A TEST FOR HETEROGENEITY OF MICROSATELLITE VARIATION

Jonathan K. Pritchard and Marcus W. Feldman

Dept. of Biological Sciences
Stanford University
Stanford, CA 94305, USA.

Abstract

Levels of variation at microsatellite loci depend on both the effective population size and the microsatellite mutation rate. In addition, the amount of variation can be affected by other forces, such as population bottlenecks or selection at linked loci. In this paper, we develop a formal statistical test for testing whether levels of microsatellite variation in different groups of loci are significantly different. We apply this test to examine differences between autosomal and Y chromosome loci, to test whether these loci have different levels of variation. We also compare autosomal and Y chromosome loci at global, and regional scales to test for Y chromosome bottleneck effects.

Introduction

In recent years, many evolutionary and ecological studies have used microsatellites as indicators of population-level processes, in part because they are frequently highly variable. Included among these have been surveys of microsatellite variation in human populations in order to investigate human history (e.g., Bowcock et al., 1994; Cooper et al., 1996; Linares et al., 1996, Tishkoff et al., 1996; Pritchard et al., 1998). Microsatellites have been used in studies of population structure and differentiation for quite a range of other organisms, including brook charr (Anger et al., 1995), bumble bees (Estoup et al., 1996), mosquitoes (Lehmann et al., 1996), *Drosophila melanogaster* (Michalakis and Veuille, 1996), polar bears (Paetkau et al., 1995), and wombats (Taylor, Sherwin and Wayne, 1994).

In the context of ecological or evolutionary studies such as these, there are a number of questions which can be addressed by comparing levels of variation in different populations, or at different loci. Here we describe two such comparisons:

i. Are the observed levels of variation at different sets of loci significantly different? For example, Goldstein et al. (1996) asked whether levels of human Y chromosome variation were significantly less than 1/4 those of human autosomal variation, as would be expected if there were a higher variance in male than female reproductive success, or if there had been a selective event in the recent history of the human Y chromosome. In the same way, one might ask whether different types of microsatellite loci (e.g., dinucleotide and tetranucleotide repeats) showed different amounts of variation, which might indicate different mutation rates, or different mutation processes (such as different levels of range constraints).

ii. Are the levels of variation observed at two or more sets of loci, in two or more pop-

ulations, consistent with a simple neutral model? Slatkin (1995) has pointed out that selection at one locus alters expected levels of variation at linked microsatellite loci. This prediction suggests a possible test for selection. If, for example, directional selection at a locus occurred in one part of a species' range, but not elsewhere, then there should be reduced variation in the neighborhood of that locus in the affected population only. We describe a linear model to test for this kind of effect. In this model, the variation at the ith locus, in the jth population, depends on both a locus effect (a function of the mutation rate for the ith locus) and a population effect (a function of the effective population size in the jth population). Selection is expected to lead to departures from this linear model.

We begin with some theoretical background regarding the distribution of the observed variance in repeat scores. We show how this variance can be converted into a random variable with convenient characteristics for hypothesis testing, including an approximately Normal distribution. We illustrate this approach by analyzing human microsatellite data, addressing each of the questions listed above.

Estimation of $\mathrm{E}(V)$, the expected variance in repeat scores

In this section we consider the theoretical sampling properties of $\overline{V}$, the mean over a series of microsatellite loci of the variance in repeat scores, under a standard model of microsatellite evolution (e.g., Zhivotovsky and Feldman, 1995).

Consider a series of microsatellite repeat loci evolving by a process of neutral mutation and drift. We assume an isolated Wright-Fisher population of haploid size N, at mutation-drift equilibrium. Mutation acts to change the number of repeats at a microsatellite locus from p to q, where p and q take integer values. Mutation is symmetric, in the sense that $\mathrm{E}(p-q)=0$; and the mutation process is independent of the absolute repeat score, so that the probability of a mutation of size $q-p$ does not depend on p. Thus, there are neither range constraints, nor mutation bias toward some focal value. Let μ be the total probability of mutation per locus per generation, and let s_i be the ith central moment of $q-p$, the size of the mutation.

