

Database characterization of Y-chromosomal 39-locus haplotypes

Natalie M. Myres¹,
 Jayne B. Ekins¹,
 Katie Hadley¹,
 Ugo A. Perego¹,
 Jacob E. Ekins¹,
 Luke A. D. Hutchison¹,
 Lara Layton¹,
 Mindy L. Lunt¹,
 Sacha S. Masek¹,
 Alison A. Nelson¹,
 Mary E. Nelson¹,
 Katie L. Pennington¹,
 Jenny L. Peterson¹,
 Amanda Sims¹,
 Trish Tolley¹,
 Alison Welch¹,
 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

Forensic analysis based on Y-chromosomal STR loci relies on haplotype frequency estimates for assessing the significance of matching haplotypes. Current frequency estimates, accepted by the forensics community, are based on the minimal 9-locus and extended 11-locus haplotypes. To augment current estimates, the present study reports frequencies for the minimal and extended haplotypes for a dataset of 7,977 male samples representing populations of the United States, Canada, Europe, Oceania, South America, and Asia. These haplotypes are further extended to a total of 39 loci, for which frequencies are reported. Paternal-line genealogy was collected for each male to provide a historical interpretation for observed haplotype stratification among populations, as measured by pairwise R_{ST} statistics.

Populations

DNA samples and pedigrees were collected with informed consent from 7,977 males of known paternal genealogy. Pedigrees were investigated using the Ancestral File database [1] to extend each subject's paternal line an average of four generations into the past. The birthplace and birth year of the most distant paternal ancestor for each subject was used to designate a geographic population assignment based on the geo-political boundaries as historically defined at the time of birth. Based on these criteria, the subjects were assigned to the following 29 world populations: Australia (58), Austria (20), Brazil (435), Canada (149), Chile (189), Denmark (213), England (998), Finland (33), France (29), Germany (453), Ireland (250), Italy (197), Japan (45), Mexico (53), Netherlands (55), Norway (101), Philippines (27), Poland (54), Portugal (42), Russia (62), Scotland (195), Spain (31), Sweden (178), Switzerland (54), Uruguay (38), USA (3836), Vanuatu (23), Wales (110), Yugoslavia (49).

	Australia	Austria	Brazil	Canada	Chile	Denmark	England	Finland	France	Germany	Ireland	Italy	Japan	Mexico	Netherlands	Norway	Philippines	Poland	Portugal	Russia	Scotland	Spain	Sweden	Switzerland	Uruguay	USA	Vanuatu	Wales	Yugoslavia	Total
No. of Individuals	59	20	435	149	189	213	998	33	29	463	250	197	46	53	55	101	27	54	42	62	195	31	178	54	38	3836	23	110	46	7977
No. of Haplotypes	20	20	298	107	130	138	401	21	25	323	141	165	36	50	43	72	24	47	37	57	119	27	110	46	30	1468	18	67	38	2722
Discriminatory Capacity	86.21%	100%	68.51%	71.81%	73.54%	64.70%	49.24%	63.64%	86.21%	71.3%	56.4%	83.76%	80%	94.34%	78.18%	71.29%	88.89%	87.04%	88.1%	91.94%	61.03%	87.1%	61.8%	85.19%	76.95%	38.01%	78.26%	60.91%	67.4%	34.12%
Haplotype Diversity	0.57	0.9	0.6102	0.5232	0.6304	0.577	0.338	0.5007	0.6306	0.6176	0.496	0.642	0.6428	0.6004	0.5904	0.5961	0.6011	0.5988	0.6172	0.6389	0.5106	0.561	0.5577	0.6007	0.6044	0.5663	0.623	0.4822	0.6104	0.57
No. of Unique Haplotypes	47	20	238	86	114	140	396	16	20	280	100	142	31	46	36	61	23	42	32	52	97	24	96	38	29	1468	14	50	38	1651
No. of Most Frequent Haplotype (h haplotypes > 5)	6	1	22	8	12	8(2)	60	6	3	16	14	6	4(2)	3	4(3)	10	2(2)	3(2)	2(5)	2(5)	14	3	11	2(6)	5	225	3	10	3(3)	466

Table 1. Haplotype frequency and diversity calculations for 29 world populations at 9-loci, 11-loci, and 39-loci

Y-STR Typing and Statistical Analysis

A complete 9-, 11-, and 36-locus Y-haplotype was generated for each sample within the Y7977 dataset from the following Y-STR loci: DYS437, DYS442, DYS444, DYS446, DYS452, DYS462, GGAATTB07, YGATAA10, YGATA-C4, DYS393, DYS394, DYS441, DYS449, DYS455, DYS458, DYS461, DYS463, DYS445, DYS454, DYS456, DYS459a/b, DYS385a/b, DYS388, DYS389I (DYS389A), DYS389II (DYS389B), DYS390, DYS391, DYS392, DYS426, DYS438, DYS439, DYS447, DYS448, DYS460, YCAIIa/b, YGATA-H4. Of the markers assayed, there are one di-, three tri-, 24 tetra-, six penta-, and one hexanucleotide repeat structures. Allele values from duplicated loci (DYS385a/b, DYS459a/b, and YCAIIa/b) were combined for analysis since alleles could not be unequivocally assigned to a defined locus [2]. Analysis of DYS389II (DYS389B) variability was considered without including DYS389I (DYS389A). Haplotype diversity and pairwise R_{ST} values were calculated using Arlequin [3] version 2.0. Haplotype frequencies were calculated using Microsoft Access.

