VNTR Allele Frequency Distributions Under the Stepwise Mutation Model: A Computer Simulation Approach

Mark D. Shriver, Li Jin, Ranajit Chakraborty and Eric Boerwinkle[1]

Center for Demographic and Population Genetics, Graduate School of Biomedical Sciences, University of Texas Health Science Center at Houston, Houston, Texas 77225

Manuscript received September 27, 1992
Accepted for publication March 26, 1993

ABSTRACT

Variable numbers of tandem repeats (VNTRs) are a class of highly informative and widely dispersed genetic markers. Despite their wide application in biological science, little is known about their mutational mechanisms or population dynamics. The objective of this work was to investigate four summary measures of VNTR allele frequency distributions: number of alleles, number of modes, range in allele size and heterozygosity, using computer simulations of the one-step stepwise mutation model (SMM). We estimated these measures and their probability distributions for a wide range of mutation rates and compared the simulation results with predictions from analytical formulations of the one-step SMM. The average heterozygosity from the simulations agreed with the analytical expectation under the SMM. The average number of alleles, however, was larger in the simulations than the analytical expectation of the SMM. We then compared our simulation expectations with actual data reported in the literature. We used the sample size and observed heterozygosity to determine the expected value, 5th and 95th percentiles for the other three summary measures, allelic size range, number of modes and number of alleles. The loci analyzed were classified into three groups based on the size of the repeat unit: microsatellites (1–2 base pair (bp) repeat unit), short tandem repeats [(STR) 3–5 bp repeat unit], and minisatellites (15–70 bp repeat unit). In general, STR loci were most similar to the simulation results under the SMM for the three summary measures (number of alleles, number of modes and range in allele size), followed by the microsatellite loci and then by the minisatellite loci, which showed deviations in the direction of the infinite allele model (IAM). Based on these differences, we hypothesize that these three classes of loci are subject to different mutational forces.

VARIABLE number tandem repeat (VNTR) loci are useful genetic markers at which alleles differ in length due to differences in the number of short sequence motifs arranged adjacent to one another. VNTR loci often have high levels of heterozygosity and large numbers of alleles, characteristics that make them informative genetic markers and useful tools in forensic science (JEFFREYS, WILSON and THEIN 1985; NATIONAL RESEARCH COUNCIL 1992) and gene mapping (NAKAMURA *et al.* 1987; SILVER 1992). Despite the usefulness of VNTR loci and their potential importance in genome organization and gene regulation, there are still many unanswered questions about their mutational mechanisms and population dynamics. VNTR allele frequency distributions can be affected by population forces (*e.g.*, genetic drift, hitchhiking, selection) and molecular level forces (*e.g.*, mutational mechanisms and mutation rates). Likely mechanisms for repeat number change at VNTR loci are slipped-strand mispairing, unequal sister chromatid exchange and genetic recombination.

Mutation-drift equilibrium models formulated to interpret genetic variation observed at the protein electrophoretic and DNA sequence level are useful tools in an investigation of the population dynamics of VNTR loci. Two models that have been used to study the population genetics of VNTR loci (JEFFREYS, Wilson and THEIN 1985; CHAKRABORTY *et al.* 1991; DEKA, CHAKRABORTY and FERRELL 1991; EDWARDS *et al.* 1992; CHAKRABORTY and DAIGER 1991; VALDES, SLATKIN and FREIMER 1993) are the infinite allele model [(IAM) WRIGHT 1949; KIMURA and CROW 1964] and the stepwise mutation model [(SMM) OHTA and KIMURA 1973; WEHRHAHN 1975; KIMURA and OHTA 1978].

The IAM and the SMM both have credibility when one considers the likely mechanisms of tandem repeat number change. The IAM stipulates that each mutation gives rise to a new allele not found in the population, resulting in an infinite number of allelic states. Unequal exchange between long tandem repeat arrays can result in a very large number of different-sized alleles, as assumed in the IAM. For the SMM, on the other hand, allelic states can be envisioned as a series of integer points on a line. In the one-step stepwise

[1] To whom correspondence should be addressed.

model, mutation events result in the conversion of an allele to the allelic state either one step to the right or to the left. Empirical evidence (reviewed in LEVINSON and GUTMAN 1987) suggests that changes in array length due to slipped-strand mispairing are usually of one repeat unit. Regardless of the mechanism of repeat number change, the IAM and the SMM provide the upper and lower limit, respectively, of the extent of variation expected. In other words, if mutation and drift are the only forces acting to determine the allele frequency distribution (*i.e.*, there is no population substructure, selection or mutational boundaries) the maximum and minimum number of alleles expected for a given heterozygosity are predicted by the IAM and the SMM, respectively (CHAKRABORTY and JIN 1992). It is therefore reasonable to test hypotheses about the mutational mechanisms and population dynamics of VNTR loci by predicting the number of alleles from the observed heterozygosity for both models and comparing the predicted with the observed number of alleles.

MATERIALS AND METHODS

Analytical formulations of the SMM and the IAM allow for estimation of the parameter $M = 4N_e v$, (N_e is the effective population size, and v is the mutation rate) from both the observed number of alleles and the heterozygosity (KIMURA and CROW 1964; OHTA and KIMURA 1973; WEHRHAHN 1975; KIMURA and OHTA 1978; CHAKRABORTY *et al.* 1991). The expected number of alleles can thus be predicted given the observed heterozygosity and sample size for both the IAM and the SMM and compared with the observed numbers of alleles. Analytical formulations of the IAM include determination of a probability distribution for the number of alleles (EWENS 1972; FUERST, CHAKRABORTY and NEI 1977). There exists no such theory for the SMM as formulated by KIMURA and OHTA (1978), the KIMURA and OHTA analytical formulation (KOAF). It is therefore not possible to judge the significance of differences between observed and expected values of the number of alleles. Neither model allows one to predict the number of modes or the allelic size range, two other important characteristics of VNTR allele frequency distributions. In addition, the analytical solution to the SMM for number of alleles is based on an approximation that may not hold when the mutation rate and population size are at levels applicable to VNTR loci in contemporary human populations.

We simulated the effects of genetic drift and mutation on VNTR allele frequency distributions. For the results presented here, the effective population size (N_e) was set to 5,000, and the population was started with 10,000 ($2N_e$) chromosomes fixed for one allele of arbitrary size (*i.e.*, no boundaries were imposed on the number of repeats per allele). One generation consisted of randomly drawing (with replacement) $2N_e$ chromosomes from the population of the previous generation. When an allele was chosen, a second random number, from a uniform distribution (0, 1) was generated to determine if the chosen allele mutates. If this number was less than v/2, the allele became one repeat unit larger; if this number was greater than 1 − v/2, the allele became one repeat unit smaller; otherwise, the size remained the same. Population sampling continued for 20,000 ($4N_e$) generations, by which time a steady state was reached for all four summary measures (data not shown). The allele frequency distributions from the final generation were stored in a data file for analysis. This process was repeated 50 independent times for each of 21 mutation rates ranging from 10^{-2}-10^{-5}.

Because the characteristics of the allele frequency distribution are dependent on the size of the sample, we sampled from the population distributions to simulate the effect of sample size. Samples of a specified size were drawn 100 times from each of the 50 frequency distributions for a total of 5000 samples drawn from each of the 21 simulation sets. This approach allowed us to get the average distribution characteristic including average number of modes, number of alleles, the heterozygosity and the range in allele size for each independent simulation replication and then for the 50 replications at each mutation rate.

Heterozygosity was calculated as one minus the sum of the squared allele frequencies. Range in allele size was calculated as the difference between the number of repeats in the largest allele and the number in the smallest. A "mode" was defined as any allele that is more frequent than both the allele one step larger and the allele one step smaller and the number of modes was counted. To estimate the significance of deviations from the expected values, the 5th and the 95th percentiles for each summary measure were then estimated as the 250th smallest and largest values, respectively.

