Rate and Directionality of Mutations and Effects of Allele Size Constraints at Anonymous, Gene-Associated, and Disease-Causing Trinucleotide Loci

Ranjan Deka, Sun Guangyun,* Diane Smelser,* Yixi Zhong,† Marek Kimmel,‡ and Ranajit Chakraborty†*

*Department of Environmental Health, University of Cincinnati; †Human Genetics Center,
University of Texas Health Science Center, Houston; and ‡Department of Statistics, Rice University

We studied the patterns of within- and between-population variation at 29 trinucleotide loci in a random sample of 200 healthy individuals from four diverse populations: Germans, Nigerians, Chinese, and New Guinea highlanders. The loci were grouped as disease-causing (seven loci with CAG repeats), gene-associated (seven loci with CAG/CCG repeats and eight loci with AAT repeats), or anonymous (seven loci with AAT repeats). We used heterozygosity and variance of allele size (expressed in units of repeat counts) as measures of within-population variability and G_{ST} (based on heterozygosity as well as on allele size variance) as the measure of genetic differentiation between populations. Our observations are: (1) locus type is the major significant factor for differences in within-population genetic variability; (2) the disease-causing CAG repeats (in the nondisease range of repeat counts) have the highest within-population variation, followed by the AAT-repeat anonymous loci, the AAT-repeat gene-associated loci, and the CAG/CTG-repeat gene-associated loci; (3) an imbalance index β, the ratio of the estimates of the product of effective population size and mutation rate based on allele size variance and heterozygosity, is the largest for disease-causing loci, followed by AAT- and CAG/CCG-repeat gene-associated loci and AAT-repeat anonymous loci; (4) mean allele size correlates positively with allele size variance for AAT- and CAG/CCG-repeat gene-associated loci and negatively for anonymous loci; and (5) G_{ST} is highest for the disease-causing loci. These observations are explained by specific differences of rates and patterns of mutations in these four groups of trinucleotide loci, taking into consideration the effects of the past demographic history of the modern human population.

Introduction

Of the different classes of microsatellite or short tandem repeat (STR) loci, trinucleotide repeat loci form a special class, since over a dozen such loci are known to cause disease (Sutherland and Richards 1995). Hundreds of trinucleotide repeat polymorphisms are interspersed throughout the human genome (McKusick 1997). The pattern of within- and between-population variation at these polymorphisms is yet to be characterized thoroughly. While all types (di-, tri-, tetra-, and pentanucleotide) of STRs have been demonstrated to be extremely useful in gene-mapping, forensic, and evolutionary studies, there are indications that these different categories of STRs differ from each other in terms of both mutation rate (Webber and Wong 1993; Chakraborty et al. 1997) and pattern of mutations (Shriver et al. 1993; Di Rienzo et al. 1994). Preliminary studies with regard to the initiatives of the Human Genome Diversity Project indicate that data from different STRs must be pooled to answer most biological and population-related questions about human diversity and the evolutionary history of humans (Chakraborty and Jin 1992; Bowcock et al. 1994). However, inference from such analyses should also take into account the between-locus variation of patterns and extents of polymorphisms at such loci (Kimmel and Chakraborty 1996; Pritchard and Feldman 1996). Interlocus variation at STRs, studied by grouping the loci by their repeat motifs, may not be enough, since variation may also be affected by genomic location and any possible functional constraint associated with the different STRs. For example, in our earlier studies (Chakraborty et al. 1997), we found indications that even with the exclusion of unstable allele sizes, the disease-causing trinucleotides have a mutation rate of 3.9–6.9 times as high as that of the tetranucleotides, while the trinucleotides without disease implications have a more moderate rate of mutations (1.2–2.0 times as high as that of the tetranucleotides).

The purpose of this research is to examine the generality of this observation and to detect additional factors contributing to mutation patterns. In particular, we investigate whether the pattern and extent of within- as well as between-population variation at trinucleotide STRs are any different depending on their genomic locations and motif compositions (e.g. AT- versus GC-rich). For this purpose, we selected three groups of trinucleotides, called ‘‘anonymous,’’ ‘‘gene-associated,’’ and ‘‘disease-causing’’ for this work, and studied their variation in four diverse human populations (Germans, Beninese, Chinese, and Papua New Guinea highlanders). We adopt a population-based study, instead of direct mutation assays, since sample sizes required for precise direct mutation assays would have been prohibitively large, and furthermore, conclusions derived from direct mutation assays are known to be restrictive (Chakraborty and Stivers 1996; see also Chakraborty, Stivers, and Zhong 1997). Since population-based studies of genetic variation are also affected by population-related factors (e.g., effective size and evolutionary history of populations), our choice of four populations represents the three major human groups (Europeans, Africans, and Asians) and one small relatively isolated group (New Guinea highlanders). This allowed us to separate the

Key words: trinucleotide diseases, genome location, stepwise mutation, allele size constraint, coalescence.

Address for correspondence and reprints: Ranajit Chakraborty, Human Genetics Center, University of Texas School of Public Health, P.O. Box 20334, Houston, Texas 77225. E-mail: rc@hgc9.sph.uth.tmc.edu.

Mol. Biol. Evol. 16(9):1166–1177. 1999

population effects from the pattern of between-locus variation of trinucleotide polymorphisms. We argue that pooling data over loci without recognizing the differences of locus type effects may yield inaccurate estimates of within- versus between-population variation at trinucleotide repeat loci (Jodice et al. 1997).

Using a generalized stepwise mutation model (Kimmel et al. 1996; Kimmel and Chakraborty 1996), we interpreted the summary measures of within- and between-population genetic variation (heterozygosity, variance of allele size, and coefficient of gene diversity, respectively) in terms of relative rates and patterns of mutation. The results are consistent with the view that in these four classes of trinucleotides, the genetic variation is mainly mutation driven, with disease-causing trinucleotides having the highest rate of mutation and the GC-rich gene-associated ones having the lowest mutation rate, with the AT-rich anonymous and gene-associated loci being at the intermediate level.

Using the relationship of mean and variance of allele size and the imbalance of heterozygosity and allele size variance (expressed by a coefficient called β; Kimmel et al. 1998), we further find that $\beta > 1$ for disease-causing and gene-associated loci (both AT- and GC-rich), but $\beta \approx 1$ for anonymous loci. Furthermore, mean and variance of allele size are positively correlated for gene-associated loci (both AT- and GC-rich), negatively correlated for anonymous loci, and not significantly correlated for disease-causing loci. Under the assumption that modern human populations have expanded following a bottleneck, the above results indicate that the mutation patterns are also different for these four groups of loci. The anonymous loci are subject to expansion-biased mutation with an upper constraint of allele size, while the effect of allele size constraint is absent for disease-causing and gene-associated loci. Expansion bias of mutations also seems to be present for both types of gene-associated loci, but not for disease-causing loci.

Analysis based on the coefficient of gene differentiation (G_{ST}) also indicates that the mutation patterns may be complex for some loci. Observed G_{ST} is not inversely correlated with the within-population heterozygosity or allele-size variance, as could be expected under an unrestricted single-step stepwise mutation model with mutation rate differences alone.

Materials and Methods

Population Samples

The studied populations belong to four geographically and racially diverse groups, viz. a German sample drawn from Northern Germany, Benin from Nigeria, a Chinese sample of Han origin, and New Guinea highlanders from the northern fringes of Papua New Guinea. A total of 50 DNA samples from unrelated individuals from each population were analyzed at each of the studied loci.