Suppose that m chromosomes are sampled from a population and, for each chromosome, the number of repeats at a particular locus is observed. Designate the number of repeats at that locus on the ith chromosome by x_i, and the observed mean number as $\overline{x} \equiv \sum x_i/m$. Then we estimate the variance $\hat{V}$ as

$$\hat{V} \equiv \frac{1}{m-1}\sum_{i=1}^{m}(x_i-\overline{x})^2. \qquad (1)$$

It has been shown previously that the expectation of $\hat{V}$ is proportional to the population size: that is, $\mathrm{E}(\hat{V}) = N\mu s_2$. This result was first obtained by Moran (1975) in the pure stepwise case where $s_2 \equiv 1$, and later by Slatkin (1995) and Zhivotovsky and Feldman (1995) in the general (symmetric) case. (The forward-recursive method, used by Moran and by Zhivotovsky and Feldman, leads to a factor of $N-1$, in place of N, in this expression.) Notice that $\hat{V}$ is precisely half the mean-square difference among individuals, a quantity which was studied by Goldstein et al. (1995) and by Pritchard and Feldman (1996).

Now consider the average of the variance $\hat{V}$ in repeat score across a series of L loci. This average will be denoted $\overline{V}$, and is calculated as

$$\overline{V} = \frac{1}{L}\sum_{i=1}^{L}\hat{V}_i, \tag{2}$$

where $\hat{V}_i$ is the variance at the ith locus. $\hat{V}_i$ is an unbiased estimator of $N\mu s_2$ (Pritchard and Feldman, 1996), hence,

$$\mathrm{E}(\overline{V}) = N\mu s_2. \tag{3}$$

Distribution of $\overline{V}$

The distribution of $\overline{V}$ is asymmetric, having a lower bound at zero, but no upper bound. In fact, simulation data indicate that the distribution tends to be concentrated to the left of its mean, with a long right-hand tail as previously noted by Goldstein et al. (1996). This skew makes the design of statistical tests based on $\overline{V}$ somewhat problematic. However, Goldstein et al. suggested that a log-transformation of $\overline{V}$ results in a random variable whose distribution is approximately Normal.

We have conducted a series of simulations (not shown) to investigate the distribution of $\ln(\overline{V})$. Our results indicate that the distribution of $\ln(\overline{V})$ generally approximates a Normal distribution. Of course, $\ln(\overline{V})$ is undefined when $\overline{V} = 0$, which means that the fit is relatively poor when there are only a few loci and $N\mu$ is small, in which case the probability that $\overline{V} = 0$ is not negligible. In practice, this does not appear to be a serious concern, provided that estimates are taken as averages over more than perhaps three or four loci, as even single microsatellite loci are rarely monomorphic.

Given that $\ln(\overline{V})$ is approximately a Normal random variable, we can describe its distribution using its expectation and variance. These depend on the number and linkage of the loci, the sample size, on $N\mu$, and on the moments of the mutation distribution. We will suggest a method of approximating the expectation and variance analytically, and for linked loci, values obtained from simulations are tabulated.

We can approximate $\mathrm{E}[\ln(\overline{V})]$ and $\mathrm{Var}[\ln(\overline{V})]$ in terms of the moments of $\overline{V}$, by the Delta method, which uses the leading terms of the Taylor Series expansion of $\ln(\overline{V})$. It is convenient to define the "bias" B, so that

$$\mathrm{E}[\ln(\overline{V})] = \ln(\mathrm{E}[\overline{V}]) + B. \tag{4}$$

Since the logarithm is a concave function, $B \leq 0$. We obtain the following approximations:

$$B \approx -\frac{\mathrm{Var}[\overline{V}]}{2(\mathrm{E}[\overline{V}])^2}, \tag{5}$$

and

$$\mathrm{Var}[\ln(\overline{V})] \approx \frac{\mathrm{Var}[\overline{V}]}{(\mathrm{E}[\overline{V}])^2} \tag{6}$$

(see e.g., exercise 10.17, and equation 10.14, respectively, of Stuart and Ord, 1987). In order to evaluate (5) and (6), we now present results concerning Var[$\overline{V}$].

Theoretical variance of $\overline{V}$.

