Estimating Per-Locus Mutation Rates

John F. Chandler

Abstract

Relative per-locus mutation rates for Y-DNA microsatellites, and also for mitochondrial DNA single-nucleotide polymorphisms, can be estimated directly from diverse collections of haplotypes without segregating population components. The calibrated average mutation rates for the FTDNA panels of 12, 25, and 37 loci are 0.00187±0.00028, 0.00278±0.00042, and 0.00492±0.00074, respectively, from a collection of haplotypes from Y-Search. The individual per-locus rates are given here.

Introduction

The task of estimating mutation rates for human DNA is troublesome, whether for individual loci or for averages over panels of loci. The mutations are so infrequent, and the generation spans so long, that direct observation of a statistically useful sample is expensive and time-consuming. Nonetheless, these rates are of great interest in genetic genealogy and related fields. Techniques for circumventing this basic difficulty to gather the necessary data include: "piggybacking" on paternity tests (Gusmão et al. 2005), testing judiciously chosen subjects with interconnected, deep-rooted pedigrees (Heyer et al. 1997), and sifting large, heterogeneous collections of measured haplotypes using statistical models to extract the mutation rates from the other information present (Zhivotovsky et al. 2004; Hutchison et al. 2004). The latter technique needs calibration, but can be applied to data collected for other purposes. Indeed, the very same DNA testing that demands knowledge of the mutation rates for interpreting the results can lead to the determination of those rates. Of particular interest is the technique introduced by Hutchison et al (2004) of sorting pairs of haplotypes by closeness and extracting information from the match profiles. This technique requires a cumbersome extra step of identifying and isolating a sub-population and depends on the dangerous assumption that the chosen sub-population is entirely characterized by a single time depth, but there is a simpler method that avoids these problems. In this paper, I develop expressions for the "mutation model curve" (MMC) of Hutchison et al. (2004) and outline a procedure for using the high-match end of the MMC for extracting mutation rates. Throughout this paper, the mutation rates are assumed to be independent of time, environment, and haplotype. Since these assumptions may be false, especially the last (Dupuy et al. 2004), the results must be taken in the context of the data considered. Of these factors, only haplotype is potentially accessible to a more detailed analysis of currently existing data.

Mathematical Model

First, we must define a function $p_j(g)$ as the probability that all loci, except locus *j*, match between two haplotypes separated by a total of *g* generations. Each haplotype consists of *N* loci, each locus *j* with a different mutation rate given by μ_j. If we assume the mutation rates are small, we may approximate mutation probabilities by an exponential function. In terms of the infinite alleles model, we would write the probability of a single locus *j* remaining at the ancestral allele after *g* generations as $\exp(-g\mu_j)$, while the probability of a mutation at that locus would be $1-\exp(-g\mu_j)$. Then p_j is just the product of the *N* individual probabilities at the *N* loci:

$$
\begin{aligned}
p_j(g) &= [1-\exp(-g\mu_j)]\prod_{i\neq j}\exp(-g\mu_i) \\
&= [\exp(g\mu_j)-1]\prod_{i}\exp(-g\mu_i)
\end{aligned}
\qquad 1
$$

where the second line of Equation 1 comes from multiplying and dividing the first by $\exp(g\mu_j)$. Note that *g*, the separation between two haplotypes, can be viewed either as the number of generations from an ancestor being compared to a particular descendant or as the "round-trip" number of generations from one person back to a shared ancestor and then forward to the contemporary being compared. Since most DNA testing is done on living individuals, the latter interpretation of *g* is more commonly applicable, but the former is equally valid. In the limit of small $g\mu$, the choice of mutation model is unimportant, and, to leading order in the small parameters, we would have for either the infinite alleles model or the stepwise mutation model:

Address for correspondence: john.chandler@alum.mit.edu

Received: September 12, 2006; accepted: October 14,2006

$$p_j(g) = g\mu_j Q(g) \tag{2}$$

where $Q(g)$ is the probability of a match on all loci in whichever model. Similarly, we may write the probability that two haplotypes match at all loci except j, k, and l as

$$p_{jkl}(g) = g^3 \mu_j \mu_k \mu_l Q(g) \tag{3}$$

and so on for any number of mismatching loci. Next we need an expression for the probability that a mutation has occurred at some number of loci, b, while the remainder have remained at the ancestral alleles. As an example, consider the case where exactly three loci (any three) out of a total of five loci, have mutated. The probability of this occurring is just the sum of the probabilities of all possible triplets of mutating loci (each given by Equation 3):

$$\begin{aligned} & g^3[\mu_1\mu_2\mu_3 + \mu_1\mu_2\mu_4 + \mu_1\mu_2\mu_5 + \\ & \mu_1\mu_3\mu_4 + \mu_1\mu_3\mu_5 + \mu_1\mu_4\mu_5 + \mu_2\mu_3\mu_4 + \\ & \mu_2\mu_3\mu_5 + \mu_2\mu_4\mu_5 + \mu_3\mu_4\mu_5]Q(g) \\ & = g^3 C(3) Q(g) \end{aligned} \tag{4}$$

where we define $C(b)$ as the sum of all products of b distinct elements of the set $\{\mu_i\}$. For the general case, the probability that b loci out of N have mutated is given by

$$M(b,g) = g^b C(b) Q(g) \tag{5}$$

Now consider the probability that a *particular* locus has mutated along with any two others, *i.e.,* the probability of a mismatch at the specified locus when all but three of the loci match. This probability is similar to that shown in Equation 4, except that the sum within the brackets includes only those terms that contain, say, μ_1. That is (continuing to use the example of b=3 and N=5), the probability is given by

$$\begin{aligned} & g^3[\mu_1\mu_2\mu_3 + \mu_1\mu_2\mu_4 + \mu_1\mu_2\mu_5 + \\ & \mu_1\mu_3\mu_4 + \mu_1\mu_3\mu_5 + \mu_1\mu_4\mu_5]Q(g) \\ & = g^3\mu_1[\mu_2\mu_3 + \mu_2\mu_4 + \mu_2\mu_5 + \\ & \mu_3\mu_4 + \mu_3\mu_5 + \mu_4\mu_5]Q(g) \\ & = g^3 \mu_1 C_1(2) Q(g) \end{aligned} \tag{6}$$

where we define a function $C_j(b-1)$ as the sum of all products of b-1 distinct elements of the set $\{\mu_i, i \neq j\}$. The limiting case for both this new function and the original unsubscripted C is defined as $C(0) = 1$. Thus, the probability of a mismatch at locus j when all but b of the loci match, given a separation of g generations, can be written as

$$M_j(b,g) = g^b \mu_j C_j(b-1) Q(g) \tag{7}$$

where b must be greater than 0, since locus j is a mismatch by definition. Clearly, the probability must vanish when b is 0. In general, we must deal with a population of haplotype pairs with a range of separations, and we thus must calculate $D_j(b)$, the overall probability of a mismatch at locus j for the whole population when all but b of the loci match. We do so by weighting the probability in Equation 7 by the fraction $f(g)$ of the population of pairs having a separation of g generations and summing over g:

$$D_j(b) = \mu_j C_j(b-1) \sum_g f(g) g^b Q(g) \tag{8}$$

Similarly, we may write $D(b)$, the total probability of b mismatches, in the same population as

$$D(b) = C(b) \sum_g f(g) g^b Q(g) \tag{9}$$

The salient feature of Equations 8 and 9 is the fact that they share a common factor encapsulating the unknown population statistics $f(g)$, and thus these equations differ only in terms that depend just on the mutation rates. It is therefore useful to define a mismatch profile function

$$\begin{aligned} P_j(b) & \equiv \frac{D_j(b)}{D(b)} \\ & = \frac{\mu_j C_j(b-1)}{C(b)} \end{aligned} \tag{10}$$

This function is the conditional probability of a mismatch on locus j, given b mismatches in all. This is directly related to the MMC as defined by Hutchison et al. (2004); their $M_{i,n}$ (probability of a match at locus i given n matches) is just $1 - P_i(N-n)$. By construction, $P_j(0) = 0$ and $P_j(N) = 1$, since no locus can mismatch if the number of mismatches is 0, and every locus must mismatch if the number is N. Of course, the MMC will depart increasingly from Equation 10 as b grows beyond the linear limits assumed in Equation 2, since the population structure will contribute differently to the non-linear terms omitted from Equations 8 and 9.

