 shows that when the same amounts of data are used, the autosomal and X-chromosomal loci are largely in agreement. Clustering from 20 markers on either autosome 5 or autosome 11 (Figure 4) revealed results very similar to those found with the X-chromosomal dataset. (These particular autosomes were chosen because exactly 20 microsatellites had been typed on them.) For these chromosomes, at $K = 6$, only American, African and (in the case of chromosome 5) Oceanic populations appear distinctly. Furthermore, a sample of 20 markers spread across all of the autosomes yielded similar results, with the Kalash appearing as a distinct group, but with the Oceanic cluster absent. The Kalash — also seen distinctly in Figure 4 from the markers on chromosome 11 — formed the sixth cluster in Rosenberg *et al.*[6] and was the only major cluster in that study that did not match a major geographical region.

Robustness

The 377 autosomal markers in the HGDP-CEPH Human Genome Diversity Cell Line Panel data[4,6,7] comprise the largest multilocus dataset presently available for studying globally distributed populations. Of interest in studies of population structure is the number of loci needed for clustering. Also considered here is the required number of sampled individuals.[6,20]

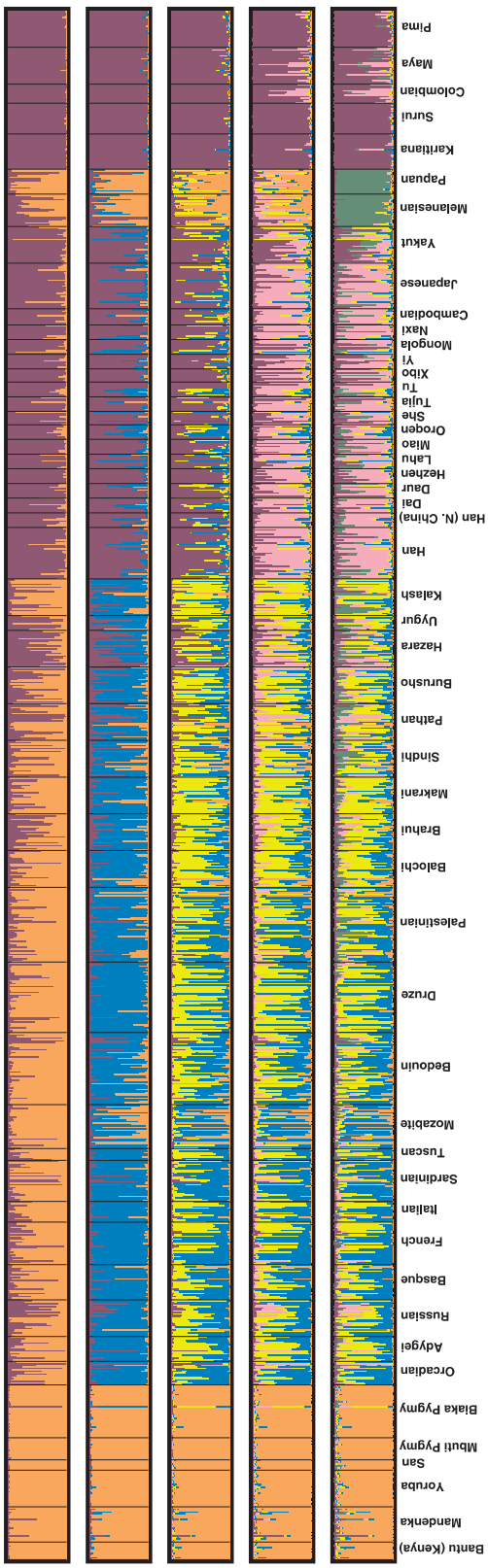


Figure 3. Estimated population structure of the X chromosome for $K = 2$ to $K = 6$ with all 20 markers included. Each individual is represented by a vertical line divided into K coloured segments to display the individual's estimated membership fraction in K clusters. Black lines separate the different populations. The figure shown for a given value of K is based on the highest-probability run (of ten) at that K and was drawn using *Distruct*[25].