In general, we can write the variance of $\overline{V}$ as

$$\mathrm{Var}(\overline{V}) = \frac{1}{L^2}\left[\sum_{i=1}^{L}\mathrm{Var}(\hat{V}_i) + 2\sum_{i<j}\mathrm{Cov}(\hat{V}_i, \hat{V}_j)\right]. \tag{7}$$

If there is moderate recombination among loci, then the covariances between $\hat{V}_i$ and $\hat{V}_j$ are negligible (see Zhivotovsky and Feldman (1995) who proved this result with the implicit assumption that the recombination rate is not extremely small.) In that case, we have

$$\mathrm{Var}(\overline{V}_u) = \frac{1}{L^2}\sum_{i=1}^{L}\mathrm{Var}(\hat{V}_i), \tag{8}$$

where $\hat{V}_i$ is the repeat-score variance at locus i and $\overline{V}_u$ indicates that $\overline{V}$ is an average over unlinked loci.

The sampling properties of $\hat{V}$ have previously been examined by Roe (1992), Zhivotovsky and Feldman (1995) and Pritchard and Feldman (1996) (who investigated $2\hat{V}$). Zhivotovsky and Feldman used a forward recursive approach to obtain the variance of $\hat{V}$ for a large sample under a general (symmetric) mutation scheme, while Pritchard and Feldman obtained $\mathrm{Var}(\hat{V})$ for a finite sample under a one-step mutation scheme. We have recently (Pritchard and Feldman, unpublished results) extended the latter result to a general mutation scheme, so that, in combination with (8), we have

$$\mathrm{Var}(\overline{V}_u) = \frac{N\mu s_4(m^2+m) + (2N\mu s_2)^2(2m^2+3m+1)}{6L(m^2-m)}. \tag{9}$$

The latter expression is not valid when there is tight linkage among loci. In order to deal with that case, we have found the following expression for $\mathrm{Var}(\overline{V}_l)$ over L completely linked loci, where $\overline{V}_l$ denotes an average over linked loci (Pritchard and Feldman, unpublished results):

$$\mathrm{Var}(\overline{V}_l) = \frac{N\mu s_4(m^2+m) + (2N\mu s_2)^2(2m^2+3m+1)}{6L(m^2-m)} + \frac{2(N\mu s_2)^2(L-1)(m^2+m+3)}{9L(m^2-m)}. \tag{10}$$

This calculation of the variance of $\overline{V}_l$ exhibits an unfortunate feature of the estimator $\overline{V}$, namely that it lacks statistical consistency for linked loci. As m and L become large, the variance of $\overline{V}$, calculated over linked loci, does not go to zero, but instead approaches $2(N\mu s_2)^2/9 \approx 0.222(N\mu s_2)^2$. Thus, for the kinds of questions considered here, it is most appropriate to use unlinked markers where possible.

We can use formula (9) or (10) (depending on the linkage situation) to estimate $\mathrm{Var}(\overline{V})$. While m and L are known *a priori*, the remaining parameters, s_2, s_4, and $N\mu$ are

typically unknown. A straightforward approach is to assume a strict stepwise mutation model, so that $s_2 = s_4 = 1$. Then, we can use $\overline{V}$ in place of $N\mu$ (eqn. 3). We will designate the estimated variance of $\overline{V}$, computed from (9) or (10) using these estimated values of s_2, s_4, and $N\mu$, by $\hat{\mathrm{V}}\mathrm{ar}(\overline{V})$.

Construction of a test statistic

These results may now be applied to construct an unbiased test statistic drawn from a Normal distribution with known variance. From (3) and (4), it follows that $\mathrm{E}[\ln(\overline{V})] = \ln(N\mu s_2) + B$. Under the approximation that $\ln(\overline{V})$ is a Normal random variable, it follows that

$$\ln(\overline{V}) - \ln(N\mu s_2) - B = \epsilon, \tag{11}$$

where ϵ is a random error term with mean 0, and variance σ^2, say. Then ϵ/σ is approximately a Standard Normal random variable. This property facilitates statistical tests using $\overline{V}$. First however, making use of (5) and (6) we propose the following estimates for B and σ^2:

$$\hat{B} = -\frac{\hat{\mathrm{V}}\mathrm{ar}(\overline{V})}{2\overline{V}^2}, \tag{12}$$

and

$$\hat{\sigma}^2 = \frac{\hat{\mathrm{V}}\mathrm{ar}(\overline{V})}{\overline{V}^2}, \tag{13}$$

where $\hat{\mathrm{V}}\mathrm{ar}(\overline{V})$ is computed as described above. Using $\overline{V}$ to estimate $N\mu$ in calculating $\hat{\mathrm{V}}\mathrm{ar}(\overline{V})$ does not in practice lead to serious error, because $N\mu$ cancels from the quadratic terms of (12) and (13). Thus, $\hat{B}$ and $\hat{\sigma}^2$ are determined in large part by L and m, which are known (see Table 1). We take advantage of this to ignore the sampling properties of $\hat{B}$ and $\hat{\sigma}^2$, in the development of a statistical test.