Equation 10 can be inverted to give an expression for the mutation rate in terms of the (observed) mismatch profiles and the C functions:

$$\mu_j = \frac{C(b)P_j(b)}{C_j(b-1)} \qquad 11$$

Of course, Equation 11 is recursive, in that the C values depend on the mutation rates, but it can serve as the basis for an iterative procedure for calculating the rates. The definition of C involves combinations of many terms—so many that the direct computation becomes prohibitive at relatively modest values of b and N. It can be shown that C obeys the recursion relation

$$\begin{aligned} C(b) &= \frac{1}{b}\sum_{k=1}^{b}(-1)^{k-1}S_k C(b-k) \\ C_j(b) &= \frac{1}{b}\sum_{k-1}^{b}(-1)^{k-1}[S_k - \mu_j^k]C_j(b-k) \end{aligned} \qquad 12$$

where

$$S_k \equiv \sum_i \mu_i^k \qquad 13$$

In the trivial case where the individual mutation rates are all the same, Equation 12 simplifies to

$$\begin{aligned} C(b) &= \binom{N}{b}\mu^b \\ C_j(b) &= \binom{N-1}{b-1}\mu^b \end{aligned} \qquad 14$$

where μ is the uniform mutation rate, and the parentheses indicate binomial coefficients, where the coefficient "N take b" stands for $N!/(b![N\text{-}b]!)$. Also, Equation 10 simplifies to

$$P_j(b) = \frac{b}{N} \qquad 15$$

Thus, in the absence of population structure, the MMC for uniform mutation rates would be straight lines from the (0, 0) to (N, 1), exactly as one would expect. This simple linear form suggests a practical iterative procedure for determining the relative mutation rates from the mismatch profiles:

1) Choose a suitable value b as large as possible, but small enough that the mismatch functions are small compared to 1.
2) From the data, find the values $P_j(b)$ for all loci j.
3) Initialize the mutation rate estimates to $\mu_j = rP_j(b)$, where r is a normalization factor chosen to give values in a convenient range. (Since we have cancelled out the population statistics, it is clear that we can obtain only relative rates from this analysis, and so the normalization is arbitrary.)
4) Use the mutation rate estimates to evaluate Equations 13, 12, and 11 to obtain a new set of estimated rates.
5) Repeat Step 4 until convergence.
6) Repeat Steps 2-5 for all values of b from 1 to the value chosen in Step 1 and take a weighted average of the mutation rate estimates. (See below for a discussion of error analysis.)

Error analysis

Analyzing data in pairs instead of singly introduces correlations between pairs with shared observations. Thus, a proper error analysis of this procedure would be complicated by the need to deal with all pairs of observations. Indeed, correlations could even bias the results. However, since the mismatch profiles are computed on restricted subsets of the pairs, the pair-to-pair correlation within each bin is greatly reduced. Thus, a simple error analysis treating each pair as an independent observable should suffice, with just one modification. Normally, the least-squares estimate of a probability p from statistics of a population of N cases carries an uncertainty of $\{p(1\text{-}p)/N\}^{0.5}$, but the relevant N here must be the lesser of the number of pairs found in a given b-bin and the total number of haplotypes, since the latter is the number of independent data. The uncertainties for the mismatch probability are scaled to mutation rate uncertainties as in Equation 11, and these in turn are used as the relative weights for the weighted average described in Step 6 above (in the usual inverse-square sense).

Calibration

The final component of this analysis is the calibration for converting the relative mutation rates to absolute rates. Pedigree-based rate determinations offer the advantage of "leverage," whereby each test subject carries an accumulation of mutations over many generations (though not so many that multiple mutations on any one locus would be an issue). However, the collection of pedigree data from volunteers who already know the results of the DNA tests leads to serious risks of selection bias in the data. At present, the only reliable large collection of mutation statistics for Y-DNA microsatellites is that collected from father-son pair studies by Gusmão et al. (2005). To make use of such statistics, we must simultaneously fit the observables k_i (number of mutations for locus i in the calibration data) and μ_i (the "observed" relative mutation rate) with a simple model by weighted-least-squares analysis:

$$\begin{aligned} k_i &= m_i n_i \\ \mu_i &= c m_i \end{aligned} \qquad \textbf{16}$$

where m_i is the absolute mutation rate for locus *i*, *c* is the calibration factor between relative and absolute rates, and n_i is the number of meioses for locus *i* in the calibration data. A third relation using μ_i' and *c*′ can be used to include a second, independent set of rate estimates in the analysis.