DNA Analysis

Table 1
Characteristics of the Studied Loci

Locus	Chromosomal Location	Repeat Motif	Range of Repeat Size
Anonymous			
D1S1589	1	AAT	10–17
D2S1352	2	AAT	5–11
D4S2361	4	AAT	7–16
D5S1453	5	AAT	3–13
D6S1006	6	AAT	13–16
D20S473	20	AAT	8–15
D21S1440	21	AAT	8–15
Gene-associated			
PLA2A	12q23–qter	AAT	10–17
AT3	1q23	AAT	5–18
D2S116	2	AAT	3–11
D6S91	6	AAT	4–19
D11S916	11	AAT	9–16
D12S86	12	AAT	5–10
D12S87	12	AAT	7–9
D22S280	22	AAT	5–9
N-Cad	18	CGG	6–14
PRPC	20q13.1	CAG	7–9
BCR	22q11	CGG	2–9
GST1	1p31	CGG	7–21
ATPase	1q22–25	CGG	7–10
B33H	3	CAG	9–18
B33L	12	CAG	6–10
Disease-causing			
HD	4p16.3	CAG	9–34
DRPLA	12pter–p12	CAG	6–24
DM	19q13.3	CTG	5–27
SCA1	6p22–23	CAG	21–37
SCA3	14q32.1	CAG	7–36
SCA6	19p13	CAG	4–14
SCA7	3p12–13	CAG	7–14

A summary of the STR loci is given in table 1. The anonymous trinucleotide repeats are from the CHLC human screening set. Primers used for amplification of these loci were obtained from Research Genetics, Inc. Among the gene-associated loci, details (primer and sequence information) of the trinucleotide repeats at N-Cadherin (N-Cad), Protective Protein (PRPC), Breakpoint cluster gene (BCR), Glutathione-S-transferase-1 (GST1), and ATPase (Na+/K+ATPase β-subunit) are given in Riggins et al. (1992). The trinucleotide repeat AT3 is located in the fifth intron of the human antithrombin gene, and to analyze the locus we used the primer information given in Waye et al. (1994). Primer sequences used for amplification of PLA2A1 (pancreatic phospholipase A-2) are given in Hammond et al. (1994). The remaining eight gene-associated loci (D2S116, D6S91, D11S916, D12S87, D22S280, B33H, B33L) were isolated from human brain cDNA libraries. The amplification conditions and sequence information for the first six loci are given in Margolis et al. (1995) and that for the remaining two in Li et al. (1993). The disease-causing loci were analyzed using published primer sequences HD (Warner, Barron, and Brock 1993), DRPLA (Li et al. 1993; Koide et al. 1994), DM (Fu et al. 1992), SCA1 (Orr et al. 1993), SCA3 (Kawaguchi et al. 1994), SCA6 (Zhuchenko et al. 1997), and SCA7 (David et al. 1997). All of the loci were analyzed following the PCR amplification protocol as given in Deka et al. (1995).

In this study, we scored the alleles in terms of their exact repeat numbers. For all of the disease-causing loci and several of the gene-associated loci (PLA2A, AT3, N-Cad, PRPC, BCR, GST1, ATPase), the repeat sizes corresponding to PCR fragment sizes have already been determined and are reported in the literature, as cited above. We used this information for assigning the repeat sizes at these loci. However, for the anonymous loci, the exact repeat sizes corresponding to the PCR product sizes have not yet been reported. We therefore determined the repeat numbers by comparing the GenBank sequences or the published sequences given in the above literature with corresponding fragment sizes (bp) of the PCR product. This was important, since with allele designations made in terms of repeat counts we can draw inferences with regard to the trend of the expansion/contraction bias of mutations and the presence of allele size constraints across loci. Note that except for the few STR loci used in forensics, this is not the usual practice in the published literature or in databases of repeat polymorphisms, which limits the possibilities for statistical analysis.

Fisher-Wright-Moran Coalescent Model

The time-continuous Moran model (Ewens 1979) assumes the population is composed of a constant number of 2N haploid individuals (chromosomes). Each individual undergoes death/birth events according to a Poisson process with intensity 1 (mean length of life of each individual is equal to 1). Upon a death/birth event, a genotype for the individual is sampled with replacement from the 2N chromosomes present at that moment, including the chromosome of the just-deceased individual.

We use the following equivalent coalescent formulation of the Fisher-Wright-Moran model for a population of 2N haploid individuals under genetic drift and mutations following a general time-continuous Markov chain:

- *Coalescent with independent branch lengths with exponential distribution with parameter* $1/(2N)$. For any two individuals from the population, the time to their common ancestor is a random variable $\tau \sim$ exponential$[1/(2N)]$.
- *Markov model of mutations with transition probabilities $P_{ij}(t)$ and intensities Q_{ij}.* If the allelic state of an individual is i at time 0, then his or her allelic state at time t (or the allelic state of his or her descendant at time t) is equal to j with probability $P_{ij}(t)$. In the finite-dimensional case, the transition matrix is equal to $P(t) = \exp(Qt)$, where Q is the transition intensity matrix satisfying the following conditions: (a) $Q_{\mathrm{ij}} \geq 0$, $i \neq j$, and (b) $\Sigma_j Q_{ij} = 0$, all i.

The above model can be used (unpublished data) to derive the equations for the joint probability distribution $R_{jk}(t)$ of allele sizes on two chromosomes drawn from the population at time t, i.e., $R_{jk}(t) = \Pr[X_1 = j, X_2 = k]$.

Measures of Variation and Imbalance at a DNA-Repeat Locus

Statistics Used to Describe a Sample of Alleles

Consider a sample of n haploid individuals or chromosomes and a locus with a denumerable set of alleles indexed by integer numbers. The expectation of the estimator of the within-population component of genetic variance

$$\frac{\hat{V}}{2} = \sum_{i=1}^{n} \frac{(X_i - \bar{X})^2}{n - 1}, \tag{1}$$

where X_i is the size of the allele at the locus in the ith chromosome present and $\bar{X}$ is the mean of X_i, is equal to $V(t)/2$ (Kimmel et al. 1996), where

$$V(t) = E(\hat{V}) = E[(X_i - X_j)^2]. \tag{2}$$

X_i and X_j are time-dependent random variables, i.e., $X_i = X_i(t)$ and $X_j = X_j(t)$, but for notational simplicity, the argument t is suppressed whenever the time dependence is clear from the context.

If p_k denotes the relative frequency of allele k in the sample, then an estimator of homozygosity has the form

$$\hat{P}_0 = \frac{n \sum_{k=1}^{n} p_k^2 - 1}{n - 1}. \tag{3}$$

Random variables X_i are not independent, but only exchangeable. Nevertheless (Kimmel *et al.* 1998), the expected value of $\hat{P}_0$ is the true homozygosity, i.e.,

$$P_0(t) = E(\hat{P}_0) = \sum_k \Pr[X_i = X_j = k]. \tag{4}$$

The Unrestricted Stepwise Mutation Model

The unrestricted stepwise mutation model (SMM) is a model in which the Markov chain describing mutations is an unrestricted random walk with the intensity of transitions (mutation events) equal to ν and the size of transitions (the change of allele size by mutation) being independent, identically distributed random variables with given distributions. In this model, the mutation–drift equilibrium is mathematically predicted (Kimmel and Chakraborty 1996). As a consequence, as $t \to \infty$, $V(t)$ converges to $V(\infty) = 4\nu N\psi''(1) = \theta\psi''(1)$, where $\psi''(1)$ is the variance of the symmetrized size of the allele size change by mutation. If the single-step SMM is assumed (mutation change equal to ∓ 1), then $\psi''(1) = 1$, and we obtain

$$V(\infty) = 4\nu N = \theta. \tag{5}$$

If the single-step SMM is assumed, $P_0(t)$ converges to a limit value which can be explicitly written as (also, see Ohta and Kimura 1973)

$$P_0(\infty) = (1 + 8N\nu)^{-½} = (1 + 2\theta)^{-½}. \tag{6}$$

Expressions (5) and (6) provide two intuitive estimators of the composite parameter θ:

$$\hat{\theta}_V = \hat{V}, \tag{7}$$

called the (allele size) variance estimator of θ, and

Table 2
Influence of Restriction in Allele Size, Mutation Rate, and Directionality on Variance, Homozygosity, and Imbalance Index

b	K	N	ν	$V^{(K)}$	$V^{(\infty)}$	$P_0^{(K)}$	$P_0^{(\infty)}$	$\beta^{(K)}$
0.5	50	10^5	0.00001	3.79	4	0.34	0.33	1.02
			0.00005	17.50	20	0.18	0.16	1.13
			0.0001	32.89	40	0.13	0.11	1.15
			0.0002	59.85	80	0.10	0.08	1.13
			0.0005	118.92	200	0.06	0.05	0.91
0.6	50	10^5	0.00001	2.36	4	0.39	0.33	0.83
			0.00005	6.14	20	0.25	0.16	0.83
			0.0001	8.04	40	0.22	0.11	0.81
			0.0002	10.90	80	0.17	0.08	0.62
			0.0005	—	200	—	0.05	—
0.5	20	10^5	0.00001	3.47	4	0.35	0.33	0.97
			0.00005	13.78	20	0.19	0.16	0.96
			0.0001	22.62	40	0.14	0.11	0.94
			0.0002	33.65	80	0.11	0.08	0.82
			0.0005	—	200	—	0.05	—