We have performed simulations to study the performance of the Taylor series approximations and find that in general the magnitude of the bias is overestimated, as is the variance. Thus the Taylor series approximations are conservative. The magnitude of the error is small for unlinked loci, but significant for linked loci. In view of the size of this error we recommend the use of simulated values (Table 1) in performing the test with linked loci. A copy of the program used for the simulations is available from JKP.

		L							
		4		8		16		32	
m	$N\mu$	B	Var	B	Var	B	Var	B	Var
8	2	-0.243	0.518	-0.203	0.393	-0.173	0.331	-0.158	0.298
	6	-0.251	0.483	-0.197	0.388	-0.168	0.333	-0.149	0.294
	10	-0.248	0.491	-0.185	0.373	-0.167	0.325	-0.152	0.296
16	2	-0.223	0.408	-0.167	0.320	-0.145	0.267	-0.120	0.240
	6	-0.221	0.392	-0.161	0.299	-0.141	0.263	-0.124	0.242
	10	-0.204	0.387	-0.149	0.302	-0.139	0.261	-0.125	0.245
32	2	-0.198	0.371	-0.158	0.290	-0.130	0.251	-0.110	0.225
	6	-0.193	0.358	-0.150	0.279	-0.127	0.238	-0.110	0.218
	10	-0.203	0.348	-0.144	0.276	-0.125	0.243	-0.107	0.219
48	2	-0.179	0.374	-0.153	0.280	-0.127	0.242	-0.111	0.219
	6	-0.186	0.342	-0.143	0.274	-0.112	0.235	-0.115	0.207
	10	-0.197	0.341	-0.145	0.268	-0.123	0.234	-0.115	0.213

Table 1. *Bias (B) and variance (Var) of* $\ln(\overline{V})$ *under the strict stepwise mutation model. The values reported are averages over* 10^4 *replicate simulations, with an accuracy of* ± 0.01 *or better.*

A test of heterogeneity, and application to human microsatellite data

We now apply these results on the log-transformed $\overline{V}$ to the analysis of human microsatellite data, with two examples of tests of heterogeneity, corresponding to the two kinds of tests described in the Introduction.

Bowcock et al., (1994), surveyed variation at autosomal microsatellite loci, typing approximately 270 chromosomes at each locus, drawn from a worldwide sample. They studied 30 dinucleotide-repeat loci, estimating $\overline{V}$ to be 10.10. One dinucleotide locus was extraordinarily variable, and if that locus is removed, then $\overline{V} = 6.83$. Pritchard et al. (1998) reported a survey of microsatellite variation in 445 human Y chromosomes, drawn from a similar set of populations. They examined six tetranucleotide loci, and two trinucleotide loci, and estimated $\overline{V} = 1.15$.

We have computed $\ln(\overline{V})$, $\hat{B}$, and $\hat{\sigma}^2$ for these data sets (Table 2), under the assumption of strict stepwise mutation. It is currently unclear whether the hyper-variable dinucleotide locus should be treated as an outlier: hence we present data both with and without this locus included (lines 1A, and 1B). In addition, from basic population genetic arguments, it is expected that the Y chromosome loci should be less variable than the autosomal loci by a factor of four. This corresponds to a predicted *difference* of ln(4) in $\ln(\overline{V})$. For this reason, we include both raw Y chromosome data, and data adjusted by addition of ln(4) (lines 2A, and 2B).

Table 2. Summary of $\overline{V}$ estimates.