Haplotype Variation

Haplotype frequency and diversity measurements (Table 1) are reported for a database of 7,977 Y-STR haplotypes complete at 9, 11, and 39-loci, representing 29 world populations. The locus values for the most frequently occurring haplotypes are shown in Table 2. When all samples are considered together, there are a total of 2,722 9-locus haplotypes with the most frequent haplotype (MFH) observed 426 times; 3,292 11-locus haplotypes with the MFH observed 337 times; and 7,265 39-locus haplotypes with the MFH observed only 9 times. The decline of haplotype frequency with increased numbers of loci is also observed in each individual population except Austria for which all haplotypes were observed only once, likely resulting from its small population sample size. Haplotype diversity for the total population steadily increases as larger numbers of loci are used, while this varies among the individual populations. In all cases, the discriminatory capacity of a haplotype increases with the number of loci, with dramatic increases seen in populations of larger sample size.

Population Structure

To evaluate the population substructure present in the database pairwise R_{ST} comparisons were calculated between individual populations (Table 3). For the 9, 11, and 39-locus haplotypes, the greatest differences were seen between Ireland and Mexico with R_{ST} values of 0.53538, 0.49974, and 0.47795 respectively. Most of the observed pairwise R_{ST} statistics tend to decrease with larger haplotype sizes. This is consistent with haplotypes that match at fewer loci sharing a most recent common ancestor (MRCA) earlier than haplotypes that match at more loci [4]. Therefore, as the number of loci defining a haplotype increases, less population substructure will be observed due to the high variation among many-locus haplotypes. R_{ST} values shown to increase with haplotype size could be the result of an elevated occurrence of recurrent mutations when considering many loci further emphasizing a need to include accurate locus-specific mutation rates in Y-haplotype analysis.

Population	9-loci	11-loci	39-loci
Australia	126	11-13	13
Austria	8	10-20	15
Brazil	22	5-23	15
Canada	8	5-23	15
Chile	12	8-25	15
Denmark	9	3-16	15
England	8	10-13	14
Finland	3	10-24	14
France	3	10-24	14
Germany	10	3-23	14
Ireland	11	5-6	14
Italy	3	8-25	14
Japan	11	11-13	14
Mexico	1	10-13	14
Netherlands	1	10-13	14
Norway	1	10-13	14
Philippines	2	11-13	14
Poland	2	11-13	14
Portugal	2	11-13	14
Russia	2	11-13	14
Scotland	2	11-13	14
Spain	11	6-21	14
Sweden	2	3-13	14
Switzerland	2	3-13	14
Uruguay	2	3-13	14
USA	2	3-13	14
Vanuatu	2	3-13	14
Wales	3	8-12	14
Yugoslavia	3	8-12	14

Table 2. Locus values of the most frequent haplotypes (MFH) observed for the total population and subpopulation samples at 9, 11, and 39 loci.

(a)	9-loci	11-loci	39-loci
Australia	126	11-13	13
Austria	8	10-20	15
Brazil	22	5-23	15
Canada	8	5-23	15
Chile	12	8-25	15
Denmark	9	3-16	15
England	8	10-13	14
Finland	3	10-24	14
France	3	10-24	14
Germany	10	3-23	14
Ireland	11	5-6	14
Italy	3	8-25	14
Japan	11	11-13	14
Mexico	1	10-13	14
Netherlands	1	10-13	14
Norway	1	10-13	14
Philippines	2	11-13	14
Poland	2	11-13	14
Portugal	2	11-13	14
Russia	2	11-13	14
Scotland	2	11-13	14
Spain	11	6-21	14
Sweden	2	3-13	14
Switzerland	2	3-13	14
Uruguay	2	3-13	14
USA	2	3-13	14
Vanuatu	2	3-13	14
Wales	3	8-12	14
Yugoslavia	3	8-12	14

Table 3. Pairwise R_{ST} values estimated for 29 world populations at 9-loci (a), 11-loci(b), and 39-loci(c). Red print indicates significance level of $p < 0.0500$.

Conclusion

Complete 9, 11, and 39-locus Y-STR haplotypes were compiled for 7,977 males representing 29 world populations. Pedigrees of known paternal genealogy were used to confidently establish a population designation specific to the Y-chromosome. Haplotype frequency estimates calculated for the total population and subpopulations show haplotype discriminatory capacity to rise sharply with increasing numbers of loci. This marked increase supports the use of larger numbers of Y-specific loci for forensic analysis where autosomal markers are difficult or impossible to obtain, such as in certain cases of rape or paternity. Pairwise R_{ST} comparisons showed significant population substructure for each of the three haplotypes. This finding supports the continued need to assess a Y-haplotype in terms of its population membership and locus-specific mutation rates, whether the haplotype is defined by few or many loci.

References

- http://www.familysearch.org/. The Church of Jesus Christ of Latter-day Saints, May 1999.
- P. Gill, C. Brenner et al. 2001. DNA Commission of the International Society of Forensic Genetics: recommendations on forensic analysis using Y-chromosome STRs. International Journal of Legal Medicine 114, 305-309.
- http://www.anthropologie.unige.ch/arlequin. Arlequin Ver. 2.000.
- B. Walsh, Estimating the Time to the Most Recent Common Ancestor for the Y chromosome or Mitochondrial DNA for a Pair of Individuals. Genetics 2001;158:897-912.