RESULTS

The relationships between the analytical expectations from the KOAF and two summary measures, heterozygosity and number of alleles, resulting from the simulations are shown in Figure 1, A and B, as a function of the mutation rate. The data shown are for a sample of 250 chromosomes. As Figure 1A shows, the heterozygosities from the simulation results match the KOAF expectations almost perfectly. In striking contrast, the number of alleles from the simulation results deviates from the analytical expectations at all but very low mutation rates (Figure 1B). Figure 1C shows the effect of increasing mutation rate on the expected number of modes and the allelic size range. The allelic size range shows a more pronounced increase in response to the mutation rate than does the number of modes.

The relationship between heterozygosity and the number of alleles for the IAM, the KOAF and the computer simulations is presented in Figure 2A. It is again evident that computer simulations of the SMM yield a larger number of alleles than the KOAF. As expected, the infinite allele model exhibits a larger expected number of alleles for a given heterozygosity. Percentiles for the simulations were calculated after simulating the effect of sampling on the summary characteristics. Figure 2B shows the 5th and 95th percentiles for the simulation SMM and the IAM. These two confidence belts show a great deal of overlap, especially at low to moderate levels of heterozygosity. In fact, some degree of overlap is seen between the two confidence belts up to a heterozygosity of 95%. However, the overlap is not complete

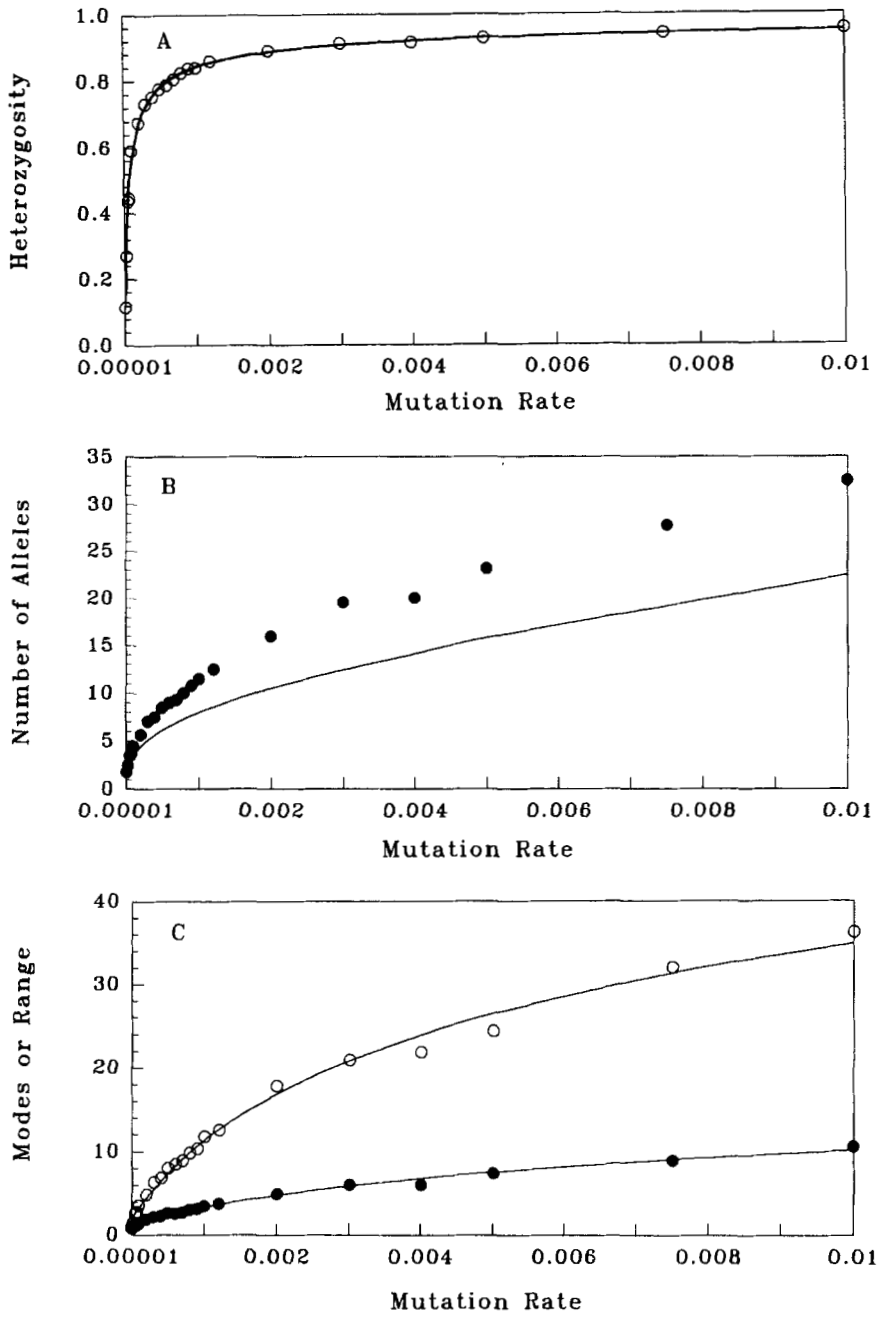

FIGURE 1.—Relationships between mutation rate and heterozygosity, number of alleles, number of modes and range in allele size. (A) Relationship between mutation rate and heterozygosity. Circles represent data from the simulations under the one-step SMM (see METHODS) considering a sample size of 250 chromosomes. The line represents the expected heterozygosity from the KOAF. (B) Relationship between mutation rate and number of alleles. Circles represent data from the simulations under the one-step SMM (see METHODS) considering a sample size of 250 chromosomes. The dashed line is the polynomial regression line fitted to the simulation results. The solid line represents the expected number of alleles from the KOAF. (C) Relationship between mutation rate and number of modes and range in allele size. Open circles represent data for range from the simulations under the one-step SMM (see METHODS) considering a population size of 250 chromosomes. Filled circles represent the data for number of modes from the simulations under the one-step SMM (see METHODS) considering a population size of 250 chromosomes. The dotted lines are polynomial regression lines fitted to the data points.

and for heterozygosities larger than 50–60%, there are areas of the confidence belts that are specific to each model. Figure 2, C and D, presents the relationships between the heterozygosity and the allelic size range and number of modes, respectively, for a sample size of 250 chromosomes. Both curves show a slow increase at low levels of heterozygosity and then a rapid increase in the number of modes and the allelic size range at higher levels. Also shown in Figure 2, C and D, are the 5th and 95th percentiles for the number of modes and the allelic size range.

To illustrate the effect that sample size has on the four summary measures, we have drawn seven samples of different size (10 to 1000) from the 50 simulations at $v = 0.001$. The results are shown in Figure 3, A and B. Figure 3A shows the effect of sample size on the expected number of alleles for both the simulations and the KOAF. It is again evident that the analytical prediction from the KOAF is less than the simulation result for number of alleles. Also notable is the difference in the shape of the curve between the KOAF and the simulation expectations. The analytical (KOAF) expectation of the number of alleles reaches a plateau quickly, showing very little increase for sample sizes >100 chromosomes. The expected number of alleles from the simulations, on the other hand, shows a steady increase even at sample sizes greater than 750 chromosomes. Figure 3B shows the effect of sample size on the allelic size range, number of modes, and heterozygosity. The range in allele size