	L	m	$\overline{V}$	$\mathrm{Var}(\overline{V})$	$\ln(\overline{V})$	$\hat{B}$	$\hat{\sigma}^2$
Dinuc.[1A]	30	270	10.12	4.60	2.31	-0.02	0.045
Dinuc.[1B]	29	270	6.83	2.18	1.92	-0.02	0.047
Y-Chr.[2A]	8	445	1.15	0.50	0.14	-0.15[3]	0.29[3]
Y-Chr.[2B]	8	445	1.15	0.50	1.53[4]	-0.15[3]	0.29[3]

1A,**1B**: Data from Bowcock et al., (1994). Data set **1B** differs from set **1A** by the deletion of one hypervariable locus. **2A**, **2B**: Data from Pritchard et al., (1998). **3**: Value estimated from simulations (Table 1). **4**: Estimate of $\ln(\overline{V})$ adjusted by addition of $\ln(4)$ to account for the predicted four-fold smaller effective population size on the Y chromosome.

We now test whether the observed values of $\overline{V}$ differ significantly between autosomal dinucleotide loci, and Y chromosome tri- and tetranucleotides (data sets 1A and 1B; 2A and 2B). Since ϵ is (approximately) a Normal random variable, we have that $\epsilon^2/\hat{\sigma}^2 \sim \chi^2_1$ which suggests a test statistic

$$X^2 = \sum_{i=1}^{k} \frac{\ln(\overline{V}_i) - \hat{B}_i - \ln(N\mu s_2)}{\hat{\sigma}_i^2} \tag{14}$$

for determining whether there is significant heterogeneity in $\overline{V}$ among k data sets. If we knew $N\mu s_2$ in advance, then X^2 would be approximately Chi-square distributed with k degrees of freedom; otherwise we might estimate a value of $N\mu s_2$ from the data, in which case X^2 would have $k-1$ degrees of freedom. Large values of X^2 suggest significant heterogeneity of $\overline{V}$.

Using the data sets listed in Table 2, we have conducted pairwise tests for heterogeneity. In effect, these tests address the question of whether the levels of variation differ significantly between two sets of loci. For each comparison, we estimated a value of $N\mu s_2$ which minimized X^2, thus making the test maximally conservative. (Minimizing X^2 in this way corresponds to obtaining the weighted least squares estimate for $N\mu s_2$, where the weights are the inverse variances of each term.)

The results are summarized in Table 3. These show that the autosomal microsatellites are much more variable than the Y chromosome microsatellites but that this difference can be explained entirely by the smaller effective population size of Y chromosomes predicted from population genetics theory. This finding is of interest, because it has been suggested elsewhere (e.g., Dorit et al., 1995) that Y chromosomes are deficient in genetic variation. Our finding agrees with the results of Goldstein et al., 1996.

Table 3. Summary of χ^2 test statistics.

	Y-Chr.[2A]	Y-Chr.[2B]
Dinuc.[1A]	**12.42**	1.26
Dinuc.[1B]	**8.08**	0.20

Summary of pair-wise χ^2 tests of the hypothesis of equality of $\overline{V}$. Significant results (≥ 3.84) are shown in bold face. Details about the data sets are given below Table 2.

A regional test of heterogeneity

As another application, we can test whether the Y chromosome has been subject to selection in particular human populations. For example, the variance in repeat scores in Y chromosomes from the Americas is much lower than elsewhere. Can this be taken as indirect evidence that there has been selection on Y chromosomes in the Americas? Our approach is to test whether there is significant heterogeneity among populations and loci, comparing autosomal dinucleotide loci with Y chromosome loci, in Africa, Australasia, and the Americas. The data, and the estimates of $\ln(\overline{V})$, $\hat{B}$, and $\hat{\sigma}^2$ are summarized in Table 4.

Table 4. $\overline{V}$ estimates by region.

	L	m	$\overline{V}$	$\mathrm{Var}(\overline{V})$	$\ln(\overline{V})$	$\hat{B}$	$\hat{\sigma}^2$
AFR (Di)[1]	29	55	7.53	2.65	2.02	-0.02	0.05
AUST(Di)[1]	29	55	5.73	1.54	1.75	-0.02	0.05
AMER(Di)[1]	29	54	5.27	1.31	1.66	-0.02	0.05
AFR (Y)[2]	8	229	1.18	0.34	1.55[3]	-0.14[4]	0.28[4]
AUST(Y)[2]	8	24	0.94	0.53	1.32[3]	-0.16[4]	0.31[4]
AMER(Y)[2]	8	40	0.50	0.10	0.69[3]	-0.16[4]	0.33[4]

1: Data from Bowcock et al., (1994), with hypervariable locus removed. **2**: Data from Pritchard et al., (1998). **3**: Estimate of $\ln(\overline{V})$ adjusted by addition of $\ln(4)$ to account for four-fold smaller N_e on Y chromosome. **4**: Values of bias and variance estimated from simulations (Table 1).

