How Can We Distinguish between Mutational "Hot Spots" and "Old Sites" in Human mtDNA Samples?

MICHAEL GURVEN[1]

Abstract New research into variation in mutation rates across nucleotide positions in human mitochondrial DNA (mtDNA) calls into question population genetics models that assume a constant mutation rate for all sites in a sequence, particularly for hypervariable control region segments I and II. Related to this research is discovering the extent to which highly polymorphic sites are really mutational "hot spots" rather than "old" sites rooted early in the phylogenetic tree. This issue is addressed through the analysis of linkage disequilibrium patterns in the mtDNAs of 10 human populations. Hot spots can be expected to show little or no disequilibrium since they can be interpreted as having randomly expressed patterns. In fact, the results suggest that many highly polymorphic sites are not old sites, but instead are hot spots. Suspected hot spots are listed and compared with hypervariable sites given by Wakeley (1993) and Hasegawa et al. (1993).

Recent interest in mutation modeling has focused on the effects of variation in mutation rates across nucleotide positions in human mitochondrial DNA (mtDNA) sequences. Since the 1960s, there has been a growing suspicion that not all sites are created equal; that, indeed, some mutate at much faster rates than others.

A particularly troublesome issue that is not addressed in the literature is the distinction between mutational "hot spots" and genuine old sites in a sample of sequenced individuals. This problem can be viewed within the context of interpreting the histogram of site-partitions in a sample, also called a frequency spectrum (Harpending et al. 1996; Gurven 1996). For example, consider a sample of 10 individuals sequenced at 100 sites. At site 1, if 1 person has A and the remaining 9 others have G, then the partition at this site is 1 and 9. Similarly, if 2 people have C at site 2 while 8 have T, this is represented as 2 and 8. Because a partition of, say, 2 and 8 is indistinguishable from a partition of 8 and 2, both of these cases would be counted as 2 in the

[1] Department of Anthropology, University of New Mexico, Albuquerque, NM 87131.

Human Biology, June 2000, v. 72, no. 3, pp. 455–471.

KEY WORDS: HUMAN mtDNA, MUTATION RATE VARIATION, ANCIENT DEMOGRAPHY, POPULATION GENETICS

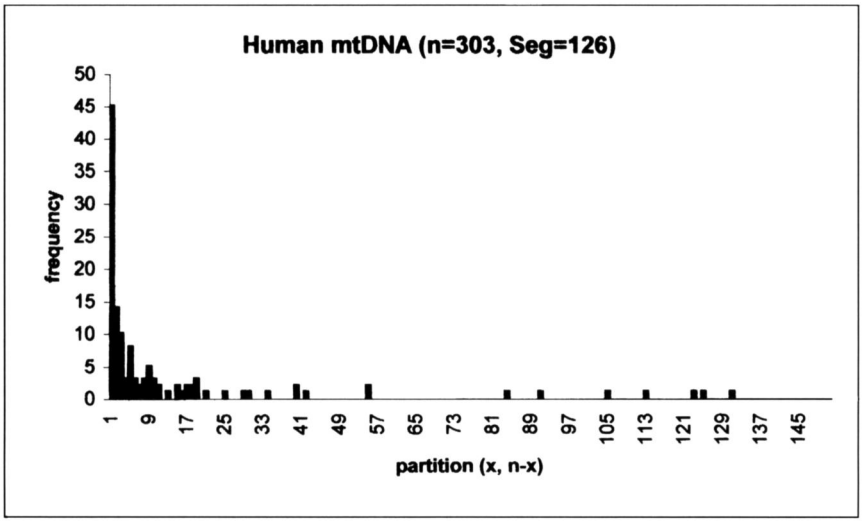

Figure 1. Frequency spectrum for human mtDNA data ($n = 303$, # *seg* $= 126$). The y-axis represents the frequency of x, $n - x$ partitions given by x-axis.

spectrum. In a sample of n individuals, the spectrum spans from 1 to $n/2$ and the sum of all $n/2$ columns equals the total number of segregating sites.

Figure 1 shows the frequency spectrum for a world human sample of 303 individuals with a total of 126 segregating sites in the mtDNA hypervariable control region (HVCR) segments I and II. The curve is generally L-shaped. This shape means that many sites are singletons, while the remaining few partition the sample into larger chunks. Although we know that a population expansion in our recent past will produce a unimodal signature in a mismatch distribution (Rogers and Harpending 1992; Harpending et al. 1993; Sherry et al. 1994), it is not clear what the signature on a frequency spectrum would be. We would expect a star phylogeny, with sites being represented as singletons (Harpending et al. 1998; Harpending et al. 1996), yet this is not entirely the case in the complete human sample, and even less so in several of the populations shown in Figure 2. We must, therefore, ask whether the mass at the right of the spectra represents "old" sites near the root of the tree that genuinely partition the sample, or instead is an artifact of some sites exhibiting much faster mutation rates than most others. How can we distinguish between these 2 possibilities? Before an attempt to answer this question is made, a brief review of recent explorations into mutation rate variation in mtDNA sequences is given.