NOTE.—$^{(K)}$ denotes the restricted single-step SMM; $^{(\infty)}$ denotes the unrestricted single-step SMM; "—" indicates that numerical calculations diverge for ν = 0.0005 due to rounding errors.

$$\hat{\theta}_{P_0} = (1/\hat{P}_0^2 - 1)/2, \qquad (8)$$

the homozygosity (heterozygosity) estimator of θ. At equilibrium,

$$\frac{E(\hat{\theta}_V)}{E(\hat{\theta}_{P_0})} \approx \frac{V(\infty)}{[1/P_0(\infty)^2 - 1]/2} = 1,$$

which leads to a parametric definition of an index $\beta(t)$, given by

$$\beta(t) = \frac{V(t)}{[1/P_0(t)^2 - 1]/2}. \qquad (9)$$

Deviation of $\beta(t)$ from 1 is a signature of disequilibrium or departure from the single-step SMM at the microsatellite locus.

The General Markov Chain Mutation Case

In the general Markov chain mutation case, explicit expressions for $V(\infty)$, $P_0(\infty)$, and $\beta(\infty)$ are not available. However, using an iterative procedure developed elsewhere (unpublished data), we can obtain the limit distribution $R_{jk} = \Pr[X_1 = j, X_2 = k]$ of allele sizes of two chromosomes from the sample and then calculate the variance, the homozygosity, and the imbalance index:

$$\frac{V(\infty)}{2} = \frac{\sum_{ij} (i - j)^2 R_{ij}}{2},$$

$$P_0(\infty) = \sum_i R_{ii},$$

$$\beta(\infty) = \frac{2V(\infty)}{[P_0(\infty)]^{-2} - 1}.$$

Modeling of Restricted Single-Step SMM

For modeling of restricted single-step SMM, the Markov chain of mutation events can be represented by a random walk on states $\{0, 1, \ldots, K\}$ with reflecting boundaries at 0 and K. Intensities of the backward and forward steps are respectively equal to νd and νb, $d + b = 1$. The overall mutation rate is ν. The matrix of transition intensities is

$$Q = \nu \begin{pmatrix} -1 & 1 & & & & \\ d & -1 & b & & \mathbf{0} & \\ & d & -1 & b & & \\ & & \ddots & \ddots & \ddots & \\ & \mathbf{0} & & d & -1 & b \\ & & & & 1 & -1 \end{pmatrix}.$$

The stationary distribution $\{\pi_i\}_{i=0,1,\ldots,K}$ of the chain has the form

$$\begin{cases} \pi_0 = b/x, & \\ \pi_i = (b/d)^i/x, & i = 1, \ldots, K - 1, \\ \pi_K = b^K/(xd^{K-1}), & \end{cases}$$

where x is a normalizing factor such that the total probability adds up to 1 (i.e., $\Sigma_{i=0}^{k} \pi_i = 1$).

Summary of Modeling Results of the Restricted SMM

Results of numerical studies using the above theory (unpublished data) are summarized in table 2 and figure 1. In mutation–drift equilibrium, restriction on allele sizes causes deviations of the imbalance index β from 1 (table 2). Factors driving β toward lower values include more severe restriction of allele size, higher mutation rate, and directionality of mutation. An intuitive explanation of these trends is that allele size constraints affect the variance of allele size more severely than they affect the heterozygosity. This, by definition of β (eq. 9) causes reduction of β. In addition, higher mutation rates drive allele sizes toward the constraints, contributing to relative reduction of β. An analogous role is played by directionality.

The same theory allows us to model the time trajectory of $\beta(t)$ following population bottleneck and subsequent growth, as we previously considered for the unrestricted SMM (Kimmel et al. 1998). Figure 1 depicts

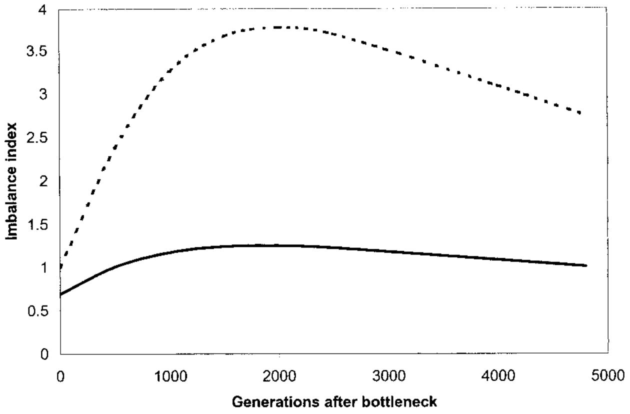

FIG. 1.—Imbalance index (β) as a function of time from bottleneck. The dashed line represents the unrestricted single-step stepwise mutation model and the solid line represents the restricted single-step stepwise mutation model with $K = 15$ possible alleles and 0.6 versus 0.4 mutation expansion bias (i.e., $b = 0.6$). In both models, the mutation rate $\nu = 10^{-4}$. The prebottleneck population size $N_{00} = 160{,}000$, and after a bottleneck (of $N_0 = 3{,}254$), the population was assumed to grow exponentially to reach $N_t = 547{,}586$ after 4,800 generations.

$\beta(t)$ for such a complex growth pattern, with a population initially of large size, N_{00}, dropping instantly to a smaller size, N_0, and then regrowing exponentially to a final size N, i.e.,

$$N(t) = \begin{cases} N_{00}, & t < 0, \\ N_0\exp(\alpha t), & t \geq 0, \end{cases}$$

where α has been selected so that $N(t) = N$ if $t = T$. We used the numerical values obtained by Rogers and Harpending (1992), who fitted distributions of pairwise differences of numbers of segregating sites in mitochondrial DNA to the data of Cann, Stoneking, and Wilson (1987). The second row of table 1 in Rogers and Harpending (1992) contains estimates concerning the world's population expansion. Correcting for the fact that Rogers and Harpending (1992) considered only females, while we consider both genders, i.e., we multiply all effective sizes by 2, we obtain expansion from $N_0 = 3{,}254$ to $N = 547{,}586$ within 120,000 years or $T = 4{,}800$ generations, assuming generation times roughly equivalent to 25 years. We combined these values with mutation rate $\nu = 10^{-4}$, typical for microsatellite loci (Weber and Wong 1993). The prebottleneck value selected was $N_{00} = 160{,}000$. Two scenarios are depicted: (1) unrestricted SMM (dashed line) and (2) restricted SMM with reflective boundaries with parameters $K = 15$ and $b = 0.6$ (continuous curve). The value $K = 15$ corresponds to a typical range of repeat counts in loci considered in this paper. The value of b is somewhat arbitrary, but the results depicted are quite robust with respect to b from interval [0.3, 0.7].

Evolutionary Scenarios and the Theoretical Relationships Between Statistical Characteristics of STR Polymorphisms

An array of possible evolutionary scenarios for microsatellite loci can be characterized using two simple indices, the imbalance index (β) and the slope of the variance-of-allele-size-vs.-mean-allele-size relationship (ρ).