We can treat this problem in the framework of the linear model. Let $\ln(\overline{V}_{i,j})$ be the value of $\ln(\overline{V})$ in the ith population and at the jth set of loci. Then we have

$$\ln(\overline{V}_{i,j}) - \ln(\hat{N}_i) - \ln(\hat{\mu}_j s_2) - \hat{B}_{i,j} = \epsilon_{i,j}. \tag{15}$$

Here, each population has a unique population size N_i, and each set of loci has a unique mutation rate μ_j. The quantity $\epsilon_{i,j}$ is a random error term which, as before, is Normally distributed, with mean 0, and variance $\sigma^2_{i,j}$. We suggest the following test statistic, X^2:

$$X^2 = \sum_{i=1}^{d}\sum_{j=1}^{n} \frac{\{\ln(\overline{V}_{i,j}) - \ln(\hat{N}_i) - \ln(\hat{\mu}_j s_2) - \hat{B}_{i,j}\}^2}{\hat{\sigma}^2_{i,j}}, \tag{16}$$

where d is the number of populations (three), and n is the number of sets of loci (two). Although there are $d+n$ parameters here, it turns out that it is not possible to obtain a unique set of estimates for all of these. Notice, for example, that addition of a constant c to all values of $\ln(\hat{N}_i)$, and subtraction of c from all values of $\ln(\hat{\mu}_j)$ would lead to identical estimates of $\mathrm{E}[\ln(\overline{V}_{i,j})]$. Therefore, we can arbitrarily set $\ln(\hat{\mu}_1) = 0$, and then estimate the remaining $d+n-1$ parameters. By a similar argument to that used above, we anticipate that X^2 will be Chi-square distributed with $dn-(d+n-1)$ degrees of freedom, provided

that the $\epsilon_{i,j}$ are independent. If anything, the $\epsilon_{i,j}$ are likely to be positively correlated. It appears that in that case, the Chi-Square distribution is conservative.

To analyze the data in Table 4., we fixed $\ln(\hat{\mu}_1) = 0$, and then estimated the remaining four parameters in our model (three population size parameters, and one mutation rate parameter). These parameters were estimated by finding the values that minimize X^2. Again, this corresponds to weighted least squares estimation. After estimating the parameters, we computed X^2 to be 0.49 which is nonsignificant. Thus, we do not find evidence for selective events at a local or regional scale on the Y chromosome and, in particular, the apparently low levels of Y chromosome variation in the Americas can be explained by stochastic effects alone.

Discussion

We have presented a statistical framework for testing hypotheses about the levels of variation at different sets of microsatellite loci. This framework can be used in a number of different contexts: for testing whether different types of loci have different mutation rates; for testing whether different parts of the genome have similar levels of variation; and for testing whether there is consistency in the levels of variation among loci and populations.

In addition, we applied our test to analyses of human microsatellite data. While we found significantly less variation among Y chromosome microsatellites than among autosomal microsatellites, this distinction disappeared once the four-fold difference in population size was considered. This argues against the hypothesis proposed elsewhere (e.g., Dorit et al., 1995), that the Y chromosome has anomalously little variation. We also looked for heterogeneity in levels of variation in different populations of humans, comparing Y chromosomal and autosomal data. Significant heterogeneity might suggest a Y chromosome bottleneck in particular human populations (caused by selection, or a high variance in male reproductive success). We found no evidence for such effects.

Acknowledgements

This work was supported by NIH grants GM28428 and GM28016 to MWF. JKP was supported by a Howard Hughes predoctoral fellowship. Thanks to M. Seielstad for allowing us to make use of unpublished microsatellite data. JKP wishes to thank the organizers of the TriNational Workshop, Drs. Uyenoyama, Takahata, and von Haeseler, for their work in hosting this excellent conference.