Early mutation models were based on the "infinite sites" assumption (Kimura 1971), i.e., each mutation results in a new site, leaving the mutation rate at any one site to be practically infinitesimal. If each site mutates at the

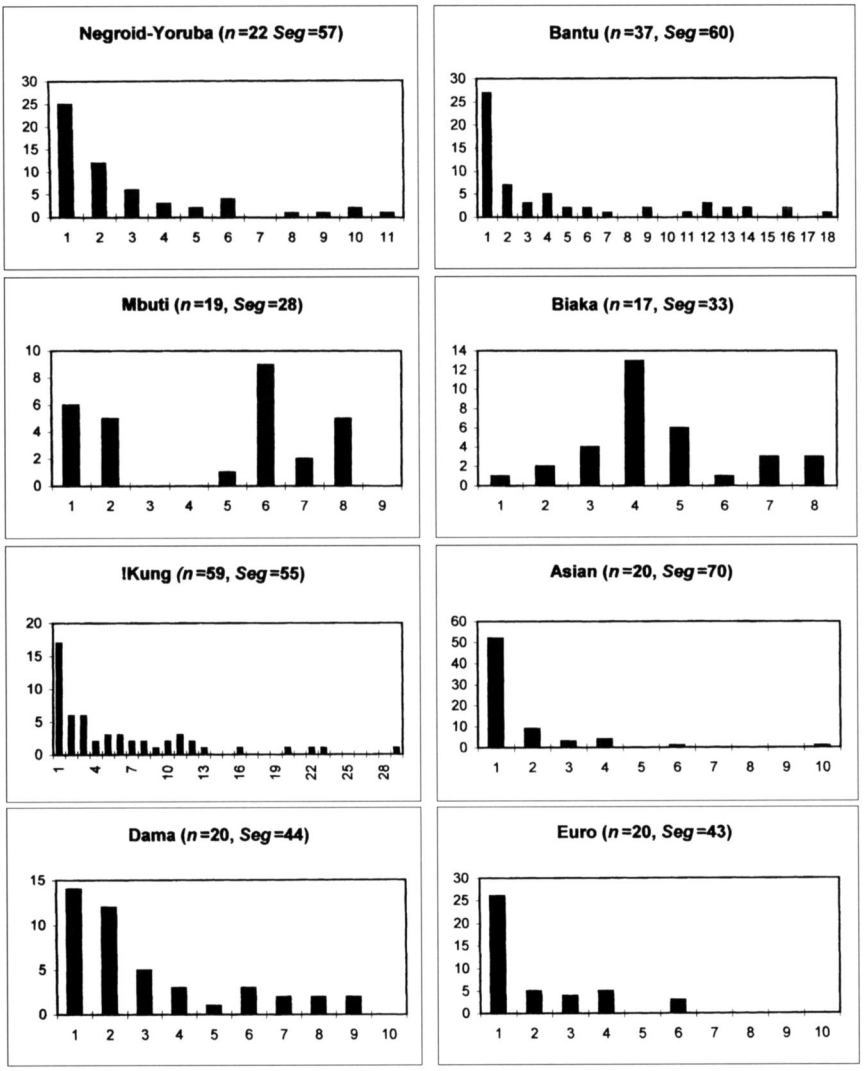

Figure 2. Frequency spectra for several human populations. Data source: Mark Stoneking Laboratory.

same rate, μ, and we have a phylogenetic tree of length T documenting the history of our sample, then the total number of hits per site, X, is a Poisson random variable with mean μT. This has been called the "one-rate model" (Wakeley 1993).

However, if we allow two different rates to divide "fast" and "slow" segregating sites, the distribution of hits per site is a mixture of two Poisson

random variables of means $\mu_1 T$ and $\mu_2 T$, with μ_1, μ_2 being the mutation rates for fast and slow sites, respectively.

Best fits to the available sequence data have been found using a gamma distribution (Yang 1990). In this case, we allow the mutation rate at each site to be drawn at random from a gamma distribution with shape parameter, α. This gives us a continuous distribution of mutation rates, μ_i. The number of hits at a particular site, i, will be Poisson-distributed with mean $\mu_i T$, while the number of hits *per* site follows a negative binomial distribution. Fast sites are distinguished from slow sites by bearing mutation rates above some arbitrary critical rate from the gamma distribution (Wakeley 1993).

The best-fit model of mutation rate variation to human mtDNA data for hypervariable control region (HVCR) segments I and II is the gamma model with $\alpha = 0.17 \pm 0.04$ (Yang and Kumar 1996). In a sample of 77 Asians, Rogers et al. (1996) provide a 95% confidence interval for α (0.09, 0.17) for the same HVCR region.[2] Kocher and Wilson (1991) have estimated $\alpha = 0.11$ for the entire control region, while Wakeley (1993) gives $\alpha = 0.47$ for HVCR I. In general, the smaller the shape parameter, the more L-shaped the distribution. This trend means that, in the control region, most sites will be either invariant or singletons, while fewer sites will display very high mutation rates. It should also be noted that none of the estimates show α to be greater than 1. In this case, the distribution would appear as a bell-shaped curve around some mean substitution rate. From the above estimates, we can see that HVCR I must have more variable sites than HVCR II.

A direct observation of hypervariability in the mtDNA control region comes from Howell et al. (1996). They detected mutations in the NADH-dehydrogenase 6 gene in relatives afflicted with Leber's hereditary optic neuropathy, which resulted in estimates of mutation rate up to 200-fold higher than those estimated from phylogenetic inference. Whether or not this increase represents a localized abnormality in mitochondrial metabolism that would be selected against and, thereby, not appear to be significant in phylogenetic analyses, is not known (Pääbo 1996). This study was also important in highlighting the realistic possibility for intraindividual variation in mtDNA. A similar result comes from Parsons et al. (1997), with implications on interpreting alternative scenarios of human evolution discussed in Loewe and Scherer (1997).

A typical method of measuring the extent of statistical association between loci is linkage disequilibrium. Different loci could have nonrandom associations due to linkage on the same chromosome. Alternatively, if they are on different chromosomes, then nonrandom associations could result from some combination of drift, selection, or nonrandom mating (Lewontin 1988).

[2]This 95% confidence interval assumes a specific population history given by their estimation of the parameter vector (θ_0, θ_1, τ). Other confidence intervals are possible by assuming different population histories (Rogers et al. 1996, Fig. 1).

In molecular sequence data, we measure the association between all possible pairs of polymorphic sites using the linkage disequilibrium (LD) test. We would expect nonrandom association when there is a recent spread of haplotypes from migration or selection, or when old haplotypes are maintained due to balancing selection (Lewontin 1995). Random associations should occur when there is no selection and polymorphism is of ancient origin.

