Direct determination of mutation characteristics of Y chromosome STR loci

Luke A. D. Hutchison[1], Natalie M. Myres[1], Jacob E. Ekins[1], Ugo A. Perego[1], Jayne B. Ekins[1], Katie Hadley[1], Lara Layton[1], Mindy L. Lunt[1], Sacha S. Masek[1], Alison A. Nelson[1], Mary E. Nelson[1], Katie L. Pennington[1], Jenny L. Peterson[1], Amanda Sims[1], Trish Tolley[1], Alison Welch[1], Scott R. Woodward[1,2].

[1] Sorenson Molecular Genealogy Foundation, Salt Lake City, Utah, USA (smgf.org)
[2] Dept. Micro and Molecular Biology, Brigham Young University, Provo, Utah, USA

We present new analysis methods for haploid STR loci, which provide significant insight into per-locus mutation models and rates, as well as population substructure. These methods do not require knowledge of genealogical relationships between individuals, but are based on all-pairs comparisons between the individuals in a dataset (as employed by "mismatch distributions"). We formulate the mutation model curve (MMC) as the probability of matching at a specific locus, given that n out of N_L loci match in total between a pair of individuals; this function shows how quickly or slowly a locus mutates away from its original value, and the degree of back-mutation that occurs. We also present a method for quantification of population substructure in the data, and for calculation of per-locus mutation rates. The mutation model curve and the calculated mutation rate are given for each of 36 Y STR loci in a predominantly European dataset of 7976 individuals. The calculated mutation rates are shown to be comparable to rates obtained by observational studies. The mutation rates for many of the 36 loci in our study have not previously been determined by observation, yet mutation rates may be predicted for these loci by calibration to loci with known rates. These analysis techniques yield additional useful information about mutation characteristics of the loci, such as inferring conditions under which identical-by-state matches are expected, and by demonstrating evidence in support of a stepwise mutation model.

Pairwise Match Histograms

Pairwise mismatch histograms (or mismatch distributions) are often used to show information about population structure. It has been proposed that mismatch distributions yield information about historical episodes of population growth and decline [1]. Various mathematical models are used to parameterize mismatch distributions [2], in order to derive quantitative information about past population events. We present here a new parameterization that provides significant insight into population structure, and allows for quantification of several useful parameters of haploid populations and haploid genetic loci.

A pairwise match histogram is the same as a mismatch distribution, only reversed from left to right, so that the number of matches is shown on the horizontal axis rather than the number of mismatches. The pairwise match histograms H and H' were computed for a dataset ("Y7976") consisting of 7976 predominantly-European males genotyped at $N_L = 36$ Y STR loci, as well as for a control dataset of the same number of computer-generated genotypes with a random allele value (1 to 10) chosen from a uniform distribution for each of 36 loci. The control dataset approximately simulates a set of completely-unrelated individuals.

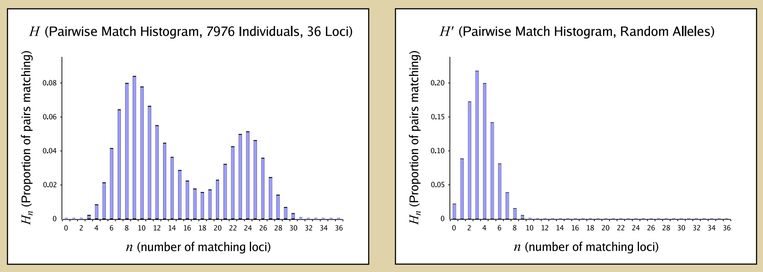

Figure 1: Pairwise match histograms H and H', for a dataset of 7976 male Y chromosome STR genotypes ("Y7976") and a control dataset of random genotypes respectively.