Table 3
Theoretical Predictions of the Imbalance Index β and the Slope of the Mean–Variance Relationship ρ, Based on the Single-Step Stepwise Mutation Model, Under Diverse Assumptions Concerning Mutation Patterns and Past Population Demography

	Constant Population		Expanding Population			
			After Bottleneck		After Equilibrium	
	β	ρ	β	ρ	β	ρ
Unrestricted						
Expansion bias . . .	1	>0	>1	>0	<1	>0
Contraction bias . .	1	<0	>1	<0	<1	<0
Unbiased	1	0	>1	0	<1	0
Constrained						
Expansion bias . . .	<1	≤0	≈1	≤0	≪1	≤0
Contraction bias . .	<1	≥0	≈1	≥0	≪1	≥0
Unbiased	<1	≈0	≈1	≈0	≪1	≈0

The imbalance index is defined in the terms of variance and heterozygosity, and therefore it is insensitive to the expansion/contraction bias. The value of the unrestricted SMM is 1 if the population is constant and in a mutation–drift equilibrium. As demonstrated by Kimmel et al. (1998), the history of past population expansion tends to increase the value of β to above 1 if the expansion was preceded by mutation–drift equilibrium, and to transiently (i.e., for several thousand generations) decrease it to below 1 if the expansion was preceded by a bottleneck.

For allele size constraints, the constant-population, mutation–drift-equilibrium value of β is less than 1 (table 2). Simulation studies for realistic values of mutation rates and allele size range demonstrate (fig. 1) that population size expansion preceded by bottleneck leads to β values around 1, while if the expansion was preceded by mutation–drift equilibrium, β descends far below 1.

The general direction of the slope ρ is insensitive to demographic factors. Rather, in the unconstrained case, it is positive if the repeat count tends to expand (on the average), negative if it tends to contract, and close to 0 if neither of these trends occurs. This is due to the fact that in the presence of expansion bias, variance and mean of the allele size grow in concert, while in the presence of contraction bias, mean decreases while variance increases.

In the presence of constraints, the above tendencies are reversed, at least if the locus is old enough, because as the allele size encounters the constraint (lower or upper), its variance starts decreasing. Table 3 includes all the discussed combinations of values of β and ρ.

Results

Descriptive Features of the Four Groups of Trinucleotide Loci

Our observations relate to 29 trinucleotide repeat loci from diverse locations in the human genome. Seven loci are disease-causing (HD, DRPLA, DM, SCA1, SCA3, SCA6, and SCA7) and all have CAG/CCG motif compositions. Fifteen loci are gene-associated, including eight loci with AAT motif compositions. Fifteen loci are gene-associated, including eight loci with AAT motif compositions (PLA2A, AT3, D2S116, D6S91, D11S916, D12S86, D12S87, and D22S280) and seven loci with CAG/CCG motif compositions (N-Cad, PRTC, BCR, GST1, ATPase, B33H, and B33L). Seven loci are anonymous (D1S1589, D2S1352, D4S2361, D5S1453, D6S1006, D20S473, and D21S1440) and all have AAT motif types. Table 4 presents the summary statistics (mean and variance of allele sizes and heterozygosity) of within-population genetic variation of these loci, studied in four diverse populations (Germans, Beninese, Chinese, and Papua New Guinea highlanders).

Although the allele frequency data of all 29 loci in the four diverse populations studied (with the allele sizes represented by copy numbers of repeat units) are not shown in detail, some of their features are worth noting. In general, the allele frequency distributions do not conform to any standard distributional form (bell shape, uniform, L shape, or J shape). The GC-rich gene-associated trinucleotides are less polymorphic, with smaller numbers of alleles and one predominant allele per locus in each population, particularly compared with the anonymous and disease-causing loci. The AT-rich gene-associated loci show allele frequency distributions most similar to those of the anonymous loci. In general, the allele sizes are overlapping in all four populations, but they show frequency differences across populations, reflecting accumulation of genetic divergence during the evolution of these populations.

The summary statistics of genetic variation, shown in table 4, also indicate that each class of loci is moderately to highly polymorphic (in terms of heterozygosity as well as within-population variance of allele sizes). The GC-rich gene-associated loci are the least polymorphic of the four groups, as reflected by their smaller heterozygosity and variance of repeat sizes within each population (also a lesser number of alleles, not shown). The disease-causing trinucleotides have the highest level of within-population variation in terms of all measures listed above, even though all subjects analyzed are disease-free, within the normal size ranges. The AT-rich anonymous and gene-associated loci are intermediate in this regard.

Rank correlations (Kendall's τ; Kendall 1947) among the number of alleles, heterozygosity, and variance of allele sizes computed for the four types of loci separately indicate that these three measures of variation are positively correlated (ranges of τ are from 0.62 to 0.68 for the anonymous loci, from 0.79 to 1.00 for AT-rich gene-associated loci, from 0.24 to 0.71 for GC-rich gene-associated loci, and from 0.78 to 0.98 for the disease-causing loci). These correlations are all significant ($P < 0.05$) with the exception of the correlations between heterozygosity and allele size variance ($\tau = 0.24$; $P \approx 0.246$) and between heterozygosity and number of alleles ($\tau = 0.33$; $P \approx 0.147$) for the seven GC-rich gene-associated loci.

Within-Population Allele Size Variance for Four Groups of Loci

Heterozygosity, as well as variance of allele sizes within populations, can be predicted from a generalized stepwise model of mutations (GSMM; stepwise mutation model with an arbitrary distribution of the size of mutational change of alleles), which has previously been shown to be an appropriate model for STR loci (Shriver et al. 1993; Di Rienzo et al. 1994; Kimmel et al. 1996). Under a GSMM, should mutation be the main factor affecting the within-population variability at an STR locus, allele size variance and heterozygosity will be positively correlated (Kimmel et al. 1998). Indeed, as mentioned earlier, the present data show exactly the same trend. Therefore, either of these two measures (heterozygosity or allele size variance) can, in theory, serve as a surrogate measure of relative mutation rate at these loci. However, as mentioned in the *Materials and Methods* section, the relationship of expected heterozygosity to mutation rate is rather involved.

The GSMM predicts that in a population that reached a steady state of variation at an STR locus, the within-population variance V of allele size is proportional to the product of effective population size N and mutation rate ν (Chakraborty et al. 1997). Thus, an analysis of variance (ANOVA) of the logarithm of the allele size variance provides estimates of relative mutation rates at different loci, as well as examining the effect of effective sizes of populations. Single-locus data on variance and heterozygosity are known to have large random errors, caused by stochastic evolutionary factors as well as by sampling (Kimmel and Chakraborty 1996; see Nei 1978 for related results on heterozygosity), and thus when data are available from multiple populations and the population size effects are not significant, stability of these sample statistics can be obtained by pooling data over populations.

Therefore, before conducting the above analyses, we investigated how the within-population variability at the 29 loci is affected by locus types (anonymous, gene-associated, and disease-causing), motif composition (AT- vs. GC-rich), population sizes (across four populations), and their interactions. Table 5 presents a summary of the results of such analysis. In order to detect locus type effects, two-way analysis of variance of the logarithm of the allele-size variance was carried out. Comparisons include anonymous versus gene-associated versus disease-causing loci, AT-rich anonymous versus gene-associated loci, AT-rich versus GC-rich gene-associated loci, and GC-rich gene-associated versus disease-causing loci (table 5). Effects of population and interaction between population and locus type were not significant in any analyses. Locus type effect was significant in the comparisons of anonymous versus gene-associated versus disease-causing loci and GC-rich gene-associated versus disease-causing loci (both at the $P < 10^{-6}$ level). The difference between AT-rich and GC-rich gene-associated loci had a P value of 0.09.