The measure of disequilibrium, D, for any pair of polymorphic sites is calculated as the difference between observed haplotypic frequencies and the expected haplotypic frequencies given the specific allelic frequencies at each site. In our samples, we considered only transitions, since there were very few transversions in the control region. As a result, there were only two nucleotide or "allelic" possibilities at each site. We arbitrarily labeled one base as B and its complement as C. Hence, when comparing sites 1 and 2, we calculate the allelic frequencies at each site, resulting in p_1, q_1 for site 1 and p_2, q_2 for site 2. The expected haplotype frequencies BB, CC, BC, CB can then be readily computed as g_{BB}, g_{CC}, g_{BC}, g_{CB}. Therefore, $D = g_{BB}g_{CC} - g_{BC}g_{CB}$. With random association, $D = g_{BB}g_{CC} - g_{BC}g_{CB} = (p_1p_2)(q_1q_2) - (p_1q_2)(p_2q_1) = 0$. Usually, we use measures of D that are normalized to the maximal conditions (Lewontin 1964, 1988), such that

$$D' = \frac{D}{D_{\max}} \quad \text{where} \quad \begin{array}{ll} D_{\max} = \min(p_1q_2, p_2q_1) & \text{if } D > 0 \text{ (coupling)} \\ D_{\max} = \min(p_1p_2, q_1q_2) & \text{if } D < 0 \text{ (repulsion)} \end{array} \quad (1)$$

D' values are therefore bound between 1 and -1. It is appropriate to regard D as a covariance measure between sites x and y, such that $D = \mathrm{E}[xy] - \mathrm{E}[x]\mathrm{E}[y]$ (Hartl and Clark 1989).

If a site is really a hot spot, we would expect for a randomized pattern between A and G or C and T due to the expected rapid-fire, back-and-forth toggling between allelic states. Usually, singletons and other sites with low absolute frequencies of the rare allele will not be hot spots. Thus, we would expect that most of the hot spots would have relatively large absolute frequencies of the rarer allele, placing such sites towards the right of a frequency spectrum. However, if those sites towards the end of the spectrum are truly old sites that deeply partition the spectrum, then we would expect significant disequilibrium ($|D'| \neq 1$) when these sites are paired with other polymorphic sites, because other individuals can be expected to have a similar shared ancestry. If those sites at the end of the spectrum are instead just hot spots, then we might expect D' values close to zero when these sites are paired with other polymorphic sites. In other words, we should expect low covariance when one or both sites have undergone significant mutation.

Of course, this assumes that the inferred phylogenetic tree linking all individuals is *not* a star phylogeny. When the tree *is* a star phylogeny, as commonly modeled in human populations undergoing rapid population ex-

Table 1. Population Summary Statistics

Group Name	*Sample Size (n)*	*# Sites Used*	*# Seg Sites (p)*	*# Singletons (s)*	*# Pairs*	*$\|D'\| \leq$ 0.05*	*% at Extreme $\|D'\| = 1$*
Herero	37	602	22	8	231	1	84.4
Bantu	37	578	60	27	1770	20	85.2
Asian	20	667	70	52	2415	3	98.4
Biaka	17	695	33	33	528	8	83.5
Mbuti	19	678	28	6	378	4	94.4
Nama	20	513	38	17	703	3	85.2
Dama	17	633	44	14	946	3	90.8
!Kung	59	568	55	17	1485	8	84.2
Europeans	20	531	43	27	903	17	93.1
Negroid/Yorubans	22	629	57	25	1596	6	91.9

pansions, detecting differences in patterns between old sites and hot spots is especially tricky. Here, we would still expect D' values to be close to zero for pairwise comparisons involving hot spots, but it is not clear to what extent D' values associated with old sites should deviate from random equilibrium. We address this latter issue through simulations.

Materials and Methods

Mitochondrial DNA HVCR I and II sequences from 10 human populations were obtained from Mark Stoneking's laboratory. These sequences were 753 base pairs (bp) in length before analysis. After removing those sites with missing data (mostly invariant), as well as transversions, the total number of sites ranged from 513 to 695 bp. Sample sizes and numbers of polymorphic sites for the test populations are given in Table 1.

D' values were calculated for all possible pairs of polymorphic sites in each of the 10 human populations. Given that there are p polymorphic sites for any particular population, there will be $p(p - 1)/2$ total pairs of polymorphic sites. Table 1 shows the number of polymorphic pairs with D' values less than or equal to 0.05, and the percentage of all pairs with maximal $D'(|D'| = 1)$ values. When the latter uninformative pairs were removed (see below), what was left were histograms of interior D' values, as shown in Figure 3.

For each of the 10 populations D^2 values were calculated for all polymorphic sites. This is the sum of the squared D' values of a given polymorphic site paired with every other polymorphic site. That is, D^2 for polymorphic site i is

$$D_i^2 = \sum_{i \neq j} D_{ij}'^2. \tag{2}$$

D^2 values were then ranked within each population from lowest to highest.