Application to Y-STRs

I have applied this procedure to a collection of 8430 Y haplotypes downloaded from Ysearch in July of 2006 as a demonstration. In the parlance of Ysearch, each haplotype consists of 37 loci, but this set includes five multi-copy loci, such that the number of independent loci is only 30. Each copy of a multi-copy locus in one haplotype must match the corresponding copy in the other haplotype if the locus is to be considered a match. This grouping of the loci avoids the necessity of guessing which copy corresponds to which in the two haplotypes being compared. About half of the collection belongs to haplogroup R1b as specified by the contributors, and the other half is an assortment of other haplogroups.

Most of the information carried by this collection is filtered out in Step 1 of the analysis, since only nearly-matching pairs are considered. In **Figure 1** and in each panel of **Figure 2**, there is a vertical line marking the lower limit of matching used in Steps 2-6, and it is apparent from **Figure 1** that only a tiny fraction of the pairs is included. The tall peak on the left of **Figure 1** represents the "typical" inter-haplogroup comparison and is thus characteristic of the particular mix of haplogroups included in the data collection. The peak for this collection is at 7/30 matching and thus corresponds to a very old population. The slightly smaller peak on the right represents the "typical" within-haplogroup comparison, mainly the comparison within R1b, which dominates this collection. It should be possible to gather collections whose histograms would show more than two peaks by focusing on distinctive clades within haplogroups. However, as noted above, the information contained in the peaks of **Figure 1** is ignored in the present analysis.

Examination of **Figure 2** shows that the "ideal" straight-line MMC indicated by Equation 15 is seldom realized. Indeed, it can be shown that, even in the absence of population structure (such as is strikingly revealed in **Figure 1**), the MMC based on Equation 10 is generally curved when the locus-specific mutation rates are not the same. The curvature is positive for loci with below-average mutation rates and negative for loci with above-average rates. Attempting to fit straight lines to the MMCs without taking this curvature into account would tend to compress the apparent dynamic range of mutation rates, thus reducing the estimates of high rates and raising the estimates of low rates.

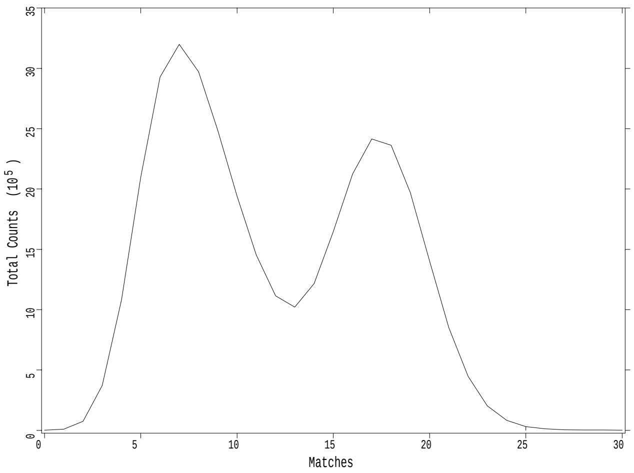

Figure 1. Histogram of matches among pairs of 8430 Y haplotypes, each consisting of 30 microsatellite loci.

The results of the analysis for this collection of Y haplotypes, combined with an independent set of 6955 20-locus Ysearch haplotypes, are presented in **Table 1**. These results include the calibration of Equation 16, and the uncertainties shown here include the contribution of the calibration and the level of agreement between the 20- and 30-locus sets. However, the error analysis here makes no provision for the uncertainties due to the assumption of constant rates.