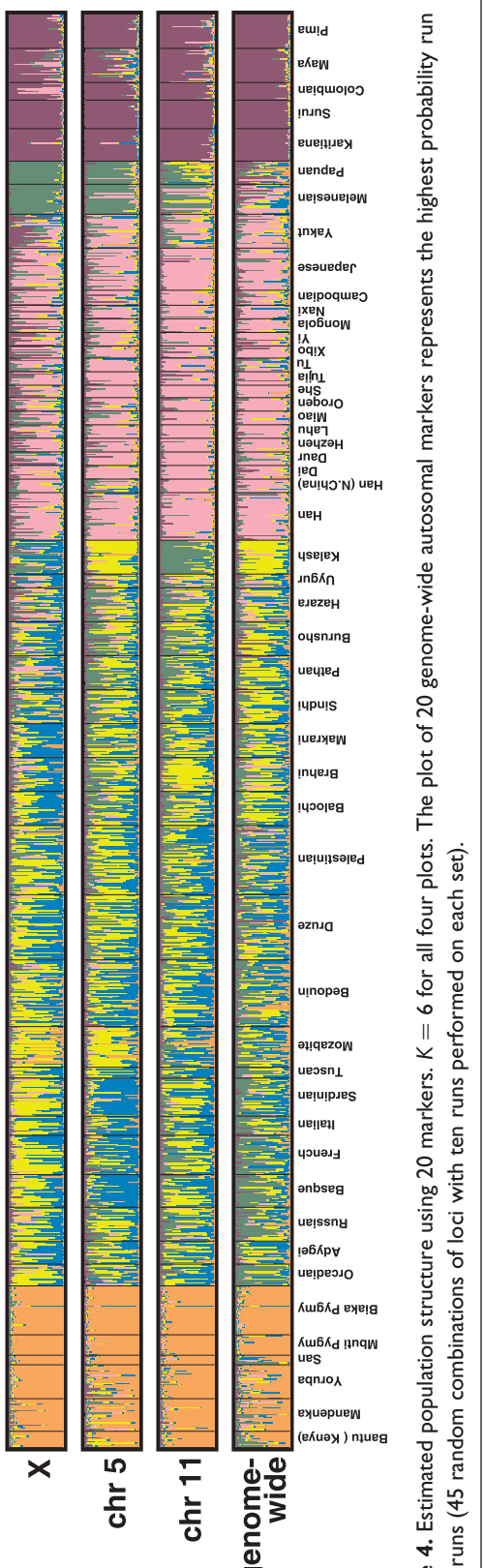


Figure 4. Estimated population structure using 20 markers. $K = 6$ for all four plots. The plot of 20 genome-wide autosomal markers represents the highest probability run of 450 runs (45 random combinations of loci with ten runs performed on each set).

Table 3. Correlation coefficients of allele frequencies. Below the diagonal: correlations for 237 X-chromosomal alleles. Above the diagonal: correlations for 4682 autosomal alleles.[6] The mean correlations across entries in the table are 0.78 (X-chromosomal) and 0.79 (autosomal).

	Africa	Europe	Middle East	C/S Asia	East Asia	Oceania	America
Africa	\	0.77	0.80	0.79	0.73	0.68	0.62
Europe	0.65	\	0.96	0.95	0.83	0.74	0.73
Middle East	0.72	0.95	\	0.95	0.83	0.74	0.71
C/S Asia	0.67	0.93	0.94	\	0.88	0.78	0.77
East Asia	0.61	0.87	0.85	0.90	\	0.81	0.80
Oceania	0.66	0.67	0.75	0.75	0.75	\	0.68
America	0.58	0.78	0.76	0.83	0.92	0.72	\

Rosenberg *et al.*[6] found that inference of membership coefficients is most successful with at least 150 markers, and this is corroborated in Figure 5. It is also seen in Figure 5 that the addition of more individuals to a subset of the entire autosomal dataset (which contains 377 markers and 1,056 individuals) did not improve population structure inference as much as did the addition of loci. Data from individuals are used to estimate allele frequencies, which can be done fairly accurately with a small number of individuals; however, as *structure* uses distinctive genotypic combinations for the construction of clusters, and multilocus combinations are more likely to be distinctive to particular groups than are single-locus types,[21] additional loci can contribute more information to cluster analysis than can the addition of more individuals to the sample (Figure 6).