The results of ANOVA, shown in table 5, can further be used to quantify the differences of the four clas-

Table 4
Statistics for Repeat Size Variations at the 17 Trinucleotide Loci in Four Human Populations

		Populations			
Locus	Statistics	Germans	Beninese	Chinese	New Guineans
Anonymous					
D1S1589	Mean	12.50	14.14	13.37	13.90
	Variance	2.31	2.40	2.13	2.26
	Expected heterozygosity	73.88	81.84	76.79	73.84
D2S1352	Mean	6.15	7.06	6.32	5.97
	Variance	2.45	3.63	1.31	2.25
	Expected heterozygosity	65.92	77.86	73.60	51.54
D4S2361	Mean	12.56	12.06	12.53	12.12
	Variance	1.10	2.55	1.26	1.78
	Expected heterozygosity	68.06	73.57	72.04	51.66
D5S1453	Mean	7.17	5.40	6.39	6.62
	Variance	5.82	4.12	4.58	2.22
	Expected heterozygosity	74.67	61.60	69.70	59.94
D6S1006	Mean	14.14	13.46	14.80	14.35
	Variance	0.97	0.59	0.18	0.29
	Expected heterozygosity	55.39	45.72	31.45	51.15
D20S473	Mean	11.55	11.95	12.22	12.03
	Variance	1.30	2.05	0.86	0.42
	Expected heterozygosity	59.94	81.58	58.98	28.13
D21S1440	Mean	9.21	9.88	9.38	9.18
	Variance	1.06	6.13	1.13	0.80
	Expected heterozygosity	68.51	65.54	64.77	64.53
AT-rich gene-associated					
PLA2A	Mean	12.63	14.43	13.24	11.35
	Variance	3.50	1.69	2.85	0.77
	Expected heterozygosity	72.70	78.52	78.59	33.80
AT3	Mean	13.35	12.58	12.66	14.35
	Variance	14.13	11.61	13.92	4.30
	Expected heterozygosity	87.49	86.95	82.12	81.08
D2S116	Mean	5.15	6.47	5.65	3.54
	Variance	2.69	1.86	1.93	1.73
	Expected heterozygosity	67.26	75.24	69.84	26.69
D6S91	Mean	13.37	13.16	14.45	14.81
	Variance	3.57	7.68	1.02	10.76
	Expected heterozygosity	76.55	87.91	71.31	78.86
D11S916	Mean	10.20	11.07	10.11	10.00
	Variance	0.59	2.32	0.11	0.00
	Expected heterozygosity	61.35	76.42	17.72	0.00
D12S86	Mean	7.01	6.56	7.55	6.20
	Variance	1.71	1.90	1.18	2.66
	Expected heterozygosity	69.47	68.10	69.17	52.46
D12S87	Mean	7.35	7.23	7.32	7.22
	Variance	0.39	0.18	0.22	0.17
	Expected heterozygosity	42.89	35.85	43.70	34.40
D22S280	Mean	7.62	7.52	7.32	7.99
	Variance	0.57	0.25	0.26	0.01
	Expected heterozygosity	50.56	50.44	43.86	2.38
GC-rich gene-associated					
N-Cad	Mean	8.28	8.37	7.97	8.60
	Variance	1.82	1.08	0.09	3.19
	Expected heterozygosity	20.40	65.45	5.92	35.54
PRTC	Mean	7.37	7.33	7.31	7.00
	Variance	0.26	0.22	0.21	0.00
	Expected heterozygosity	47.25	44.44	42.92	0.00
BCR	Mean	6.14	6.21	6.45	6.17
	Variance	2.65	0.49	0.97	0.34
	Expected heterozygosity	66.79	29.80	49.08	18.36
GST1	Mean	10.22	10.28	10.19	10.54
	Variance	0.76	1.07	0.40	4.88
	Expected heterozygosity	49.43	26.85	21.92	22.81
ATPase	Mean	8.01	7.85	8.36	7.91
	Variance	0.83	0.61	0.64	0.08
	Expected heterozygosity	56.12	57.80	46.24	16.55
B33H	Mean	13.70	10.94	12.75	13.40
	Variance	3.83	2.85	1.86	2.97
	Expected heterozygosity	77.68	81.20	62.98	72.34
B33L	Mean	8.03	8.00	7.98	8.00
	Variance	0.05	0.31	0.04	0.00
	Expected heterozygosity	9.68	44.77	1.96	0.00

Table 4
Continued

		POPULATIONS			
LOCUS	STATISTICS	Germans	Beninese	Chinese	New Guineans
Disease-causing					
HID	Mean	18.69	17.36	17.49	15.95
	Variance	14.60	7.57	2.86	17.01
	Expected heterozygosity	80.96	83.27	46.30	36.28
DRPLA	Mean	14.18	12.31	15.17	11.09
	Variance	7.79	3.73	10.41	12.02
	Expected heterozygosity	77.64	79.94	82.10	77.66
DM	Mean	12.09	9.56	10.49	12.17
	Variance	39.15	13.08	16.19	3.92
	Expected heterozygosity	83.94	82.85	81.34	80.12
SCA1	Mean	30.10	28.87	28.30	32.81
	Variance	2.05	5.95	2.58	6.86
	Expected heterozygosity	71.33	84.03	72.08	83.55
SCA3	Mean	14.55	17.63	14.39	20.57
	Variance	24.69	40.76	51.96	20.51
	Expected heterozygosity	82.67	90.38	71.29	89.20
SCA6	Mean	11.78	11.11	11.93	7.00
	Variance	2.96	1.87	4.73	0.00
	Expected heterozygosity	73.70	61.67	70.95	0.00
SCA7	Mean	10.44	10.41	10.25	10.86
	Variance	0.90	1.37	0.68	1.20
	Expected heterozygosity	34.91	55.47	33.80	61.27

ses of trinucleotide loci in terms of their mutation rates. Following the method described in Chakraborty et al. (1997), figure 2*a* shows the estimated relative mutation rates (in logarithmic scale) with ∓ 1 SD bounds for the four classes of trinucleotides. Pairwise comparisons of data shown in this figure imply that, on average, the seven disease-causing loci have mutation rates 4.09 times as high as those of the anonymous trinucleotides. The seven anonymous trinucleotides, on average, have a mutation rate 2.47 times as high as those of the seven GC-rich gene-associated loci. On the other hand, the eight AT-rich gene-associated loci have a mutation rate only 1.23 times as low as that of the anonymous loci. This makes the GC-rich gene-associated trinucleotides the least mutable and disease-causing trinucleotides the most mutable of these four classes, with a mutation rate ratio of 10.11 between them. This finding is significant since at present there are no such data suggesting mutation rate variation across repeat loci of the same motif size classified by genomic locations.

Table 5
Two-Way Analysis of Variance of the Logarithm of Within-Population Variance of Repeat Sizes

Sources	df	SS	MS	*F* ratio	*P*
Anonymous (7 loci) vs. gene-associated (15 loci) vs. disease-causing (8 loci)					
Locus type . . .	2	68.65	34.33	19.17	$<10^{-6}$
Population . . .	3	5.42	1.81	1.01	0.39
Interaction . . .	6	2.99	0.50	0.28	0.95
Residual	100	179.04	1.79		
Total	111	256.10			
Anonymous (7 loci, AT-rich) vs. AT-rich gene-associated (8 loci)					
Locus type . . .	1	0.69	0.69	0.39	0.54
Population . . .	3	6.13	2.05	1.16	0.34
Interaction . . .	3	0.40	0.13	0.08	0.97
Residual	51	90.15	1.77		
Total	58	97.37			
AT-rich gene-associated (8 loci) vs. GC-rich gene-associated (7 loci)					
Locus type . . .	1	6.94	6.94	2.95	0.09
Population . . .	3	4.96	1.66	0.70	0.56
Interaction . . .	3	3.18	1.06	0.45	0.72
Residual	49	115.39	2.36		
Total	56	130.47			
GC-rich gene-associated (7 loci) vs. disease-causing (8 loci)					
Locus type . . .	1	70.09	70.09	40.04	$<10^{-6}$
Population . . .	3	3.29	1.10	0.63	0.60
Interaction . . .	3	1.75	0.58	0.33	0.80
Residual	45	78.77	1.75		
Total	52	153.91			

NOTE.—In all analyses populations are: Germans, Chinese, Beninese, and New Guineans. SS = sum of squares, MS = mean sum of squares.