The average mutation rate for these loci, considered as 37 loci in Ysearch terms, is 0.00492±0.00074 mutation per locus per generation. In contrast, the average rates for the first 12 and the first 25 are 0.00187±0.00028 and 0.00278±0.00042, respectively. The extreme cases, CDYa and CDYb, are thus about seven times as fast as the average for all 37, while DYS426 is about 1/50 as fast. The uncertainties for the average rates are the statistical standard deviations, scaled such that χ^2 per degree of freedom is unity. They include the effect of the correlations introduced by scaling each whole set of relative mutation rates with a common calibration factor and by multiple-counting the multi-copy loci. The uncertainties are thus about triple the corresponding

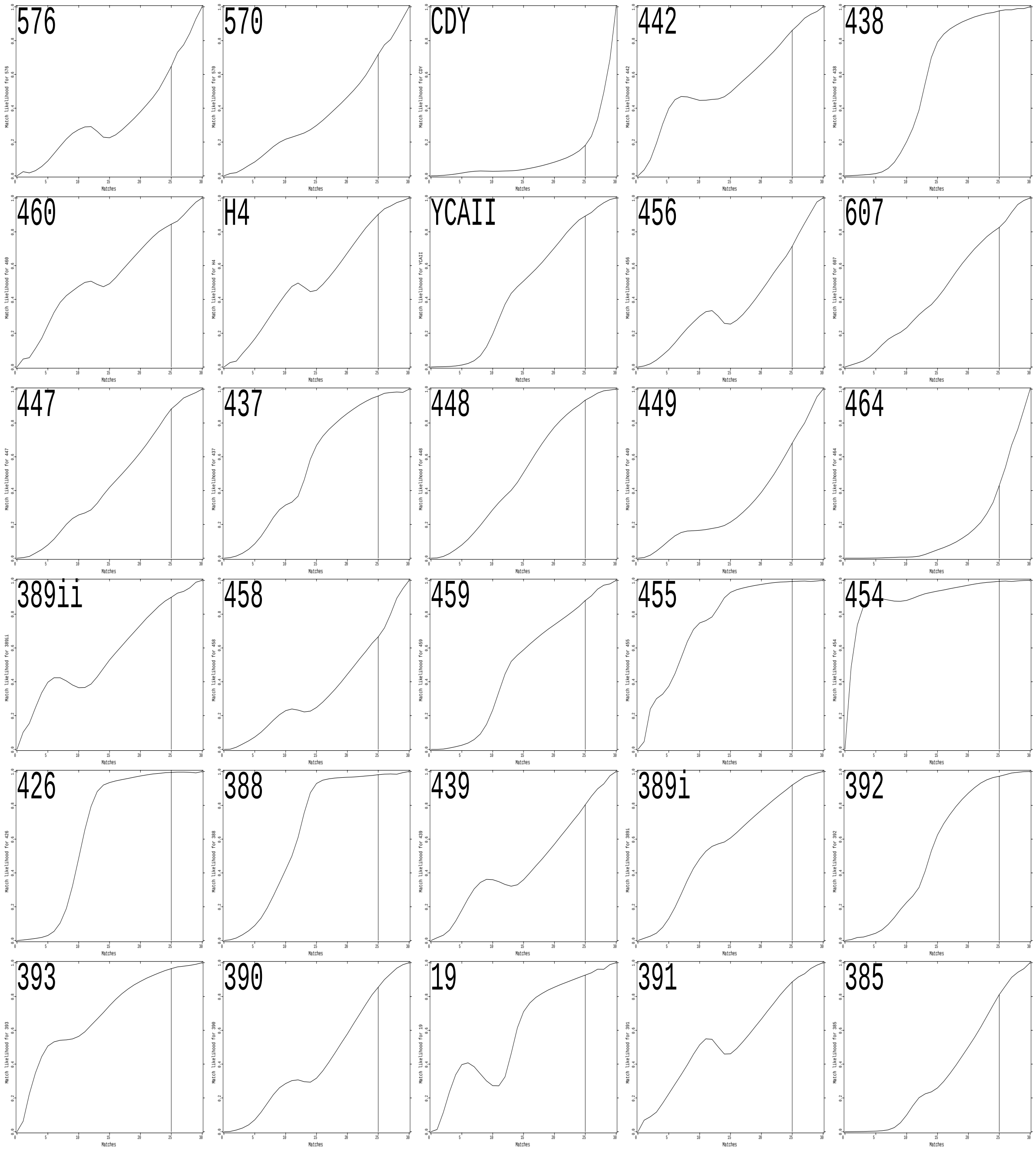

Figure 2. Mutation model curves for 30 Y microsatellite loci computed from 8430 haplotypes. Each panel plots the probability of a match for the specified locus over the range from 0 to 30 matching loci. The panel labels are abbreviated by omitting "DYS."

normalized root-sum-squares of the individual per-locus standard deviations. Not surprisingly, the 12-locus average has the smallest uncertainty: 11 of the 12 are "anchored" by calibration data. Also, the per-locus standard deviations have a component which varies as the square-root of the mutation rate, and thus the composite uncertainties should also be smaller for smaller average rates.