Oceania appears in Figure 6 as a distinct cluster with only ten loci and between 35 and 100 individuals per region. Because the Oceanic populations together contain 39 individuals, increasing the number of individuals beyond 35 per region meant that every Melanesian and Papuan was included in the subset run of *structure*. Thus, the distinctive allele frequencies of these populations identify this particular genetic cluster, despite the use of only ten loci.

Discussion

The same techniques as Rosenberg *et al.*[6] and Zhivotovsky *et al.*[7] were used to analyse genetic structure as inferred from 20 microsatellite markers on the X chromosome. Multidimensional scaling (Figure 2) did not reveal major departures from the patterns exhibited by the autosomal data. As was also observed on the autosomes, both America and Oceania are the regions exhibiting the lowest heterozygosity (0.57 and 0.64, respectively) on the X chromosome.

Seielstad *et al.*[3] used a migration model to attribute differences in F_{st} across genetic systems to a difference in male and female migration rates. By contrast, a divergence model was used here and it was found that the differences observed in F_{st} values can, in many cases, be explained by the smaller effective population size of X chromosomes compared with autosomes. This is similar to what was observed by Jorde *et al.*,[1] who reported higher G_{st} values in Y-chromosome restriction-site polymorphisms and mtDNA compared with autosomal systems, and found that this difference was expected because of the lower effective population size of the uniparentally-inherited portions of the genome. In those regions here where the smaller number of X chromosomes does not provide a sufficient explanation (Africa, Eurasia, Europe, Central/South Asia and East Asia), the assumptions of the divergence model — especially that of constant population size — may be responsible for the disagreement.

Upon closer examination of these differences in observed F_{st}, the data here provide some support for the idea that genetic drift occurs faster in females than in males, or, equivalently, that the female effective population size is smaller than that of males. Many factors could potentially explain this observation; a larger correlation in females between reproductive success in parents and offspring or a smaller generation time in females[22] may increase the rate of drift in females compared with that in males.

The use of X-chromosomal data revealed clustering similar to that obtained using autosomal data, but with less resolution (Figures 3–5). In America, Africa and Oceania, inferred clusters corresponded closely with predefined populations using both the autosomal and X-chromosomal loci, but the pattern of admixture observed by Rosenberg *et al.*[6] is not exactly the same as that revealed by the X chromosome, due to reduced resolution of clusters.

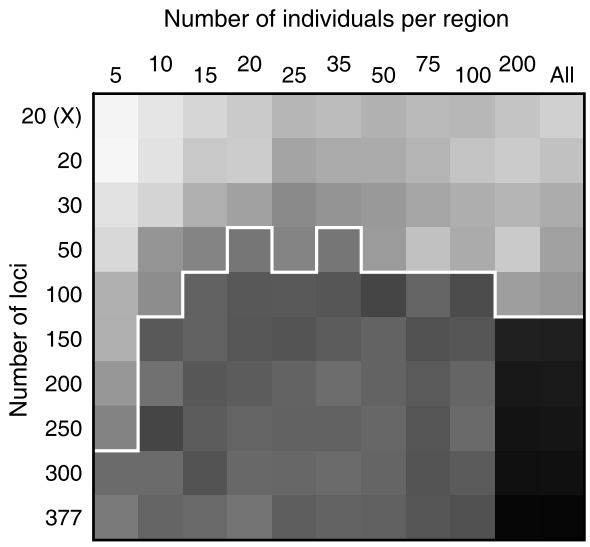


Figure 5. Grid of similarity coefficients. Each square in the grid represents a similarity coefficient[6] between 0 and 1 (0 corresponds to white, 1 to black) that measures the similarity between a *structure* run with a subset of the autosomal dataset and the entire dataset used by Rosenberg *et al.*[6] The subsets varied in the number of loci used (labelled on the vertical axis) and the number of individuals used per region (on the horizontal axis). For region *i*, *N* individuals per region corresponds to $min(N, S_i)$ individuals, where S_i is the total sample size of region *i*. Thus, five individuals per region corresponds to 35 individuals overall, ten per region to 70, 15 per region to 105, 20 per region to 140, 25 per region to 175, 35 per region to 245, 50 per region to 339, 75 per region to 489, 100 per region to 639 and 200 per region to 1,005 individuals total. Ten runs of each subset of the data were performed and the median similarity coefficient between the best subset run and ten runs of the entire dataset was used to generate a given square. Those values below the white line have a similarity coefficient of 50% or higher with the entire dataset. Using 150 loci or more, and 200 or more individuals per region, runs had similarity coefficients ranging from 0.87 to 0.98.