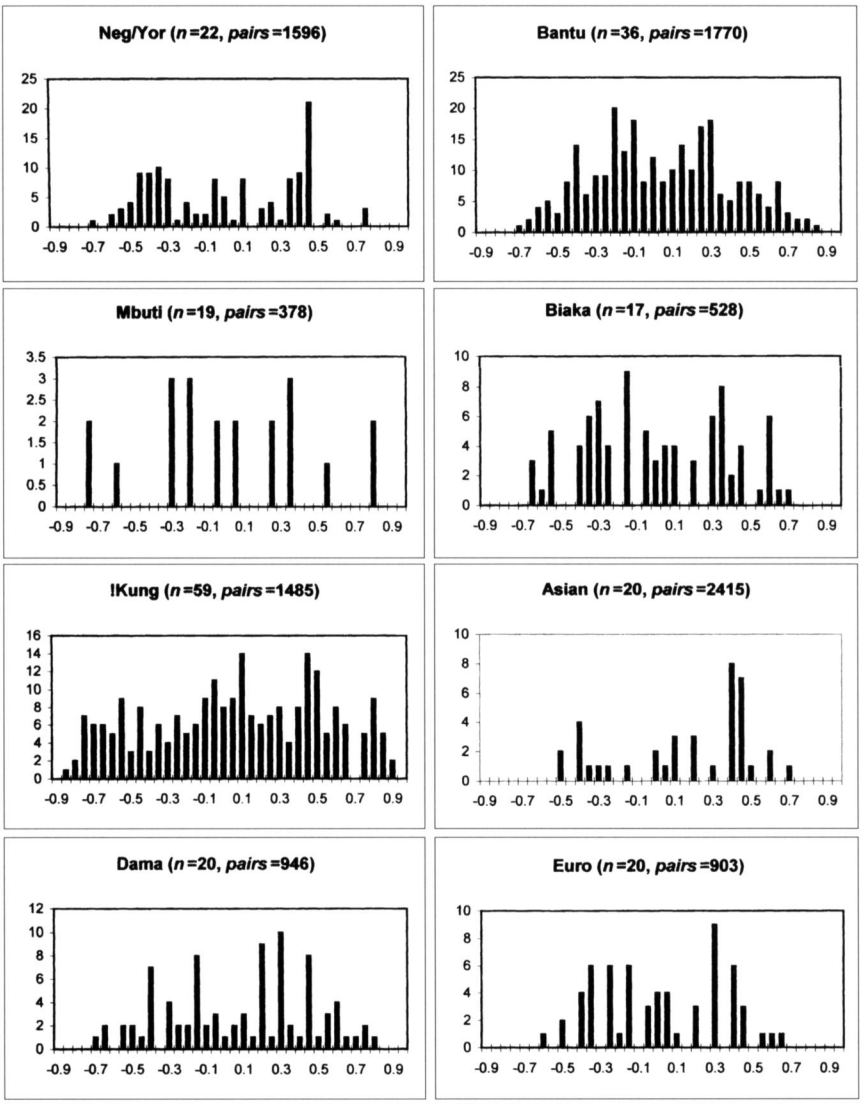

Figure 3. Distribution of interior D' values in several human populations. All cases of $|D'| = 1$ were eliminated from histograms.

The seven polymorphic sites with the lowest values are listed in Table 2. Sites are numbered according to those given in the Cambridge reference (Anderson et al. 1981). Since singletons were included in the analysis, D^2 cannot be very small. Given s singletons in a population, the maximal D^2 value for any site

Table 2. Potential Hot Spots as Indicated by Low D^2 Values in 10 Human Populations

Population	*Site #*	D^2	*Deviation*	*Spectrum*	*LD Rank*	*Wakeley*	*Hasegawa et al.*
Herero	263	11.148	0.758	6	1	N/A	N/A
	16,148	14.902	0.469	2	2	No	Yes
	16,209	15.444	0.427	4	3	No	Yes
	16,311	15.970	0.387	10	4	Yes[a]	Yes[a]
	189	16.885	0.317	8	5	N/A	N/A
	195	17.031	0.305	4	6	N/A	N/A
Bantu	16,311	33.088	0.810	13	1	Yes[a]	Yes[a]
	16,129	34.960	0.751	13	2	N/A	Yes[a]
	16,278	36.490	0.703	12	3	Yes	Yes
	247	36.693	0.697	16	4	N/A	N/A
	152	36.707	0.697	16	5	N/A	N/A
	263	38.242	0.649	14	6	N/A	N/A
	16,230	39.719	0.603	12	7	No	No
Asian	152	60.763	0.485	4	1	N/A	N/A
	16,362	62.911	0.358	4	2	Yes[a]	Yes[a]
	16,223	63.306	0.335	10	3	Yes[a]	Yes[a]
	16,126	63.730	0.310	2	4	N/A	Yes
	16,172	63.846	0.303	2	5	Yes	Yes
	16,311	63.879	0.301	3	6	Yes[a]	Yes[a]
	16,278	63.984	0.295	2	7	Yes	Yes
	16,189	64.638	0.257	6	8	Yes[a]	Yes[a]
	16,129	64.676	0.254	3	9	N/A	Yes[a]
Biaka	306	6.696	0.816	7	1	N/A	N/A
	203	11.971	0.646	2	2	N/A	N/A
	297	17.516	0.467	6	3	N/A	N/A
	16,293	21.320	0.345	7	4	Yes	Yes
	93	24.880	0.230	7	5	N/A	N/A
	16,274	25.938	0.196	8	6	Yes	Yes
	182	25.952	0.195	8	7	N/A	N/A
Mbuti	195	21.389	0.267	5	1	N/A	N/A
	16,172	21.891	0.243	7	2	Yes	Yes
	182	22.217	0.228	6	3	N/A	N/A
	16,294	23.717	0.156	8	4	Yes[a]	Yes[a]
	16,148	23.717	0.156	8	5	No	No
	16,274	23.717	0.156	2	6	Yes	Yes
	247	23.717	0.156	8	7	N/A	N/A
Nama	16,311	21.019	0.799	7	1	Yes[a]	Yes[a]
	146	24.502	0.625	6	2	N/A	N/A
	16,230	24.592	0.620	8	3	No	No
	16,172	25.162	0.592	2	4	Yes	Yes
	16,278	26.352	0.532	6	5	Yes	Yes
	263	27.357	0.482	5	6	N/A	N/A
	16,294	27.465	0.477	7	7	N/A	N/A
Dama	16,311	24.098	0.652	8	1	Yes[a]	Yes[a]
	152	30.708	0.424	5	2	N/A	N/A
	16,304	31.138	0.409	7	3	Yes	Yes
	16,129	32.562	0.360	9	4	N/A	Yes[a]
	16,278	34.140	0.306	8	5	Yes	Yes
	283	34.255	0.302	3	6	N/A	N/A
	195	35.139	0.271	9	7	N/A	N/A