Bimodality Demonstrates Substructure

Because population substructure, by its nature, changes the proportion of haplotype pairs matching at a given number of loci, the shape of a pairwise match histogram reveals information about population substructure in the data. Interestingly, H is strongly bimodal in shape for the Y7976 dataset, indicating that a large proportion of pairs of haplotypes match at a majority of their loci (approximately 24 out of 36), and that another large proportion of pairs match at relatively few loci (approximately 9 out of 36). The strongly bimodal shape of H is surprising, because if the probabilities of match at each locus were independent (indicating the absence of substructure in the data), we would expect H to take on a simple binomial distribution as $H \sim \text{Binomial}(n, N_L, p)$, with p equal to the likelihood of match at a locus. The control dataset H' is strongly binomially distributed, with an RMS error (compared to a binomial distribution with $p = 0.1$) of 0.006%, well within the limits of sampling error. The expected value of the number of matches is very close to the theoretical value of $E(H') = p \cdot N_L = 3.6$.

We thus propose a new parameterization of pairwise match histograms as the linear combination of a set of component binomial distributions, each of which indicates a different level of substructure, and whose parameters are recovered using *Binomial Mixture Modeling*. The results for the Y7976 dataset are shown in Figure 2. The pairwise match histogram of the Y7976 data was fit with two binomial distributions, B and W. These two distributions correspond to "between-" and "within-population" relatedness levels, and correspond to pairs of samples that match at a smaller and larger number of total loci (n) respectively.

The distribution of pairs matching at higher n share a more recent common ancestor.

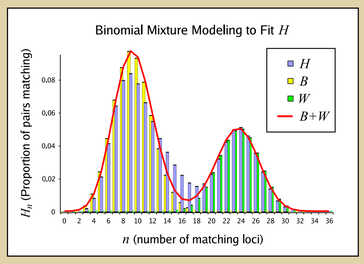

Figure 2: Binomial fitting of pairwise match histogram for Y7976 dataset.

Verification of Substructure Hypothesis

To test the hypothesis that different peaks in the pairwise match histogram correspond to different levels of relatedness, and to test the interpretation of B and W as representing between- and within-cluster pairs, a second, 20-locus Y-STR dataset was obtained from an independent source (ysearch.org [3]) , in which each of 2506 samples was also classified into one of 47 different YCC binary haplogroups (Figure 3).

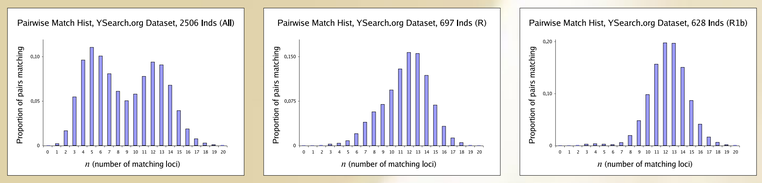

Figure 3: Pairwise match histograms for a control dataset of 2506 Y-STR genotypes from ysearch.org; the corresponding subset that fell in major clade R; and the further subset who were specifically members of haplogroup R1b.

The pairwise match histogram for the entire dataset of 2506 individuals was bimodal (Figure 3, left). However, the histogram of the subset of 697 individuals who fell in major clade R (Figure 3, center) was unimodal, and the smaller subset of 628 individuals who were specifically members of haplogroup R1b (Figure3, right) was binomially distributed. This validates the hypothesis that B indicates pairwise relationships with a level of substructure resulting from sharing a recent common ancestor, and that W indicates pairwise relationships resulting from a more distant common ancestor.

Mutation Model Curves (MMCs)

Y chromosome STR loci vary significantly in mutation rate [5]. The actual mutational characteristics of STR loci have only approximately been modeled (e.g. there exist the "Infinite Alleles" and "Stepwise Mutation" models). To better understand mutational characteristics of individual loci, we define the "mutation model curve" (MMC) for each locus as the likelihood that a pair of individuals match at locus i, given that they match at n loci in total (Figure 4), or $M_{i,n} = P(\text{match at } i \,|\, n \text{ matches})$.