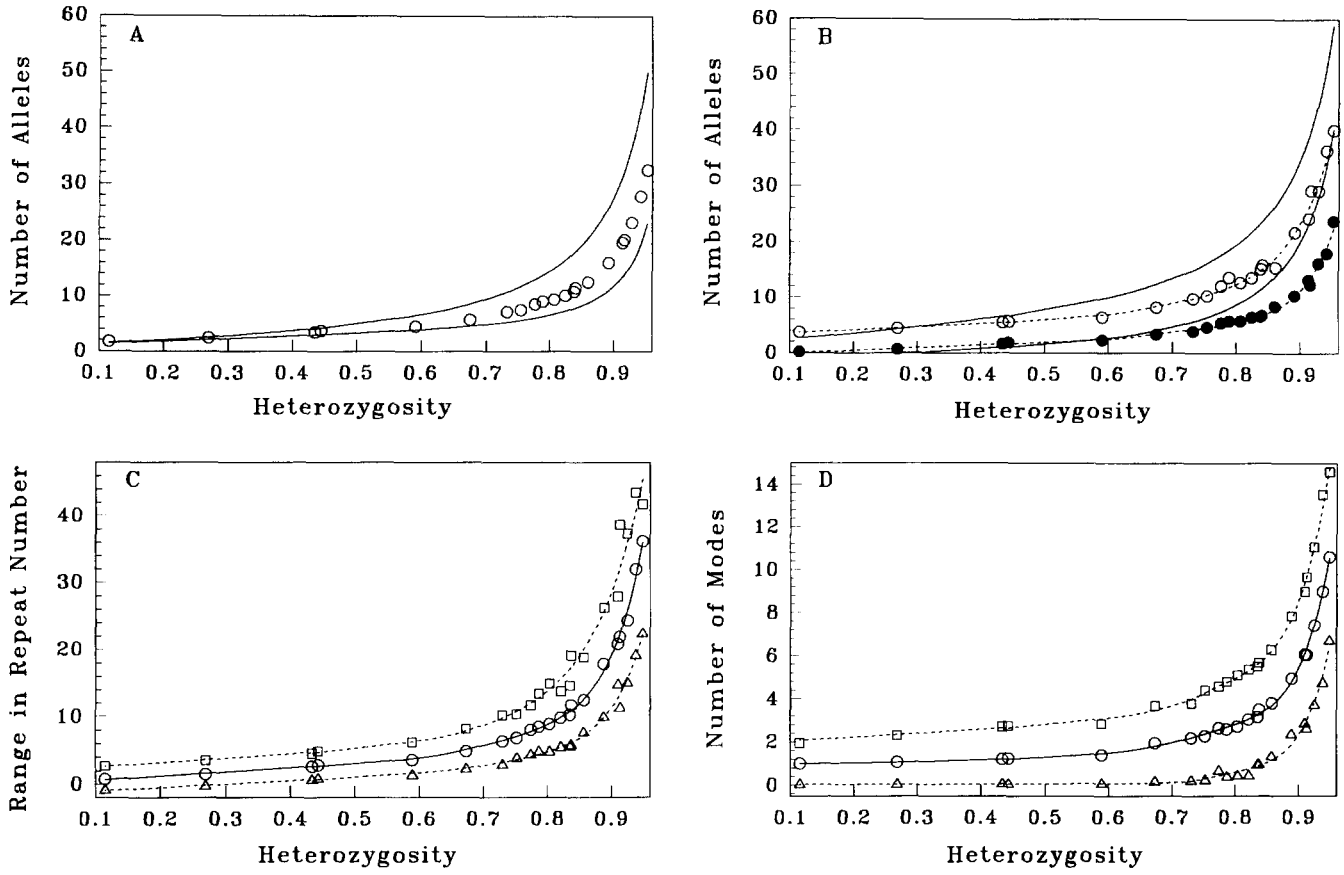

FIGURE 2.—Relationships between heterozygosity and number of alleles, number of modes and range in allele size. (A) Expected numbers of alleles under three models. Circles represent data from the simulations under the one-step SMM (see METHODS) considering a sample size of 250 chromosomes. Lower solid line represents the expectation of the Kimura and Ohta SMM for a sample of 250 chromosomes. Upper solid line represents the expectation for the infinite allele model for a sample of 250 chromosomes. (B) Percentiles of the distribution of number of alleles under the IAM and from the one-step SMM simulation. Open and filled circles represent the 95th and 5th percentiles, respectively, of the simulation results for a sample of 250 chromosomes. Dashed lines are polynomial regression lines fitted to these data points. The upper and lower solid lines are the 95th and 5th percentile lines, respectively, under the IAM for a sample of 250 chromosomes. (C) Relationship between heterozygosity and range in repeat number from simulations under the one-step SMM (see METHODS). Circles represent the mean numbers of modes considering a sample of 250 chromosomes. Triangles and squares represent the 5th and 95th percentiles, respectively. Lines are polynomial regression lines fitted to the data points. (D) Relationship between heterozygosity and number of modes from the simulations under the one-step SMM (see METHODS). Circles represent the mean range in repeat number considering a sample size of 250 chromosomes. Triangles and squares represent the 5th and 95th percentiles, respectively. Lines are polynomial regression lines fitted to the data points.

and the number of alleles are more sensitive to sample size than the heterozygosity and the number of modes. The number of modes and the heterozygosity reach plateaus at smaller numbers of chromosomes sampled.

We then addressed the question of how robust were our results to variation in effective population size by doing a limited set of simulations at different effective population sizes. Simulations were performed with N_e = 1,000, 2,500 and 10,000 for three different mutation rates. The results for the number of alleles are presented graphically in Figure 4. The summary measures observed in these additional simulations are very close to the simulation results when the effective population size is 5,000, and they show no evidence of any systematic bias. The conclusion for the allelic size range and the number of modes was very similar to Figure 4 (data not shown).

From the sample sizes and observed heterozygosities of a total of 54 VNTR loci reported in the literature, we estimated the expected value, and the 5th and 95th percentiles for the number of alleles, number of modes and allelic size range. These data were then compared with the observed values of the number of alleles, the number of modes and the range in allele size. Our comparisons were confined to substantial samples randomly drawn from large homogeneous populations. Unambiguous typing and discrete determination of the number of repeats per allele were also important criteria in the choice of loci. The loci were grouped by repeat unit size into three commonly used classes, microsatellites with 1–2 bp repeats, short tandem repeats (STRs) with 3–5 bp repeats, and minisatellites with 15–70 bp repeats. Tables 1–3 give the expected number of alleles and