Table 2. Continued

Population	*Site #*	D^2	*Deviation*	*Spectrum*	*LD Rank*	*Wakeley*	*Hasegawa et al.*
!Kung	189	33.869	0.544	8	1	N/A	N/A
	16,214	34.724	0.521	9	2	Yes	No
	198	35.346	0.504	23	3	N/A	No
	207	35.927	0.488	4	4	N/A	N/A
	16,129	36.404	0.476	29	5	N/A	Yes
	195	36.558	0.471	5	6	N/A	N/A
	16,311	36.628	0.470	11	7	Yes[a]	Yes
European	195	30.543	0.764	6	1	N/A	N/A
	16,189	31.527	0.698	6	2	Yes[a]	Yes[a]
	146	32.510	0.633	4	3	N/A	N/A
	16,294	32.972	0.602	6	4	Yes[a]	Yes[a]
	263	33.809	0.546	4	5	N/A	N/A
	16,278	34.260	0.516	3	6	Yes	Yes
	16,187	34.656	0.490	4	7	Yes	Yes
	16,223	34.933	0.471	4	8	Yes[a]	Yes[a]
Negroid/	16,278	42.007	0.518	10	1	Yes	Yes
Yoruba	152	42.449	0.502	11	2	N/A	N/A
	182	42.653	0.494	6	3	N/A	N/A
	16,294	43.038	0.480	4	4	Yes[a]	Yes[a]
	195	43.428	0.466	8	5	N/A	N/A
	16,320	44.763	0.416	3	6	Yes	No

"Yes" and "No" refer to whether or not the site appeared in the Wakeley (1993) or Hasegawa et al. (1993) lists. "N/A" means the site did not appear in either of these lists.

a. ≥9 substitutions at site, representing the most hypervariable of sites in the Wakeley (1993) and Hasegawa et al. (1993) lists.

will be $p - 1$ because all singletons have $|D'| = 1$. The minimal D^2 value will be s, the number of singletons. For each of the listed sites, i, the deviation of D_i^2 from the maximal D^2 was calculated, normalized to the total interval of possible D^2 values, or $|p - 1 - D_i^2|/(p - 1 - s)$. The partitioning of the frequency spectrum is also given for each of the polymorphic sites in the list.

Finally, it was determined whether the sites were featured in published lists of expected hot spots in HVCR I. Wakeley (1993) lists hot spots for sites 16,130–16,379 using Caucasian data from Di Rienzo and Wilson (1991). Hasegawa et al. (1993) lists hot spots for sites 16,024–16,400 using data from world populations from Di Rienzo and Wilson (1991), Horai and Hayasaka (1990), and Vigilant (1990).

Results

Immediately apparent from Table 1 is that most of the D' values hover at the maximal ± 1. As mentioned above, most of these values resulted from

comparisons involving one or two singletons. If a given population contains p polymorphic sites, s of which are singletons, there should be $(2p - s - 1)s/2$ pairwise comparisons involving at least 1 singleton. This should represent the minimal number of D' values equal to $+1$ or -1. The minimal fraction of all D' values that are at their maximal values because of the singleton effect will be $(2p - s - 1)s/(p(p - 1))$. Since singletons accounted for a large proportion of polymorphic sites in each population, most of the pairings resulted in uninformative D' values. A similar pattern was revealed by Merriwether et al. (1991) in a sample of 3065 mtDNAs from 62 geographical regions, which exhibited a total of 81 polymorphic sites.

Inspection of the patterning of interior D' values in the different populations revealed that there were generally only a few cases of D' values being close to zero (Figure 3). This does not mean that evidence for hot spots is lacking. Given the number of sites that were singletons compared to total numbers of polymorphic sites, it seems inevitable that D' values will deviate from zero for some of the comparisons. Therefore, we ranked the D^2 values for each site from lowest to highest, assuming that hot spots should on average have a more random association with other sites than would old sites. This assumption made the highest ranked sites with the lowest summed D^2 values the best candidates for being hot spots.

All of the "hottest" sites from both the Wakeley (1993) and Hasegawa et al. (1993) lists appear throughout Table 2. Almost all populations contained at least one of these rapidly mutating sites. By comparing results for the separate populations, we observed that either some hot spots were present in certain populations but not in others, or that hot spots were expressed differently in each of the populations, perhaps attributable in part to sampling bias. For example, the Asian and European samples showed a closer match to the hottest sites listed in Wakeley (1993), probably because the sample he used consisted primarily of Caucasians. In addition, fewer sites were detected as hot from HVCR II than from HVCR I, agreeing with the assertion made by Wakeley (1993) that the former contains fewer hypervariable sites than the latter.

Many of the sites detected as hot were those which appeared at the right of the respective population's frequency spectrum. Yet certainly not *all* sites at the right of the spectrum were hot. The sites remaining in the mass towards the right of the spectra must, therefore, be old sites.

