Allocation of YSTR Microvariant Alleles to Y-Chromosome Binary Haplogroups

Natalie M. Myres[1], Kathleen H. Ritchie[1], Robert Hughes[1], Alice A. Lin[2], Scott R. Woodward[1], Ugo A. Perego[1,3], Peter A. Underhill[2]

1) Sorenson Molecular Genealogy Foundation (SMGF), Salt Lake City, UT, USA. 2) Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California, USA. 3) Dip. di Genetica e Microbiologia, Università di Pavia, Pavia, Italy.

Abstract

Y-chromosome short tandem repeat (YSTR) loci are used extensively in studies of population substructure, temporality of population dynamics, and forensic identification. The occurrence of non-consensus YSTR alleles, such as unusually short alleles or partial insertion/deletion events (microvariants), have been used successfully as indicators of common ancestry among YSTR haplotypes, exposing further levels of phylogenetic substructure with restricted geographic distributions. However, the high variability of STR loci can potentially lead to false associations due to homoplasy (ie, recurrent mutation). Thus, YSTR haplotypes are best interpreted within the context of the binary marker defined Y-chromosome phylogeny. To identify YSTR microvariant alleles potentially useful for elucidating further phylogenetic substructure within binary haplogroups, we have assessed the haplogroup affiliation of microvariant alleles found at informative frequencies in the SMGF public YSTR database for the following YSTR loci: DYS385, DYS392, DYS441, DYS446, DYS447, DYS449 and DYS464. We report haplogroup affiliations for each variant allele and geographic origins of representative samples.

Introduction

The non-recombining portion of the Y-chromosome is particularly useful in modern human evolutionary studies due to its patrilineal inheritance and sensitivity to genetic drift. Two marker systems widely used in Y-chromosome analysis include single nucleotide polymorhisms (SNP) markers and short tandem repeat (YSTR) loci. Because of their low mutation rates, SNP markers have formed the basis for a detailed and well developed Y-chromosome phylogeny. In contrast, the exceptionally high mutation rates of YSTR markers make them useful for evaluating the diversity and temporality of SNP-defined Y-chromosome haplogroups.

Non-consensus YSTR mutations have been successfully used in previous studies to infer common ancestry among Y-chromosome haplotypes. Examples include large scale deletions (Cadenas 2007), unusually short alleles (Cinnioglu 2004), null alleles, and partial deletion/insertion (microvariant) alleles (Myres 2007). Such non-conforming alleles show decreased mutagenicity and hence can be informative for identifying potential phylogenetic substructure within the binary Y-chromosome gene tree. However, recurrent mutations can also lead to false associations when YSTR mutations are not evaluated in combination with binary markers. Herein we investigate an array of YSTR microvariants from a set of 37 YSTR loci to determine those which may be of common descent. Candidate alleles are evaluated for SNP haplogroup membership to determine which reveal potentially informative branches within the Y-chromosome tree.

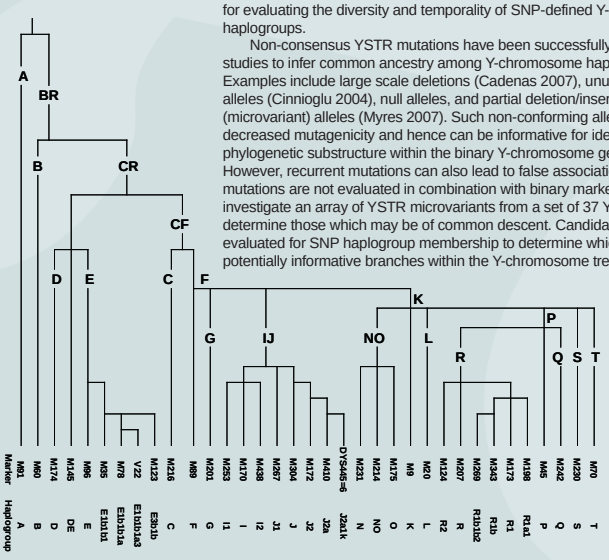

Figure 1:
Basic phylogenetic relationships and nomenclature for markers included in this study.

Methods

Samples were collected by the Sorenson Molecular Genealogy Foundation (SMGF) according to approved informed consent protocols. Over 30,000 DNA samples with geographic ancestry representing over 130 countries were genotyped at up to 37 YSTR loci. YSTR haplotypes were generated using custom designed multiplex panels of fluorescent labeled primers, capillary electrophoresis analyzers with internal size standard, and quantitative fragment analysis software.