The above results are based on averages of loci, grouped in four classes, in which nonsignificant population effects are ignored. In figure 2*b*, we show the results of a nonparametric analysis, treating the data from all four populations as replicate observations. For within-population variance (in logarithmic scale), the cumulative distribution function of the disease-causing loci is significantly farther to the right than that of the anonymous loci, which in turn is to the right of the cumulative distribution function of the GC-rich gene associated loci ($P < 0.001$ and $P < 0.05$, respectively, by the one-sided, two-sample Kolmogorov-Smirnov test). The distribution of the logarithm of the allele size variance for the AT-rich gene-associated loci interpolates between those of the anonymous and GC-rich gene-associated loci, not differing significantly from any of them. This is consistent with the trends of mutation rate estimates based on ANOVA (fig. 2*a*). The shift of the cumulative distribution functions is in the same direc-

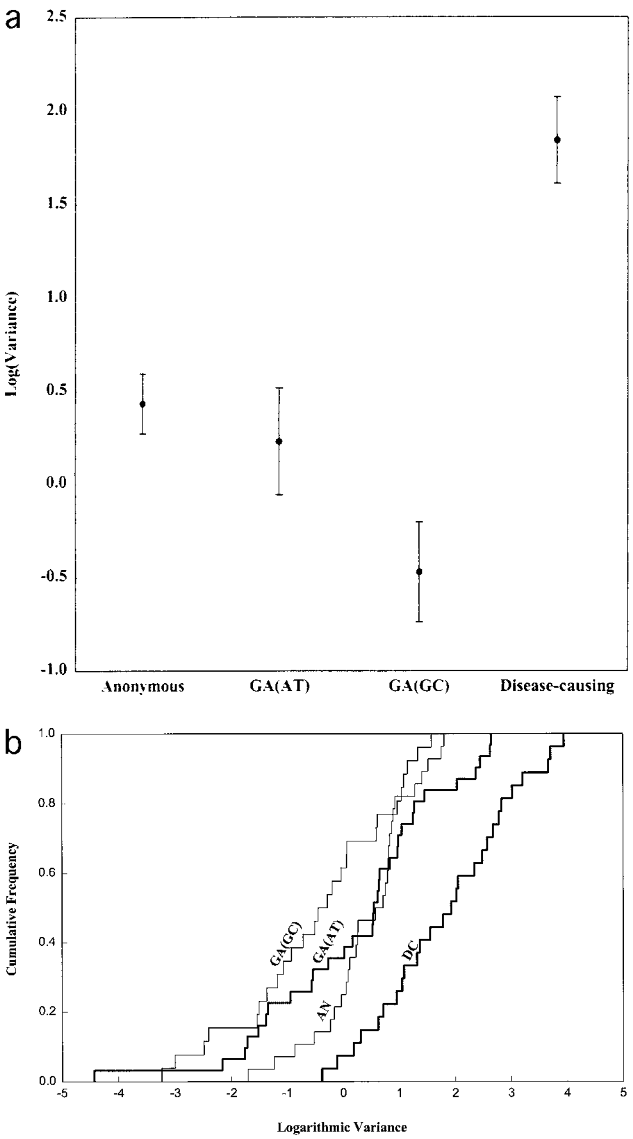

FIG. 2.—*a,* Estimates of the logarithms of the relative mutation rates, i.e., logarithms of within-population allele size variances, denoted Log(Variance), for the four groups of triplet repeats. Approximate 95% confidence intervals based on locus-to-locus distribution of variances and heterozygosities were obtained using the delta method. *b,* Cumulative frequencies of the logarithms of within-population allele variances of the four groups of triplet repeats.

tion. The stochastic ordering (GC-rich gene-associated < anonymous ≈ AT-rich gene-associated < disease-causing) is consistent with mutation rates ordered in the same direction.

Imbalance of Heterozygosity and Allele Size Variance

As discussed in the *Materials and Methods* section, the imbalance index β (the ratio of the estimate of 4*Nν* based on allele size variance to that based on heterozygosity) is affected by rate and pattern of mutations, as well as by the past demographic history of populations.

In figure 3, we present the empirical data of β by plotting the estimated ln β values with one-standard-deviation error bars for the four groups of loci within each population as well as averaged over four populations. These computations show that for the pooled data (averaged over the four populations), we observe the trend: β(anonymous) < β(AT-rich gene-associated) ≈ β(GC-rich gene-associated) < β(disease-causing). For the individual populations, the imbalance index was the lowest for the anonymous loci, except for the Nigerians. The disease-causing loci showed the highest imbalance index in all populations except for the New Guineans. Also, the pooled data indicates β values significantly larger than 1 (i.e., ln β > 0) for all groups of loci except for the anonymous loci, for which β was not significantly different from 1.

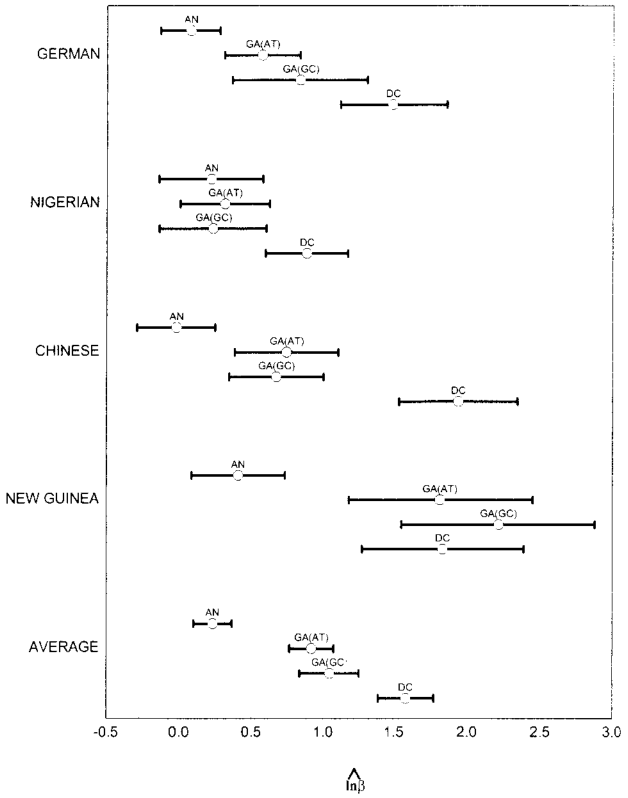

FIG. 3.—Estimates of the logarithms of the imbalance index (ln β) defined in equation (9) for the four groups of triplet repeats in four human populations, along with those averaged over four populations. Abbreviations: AN = anonymous; GA(AT) = gene-associated AT-rich; GA(GC) = gene-associated GC-rich; DC = disease-causing loci.

In summary, the examination of the imbalance index (β) suggests that the allele size constraint is not the major determining factor for a lower coefficient of gene differentiation in the case of the GC-rich gene-associated loci. However, the anonymous loci may have a constraint in the direction of large allele size, preventing their coefficient of gene differentiation from becoming large and keeping their imbalance index close to one (see *Discussion*).