Another unsurprising result is that the fits agree rather well with the calibration data – the exceptions being mainly those loci with the fewest father-son pairs, such as DYS388 and DYS437. These fits, however, are at odds with previous analyses using the method of Hutchison et al. (2004), in that the range of rates is now wider because of the proper accounting for MMC curvature, as noted above. Comparison with analyses of

Table 1. Calibration data and final results for absolute mutation rates (per copy for multi-copy loci, indicated by asterisks). Weighted fit to 6955 20-locus and 8340 30-locus haplotypes.

Locus	Calibration data	Best fit	Std. dev.	Locus	Calibration data	Best fit	Std. dev.
DYS393	0.00075	0.00076	0.00014	DYS447		0.00264	0.00041
DYS390	0.00227	0.00311	0.00048	DYS437	0.00222	0.00099	0.00017
DYS19	0.00168	0.00151	0.00025	DYS448		0.00135	0.00020
DYS391	0.00351	0.00265	0.00041	DYS449		0.00838	0.00128
DYS385 *	0.00224	0.00226	0.00035	DYS464 *		0.00566	0.00087
DYS426		0.00009	0.00002	DYS460	0.00450	0.00402	0.00069
DYS388	0.00057	0.00022	0.00004	Y-GATA-H4	0.00290	0.00208	0.00033
DYS439	0.00530	0.00477	0.00073	YCAII *	0.00000	0.00123	0.00019
DYS389i	0.00188	0.00186	0.00028	DYS456		0.00735	0.00115
DYS392	0.00061	0.00052	0.00010	DYS607		0.00411	0.00066
DYS389ii	0.00226	0.00242	0.00041	DYS576		0.01022	0.00167
DYS458		0.00814	0.00124	DYS570		0.00790	0.00138
DYS459 *		0.00132	0.00021	CDY *		0.03531	0.00549
DYS455		0.00016	0.00004	DYS442		0.00324	0.00051
DYS454		0.00016	0.00003	DYS438	0.00044	0.00055	0.00012

volunteer-driven, pedigree-based data is unprofitable because of the unknown biases associated with the latter.

Application to mtDNA

The same analysis can be applied to mitochondrial DNA, but there are too many base pairs in the standard HVR sequences (1143 in the HVR1+HVR2 test offered by Family Tree DNA) to treat them all as separate loci here. Dealing with compound loci consisting of 20 base pairs each will keep the number of loci down to a manageable 58. I downloaded 2717 such haplotypes from Mitosearch in April of 2006 and treated them in the same way as the Y DNA haplotypes above, except that I have no corresponding calibration data from direct observation. The results are therefore expressed with an arbitrary scale factor to give an average rate of 0.0001 mutation per generation per (compound) locus. It is worth noting that the histogram of matches, shown in Figure 3, is very different from that in Figure 1. There is only one peak in the distribution, and the "tail" of that peak is still significantly above zero all the way to perfect matches. In other words, perfect matches are not at all uncommon, and there is no obvious separation of population components.

The results of the mtDNA analysis are shown in Table 2, with each compound locus labeled by the position of its last base pair. Within the resolution of this analysis, four of the loci show no mutability at all, and seven others are at or below 1% of the average rate. (Of course, two loci are shorter than 20 base pairs, being 9 and 14, respectively, and both of these figure in the extreme low rate lists.) At the high end, the locus 301-320 is about 14 times as fast as the average.