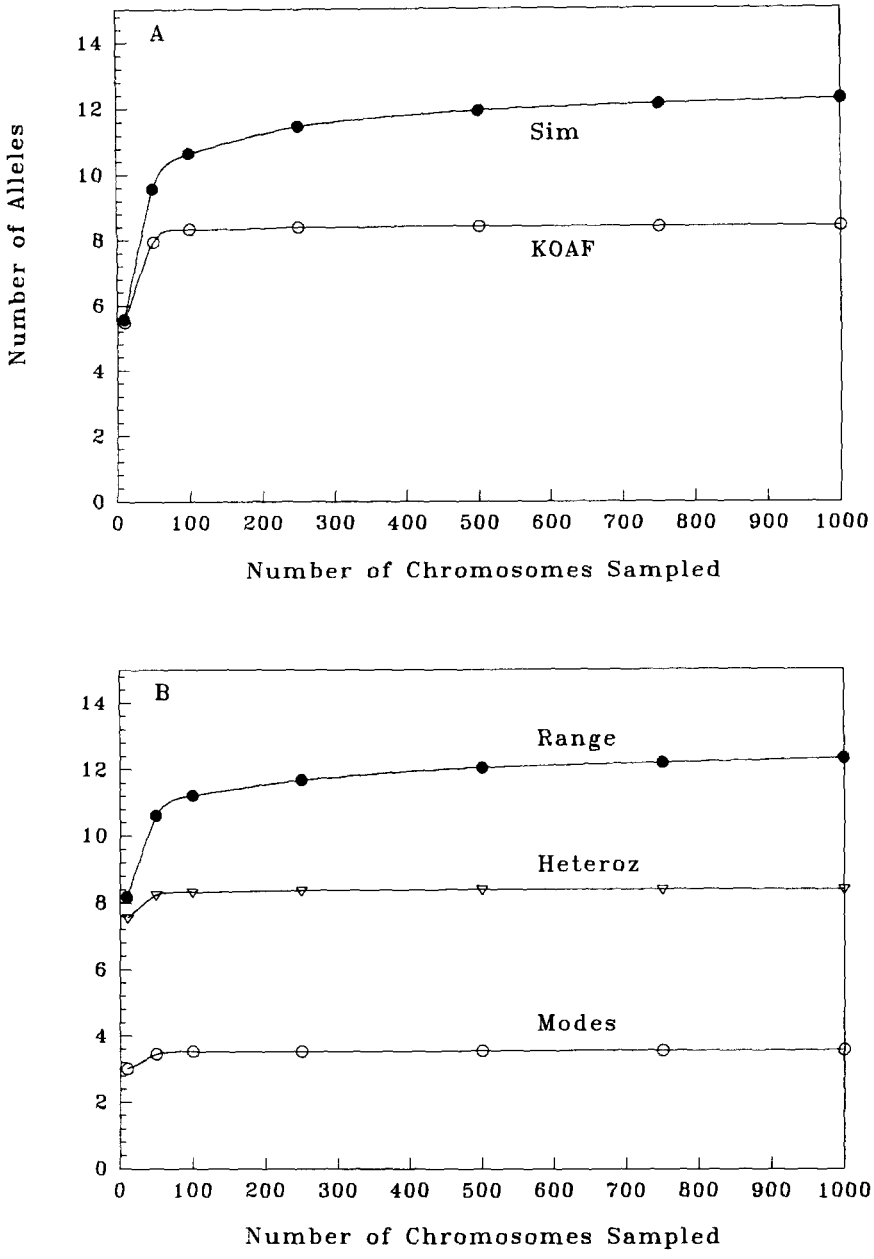

FIGURE 3.—Relationships between the sample size and number of alleles, number of modes and range in allele size. (A) Relationship between sample size and the number of alleles under and for the one-step simulation data (filled circles) and the KOAF (open circles) for number of alleles for $v = 0.001$. Lines are the polynomial regression lines fitted to the data points. (B) Relationship between sample size and the range (filled circles), heterozygosity (triangles), multiplied by 10 to fit on the same scale, and number of modes (open circles) from the one-step SMM simulation data. The mutation rate for this simulation data set was 0.001. The lines are the polynomial regression lines fitted to the data points.

probabilities and percentiles from the IAM and the computer simulations, respectively, for the three classes of loci. The data in Table 1 indicate that 77% (24/31) of the microsatellite loci fit both models, 6% (2/31) fit the simulation SMM only, and 16% (5/31) fit the IAM only. For the STR loci (see Table 2), 75% (9/12) fit both models and 25% (3/12) fit the simulation SMM only. There were no STR loci that had significantly more alleles than predicted from the simulation results. Table 3 shows the results for the minisatellite data; 45% (5/11) of the minisatellite loci fit both models, 27% (3/11) fit the IAM model only and 18% (2/11) fit the simulation SMM only. Only one locus, D2S44, had significantly more alleles than predicted from the simulation and significantly fewer than expected under the IAM and thus fell between the two models.

Tables 4–6 show the observed values, simulation expectations and 5th and 95th percentiles for the number of modes and allelic size range for the three classes of tandem repeat loci. Whereas none of the microsatellite loci (see Table 4) had more modes than the number of modes at the 95th percentile, 19% (6/31) of the microsatellite loci had an observed number of modes that was equal to the number at the 95th percentile. Nine of the 31 microsatellite loci had observed values of the size range that were outside the 95th percentile from the simulations. Three microsatellite loci had a size range equal to the value at the 95th percentile. In striking contrast, none of the STR loci (see Table 5) demonstrated significantly more modes or larger size ranges than predicted from the simulation results. Six of the 11 or 55% of the minisatellite loci (see Table 6) had significantly more

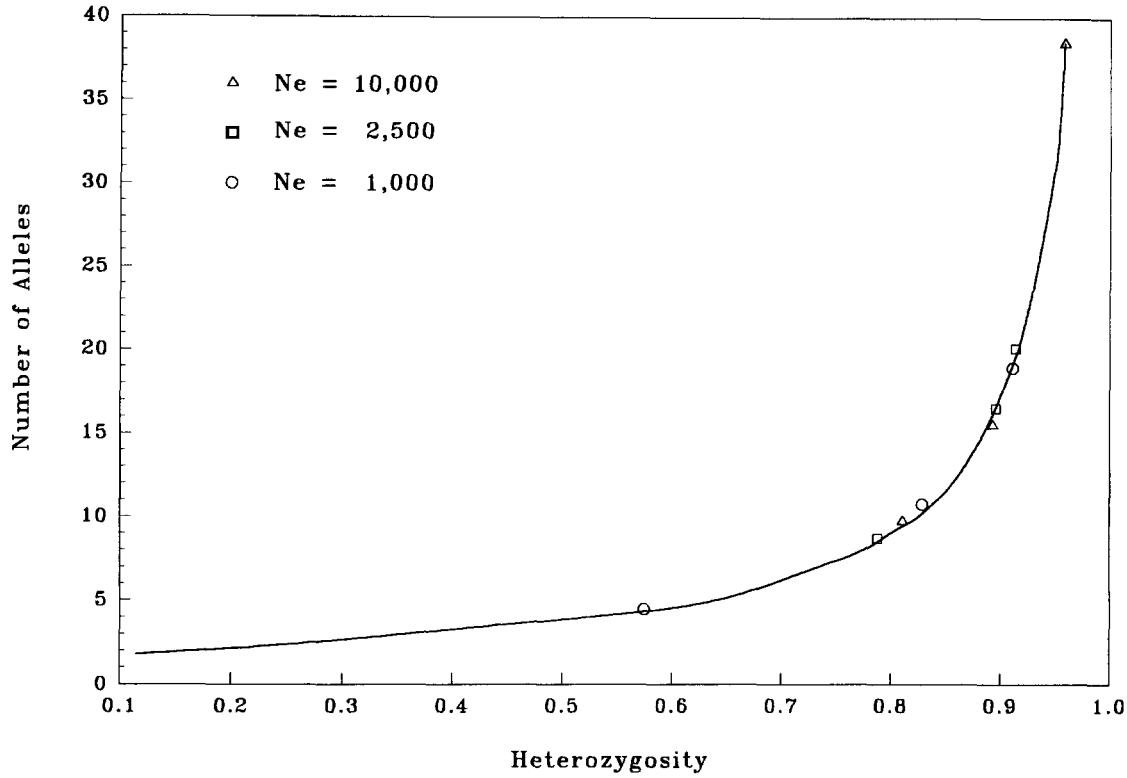

FIGURE 4.—Relationship between heterozygosity and the number of alleles for different values of N_e. The line is a polynomial regression line fitted to the average number of alleles from the simulation results when N_e was equal to 5000. The points are simulation results of average number of alleles for three different effective population sizes.

modes than expected from our simulation model. Seven of the 11 minisatellite loci had a range in allele size that was significantly greater than expected.

DISCUSSION

We have performed multiple computer simulations of VNTR evolution under a one-step stepwise mutation model in finite populations to estimate the equilibrium distribution of the number of alleles, the allelic size range, the number of modes in the allele frequency distribution and the heterozygosity. As expected, increases in both population size and mutation rate resulted in increases in all four summary measures. Two of these summary measures, heterozygosity and number of alleles, can be analytically derived from the IAM and the SMM. The simulation results agreed with the prediction of the SMM as formulated by KIMURA and OHTA [1978, (KOAF)] for the heterozygosity but not for the number of alleles. As expected, the simulation results, which were derived under a stepwise mutational mechanism, gave heterozygosities and numbers of alleles that were much lower than predicted by the IAM. Comparison of the simulation results to VNTR distributions reported in the literature showed differences between loci classified by repeat unit length. In general, STR loci were most similar to the simulation results under the SMM for the three summary measures (number of alleles, number of modes and range in allele size), followed by the microsatellite loci and then by the minisatellite loci, which showed deviations in the direction of the IAM. Based on these differences, we hypothesize that these three classes of loci are subject to different mutational forces.