Simulations. Because it was important to know how well the LD method detected hot spots, test populations were simulated using the Treesim program donated by Paul Lewis. This program uses a Jukes-Cantor (1969) model for allowing a constant mutation rate for all sites, and the Hasegawa, Kishino, and Yano (1985) model for assigning gamma-distributed rates to each site. These models permit the simulation of populations where sites at the right of a spectrum are only old sites, and of populations where sites at the right are

Table 3. Poisson-distributed Mutation Rate Simulations

	Low Mutation Rate ($\mu = 3.3 \times 10^{-7}$)		*Intermediate* ($\mu = 7.8 \times 10^{-7}$)		*High Mutation Rate* ($\mu = 2.2 \times 10^{-6}$)	
LD Rank	*Deviate*	*Spectrum*	*Deviate*	*Spectrum*	*Deviate*	*Spectrum*
1	0.19	1	0.26	2	0.31	3
2	0.19	1	0.23	2	0.26	2
3	0.02	1	0.16	2	0.25	3
4	0.02	1	0.14	2	0.23	2
5	0.00	1	0.14	2	0.21	3
6	0.00	1	0.13	2	0.20	3
7	0.00	1	0.09	2	0.18	2
8	0.00	1	0.06	2	0.17	2
9	0.00	1	0.05	2	0.17	2
10	0.00	1	0.01	1	0.16	2

LD Rank, sites ranked from lowest to highest D^2; *deviate,* deviation from maximal D^2 score; *spectrum,* frequency spectrum partition value for the listed site. For each mutation rate 10 simulations were run, with $n = 25$, site length of 600, and a time depth of 60,000 years for each simulation.

due to the presence of hot spots. In the latter case, since it is possible to discern from the simulations the rate at each site, we can determine whether the D^2 value is useful in distinguishing between high and low variable sites with respect to their position in the spectrum. All simulated populations consisted of 25 individuals, each with 600 sites and a transition/transversion ratio of 20. The estimated time depth of all trees was 60,000 years. Tables 3 and 4 report the average deviation score, frequency spectrum partitioning, and rank of the mutation rate (for gamma simulations) for each of the 10 sites with the lowest D^2 values.

Since old sites are expected to give nonrandom associations when the phylogenetic tree is complex, simulations were done assuming a star phylogeny, as is expected for human populations undergoing a recent population expansion. A constant mutation rate for all sites resulting in a Poisson-distributed number of hits under a star phylogeny (with a finite sites assumption) resulted in most polymorphisms being singletons (95%, 87%, and 65%, respectively, for the 3 sets of simulations: Table 3). Frequency spectra showed all the mass concentrated at the far left (graphs not shown), which was plausible because this model is the base from which deviation in the human data requires explanation. LD tests on these simulated populations showed most D' values fixed at $+1$ or -1, as expected, and most D^2 values fixed at $p - 1$.

The largest deviation score from the maximal D^2 value was always significantly less than 0.5, while the decline in deviations was also very rapid because of the small number of pairwise comparisons that were significant (Table 3). Compared with real human mtDNA data, where deviations are usually higher (except in the Mbuti sample where the highest deviation was

Table 4. Gamma-distributed Mutation Rate Simulations

	$\alpha = 0.16$			$\alpha = 0.50$			$\alpha = 1.0$		
LD Rank	*Deviate*	*Spectrum*	*Rate*	*Deviate*	*Spectrum*	*Rate*	*Deviate*	*Spectrum*	*Rate*
1	0.60	10	8	0.43	6	19	0.34	3	108
2	0.53	8	8	0.39	5	45	0.32	4	33
3	0.51	7	19	0.37	6	24	0.29	3	140
4	0.46	5	8	0.35	4	35	0.27	3	140
5	0.43	5	17	0.33	3	54	0.26	3	78
6	0.42	5	22	0.32	3	46	0.24	3	62
7	0.42	4	28	0.31	4	12	0.24	2	104
8	0.41	4	35	0.31	3	62	0.23	2	71
9	0.40	4	26	0.28	3	56	0.22	2	146
10	0.38	4	35	0.28	3	76	0.22	3	77

LD Rank, sites ranked from lowest to highest D^2; *deviate,* deviation from maximal D^2 score; *spectrum,* frequency spectrum partition value for the listed site; *rate* = rank of the actual mutation rate for the listed site ($\leq$30 is fastest 5%, $\leq$60 is fastest 10%). For each value of the shape parameter, α, 10 simulations were run, with $n = 25$, site length of 600, and a time depth of 60,000 years for each simulation.

0.27), the above results were consistent for both low, intermediate, and high mutation rates.[3] Hence, increasing the mutation rate in a one-rate model merely increases the total number of polymorphic sites without significantly altering the distributions of D' or D^2.

Simulations were also done using gamma-distributed rates, with the α-shape parameter set at 0.16, 0.50, and 1.0 (Table 4). The frequency spectra of these populations were more L-shaped, similar to those of many human populations (Figure 4). However, polymorphic sites were observed at the far right of the spectra only in the $\alpha = 0.16$ simulations.

In the simulations where $\alpha = 0.16$ or 0.50, sites with the lowest D^2 values were always hot spots, with mutation rates in the top 5% or 10% of the gamma distribution. These sites were also located at the right edge of the spectrum. However, it was not always the case that sites with the highest mutation rates had the lowest D^2 values. This could happen for two reasons. First, as observed in some of these simulations, hot spots were at a higher risk of producing a transversion. Since transversions were discarded from all analyses, potential hot sites were thrown away. However, in human data, the number of transversions is so low that we need not worry about them. The other reason could be that hot spots were occasionally represented as partitioning the sample into 2 and 23, or were represented as singletons. We have

[3]The per site mutation rate per year in the low, intermediate, and high rate scenarios was 3.3×10^{-7}, 7.8×10^{-7}, and 2.2×10^{-6}, respectively.