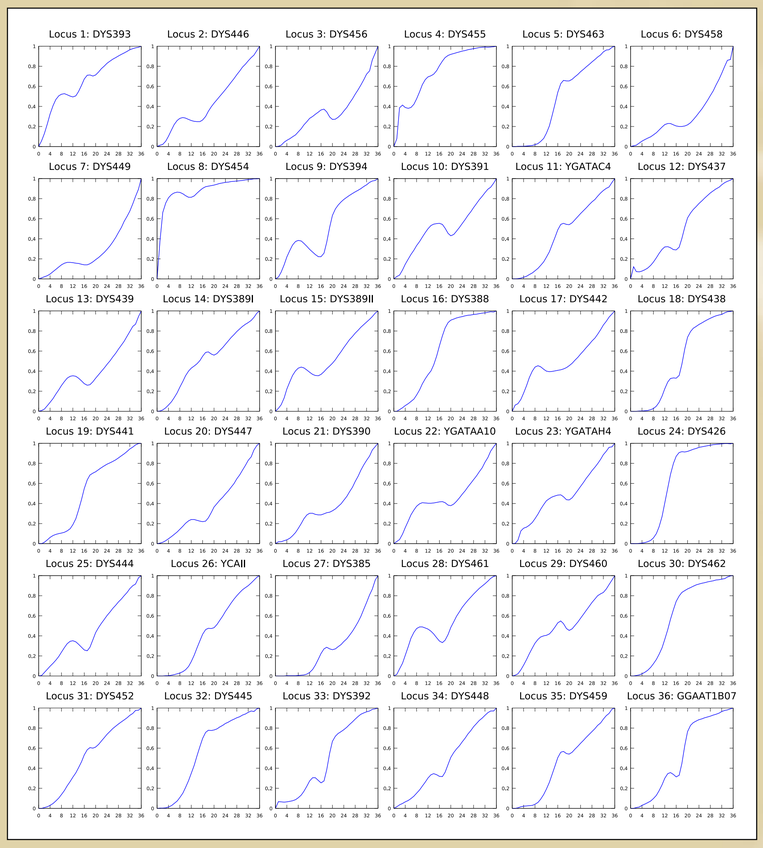

Figure 4: Mutation model curves (MMCs) for each of the 36 loci in the Y7976 dataset. The shape of the curve indicates the mutation rate, and the degree of recurrent mutation that occurs.

The shape of an MMC yields information about how quickly a locus mutates away from an initial ancestral allele (by the steepness of dropoff of the curve at high n), and about the degree of recurrent mutation (back mutation) that occurs (indicated by local maxima in the curve at smaller values of n). MMCs give the first real insight into the relative differences in mutational characteristics between STR loci.

Calculation of Per-Locus Mutation Rates

The actual rate at which a locus mutates away from a given ancestral allele may be calculated directly from the MMC. By considering only the region of the MMC that is not involved in recurrent mutations, we can eliminate "identical-by-state" (IBS) matches from consideration. Comparing H for the Y7976 dataset with the MMC for each locus, we can see that there is very little if any recurrent mutation (or no local maxima in the MMC) within the range of the predominant "within-cluster" population component in our dataset (i.e. where n >= 18 (Figure 2)). By weighting the MMC by the within-cluster distribution W (i.e. weighting by the likelihood a pair falls in the same population), and summing over n, we can obtain an estimate $q_i = P(\text{match at } i \,|\, \text{same pop})$ of the likelihood that two random individuals from the same population will match at the given locus:

$$q_i \approx \sum_{n=0}^{N_L} M_{i,n} \cdot W_n \,.$$

The probability of match at a locus after t generations, given the locus mutation rate μ, is $q(t) = (1-\mu)^{2t} \simeq \mathrm{e}^{-2\mu t}$. Rearranging this, we can calculate an approximate population depth for calibration purposes, t_c, using the mean $\overline{\mu}_K$ of a set of known mutation rates for a subset of the loci for which the mutation rate has been determined empirically (by observing mutations across generations). The approximate population depth can then be used to predict a mutation rate for loci that were not involved in the observational study.

$$t_c = \overline{t_i} = -\frac{1}{N_L} \sum_{i=1}^{N_L} \ln q_i / 2\overline{\mu}_K$$

$$\mu_i = -\ln(q_i)/2t_c$$

We used the rates from [5] (Kayser) for calibration purposes. Mutation rates calculated are given in Table 1, and Heyer's rates [6] are given for comparison. The rates for the ysearch.org dataset are also shown, calculated using the same method. The calculated mutation rates agree with the expected values for several loci with known extreme rates (e.g. DYS454 and DYS455), and in general are congruous with the shape of each MMC.