The analytical prediction of the average heterozygosity based on the KOAF (OHTA and KIMURA 1973; KIMURA and OHTA 1978; WEHRHAHN 1975) agreed with the simulation results for all mutation rates studied. However, the number of alleles found in these simulation studies was larger than predicted by the KOAF when the mutation rate was high enough to produce high levels of heterozygosity. An approximation was needed to derive the analytical formulation of the number of alleles from the average heterozygosity (KIMURA and OHTA 1978), and it is possible that this approximation was an oversimplification. CHAKRABORTY *et al.* (1991) and EDWARDS *et al.* (1992) have examined the fit of minisatellite and STR loci, respectively, to the IAM and the KOAF. In light of the deviation of the average number of alleles obtained in the simulations from the KOAF expectations, we have reexamined the fit of these loci to the SMM using the simulation results by judging the significance of differences between the observed number of alleles and the number of alleles expected from the one-step SMM. In addition, the simulation results

TABLE 1

Microsatellite loci: observed number of alleles and IAM and SMM expectations

Locus	n[a]	H[b]	k[c]	E(IAM)[d]	P(IAM)[e]	E(k)[f]	5th%[g]	95th%[h]	Fit[i]
Mfd1[j]	86	0.575	5	5.19	0.591	4.52	1.64	5.76	Both
Mfd2[j]	86	0.323	4	2.64	0.924	2.76	1.49	5.17	Both
Mfd3[j]	86	0.702	6	7.77	0.304	6.28	3.84	8.54	Both
Mfd4[j]	86	0.486	6	4.06	0.925	3.33	1.18	5.17	Both
Mfd5[j]	154	0.816	11	14.40	0.183	9.52	5.97	13.05	Both
Mfd6[j]	154	0.562	7	5.55	0.840	3.48	1.45	5.57	IAM
Mfd7[j]	154	0.529	6	5.01	0.796	3.21	1.23	5.24	Both
Mfd8[j]	154	0.627	8	6.74	0.795	4.66	2.41	6.89	Both
Mfd9[j]	154	0.738	9	10.04	0.438	7.36	4.54	10.04	Both
Mfd10[j]	154	0.443	6	3.94	0.933	3.52	1.53	5.45	Both
c-act[k]	74	0.861	12	15.53	0.165	11.82	7.60	15.91	Both
TNFa[l]	188	0.873	12	21.32	0.008	13.61	8.76	18.43	SIM
TNFb[l]	187	0.730	6	10.09	0.086	7.29	4.48	9.93	Both
X75b-VSSM[m]	392	0.865	12	23.84	0.002	13.23	8.64	18.27	SIM
D16S291[n]	238	0.793	9	13.99	0.075	8.95	5.55	12.46	Both
D16S292[n]	158	0.742	10	10.24	0.555	7.46	4.65	10.26	Both
D16S298[n]	156	0.801	11	13.36	0.282	8.98	5.62	12.31	Both
D16S299[n]	158	0.725	8	9.56	0.358	7.07	4.36	9.75	Both
D16S300[n]	156	0.596	7	6.12	0.757	4.02	1.91	6.22	Both
D16S304[n]	152	0.611	12	6.40	0.995	4.31	2.15	6.55	IAM
D16S308[n]	92	0.767	10	10.15	0.569	7.71	4.79	10.57	Both
D16S301[n]	154	0.639	6	7.01	0.432	4.96	2.67	7.27	Both
D16S305[n]	108	0.809	15	12.71	0.821	9.05	5.64	12.50	IAM
D16S303[n]	212	0.434	6	4.03	0.924	3.69	1.72	5.59	Both
D16S285[o]	138	0.825	12	14.68	0.254	9.85	6.23	13.61	Both
D16S186[p]	224	0.613	10	6.89	0.876	4.39	2.13	6.66	IAM
D16S287[q]	158	0.790	10	12.68	0.240	8.64	5.42	11.84	Both
D8S135[r]	160	0.667	6	7.76	0.311	5.67	3.23	8.07	Both
angio-CA[s]	200	0.778	11	12.56	0.377	8.45	5.23	11.44	Both
PLAT[t]	290	0.773	18	13.23	0.947	8.51	5.27	11.72	IAM
ApoCII[u]	324	0.857	11	21.53	0.586	12.32	8.02	17.04	Both

[a] Number of chromosomes sampled; [b] heterozygosity; [c] observed number of alleles; [d] expected number of alleles under the IAM; [e] probability that the observed number of alleles is greater than the that expected under the IAM; [f] expected number of alleles based on the simulation results; [g] 5th percentile from the simulations; [h] 95th percentile from the simulations; [i] fit of this locus to the IAM, SIM or both; [j] WEBER and MAY 1989; [k] LITT and LUTY 1989; [l] JONGENEEL *et al.* 1991; [m] JANSEN *et al.* 1992; [n] THOMPSON *et al.* 1992; [o] KONRADI *et al* 1991; [p] PHILLIPS *et al* 1991a; [q] PHILLIPS *et al.* 1991b; [r] WOOD and SCHERTZER 1991; [s] KOTELEVTSEV *et al.* 1991; [t] SADLER, BLANTON and DAIGER 1991; [u] CHAKRABORTY *et al.* 1991.

TABLE 2

STR loci: observed number of alleles and IAM and SMM expectations

Locus	n[a]	H[b]	k[c]	E(IAM)[d]	P(IAM)[e]	E(k)[f]	5th %[g]	95th %[h]	Fit[i]
HUMHPRTB[j]	417	0.751	9	12.78	0.146	8.13	4.99	11.08	Both
HUMTHO1[j]	372	0.767	7	13.48	0.023	8.47	5.22	11.49	SIM
HUMRENA4[j]	354	0.364	4	3.52	0.757	4.38	2.17	5.98	Both
HUMFABP[j]	574	0.647	7	8.95	0.303	5.45	3.10	8.01	Both
HUMARA[j]	240	0.881	16	24.08	0.029	14.77	9.38	20.16	SIM
ACTBP2[k]	78	0.930	21	27.59	0.053	22.77	16.65	28.79	Both
Human TH[l]	70	0.785	5	10.29	0.025	8.00	4.90	11.12	SIM
CD4 gene[m]	200	0.689	7	8.74	0.327	6.30	3.71	8.73	Both
ACPP[n]	150	0.661	6	7.52	0.347	5.49	3.11	7.84	Both
MYCN[o]	118	0.415	3	3.52	0.533	3.64	1.62	5.59	Both
GABARBI[p]	178	0.715	7	9.43	0.237	6.91	4.23	9.47	Both
PLA2A1[q]	336	0.725	7	11.02	0.106	7.43	4.52	10.14	Both

[a] Number of chromosomes sampled; [b] heterozygosity; [c] observed number of alleles; [d] expected number of alleles under the IAM; [e] probability that the observed number of alleles is greater than the that expected under the IAM; [f] expected number of alleles based on the simulation results; [g] 5th percentile from the simulations; [h] 95th percentile from the simulations; [i] fit of this locus to the IAM, SIM or both; [j] EDWARDS *et al.* 1992; [k] POLYMEROPOULOS *et al.* 1991a; [l] POLYMEROPOULOS *et al.* 1991b; [m] EDWARDS *et al.* 1991; [n] DOAK *et al.* 1991; [o] FOUGEROUSSE *et al.* 1992; [p] DEAN *et al.* 1991; [q] Li Jin, personal communication.