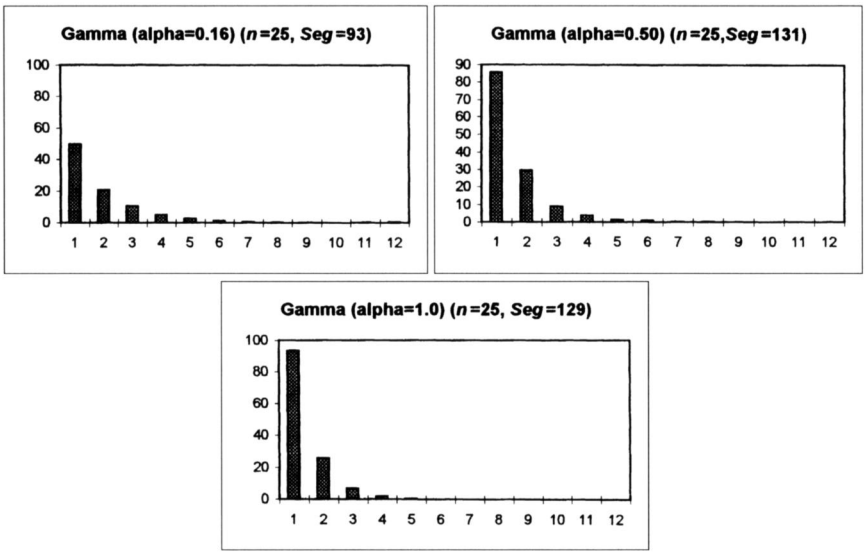

Figure 4. Frequency spectra for the gamma-based (α = 0.16, 0.50, 1.0) simulations. Spectra represent averages of 10 simulations of 600 sites for n = 25 individuals.

already seen how singletons (k = 1) respond to disequilibrium analysis. The verdict for k = 2 and k = 3 is not much better, since there is only one possible interior D' value when k = 2 and only three for when k = 3 (Lewontin 1995). Thus, it seems that the LD method is not reliable for detecting hot spots that may appear in the left portion of the spectrum. Indeed, the LD method was successful in detecting all hot spots with spectrum partition values of 6 or greater, while its success at finding hot spots for other sites further to the left of the spectrum was relatively poor. The average rank for mutation rates among sites with the same spectrum partition values (either 3, 4, or 5) that appeared in the "top 10" lists of Table 4 was 24, 24, and 122 for α = 0.16, 0.50, and 1.0, respectively, compared with 18, 59, and 48, among sites not listed in the top 10.

The pattern of deviation scores in the gamma-simulated samples with α = 0.16 matched the human pattern more closely than any of the Poisson-simulated samples. All these simulations had maximal deviations above 0.5, which then declined slowly, as would be expected if variable mutation rates were driving disequilibrium.

Discussion

The goal of this study was to determine how robust the LD test was in detecting hot spots when the potential for old sites exists. It is reasonable to

believe that the LD measures applied above are useful in detecting hot spots that appear towards the right of the frequency spectrum. With a one-rate model, a sample will be dominated by singletons with little or no mass in the right of the spectrum, leaving a narrow range of potential D^2 values. Since no D' values are close to zero in one-rate modeled samples, D^2 values are usually close to the maximal D^2 values and deviations are, therefore, lower than in gamma-simulated samples. Based on these results, the LD measures are at minimum able to distinguish between the presence and absence of rate variation.

Deviation scores decline gradually for populations with L-shaped frequency spectra. For the Biaka and Dama, the decline looks to be exponentially decreasing. If fewer sites in these populations have high deviation scores while the rest of the sites drop off rapidly, the spectrum could be dominated by old sites rather than hot spots. Although the spectrum for the Mbuti is far from L-shaped, the decline in deviations is still gradual. This decline could reflect an accumulated mass in the left *and* right portions of the spectrum (Figure 2). Out of the 9 sites at position 6 of the Mbuti spectrum, all but 1 had maximal D^2 values, suggesting that most of these are old sites. As old sites with extreme disequilibrium, they 'act' as singletons from the perspective of the D^2 test. Consequently, the low deviation scores in the Mbuti sites imply that there is either a low level of rate variation or else some anomalous sampling bias in the data. Another possibility is that the Pygmy populations do not necessarily follow star phylogenies, but rather their population sizes may have been constant over recent time (Slatkin 1994). This would cause branches near the root of the phylogenetic tree to be long, and produce high levels of linkage disequilibrium.

The LD measures are useful in demonstrating rate variation among sites, and future research in this area may prove promising. A clear advantage is that using the LD test does not assume *a priori* that rate variation is gamma-distributed, where the shape parameter, α, is estimated by inferring the number of changes at a site through some tree-building scheme (e.g., neighbor-joining, parsimony). Relying on tree reconstruction means that one runs the risk of underestimating the number of changes at a site when hot spots are common, thereby overestimating α (Yang 1996).

This analysis also provides some insights into heterogeneity between human populations in the patterning of hot spots. From Table 2, it is evident that not all sites labeled "hot" in the Wakeley (1993) and Hasegawa et al. (1993) lists appear as the hottest sites in each population. Hot spots in HVCR II might appear more frequently in African populations, possibly because hot spots in HVCR I are "slower" than those in non-African populations, thus making sites in HVCR II appear fast. On the other hand, it is more likely that a hot spot could be represented differently in different populations. A rapidly firing hot spot leaves a randomized pattern of A and G or C and T at a site. If we observe $k = 5$ at a particular hot spot in population A and observe k

$= 3$ or $k = 7$ at the same site in population B, we could assume that a single mutation rate accounts for the discrepancy rather than a different mutation rate at the site for each population.

Likewise, we could expect the same pattern if we were to draw 2 or more samples from the *same* population. We might expect positions of hot spots in the frequency spectra of 2 populations to be similar if the 2 populations have recently diverged. Alternatively, 2 populations that diverged further in the past, or a single population stable over a long period of time, might show hot spots at further distances on their respective spectra. Greater understanding of the biological mechanisms of hot spots is necessary. Wakeley (1994, 1996) suggests that biases in the transition/transversion ratio, typical in most animal mitochondrial DNA, is a critical component to understanding patterns of mutation, and suggests that selection at some level is not an unreasonable cause.