References

Anger, B., L. Bernatchez, A. Angers, L. Desgroseillers. 1995. Specific microsatellite loci for brook charr reveal strong population subdivision on a microgeographic scale. Journal Of Fish Biology. **47**:177-185.

Bowcock, A. M., A. R. Linares, J. Tomfohrde, E. Minch, J. R. Kidd and L. L. Cavalli-Sforza. 1994. High resolution of human evolutionary trees with polymorphic microsatellites. Nature **368**:455-457.

Cooper, G., W. Amos, D. Hoffman, and D. C. Rubinsztein. 1996. Network analysis

of human Y microsatellite haplotypes. Human Molecular Genetics **4**:1759-1766.

Dorit, R.L., H. Akashi, and W. Gilbert. 1995. Absence of polymorphism at the ZFY locus on the human Y chromosome. Science. **268**:1183-1185.

Estoup, A., C. Tailliez, J-M. Cornuet, and M. Solignac. 1995. Size homoplasy and mutational processes of interrupted microsatellites in two bee spices, Apis mellifera and Bombus terrestris (Apidae). Molecular Biology And Evolution. **12**:1074-1084.

Goldstein, D. B., A. R. Linares, L. L. Cavalli-Sforza, and M. W. Feldman. 1995. An evaluation of genetic distances for use with microsatellite loci. Genetics **139**:463-471.

Goldstein, D.B., L.A. Zhivotovsky, K. Nayar, A.R. Linares, L.L. Cavalli-Sforza, M.W. Feldman. 1996. Statistical properties of the variation at linked microsatellite loci: Implications for the history of human Y chromosomes. Molecular Biology And Evolution. **13**:1213-1218.

Lehmann, T., W. A. Hawley, L. Kamau, D. Fontenille, F. Simard, and F. H. Collins. 1996. Genetic differentiation of Anopheles gambiae populations from East and West Africa: Comparison of microsatellite and allozyme loci. Heredity. **77**:192-200.

Linares, A. R., K. Nayar, D. B. Goldstein, J. M. Hebert, M. T. Seielstad, P. A. Underhill, A. A. Lin, M. W. Feldman, and L. L. Cavalli Sforza. 1996. Geographic clustering of human Y-chromosome haplotypes. Ann. Hum. Genet. **60**:401-408.

Michalakis, Y., and M. Veuille. 1996. Length variation of CAG-CAA trinucleotide repeats in natural populations of Drosophila melanogaster and its relation to the recombination rate. Genetics. **143**:1713-1725.

Moran, P. A. P. 1975. Wandering distributions and the electrophoretic profile. Theor. Pop. Biol. **8**:318-330.

Paetkau, D., W. Calvert, I. Stirling, and C. Strobeck. Microsatellite analysis of population structure in Canadian polar bears. Molecular Ecology. **4**:347-354.

Pritchard, J. K., and M. W. Feldman, 1996. Statistics for microsatellite variation based on coalescence. Theor. Pop. Biol. **50**:325-344.

Pritchard, J.K., M.T. Seielstad, A. Perez-Lezaun, and M.W. Feldman, 1998. The ages of human populations: a study of Y chromosome microsatellites. Submitted.

Roe, A. 1992. "Correlations and Interactions in Random Walks and Population Genetics." Ph.D. Thesis, University of London, London, UK.

Slatkin, M. 1995. A Measure of Population Subdivision Based on Microsatellite Allele Frequencies. Genetics. **139**:457-462.

Stuart, A. and Ord, K. 1987. "Kendall's Advanced Theory of Statistics." John Wiley & Sons, New York.

Taylor, A. C., W. B. Sherwin, and R. K. Wayne. 1994. Genetic variation of microsatellite loci in a bottlenecked species: The northern hairy-nosed wombat Lasiorhinus krefftii. Molecular Ecology. **3**:277-290.

Tishkoff S. A., E. Dietzsch, W. Speed, A.J. Pakstis, J. R. Kidd, K Cheung, et al., (1996). Global patterns of linkage disequilibrium at the CD4 locus and modern human origins. Science **271**:1380-1387.

Zhivotovsky, L.A., and M.W. Feldman. 1995. Microsatellite variability and genetic distances. Proc. Natl. Acad Sci. USA **92**:11549-11552.