TABLE 3

Minisattellite loci: observed number of alleles and IAM and SMM expectations

Locus	n[a]	H[b]	k[c]	E(IAM)[d]	P(IAM)[e]	E(k)[f]	5th %[g]	95th %[h]	Fit[i]
D17S5[j]	302	0.849	14	20.13	0.067	11.92	7.67	16.22	Both
D2S44[j]	302	0.975	67	90.13	<.001	49.02	32.96	56.65	Neither
D9S7[j]	272	0.846	15	20.81	0.083	11.44	7.30	15.93	Both
D14S13[j]	164	0.948	30	43.56	0.004	30.61	22.00	38.59	SIM
D19S20[j]	168	0.775	13	13.13	0.562	8.29	5.14	11.35	IAM
D16S83[j]	156	0.871	15	20.13	0.103	13.29	8.63	18.02	Both
D1S74[j]	156	0.931	22	34.46	0.004	24.73	17.35	31.47	SIM
D3S42[j]	168	0.750	15	10.73	0.950	7.64	4.75	10.53	IAM
Apo B[k]	480	0.768	12	14.13	0.321	8.57	5.31	11.67	Both
D4F35S1[l]	240	0.862	17	20.81	0.194	12.61	7.84	17.48	Both
D1S80[m]	198	0.791	16	13.32	0.571	8.78	5.45	11.93	IAM

[a] Number of chromosomes sampled; [b] heterozygosity; [c] observed number of alleles; [d] expected number of alleles under the IAM; [e] probability that the observed number of alleles is greater than the that expected under the IAM; [f] expected number of alleles based on the simulation results; [g] 5th percentile from the simulations; [h] 95th percentile from the simulations; [i] fit of this locus to the IAM, SIM or both; [j] ODELBERG *et al.* 1989; [k] CHAKRABORTY *et al.* 1991; [l] MARIAT, LAUTHIER and VERGNAUD 1991; [m] BUDOWLE *et al.* 1991.

enabled us to predict the allelic size range and number of modes and determine the significance of deviations of these two characteristics of VNTR allele frequency distributions.

Because our simulation results are dependent on the effective population size, it is important to consider the robustness of our conclusions with respect to this parameter. Estimation of the effective population size is difficult because not only does one need good estimates of the mutation rate and the extent of variation in the population, an accurate mutation-drift model is needed. Previous studies (*e.g.*, CHAKRABORTY and NEEL 1989; HARTL 1980; NEI and GRAUR 1984) have estimated the effective population size of various human populations at between 2,500 and 10,000. For the simulation results described here, the effective population size was set equal to 5,000 individuals. We addressed the question of how robust were our conclusions to variation in effective population size by doing a limited set of simulations at different effective population sizes. The summary measures observed in these additional simulations are very close to the simulation results when the effective population size is 5,000 (see Figure 4), and given that they are sometimes smaller and sometimes larger than the predictions, we conclude that there is no systematic bias within this range of effective population sizes.

We have studied the effects of varying sample sizes by repeated sampling from the same set of 50 independent simulations at the population level (with N_e = 5000) and, therefore, samples from the same simulation are not independent. It is tedious to demonstrate the sample size effect by separate independent simulations for each combination of parameter values. Nevertheless, we conducted independent population simulations for selected parameter values. For example, with 500 independent replications (N_e = 5000), a mutation rate 0.001 and a sample size of 250 chromosomes, the average number of alleles was 11.09 with a confidence belt (7, 15), the mean number of modes was 3.42 with the confidence belt (1, 6) and the average range was 11.32 with a confidence belt (6, 17). None of these values were substantially different from the results shown in Figure 2, A–D. Therefore, the nonindependence from sampling each simulation multiple times does not change the conclusions of this study.

At low levels of heterozygosity there is a great deal of overlap between the confidence intervals of the number of alleles based on the IAM and those observed from the simulations (see Figure 2B). However, at higher levels of heterozygosity (H > 50%), the confidence intervals have large areas that do not overlap. Therefore, when heterozygosity is high, loci evolving by IAM-like mechanisms can be distinguishable from loci that mutate by stepwise processes. We examined the allele frequency distributions of a total of 31 microsatellites, 12 STRs and 11 minisatellites. From the observed heterozygosity and sample size at each locus, we estimated the number of alleles, number of modes, and range in allele size using the simulation SMM results. There were differences among the three groups of loci in terms of their agreement with the simulation results. None of the STR loci had significantly more alleles than the number expected given the simulation results, whereas 16% of the microsatellite loci and 36% of the minisatellite loci had more alleles than expected. A similar trend was evident in the fit of the VNTR loci to the simulated data for number of modes and allelic size range. Minisatellite loci as a group had more modes than expected from the SMM simulations. Microsatellite and STR loci did not have significantly more modes than expected, although a number of microsatellite loci (19%) were borderline significant for excess modes. For the allelic size range, however, both microsatellite and

TABLE 4

Microsatellite loci: observed size range and number of modes and predictions from simulation results

Locus	m[a]	E(m)[b]	5th%[c]	95th%[c]	r[e]	E(r)[f]	5th%[g]	95th%[h]
Mfd1	3[?]	1.26	0	2.71	10*	2.61	0.76	5.43
Mfd2	1	1.43	0.28	2.85	3	2.76	0.36	4.34
Mfd3	3	2.15	0.28	4.00	7	5.69	2.77	8.48
Mfd4	2	1.10	0	2.53	6*	2.08	0.27	4.66
Mfd5	4	2.95	0.65	5.24	18*	9.35	5.09	14.49
Mfd6	2	1.15	0	2.57	6	2.47	0.51	6.35
Mfd7	3[?]	1.06	0	2.47	6[?]	2.17	0.28	5.81
Mfd8	2	1.56	0.02	3.09	7	3.80	1.47	7.37
Mfd9	3	2.39	0.39	4.32	10[?]	6.81	3.55	9.74
Mfd10	3[?]	1.24	0.08	2.73	7*	2.60	0.53	4.58
c-actin	4	4.08	1.55	6.71	14	12.90	6.98	19.77
TNFa	3	4.22	1.53	6.92	12	14.31	7.97	22.05
TNFb	2	2.34	0.38	4.23	6	6.67	3.51	9.62
X75b-VSSM	4	3.86	1.27	6.43	11	13.47	7.75	21.42
D16S291	2	2.71	0.57	4.84	8	8.52	4.58	12.95
D16S292	2	2.41	0.41	4.33	9	6.92	3.63	9.85
D16S298	2	2.80	0.57	4.99	10	8.68	4.72	12.98
D16S299	2	2.31	0.37	4.16	7	6.48	3.34	9.32
D16S300	2	1.34	0.00	2.82	6	3.08	0.94	6.86
D16S304	3[?]	1.45	0.00	2.96	14*	3.41	1.18	7.11
D16S308	2	2.56	0.49	4.65	15*	7.41	3.84	11.04
D16S301	1	1.66	0.06	3.26	5	4.14	1.17	7.60
D16S305	5	2.91	0.63	5.17	14[?]	8.93	4.89	13.59
D16S303	2	1.28	0.12	2.78	7*	2.78	0.73	4.50
D16S285	3	3.08	0.77	5.41	11	9.80	5.51	15.24
D16S186	3[?]	144	0.00	2.95	24*	3.5	1.15	7.25
D16S287	2	2.71	0.52	4.84	12	8.27	4.48	12.18
D8S135	1	1.89	0.18	3.53	5	4.92	2.24	8.11
angio-CA	2	2.61	0.49	4.67	10	7.97	4.29	11.82
PLAT	5[?]	2.58	0.51	4.62	22*	7.95	4.27	11.68
Apo CII	5	3.62	1.06	6.12	18	12.46	7.23	20.31

[a] Observed number of modes; [b] expected number of modes from the simulation results; [c] 5th percentile based on the simulations; [d] 95th percentile based on the simulations; [e] observed range in allele size; [f] expected range in allele size from the simulation results; [g] 5th percentile based on the simulations, [h] 95th percentile based on the simulations; [i] repeat unit size has not been determined and was assumed to be the difference between the two most closely spaced alleles; * observed value is greater than 95th percentile of the simulation results; [?] observed value is equal to value at 95th percentile of the simulation results.