Note that the Oceanic (Melanesian and Papuan) populations in Figure 3 appear most similar to the African populations for $2 \leq K \leq 4$, and then appear as their own genetic cluster at $K = 6$. This contrasts with the analysis of Wilson *et al.*,[23] whose analysis of 23 X-linked microsatellites using *structure* showed the Oceanic population combining with the Chinese population at $K = 3$. A possible explanation for the results here may be a migration from Africa to Oceania separate from the primary migration out of Africa to other regions.[24]

While choosing representative individuals from various populations is an important factor in the success of studies concerned with inference of population structure, the robustness of *structure* is much more dependent on the number of microsatellite markers used (Figures 5 and 6). In common with Rosenberg *et al.*,[6] it is observed here that ancestry inference is most successful with at least 150 loci (Figure 5). Bamshad *et al.*[20] reported that correct assignment to the continent of origin with a mean accuracy of at least 90 per cent required a minimum of 60 loci and reached 99–100 per cent accuracy when more than 100 loci were used.

In contrast to this study, Bamshad *et al.*[20] considered a sample correctly assigned if the cluster with the greatest membership coefficient for an individual was the same as the predefined assignment. The criterion here compares the membership coefficients across all *K* clusters calculated when using *structure* on a subset of the data, with assignment made based on the full dataset. Thus, it is a measure of how well the results with smaller amounts of data match those with larger datasets, rather than a measure of 'correct assignment'. The difference in these criteria is likely to account for the smaller amount of genetic data regarded as sufficient by Bamshad *et al.*[20] The similarity coefficient *C* may be more sensitive to differences in membership coefficients between two runs and can be viewed as a conservative measure of similarity for the runs: visual similarity between graphs of estimated membership coefficients (Figure 6) can be achieved even with fairly small values of *C* (Figure 5). In Figure 6, for example, the plot using 100 loci and a maximum of 200 individuals per region is quite similar to the plot of the full data, while the similarity coefficient[6] between the *structure* runs of that particular subset and the entire dataset is 0.379. *C* does not make use of the 'correct' predefined structure, and, thus, unlike the criterion used by Bamshad *et al.*[20], is unaffected by errors among the predefined labels.

While most studies to date have lacked the power to make strong inferences about population structure (due to the very recent availability of datasets with individuals assayed for large numbers of loci), future studies should choose an appropriate number both of individuals per region and of loci for these analyses. Note, however, that the sampling scheme may affect the estimated structure. For example, finer distinctions among populations of interest become visible when individuals who are more distantly related to those populations are omitted from analysis.[6]

Although differences between the population structure based on the autosomes and X-linked loci may be expected due to differences in male and female demography, the differences between the results here and those of Rosenberg *et al.*[6] were largely due to the smaller number of X chromosomes in a population compared with autosomes, and to the smaller amount of data available from the X chromosome. From these results, it might be inferred that sex-biased demographic processes have not had a great influence on human population structure.