Relationship of Mean and Variance of Allele Sizes Within Populations

Another factor that may influence allele size distribution is the contraction/expansion bias of mutations, i.e., *b* ≠ 0.5. For example, in the absence of allele size

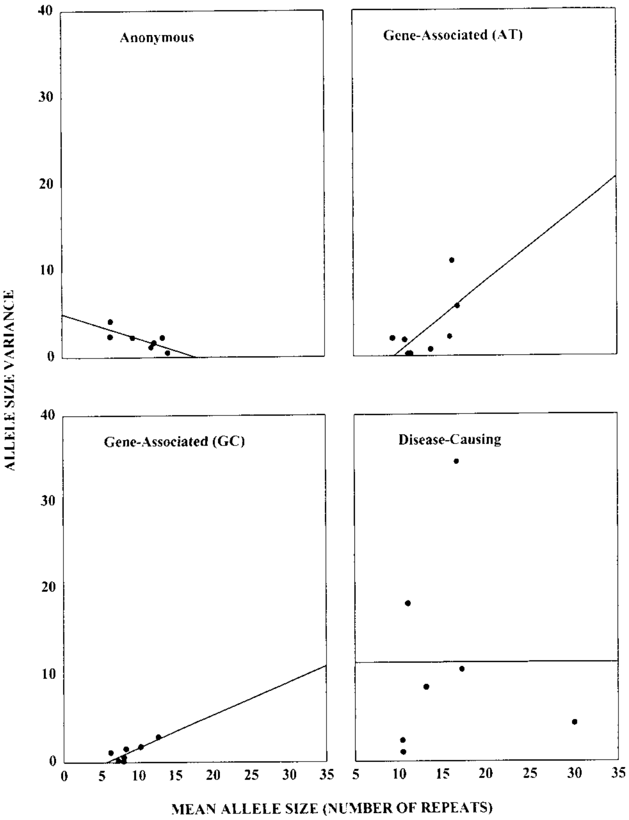

FIG. 4.—Variance of allele sizes plotted against the mean allele sizes for the four groups of triplet repeats. The data points are means and variances for each locus averaged over the four populations. *a,* Seven anonymous loci (AT-rich). *b,* Eight AT-rich gene-associated loci. *c,* Seven GC-rich gene-associated loci. *d,* Seven disease-causing loci (GC-rich).

constraints, an expanding trend of mutations implies that loci with higher mutation rates display higher mean allele sizes. Statistically, this translates to a positive correlation between mean and variance of allele size (i.e., $\rho > 0$, see *Evolutionary Scenarios and the Theoretical Relationships Between Statistical Characteristics of STR Polymorphisms* section for further details).

For the pooled data, figure 4 shows the relationship between mean and variance of allele sizes, with the four groups of loci plotted in four different panels (see fig. 4 legend). Different relationship trends are obvious from these panels. Both AT- and GC-rich gene-associated loci depict significant positive correlations of mean and variance of allele sizes (i.e., $\rho > 0$). The disease-causing loci show a slight increasing positive trend, although this trend is not significant (i.e., $\rho \approx 0$). The anonymous loci show a significantly negative trend (i.e., mean allele size decreasing with increasing variance, $\rho < 0$). Considered in isolation of the mutation rate differences discussed above, this trend would imply a contraction bias.

Gene Diversity Analysis

The differences in the patterns of mutations for these four classes of trinucleotide loci may be more complex than the ordering of mutation rates inferred from the ANOVA results discussed earlier. This possibility was examined by a gene diversity analysis of four groups of loci across four diverse populations. In table 6, we present a summary of this analysis for four classes of loci. The total heterozygosity or allele size variance in the pooled population (H_T or V_T, respectively) is decomposed into heterozygosity or variance within populations (H_S or V_S) and genetic diversity between populations (D_{ST} or V_{ST}). The coefficient of gene diversity can be defined as $G_{ST}(H) = D_{ST}/H_T$ based on heterozygosity, and as $G_{ST}(V) = V_{ST}/V_T$ based on allele size variance. Should mutation rate differences by the only factor distinguishing the four classes of trinucleotides, we would expect a decreasing order of $G_{ST}(H)$ or $G_{ST}(V)$ as within-population polymorphism (H_S or V_S) increases (Chakraborty and Jin 1992; Jin and Chakraborty 1995; unpublished data). The results shown in table 6 do not exactly conform to this trend. The levels of polymorphism (H_S or V_S) within populations are in the order of mutation rates, discussed earlier. However, the disease-causing loci, instead of showing the smallest coefficients of gene diversity ($G_{ST}(H)$ or $G_{ST}(V)$), exhibit the largest. This indicates that in addition to mutation rate differences, there may be additional factors that dictate the patterns of polymorphisms observed for these four classes of loci.

Discussion and Conclusions

Data presented above indicate that the within-population variance of allele sizes at seven disease-causing trinucleotides is, on average, larger than that of the seven anonymous trinucleotides. Furthermore, both AT- and GC-rich gene-associated trinucleotides display the

Table 6
Gene Diversity Analysis of Four Groups of Trinucleotides in Four Diverse Populations

	Anonymous (7 loci)	GC-Rich Gene-Associated (7 loci)	AT-Rich Gene-Associated (8 loci)	Disease-Causing (8 loci)
Gene diversity statistics based on heterozygosity				
H_T	0.718 ∓ 0.026	0.424 ∓ 0.078	0.665 ∓ 0.066	0.798 ∓ 0.056
H_S	0.636 ∓ 0.036	0.384 ∓ 0.069	0.586 ∓ 0.065	0.689 ∓ 0.058
$G_{ST}(H)$	0.113 ∓ 0.026	0.096 ∓ 0.013	0.120 ∓ 0.025	0.137 ∓ 0.042
Gene diversity statistics based on allele-size variance				
V_T	2.25 ∓ 0.47	1.34 ∓ 0.51	3.47 ∓ 1.33	13.86 ∓ 4.94
V_S	2.07 ∓ 0.44	1.16 ∓ 0.38	3.02 ∓ 1.30	11.34 ∓ 4.43
$G_{ST}(V)$	0.082 ∓ 0.020	0.132 ∓ 0.078	0.132 ∓ 0.060	0.182 ∓ 0.050

smallest allele size variance, although not significantly so. These differences are not due to differences in motif compositions (e.g., the AT- and GC-rich gene-associated loci did not show significant differences in allele size variance; table 5 and fig. 2). Nor are these differences contributed by population-related factors (table 5).

This trend, noted for the first time for the trinucleotide repeat polymorphism data in this study, can be explained as the mutation rate difference, with the GC-rich gene associated trinucleotides being the least mutable and disease-causing trinucleotides being the most mutable, with a mutation rate ratio of 10.11 between them. However, our examination of interpopulation differences in allele frequency distributions and the congruence of estimates of $\theta = 4N\nu$ based on within-population allele size variance and per-locus heterozygosity suggests that the distinctions of mutation patterns of these four classes of trinucleotides may be more complex than a simple mutation rate difference.

In fact, the results regarding the patterns of the relationship of mean and variance of allele size (as expressed by its slope ρ), as well as those of the imbalance index (β) for the four groups of loci interpreted in conjunction with each other, reveal such complexity. Recall that for anonymous loci mean and variance of allele sizes are negatively correlated ($\rho < 0$), while $\beta \approx 1$. As depicted in table 3, for a population of constant effective size in mutation–drift equilibrium, these two features would be consistent with an unrestricted stepwise mutation model with a contraction bias of mutations. However, for a population that is expanding following a bottleneck, stepwise mutations with expansion bias and an upper allele size constraint also can produce the same patterns of β and ρ.

For both GC- and AT-rich gene-associated loci, our observations are $\beta > 1$ and $\rho > 0$. In terms of table 3, these values are consistent with an unrestricted SMM with expansion bias of mutations in a population expanding following a bottleneck.

For the disease-causing loci, β was the largest ($\beta > 1$) and ρ was not significantly different from 0. These results are best explained by an unrestricted SMM with no contraction/expansion bias of mutations in a population expanding following a bottleneck.

In a sense, the results of the gene diversity analysis also reflect a complexity of differences of mutation patterns across the four groups of loci. A simple mutation rate difference across loci does not explain the trends of G_{ST}, which is highest for the disease-causing loci, since under the unrestricted single-step SMM, loci with the largest mutation rates should exhibit the smallest variances (Jin and Chakraborty 1995). The lower G_{ST} of the anonymous loci may be a reflection of their upper allele size constraint. However, since no allele size constraint effect is visible for gene-associated and not disease-causing loci, the greater heterozygosity and allele size variance in disease-causing loci could have resulted from occasional multistep mutations.

Finally, we note that although we used the estimates of the effective population sizes based on mtDNA studies of Rogers and Harpending (1992) to illustrate the impact of population size expansion following a bottleneck, the general trend, manifested by $\beta > 1$, is robust with respect to changes in parameter values (Kimmel et al. 1998).