TABLE 5

STR loci: observed size range and number of modes and predictions from simulation results

Locus	m[a]	E(m)[b]	5th%[c]	95th%[d]	r[e]	E(r)[f]	5th%[g]	95th%[h]
HUMHPRTB	2	2.45	0.41	4.40	9	7.47	3.99	10.80
HUMTHO1	2	2.55	0.45	4.52	6	7.88	4.23	11.38
HUMRENA4	2	1.54	0.37	3.12	4	3.51	1.01	4.55
HUMFABP	2	1.70	0.10	3.23	6	4.58	2.13	7.87
HUMARA	4	4.51	1.77	7.33	16	15.56	9.03	24.49
ACTBP2	5	8.86	5.14	12.68	21	29.17	17.30	39.90
Human TH	2	2.72	0.53	4.88	4	7.87	4.00	12.19
CD4 gene	2	2.05	0.26	3.77	6	5.57	2.74	8.64
ACPP	2	1.84	0.16	3.52	5	4.73	2.13	7.79
MYCN	1	1.35	0.18	2.87	2	2.82	0.60	4.37
GABARB1	1	2.25	0.35	4.10	6	6.27	3.23	9.29
PLA2A1	2	2.30	0.34	4.16	6	6.72	3.51	9.87

[a] Observed number of modes; [b] expected number of modes from the simulation results; [c] 5th percentile based on the simulations; [d] 95th percentile based on the simulations; [e] observed range in allele size; [f] expected range in allele size from the simulation results; [g] 5th percentile based on the simulations; [h] 95th percentile based on the simulations; [i] repeat unit size has not been determined and was assumed to be the difference between the two most closely spaced alleles; * observed value is greater than 95th percentile of the simulation results, [7] observed value is equal to value at 95th percentile of the simulation results.

TABLE 6

Minisatellite loci: observed size range and number of modes and predictions from simulation results

Locus	m[a]	E(m)[b]	5th%[c]	95th%[d]	r[e]	E(r)[f]	5th%[g]	95th%[h]
D17S5	4	3.56	1.15	5.97	14	12.01	6.68	18.20
D2S44	46*	16.21	11.86	21.60	249*	55.01	32.98	51.16
D9S7	5	3.39	0.95	5.81	22*	11.47	6.50	18.34
D14S13	25*	10.63	6.35	14.97	213*	35.61	21.59	47.12
D19S20	10*	2.61	0.48	4.68	22*	7.84	4.23	11.79
D16S83	7[?]	4.19	1.53	6.86	19	14.06	8.02	22.25
D1S74	20*	8.45	4.74	12.27	62*[i]	28.44	17.22	39.63
D3S42	11*	2.46	0.43	4.43	88*[i]	7.16	3.82	10.60
Apo B	3	2.54	0.46	4.55	11	7.96	4.38	11.41
D4F35S1	8*	3.79	1.27	6.35	18	12.94	7.41	20.55
D1S80	5[?]	2.70	0.52	4.84	16*	8.37	4.50	12.52

[a] Observed number of modes; [b] expected number of modes from the simulation results; [c] 5th percentile based on the simulations; [d] 95th percentile based on the simulations; [e] observed range in allele size; [f] expected range in allele size from the simulation results; [g] 5th percentile based on the simulations; [h] 95th percentile based on the simulations; [i] repeat unit size has not been determined and was assumed to be the difference between the two most closely spaced alleles; * observed value is greater than 95th percentile of the simulation results; [?] observed value is equal to value at 95th percentile of the simulation results.

minisatellite loci showed numerous differences from the SMM simulation predictions. The STR loci fit the allelic size range expectations from the SMM simulations well and showed no significant differences. Overall, none of the STR, 35% of the microsatellite and 73% of the minisatellite loci were different from the simulations for at least one of the three summary measures.

We chose loci from the literature that were typed in large homogeneous populations to avoid the effects of population substructure and recent decreases in population size on the allele frequency distributions. Because there were no significant deviations for any summary measure predictions for the STR loci, we conclude that these loci have a mutational mechanism very much like the one-step SMM simulated here. Of the hypothesized mechanisms of VNTR mutation, slipped-strand mispairing, also known as replication slippage, is most like the SMM. Given our results, this mechanism is therefore the most likely mechanism of mutation at STR loci. Thirty-five percent of the microsatellite loci analyzed had significant deviations from the simulation expectations for the number of alleles or their size range. In addition, 19% of this class of loci were borderline significant for deviation from the expected number of modes. If the one-step SMM accurately describes their mutational mecha-

nism, such results are unlikely by chance alone. Therefore, we hypothesize that the mutational mechanism of microsatellite loci is close to but is not exactly a one-step SMM type mechanism. One explanation for the small, but significant, excesses in number of alleles and allelic size range for this class of loci is that the mutational mechanism is slipped-strand mispairing with a very low frequency of multistep mutations.

The minisatellite loci are a heterogeneous group. Some (D2S44, D14S13, D19S20, D1S74, D3S42 and D1S80) show deviations from the simulation SMM results for more than one summary measure, whereas others (D17S5, D16S83 and Apo B) show no significant differences between the observed summary measures and their simulation SMM results. As in the case of microsatellites, a multistep SMM is plausible for the loci that show small deviations from the one-step SMM. However, when the range in allele size is very large (*e.g.*, D2S44, D14S13) unequal exchange or multiple mechanisms as have been proposed by DEKA, CHAKRABORTY and FERRELL (1991) and FORNAGE *et al.* (1992) are more likely.

It is important to point out that multimodality, by itself, is not sufficient evidence for multiple mutational mechanisms as proposed by DEKA, CHAKRABORTY and FERRELL (1991), FORNAGE *et al.* (1992) and SHRIVER *et al.* (1991). Our results show (see Figure 2C) that the expected number of modes is greater than one for heterozygosities as low as 45% under the one-step SMM. This observation is also consistent with VALDES, SLATKIN and FREIMER's findings (1993), although their analysis does not examine the same four summary measures as we studied here. Presumably, the number of modes observed in our simulations and in nature are largely the result of random changes in allele frequencies caused by genetic drift. We conclude that the observed number of alleles alone is insufficient to characterize VNTR loci and that the number of modes and the range are essential allele frequency distribution characteristics for the classification of VNTR loci.

More generally, VNTR loci are being used as genetic markers to address several population genetic questions (FLINT *et al.* 1989; DEKA, CHAKRABORTY and FERRELL 1991). One assumption needed to make certain inferences from population allele frequency distributions (*e.g.*, estimation of the effective population size and the mutation rate) is the choice of a particular mutation-drift model. STR loci stand out among VNTR loci because STR loci fit the one-step SMM model. It is apparent that because of the close fit of the STR loci to the SMM in large panmictic populations, they offer advantages in population studies that microsatellite and minisatellite loci do not. This theoretical advantage is compounded by the relative ease of typing STR loci over minisatellite and microsatellite loci (EDWARDS *et al.* 1992). Using computer simulations we have shown that a simple mutation-drift model can explain many characteristics of VNTR allele frequency distributions. This approach was efficient and informative and can be applied to other questions about the evolution of VNTR loci.

This research was partially supported by NIJ-92-IJ-CX-K024 and NIH-1R01-GM41399.

LITERATURE CITED

BUDOWLE, B., R. CHAKRABORTY, A. M. GIUSTI, A. J. EISENBERG and R. C. ALLEN, 1991 Analysis of the VNTR locus D1S80 by the PCR followed by high-resolution PAGE. Am. J. Hum. Genet. **48:** 137–144.

CHAKRABORTY, R., and S. P. DAIGER, 1991 Polymorphisms at VNTR loci suggest homogeneity of the white population of Utah. Hum. Biol. **63:** 571–587.

CHAKRABORTY, R., and L. JIN, 1992 Heterozygote deficiency, population substructure and their implications in DNA fingerprinting. Hum. Genet. **88:** 267–272.

CHAKRABORTY, R., and J. NEEL, 1989 Description and validation of a method for simultaneous estimation of effective population size and mutation rate from human population data. Proc. Natl. Acad. Sci. USA **86:** 9407–9411.

CHAKRABORTY, R., M. FORNAGE, R. GUEGUEN and E. BOERWINKLE, 1991 Population genetics of hypervariable loci: analysis of PCR based VNTR polymorphism within a population, pp. 127–143 in *DNA Fingerprinting Approaches and Applications*, edited by T. BURKE, G. DOLF, A. J. JEFFREYS and R. WOLFF. Birkhauser Verlag, Basel/Switzerland.

DEAN, M., S. LUCAS-DERSE, S. J. O'BRIEN, E. F. KIRKNESS, C. M. FRASER and D. GOLDMAN, 1991 Genetic mapping of the β1 GABA receptor gene to human chromosome 4, using a tetranuleotide repeat polymorphism. Am. J. Hum. Genet. **49:** 621–626.