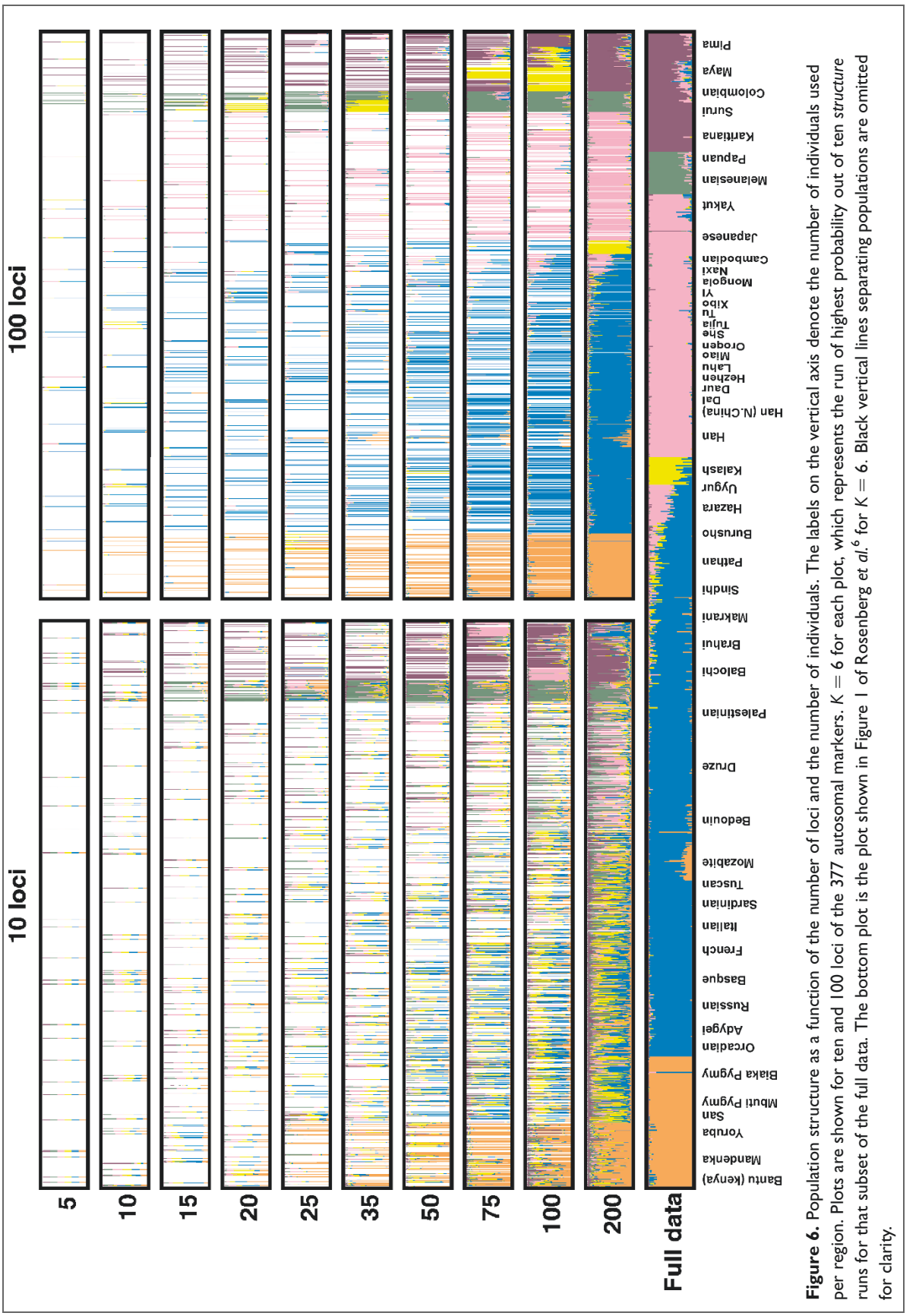


Figure 6. Population structure as a function of the number of loci and the number of individuals. The labels on the vertical axis denote the number of individuals used per region. Plots are shown for ten and 100 loci of the 377 autosomal markers. $K = 6$ for each plot, which represents the run of highest probability out of ten *structure* runs for that subset of the full data. The bottom plot is the plot shown in Figure 1 of Rosenberg *et al.*[6] for $K = 6$. Black vertical lines separating populations are omitted for clarity.

Acknowledgments

This research was supported in part by NIH Grants GM28428 and GM28016. Sohini Ramachandran is also supported by a NDSEG fellowship. Noah A. Rosenberg is supported by an NSF Postdoctoral Fellowship in Biological Informatics.