In summary, this study reflects that within the class of trinucleotides, the mutation patterns of loci are not uniform. We showed that gene differentiation between populations, as well as relationships between different measures of within-population genetic variation (i.e., heterozygosity vs. allele-size variance, studied by the imbalance index, figure 3), help to examine the complexity of the pattern and rates of mutations at the repeat loci. The relationship of the mutation pattern differences with the genomic context of repeat loci may extend to other types of loci as well. This has to be investigated in further detail.

Acknowledgments

This work was supported by U.S. Public Health Service Research grants GM 41399 (to R.C.), GM 53545 (to R.C. and M.K.) and GM 45861 (to R.D. and R.C.) from the U.S. National Institutes of Health and funding from the Keck's Center for Computational Biology at the Rice University (to M.K.). We thank Dr. Wolfgang Stephan and two anonymous reviewers for their constructive suggestions for improving the presentation of this work.

LITERATURE CITED

Bowcock, A. M., A. Ruiz-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.

Cann, R. L., M. Stoneking, and A. C. Wilson. 1987. Mitochondrial DNA and human evolution. Nature **325**:31–36.

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

Chakraborty, R., M. Kimmel, D. N. Stivers, L. J. Davison, and R. Deka. 1997. Relative mutation rates at di-, tri, and tetranucleotide microsatellite loci. Proc. Natl. Acad. Sci. USA **94**:1041–1046.

Chakraborty, R., and D. N. Stivers. 1996. Paternity exclusions by DNA markers: effects of paternal mutations. J. Forensic Sci. **41**:667–673.

Chakraborty, R., D. N. Stivers, and Y. Zhong. 1997. Estimation of mutation rates from parentage testing data: applications to STR and VNTR loci. Mutat. Res. **354**:41–48.

David, G., N. Abbas, G. Stevanin et al. (19 co-authors). 1997. Cloning of the SCA7 gene reveals a highly unstable CAG repeat expansion. Nat. Genet. **17**:65–70.

Deka, R., L. Jin, M. D. Shriver, L. M. Yu, S. DeCroo, J. Hundrieser, C. H. Bunker, R. E. Ferrell, and R. Chakraborty. 1995. Population genetics of dinucleotide (dC-dA)n.(dG-dT)n polymorphisms in world populations. Am. J. Hum. Genet. **56**:461–474.

Di Rienzo, A., A. C. Peterson, J. C. Garza, A. M. Valdes, M. Slatkin, and N. B. Freimer. 1994. Mutational processes of simple-sequence repeat loci in human populations. Proc. Natl. Acad. Sci. USA **91**:3166–3170.

Ewens, W. J. 1979. Mathematical population genetics. Springer-Verlag, Berlin.

Fu, Y.-H., A. Pizzuti, R. G. Fenwick Jr. et al. (15 co-authors). 1992. An unstable triplet repeat in a gene related to myotonic muscular dystrophy. Science **255**:1256–1258.

Hammond, H. A., L. Jin, Y. Zhong, C. T. Caskey, and R. Chakraborty. 1994. Evaluation of 13 short tandem repeat loci for use in personal identification applications. Am. J. Hum. Genet. **55**:175–189.

Jin, L., and R. Chakraborty. 1995. Population substructure, stepwise mutations, heterozygote deficiency and their implications in DNA forensics. Heredity **74**:274–285.

Jodice, C., B. Giovannone, V. Calabresi, M. Bellocchi, L. Terrenato, and A. Novelletto. 1997. Population variation analysis at nine loci containing expressed trinucleotide repeats. Ann. Hum. Genet. **61**:425–438.

Kawaguchi, Y., T. Okamota, M. Taniwaki et al. (13 co-authors). 1994. CAG expansions in a novel gene for Machado-Joseph disease at chromosome 14q32.1. Nat. Genet. **8**:221–228.

Kendall, M. G. 1947. The advanced theory of statistics. Vol. 1. Charles Griffin, London.

Kimmel, M., and R. Chakraborty. 1996. Measures of variation at DNA repeat loci under a general stepwise mutation model. Theor. Popul. Biol. **50**:345–367.

Kimmel, M., R. Chakraborty, J. P. King, M. Bamshad, W. S. Watkins, and L. B. Jorde. 1998. Signatures of population expansion in microsatellite data. Genetics **148**:1921–1930.

Kimmel, M., R. Chakraborty, D. N. Stivers, and R. Deka. 1996. Dynamics of repeat polymorphisms under a forward-backward mutation model: within- and between-population variability at microsatellite loci. Genetics **143**:549–555.

Koide, R., T. Ikeuchi, O. Onodera et al. (13 co-authors). 1994. Unstable expansion of CAG repeat in hereditary dentatorubral-pallidoluysian atrophy (DRPLA). Nat. Genet. **6**: 9–13.

Li, S.-H., M. G. McInnis, R. L. Margolis, S. E. Antonarakis, and C. A. Ross. 1993. Novel triplet repeat containing genes in human brain: cloning, expression, and length polymorphisms. Genomics **16**:572–579.

McKusick, V. A. 1997. Mendelian inheritance in man. 12th edition. Johns Hopkins, Baltimore.

Margolis, R. L., T. S. Breschel, S.-H. Li, A. S. Kidwai, M. G. McInnis, and C. A. Ross. 1995. Polymorphic (AAT)n trinucleotide repeats derived from a human brain cDNA library. Hum. Genet. **96**:495–496.

Nei, M. 1978. Estimation of average heterozygosity and genetic distance from a small number of individuals. Genetics **89**:583–590.

Ohta, T., and M. Kimura. 1973. A model of mutation appropriate to estimate the number of electrophoretically detectable alleles in a finite population. Genet. Res. **22**:201–204.

Orr, H. T., M. Chung, S. Banfi, T. J. Kwiatkowski Jr., A. Servadio, A. L. Beaudet, A. E. McCall, L. A. Duvick, L. P. W. Ranum, and H. Y. Zoghbi. 1993. Expansion of an unstable trinucleotide CAG repeat in spinocerebellar ataxia type 1. Nat. Genet. **4**:221–226.

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

Riggins, G. J., L. K. Lockey, J. L. Chastain, H. A. Leiner, S. L. Sherman, K. D. Wilkinson, and S. T. Warren. 1992. Human genes containing polymorphic trinucleotide repeats. Nat. Genet. **2**:186–191.

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

Rubinsztein, D. C., W. Amos, J. Leggo, S. Goodburn, R. S. Ramesar, J. Old, R. Bontrop, R. McMahon, D. E. Barton, and M. A. Ferguson-Smith. 1994. Mutational bias provides a model for the evolution of Huntington's disease and predicts a general increase in disease prevalence. Nat. Genet. **7**:525–530.

Shriver, M. D., L. Jin, R. Chakraborty, and E. Boerwinkle. 1993. VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics **134**:186–191.

Sutherland, G. R., and R. I. Richards. 1995. Simple tandem DNA repeats and human genetic disease. Proc. Natl. Acad. Sci. USA **92**:3636–3641.

Warner, J. P., L. H. Barron, and D. J. H. Brock. 1993. A new polymerase chain reaction (PCR) assay for the trinucleotide repeat that is unstable and expanded on Huntington's disease chromosomes. Mol. Cell. Probes **7**: 235–239.

Wave, J. S., B. Eng, H. Y. Ni, M. A. Blajchman, and G. Carmody. 1994. Trinucleotide repeat polymorphism within the human antithrombin gene (AT3): allele frequency data for three population groups. Mol. Cell. Probes **8**:149–154.

Webber, J. L., and C. Wong. 1993. Mutation of human short tandem repeats. Hum. Mol. Genet. **2**:1123–1128.

Zhuchenko, O., J. Bailey, P. Bonnen, T. Ashizawa, D. W. Stockton, C. Amos, W. B. Dobyns, S. H. Subramony, H. Y. Zoghbi, and C. C. Lee. 1997. Autosomal dominant cerebellar ataxia (SCA6) associated with small polyglutamine expansions in the a1A-voltage-dependent calcium channel. Nat. Genet. **15**:62–69.

Wolfgang Stephan reviewing editor

Accepted May 13, 1999