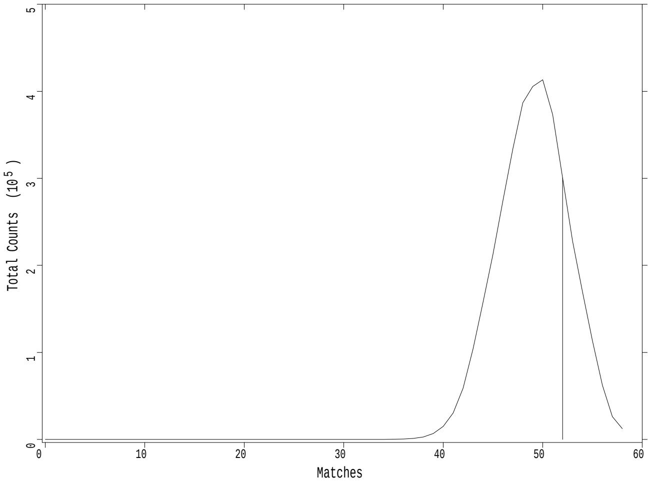

Figure 3. Histogram of matching among pairs of 2717 mtDNA haplotypes

Conclusion

In the Spring of 2006, Family Tree DNA introduced an additional panel consisting of 30 Y DNA microsatellite loci beyond the 37 analyzed here. At the time of this writing, there are hundreds, though not thousands, of completed tests delivered to customers and potentially available for analysis. In the near future, it should be possible to apply the method described here to these extended haplotypes and estimate the rates of all tested loci.

Web Resources

www.ysearch.org	Y-STR database
www.mitosearch.org	mtDNA database

References

Dupuy, BM, et al (2004) Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. Hum Mutat 23:117-124.

Gusmão, P, et al (2005) Mutation rates at Y chromosome specific microsatellites. Hum Mutat 26:520-528.

Heyer, E, et al (1997) Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees. Hum Mol Genet 6:799–803.

Hutchison, LAD, et al (2004) Direct determination of mutation characteristics of Y chromosome STR loci. Poster presented at the American Society of Human Genetics 2004 Annual Meeting, October 2004, Toronto.

Zhivotovsky, LA, et al (2004) The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet 74:50-61

Table 2 Normalized mutation rates for compound mtDNA loci

Last base	Best Fit	Std. Dev.	Last base	Best Fit	Std. Dev.
16020	0.00000000	0.00000000	20	0.00000138	0.00000026
16040	0.00000005	0.00000002	40	0.00000078	0.00000011
16060	0.00001345	0.00000133	60	0.00000637	0.00000065
16080	0.00002193	0.00000356	80	0.00025204	0.00008546
16100	0.00008002	0.00000897	100	0.00002812	0.00000395
16120	0.00002868	0.00000475	120	0.00001461	0.00000164
16140	0.00022024	0.00001290	140	0.00000426	0.00000049
16160	0.00004133	0.00000467	160	0.00044132	0.00001566
16180	0.00011192	0.00000279	180	0.00000350	0.00000037
16200	0.00025233	0.00001461	200	0.00015501	0.00001005
16220	0.00007979	0.00000200	220	0.00002686	0.00000293
16240	0.00020343	0.00001573	240	0.00003063	0.00000322
16260	0.00007544	0.00000295	260	0.00001839	0.00000087
16280	0.00015686	0.00000603	280	0.00031359	0.00006351
16300	0.00035193	0.00003702	300	0.00001306	0.00000066
16320	0.00036923	0.00004056	320	0.00143273	0.00011906
16340	0.00002024	0.00000236	340	0.00001092	0.00000055
16360	0.00006397	0.00000324	360	0.00000021	0.00000002
16380	0.00005363	0.00000566	380	0.00000137	0.00000050
16400	0.00003918	0.00000701	400	0.00000210	0.00000025
16420	0.00000000	0.00000000	420	0.00000496	0.00000103
16440	0.00000100	0.00000007	440	0.00000006	0.00000003
16460	0.00000000	0.00000000	460	0.00004121	0.00000543
16480	0.00000540	0.00000086	480	0.00007197	0.00001360
16500	0.00001821	0.00000436	500	0.00005288	0.00001276
16520	0.00048480	0.00005062	520	0.00001667	0.00000081
16540	0.00001170	0.00000168	540	0.00014729	0.00000356
16560	0.00000002	0.00000001	560	0.00000237	0.00000030
16569	0.00000000	0.00000000	574	0.00000058	0.00000009