Significantly, this method allows us to take a few rates that are known by observation, and extrapolate to provide mutation rates for loci which have never before been studied by direct observation. This is immensely useful, as observational studies can be prohibitively expensive to perform.

Locus Number [i]	Locus Name	Calculated mutation rate [μ_i]. Y7976 dataset	Calculated mutation rate [μ_i]. ysearch.org dataset	Kayser mutation rate (95% CI) [5]	Heyer mutation rate (95% CI) [6]
1	DYS393	.001119	.001434	0 (–.00715)	0 (–1.31)
2	DYS446	.000143			
3	DYS456	.005342			
4	DYS455	.000312	.000460		
5	DYS463	.001623			
6	DYS458	.005804	.006298		
7	DYS449	.006464	.007839		
8	DYS454	.000235	.000466		
9	DYS394/19	.001548	.001453	.00201 (.00026–.00682)	.0044 (.0001–.0233)
10	DYS391	.003155	.004145	.00482 (.00061–.0164)	.0044 (.0001–.0233)
11	YGATAC4	.002355			
12	DYS437	.001755	.002147		
13	DYS439	.004181	.004947		
14	DYS389I	.002184	.002230	.00235 (.00006–0.0125)	
15	DYS389II	.002584	.002724	.00471 (.0006–.016)	
16	DYS388	.000379	.000245		
17	DYS442	.003066			
18	DYS438	.001018			
19	DYS441	.001323			
20	DYS447	.003891	.004004		
21	DYS390	.004405	.005320	.00858 (.00254–.0203)	0 (–1.31)
22	YGATAA10	.003794			
23	YGATAH4	.003044			
24	DYS426	.000271	.000255		
25	DYS444	.002956			
26	YCAII	.002496		0 (– ?)	
27	DYS385	.005592	.005678	.0021 (.00027–.00715)	.0044 (.0001–.0233)
28	DYS461	.002327			
29	DYS460	.002942			
30	DYS462	.000532			
31	DYS452	.001743			
32	DYS445	.000952			
33	DYS392	.001488	.001587	0 (–.00715)	.0044 (.0001–.0233)
34	DYS448	.002375	.002711		
35	DYS459	.002353	.002589		
36	GGAAT1B07	.000925			

Table 1: Mutation rates calculated from the MMCs of the Y7976 dataset, calibrated against Kayser's rates, and with Heyer's rates shown for comparison.

Conclusions

New methods have been presented for determining parameters of haploid populations, as well as for determining mutational properties of haploid STR loci. These methods do not require prior knowledge of relatedness between individuals in the dataset, and yet provide a means to calculate relative mutation rates for loci based on statistical parameters of individuals who share a recent common ancestor, or fall in the same haploid population. By calibration of a few loci in the dataset to the results of an observational study, mutation rates for other loci may be predicted (even in the absence of further observational data).

References

[1] Alan R. Rogers and Henry Harpending. *Population Growth Makes Waves in the Distribution of Pairwise Genetic Differences*. Molecular Biology and Evolution. 9(3):552–569, 1992.

[2] Setfan Schneider and Laurent Excoffier. *Estimation of Past Demographic Parameters From the Distribution of Pairwise Differences When the Mutation Rates Vary Among Sites: Application to Human Mitochondrial DNA*. Genetics. 152:1079–1089, 1999.

[3] *YSearch.org website*, May 2004. http://www.ysearch.org/ . FamilyTree DNA.

[4] YCC (Y Chromosome Consortium). *A nomenclature system for the tree of human Y-chromosomal binary haplogroups*. February 2002. Genome Research 12 (2), 339–348.

[5] Manfred Kayser, Lutz Roewer et al. *Characteristics and frequency of germline mutations at microsatellite loci from the human Y chromosome, as revealed by direct observation in father/son pairs*. American Journal of Human Genetics, 66:1580–1588, 2000.

[6] Evelyne Heyer, Jack Puymirat, Patrick Dieltjes, Egbert Bakker, and Peter de Knijff. *Estimating Y chromosome specific microsatellite mutation frequencies using deep rooting pedigrees*. Human Molecular Genetics, 6(5):799–803, 1997.

Presented at the American Society of Human Genetics 2004 Annual Meeting, October 2004, Toronto, Ontario, Canada