DEKA, R., R. CHAKRABORTY and R. E. FERRELL, 1991 A population genetic study of six VNTR loci in three ethnically defined populations. Genomics **11:** 83–92.

DOAK, S., S. JORDAN, P. MCWILLIAM, and P. HUMPHRIES, 1991 Tetranucleotide repeat polymorphism at the ACPP locus. Nucleic Acids Res. **19:** 4793.

EDWARDS A., H. A. HAMMOND, L. JIN, C. T. CASKEY and R. CHAKRABORTY, 1992 Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups. Genomics **12:** 241–253.

EDWARDS, M. C., P. R. CLEMENS, M. TRISTAN, A. PIZZUTI and R. A. GIBBS, 1991 Pentanucleotide repeat length polymorphism at the human CD4 locus. Nucleic Acids Res. **19:** 4791.

EWENS, W. J., 1972 The sampling theory of selectively neutral alleles. Theor. Popul. Biol. **3:** 87–112.

FLINT, J, A. J. BOYCE, J. J. MARTINSON and J. B. CLEGG, 1989 Population bottlenecks in Polynesia revealed by minisatellites. Hum. Genet. **83:** 257–263.

FORNAGE, M., L. CHAN, G. SIEST and E. BOERWINKLE, 1992 Allele frequency distribution of the$(TG)_n(AG)_n$ microsatellite in the apolipoprotein C-II gene. Genomics **12:** 63–68.

FOUGEROUSSE, F., R. MELONI, C. ROUDAUT and J. S. BECKMANN, 1992 Tetranucleotide repeat polymorphism at the human N-MYC gene (MYCN). Nucleic Acids Res. **20:** 1165.

HARTL, D. L., 1980 *Principles of Population Genetics.* Sinauer Associates, Sunderland, Mass.

JANSEN, G., P. J. DE JONG, C. AMEMIYA, C. ASLANIDIS, D. J. SHAW, H. G. HARLEY, D. BROOK, R. FENWICK, R. G. KORNELUK, C. TSILFIDIS, G. SHUTLER, R. HERMENS, N. G. M. WORMSKAMP, H. J. M. SMEETS and BEWIERINGA, 1992 Physical and genetic

characterization of the distal segment of the myotonic dystrophy are on 19q. Genomics **13:** 509–517.

JEFFREYS, A. J., V. WILSON and L. THEIN, 1985 Hyper variable "minisatellite" regions in human DNA. Nature **314:** 67–73.

JONGENEEL, C. V., L. BRIANT, I. A. UDALOVA, A. SEVIN, S. A. NEDOSPASOV and A. CAMBON-THOMSEN, 1991 Extensive genetic polymorphism in the human tumor necrosis factor region and relation to extended HLA haplotypes. Proc. Natl. Acad. Sci. USA **88:** 9717–9721.

KIMURA, M., and J. F. CROW, 1964 The number of alleles that can be maintained in a finite population. Genetics **49:** 725–738.

KIMURA, M., and T. OHTA, 1978 Stepwise mutation model and distribution of allelic frequencies in a finite population. Proc. Natl. Acad. Sci. USA **75:** 2868–2872.

KONRADI C., L. OZELIUS, W. YAN, J. F. GUSELLA and O. BREAKEFIELD, 1991 Dinucleotide repeat polymorphism (D16S285) on human chromosome 16. Nucleic Acids Res. **19:** 5449.

KOTELEVTSEV, Y. V., E. CLAUSER, P. CORVOL and F. SOUBRIER, 1991 Dinucleotide repeat polymorphism in the human angiotensinogen gene. Nucleic Acids Res. **19:** 6978.

LEVINSON, G. and G. A. GUTMAN, 1987 Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol. Biol. Evol. **4:** 203–221.

LITT M., and J. A. LUTY, 1989 A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. Am. J. Hum. Genet. **44:** 397–401.

MARIAT, D., V. LAUTHIER and G. VERGNAUD, 1991 A VNTR isolated by size selection of human DNA fragments detects RFLPs at the extremity of 1p and 4q. Nucleic Acids Res. **19:** 4572.

NAKAMURA, Y., M. LEPPERT, P. O'CONNELL, R. WOLFF, T. HOLM, M. CULVER, C. MARTIN, E. FUJIMOTO, M. HOFF, E. KUMLIN and R. WHITE, 1987 Variable number of tandem repeat (VNTR) markers for human gene mapping. Science **235:** 1616–1622.

NATIONAL RESEARCH COUNCIL, 1992 *DNA Technology in Forensic Science*. National Academy Press, Washington, D.C.

NEI, M., and D. GRAUR, 1984 Extent of protein polymorphism and the neutral mutation theory. Evol. Biol. **17:** 73–118.

ODELBERG, S. J., R. PLAETKE, J. R. ELDRIDGE, L. BALLARD, P. O'CONNELL, Y. NAKAMURA, M. LEPPERT, J.-M. LALOUEL and R. WHITE, 1989 Characterization of eight VNTR loci by agarosegel electrophoresis. Genomics **5:** 915–924.

OHTA, T., and M. KIMURA, 1973 The model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a genetic population. Genet. Res. **22:** 201–204.

PHILLIPS, H. A., P. HARRIS, R. I. RICHARDS, G. R. SUTHERLAND and J. C. MULLEY, 1991a Dinucleotide repeat polymorphisms at the D16S164, D16S168, and D16S186 loci at 16q21-q22.1. Nucleic Acids Res. **19:** 6964.

PHILLIPS, H. A., V. J. HYLAND, K. HOLMAN, D. F. CALLEN, R. I. RICHARDS and J. C. MULLEY, 1991b Dinucleotide repeat polymorphism at D16S287. Nucleic Acids Res. **19:** 6664.

POLYMEROPOULOS, M. H., D. S. RATH, H. XIAO and C. R. MERRIL, 1991a Tetranucleotide repeat polymorphism at the human beta-actin related pseudogene H-beta-Ac-psi-2 (ACTBP2) Nucleic Acids Res. **19:** 1432.

POLYMEROPOULOS, M. H., H. XIAO, D. S. RATH and C. R. MERRIL, 1991b Tetranucleotide repeat polymorphism at the human tyrosine hydroxylase gene (TH). Nucleic Acids Res. **19:** 3753.

SADLER, L. A., S. H. BLANTON and S. P. DAIGER, 1991 Dinucleotide repeat polymorphism at the human tissue plasminagen activator gene (PLAT). Nucleic Acids Res. **19:** 6058.

SHRIVER, M. D., S. P. DAIGER, R. CHAKRABORTY and E. BOERWINKLE, 1991 Multimodal distribution of length variation in VNTR loci detected using PCR. Crime Lab. Digest **18:** 144–147.

SILVER, L. M., 1992 Bouncing off microsatellites. Nature Genetics **2:** 8–9.

THOMPSON, A. D., Y. SHEN, K. HOLMAN, G. R. SUTHERLAND, D. F. CALLEN and R. I. RICHARDS, 1992 Isolation and characterization of $(AC)_n$ microsatellite genetic markers from human chromosome 16. Genomics **13:** 402–408.

VALDES, A. M., M. SLATKIN and N. B. FREIMER, 1993 Allele frequencies at microsatellite loci: the stepwise mutation model revisited. Genetics, **133:** 737–749.

WEBER, J. L., and P. E. MAY, 1989 Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. Am. J. Hum. Genet. **44:** 388–396.

WEHRHAHN, C. F., 1975 The evolution of selectively similar electrophoretically detectable alleles in finite natural populations. Genetics **80:** 375–394.

WOOD, S., and M. SCHERTZER, 1991 Dinulceotide repeat polymorphism at the D8S135 locus on chromosome 8p. Nucleic Acids Res. **19:** 6664.

WRIGHT, S., 1949 Genetics of populations. Encyclopedia Britannica, 14 Ed. **10:** 111, 112.

Communicating editor: A. G. CLARK