References

1. Jorde, L.B., Watkins, W.S., Bamshad, M.J. *et al.* (2000), 'The distribution of human genetic diversity: A comparison of mitochondrial, autosomal and Y-chromosome data', *Am. J. Hum. Genet.* Vol. 66, pp. 979–988.
2. Oota, H., Settheetham-Ishida, W., Tiwaweck, D., *et al.* (2001), 'Human mtDNA and Y-chromosome variation is correlated with matrilocal versus patrilocal residence', *Nature Genet.* Vol. 29, pp. 20–21.
3. Seielstad, M.T., Minch, E. and Cavalli-Sforza, L.L. (1998), 'Genetic evidence for a higher female migration rate in humans', *Nature Genet.* Vol. 20, pp. 278–280.
4. Cann, H.M., de Toma, C., Cazes, L. *et al.* (2002), 'A human genome diversity cell line panel', *Science* Vol. 296, pp. 261–262.
5. Pritchard, J.K., Stephens, M. and Donnelly, P. (2000), 'Inference of population structure using multilocus genotype data', *Genetics* Vol. 155, pp. 945–959.
6. Rosenberg, N.A., Pritchard, J.K., Weber, J.L. *et al.* (2002), 'Genetic structure of human populations', *Science* Vol. 298, pp. 2381–2385.
7. Zhivotovsky, L.A., Rosenberg, N.A. and Feldman, M.W. (2003), 'Features of evolution and expansion of modern humans, inferred from genome-wide microsatellite markers', *Am. J. Hum. Genet.* Vol. 72, pp. 1171–1186.
8. Sabatti, C. and Risch, N. (2002), 'Homozygosity and linkage disequilibrium', *Genetics* Vol. 160, pp. 1707–1719.
9. Weir, B. (1996), *Genetic Data Analysis II*, Sinauer Press, Sunderland, MA.
10. Barbujani, G., Magagni, A., Minch, E. *et al.* (1997), 'An apportionment of human DNA diversity', *Proc. Natl. Acad. Sci. USA* Vol. 94, pp. 4516–4519.
11. Lewontin, R.C. (1972), 'The apportionment of human diversity', *Evol. Biol.* Vol. 6, pp. 381–398.
12. Lewis, P. O. and Zaykin, D. V. (2001), 'Genetic Data Analysis: Computer program for the analysis of allelic data', *http://lewis.eeb.uconn.edu/lewishome/software.html.*
13. Slatkin, M. (1991), 'Inbreeding coefficients and coalescence times', *Genet. Res.* Vol. 58, pp. 167–175.
14. Slatkin, M. (1995), 'A measure of population subdivision based on microsatellite allele frequencies', *Genetics* Vol. 139, pp. 457–462.
15. Pérez-Lezaun, A., Calafell, F., Seielstad, M. *et al.* (1997), 'Population genetics of Y-chromosome short tandem repeats in humans', *J. Mol. Evol.* Vol. 45, pp. 265–270.
16. Ewens, W.J. (1969), *Population Genetics*, Methuen & Co., London, UK.
17. Hartl, D.L. and Clark, A.G. (1997), *Principles of Population Genetics*, Sinauer Press, Sunderland, MA.
18. Nordborg, M. and Krone, S. (2002), 'Separation of time scales and convergence to the coalescent in structured populations', In: Slatkin, M. and Veuille, M., (eds), *Modern Developments in Theoretical Population Genetics* Oxford University Press, Oxford, UK.
19. Rosenberg, N.A., Pritchard, J.K., Weber, J.L. *et al.* (2003), 'Response to comment on 'Genetic structure of human populations'', *Science* Vol. 300, pp. 1877.
20. Bamshad, M.J., Wooding, S., Watkins, W.S. *et al.* (2003), 'Human population genetic structure and inference of group membership', *Am. J. Hum. Genet.* Vol. 72, pp. 578–589.
21. Edwards, A.W.F. (2003), 'Human genetic diversity: Lewontin's fallacy', *Bioessays* Vol. 25, pp. 798–801.
22. Helgason, A., Hrafnkelsson, B., Gulcher, J.R. *et al.* (2003), 'A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: Evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes', *Am. J. Hum. Genet.* Vol. 72, pp. 1370–1388.
23. Wilson, J.F., Weale, M.E., Smith, A.C. *et al.* (2001), 'Population genetic structure of variable drug response', *Nature Genet.* Vol. 29, pp. 265–269.
24. Disotell, T.R. (1999), 'Human evolution: the southern route to Asia', *Curr. Biol.* Vol. 9, pp. R925–R928.
25. Rosenberg, N.A. (2004), 'Distruct: A program for the graphical display of population structure', *Mol. Ecol. Notes*, http://www.blackwell-synergy.com/ links/doi/10.1046/j.1471-8286.2003.00566.x/full.