If we discover that some hot spots are universal to all human populations while other hot spots are specific to different regional or ethnic populations (e.g., Table 2), we will have to consider phylogenetic context in interpreting the origins and distribution of these variable hot spots. For example, a detailed understanding of phylogenetic relationships may help us specify whether the frequent occurrence of 182 and 189 as hot spots in several of the African populations is an artifact of mtDNA diversification in geographically isolated populations or whether it reflects selective processes. For these types of analyses, we would need to be careful in how we classify our study populations. "Biaka" and "Mbuti," for example, are discrete ethnic populations, while "Asian" and "European" are composite groupings that may include multiple ethnic subpopulations with different mtDNA histories. Grouping multiple ethnic subpopulations into a single population might obscure subtle variation in the pattern of hot spots within these subpopulations.

The appeal of using LD to detect qualitative differences between sites is its overall simplicity. However, as mentioned above, it is not completely independent of phylogenetic analysis. Knowledge of the demographic history of a study population is necessary for interpreting the results of these types of analyses. Populations that have undergone an expansion might not show much linkage unless some linkage already existed in the founding population before the time of expansion, when most of the coalescent events occur (Slatkin 1994). A population at constant size is more likely to have coalescent events occur throughout the tree, which should result in more intermediate LD values.

Acknowledgements I thank Henry Harpending, Paul Lewis, Steve Sherry, and Mark Stoneking for their excellent advice and assistance. I am also grateful to Alan Rogers and Tad Schurr for their many helpful comments on earlier versions of this manuscript.

Received 18 September 1998; revision received 15 April 1999.

Literature Cited

Anderson, S., A.T. Bankier, B.G. Barrell et al. 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290:457–465.

Di Rienzo, A., and A.C. Wilson. 1991. Branching pattern in the evolutionary tree for human mitochondrial DNA. *Proc. Natl. Acad. Sci. USA* 88:1597–1601.

Gurven, M.D. 1996. What can we infer from the mtDNA evidence about human demographic history? B.A. thesis, Department of Anthropology, The Pennsylvania State University.

Harpending, H.C., S.T. Sherry, A.R. Rogers et al. 1993. The genetic structure of ancient human populations. *Curr. Anthropol.* 34:483–496.

Harpending, H.C., J. Relethford, and S.T. Sherry. 1996. Methods and models for understanding human diversity. In *Molecular Biology and Human Diversity,* A. Boyce and N. Mascie-Taylor, eds. Cambridge, England: Cambridge University Press, 283–299.

Harpending, H.C., M.A. Batzer, M.D. Gurven et al. 1998. Genetic traces of ancient demography. *PNAS* 95:1961–1967.

Hartl, D.L., and A.G. Clark. 1989. *Principles of Population Genetics*. Cambridge, MA: Sinauer Associates.

Hasegawa, M., A. Di Rienzo, T.D. Kocher et al. 1993. Toward a more accurate time scale for the human mitochondrial DNA tree. *J. Mol. Evol.* 37:347–354.

Hasegawa, M., H. Kishino, and T. Yano. 1985. Dating of the human-ape splitting by a molecular clock of mtDNA. *J. Mol. Evol.* 22:160-174.

Horai, S., and K. Hayasaka. 1990. Intra specific nucleotide sequence differences in the major noncoding region of human mitochondrial DNA. *Am. J. Hum. Genet.* 46:828–842.

Howell, N., I. Kubacka, and D.A. Mackey. 1996. How rapidly does the human genome evolve? *Am. J. Hum. Genet.* 59:501–515.

Jukes, T.H., and C.R. Cantor. 1969. Evolution of protein molecules. In *Mammalian Protein Metabolism III,* H.N. Munro, ed. New York: Academic Press, 21–132.

Kimura, M. 1971. Theoretical foundation of population genetics at the molecular level. *Theor. Popul. Biol.* 2:174–208.

Lewontin, R.C. 1964. The interaction of selection and linkage 1. General considerations: heterotic models. *Genetics* 49:49–67.

Lewontin, R.C. 1988. On measures of gametic disequilibrium. *Genetics* 120:849–852.

Lewontin, R.C. 1995. The detection of linkage disequilibrium in molecular sequence data. *Genetics* 149:377–388.

Loewe, L., and S. Scherer. 1997. Mitochondrial Eve: The plot thickens. *TREE* 12:422–423.

Merriwether, D.A., A.G. Clark, S.W. Ballinger et al. 1991. The structure of human mitochondrial DNA variation. *J. Mol. Evol.* 33:543–555.

Pääbo, S. 1996. Mutational hot spots in the mitochondrial microcosm. *Am. J. Hum. Genet.* 59:493–496.

Parsons, T.J., D.S. Muniec, K. Sullivan et al. 1997. A high observed substitution rate in the human mitochondrial DNA control region. *Nat. Genet.* 15:363–368.

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

Rogers, A.R., A.E. Fraley, M.J. Bamshad et al. 1996. Mitochondrial mismatch analysis is insensitive to the mutational process. *Mol. Biol. Evol.* 13:895–902.

Sherry, S.T., A.R. Rogers, H.C. Harpending et al. 1994. Mismatch distributions of mtDNA reveal recent human population expansions. *Hum. Biol.* 6:761–775.

Slatkin, M. 1994. Linkage disequilibrium in growing and stable populations. *Genetics* 137:331–336.

Vigilant, L. 1990. Control region sequences from African populations and the evolution of human mitochondrial DNA. Ph.D. diss. Berkeley, CA: University of California.

Wakeley, J. 1993. Substitution rate variation among sites in hypervariable region 1 of human mitochondrial DNA. *J. Mol. Evol.* 37:613–623.

Wakeley, J. 1994. Substitution rate variation among sites and the estimation of transition bias. *Mol. Biol. Evol.* 11:436–442.

Wakeley, J. 1996. The excess of transitions among nucleotide substitutions: New methods of estimating transition bias underscore its significance. *TREE* 11:158–163.

Yang, Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *TREE* 11:367–372.

Yang, Z., and S. Kumar. 1996. Approximate methods for estimating the pattern of nucleotide substitution and the variation of substitution rates among sites. *Mol. Biol. Evol.* 13:650–659.