YSTR haplotypes in the online SMGF database (www.smgf.org) were queried for partial repeat alleles (microvariants). Microvariant alleles found at appreciable frequencies in regions outside of North America were further assessed to determine binary haplogroup affiliation. A hierarchical genotyping approach was employed using Taqman assays for the following SNP markers: M9, M181, M35, V22, M201, M123, M258, M253, M172, M214, M343, M269, M198, and M242 (Underhill 2001, Cinnioglu 2004). The basic phylogenetic relationships of these markers are shown in Figure 1. Network analysis was performed using the median joining algorithm in Network (http://www.fluxus-engineering.com) software.

Results

The SMGF Y-chromosome database, containing 30,362 haplotypes of up to 37 YSTR loci, was screened for partial repeats to reveal a total of 2009 samples carrying microvariant alleles. Variants occur in 22 of the 37 YSTR loci examined and are present in loci composed of di, tri, tetra, penta-, and hexamer nucleotide repeat motifs. Of the 27 variant types observed, 19 demonstrate frequencies of at least .01% (Table 1). DYS449, a tetrameric repeat locus, shows the highest number of variant allele types (0.1, 0.2, and 0.3).

The 0.2 variant of locus DYS458 (DYS458.2) is the most frequent microvariant in the SMGF Y-database, occurring at 1.26% (Table 1). We previously reported that DYS458.2 evolved independently on at least two distinct binary haplogroup backgrounds, J1-M267 and R1b1b2a1-M405 (Myres 2007). In addition to DYS458.2, microvariants DYS385.2, DYS392.2, DYS441.1, DYS447.2, DYS449.2, DYS464.1, and DYS464.3 display high frequencies that remain at informative levels when considering populations from outside of North American (Table 2). Samples representing each of these variants were genotyped using binary SNP markers to determine their Y-chromosome haplogroup affiliations.

The haplogroup membership(s) of each microvariant is shown in Table 3. At the level of haplogroup resolution investigated in this study, most micorvariants partition into more than one haplogroup, usually 3 or 4, while only DYS392.2 is limited to a single lineage. The distribution

STR Locus	Repeat Motif	Sample Size	Variant Allele				
			0.1 (%)	0.2 (%)	0.3 (%)	0.4 (%)	0.5 (%)
DYS385*	tetramer	27330	0	0.07	0	0	0
DYS388	trimer	27357	0	0.01	0	0	0
DYS389I	tetramer	26670	0	0	0	0	0
DYS389B	tetramer	26123	0	0	0	0	0
DYS390	tetramer	27631	0	0	0	0	0
DYS391	tetramer	27615	0	0	0	0	0
DYS392	trimer	27136	0.01	0.03	0	0	0
DYS393	tetramer	27655	0	0	0	0	0
DYS19	tetramer	28102	0	0.01	0	0	0
DYS426	trimer	27522	0	0.01	0	0	0
DYS437	tetramer	27878	0.01	0	0	0	0
DYS438	pentamer	27574	0	0	0	0	0
DYS439	tetramer	27533	0	0	0	0	0
DYS441	tetramer	27101	0.04	0	0.01	0	0
DYS442	tetramer	28189	0.01	0	0	0	0
DYS444	tetramer	27890	0	0	0.01	0	0
DYS445	tetramer	27921	0.03	0	0	0	0
DYS446	pentamer	27974	0.04	0	0	0	0
DYS447	pentamer	25633	0	0.01	0	0.09	0
DYS448	hexamer	27295	0	0.09	0	0.02	0
DYS449	tetramer	27587	0.01	0.1	0.01	0	0
DYS452	pentamer	26909	0	0	0	0	0
DYS454	tetramer	27615	0	0	0.01	0	0
DYS455	tetramer	27602	0	0	0	0	0
DYS456	tetramer	27936	0	0	0	0	0
DYS458	tetramer	27664	0.01	1.26	0	0	0
DYS459*	tetramer	27838	0	0	0	0	0
DYS460	tetramer	27112	0	0	0	0	0
DYS461	tetramer	27459	0	0	0	0	0
DYS462	tetramer	28110	0	0	0	0	0
DYS463	pentamer	25617	0	0	0	0	0
DYS464*	tetramer	26789	0.03	0	0.3	0	0
GGAAT1B07	pentamer	28104	0	0	0	0	0
YCAII*	dimer	27581	0.01	0	0	0	0
YGATAA10	tetramer	28194	0	0	0	0	0
YGATAC4	tetramer	28275	0	0	0.04	0	0
YGATAH4	tetramer	27663	0	0	0	0	0

Table 1:
YSTR microvariant frequency by locus.
* Multi-copy locus frequencies normalized according to copy number.

Results cont.

of most microvariant types across multiple Y-haplogroups indicates that parallel partial repeat mutations arose independently on different haplogroup backgrounds.

DYS385.2 allocates into the four haplogroups D-M174, R1b-M343*(xM269), R1b1b2-M269, and Q-M242, indicating this variant likely has four separate origins and may represent further phylogenetic substructure within each haplogroup. DYS385.2 samples which are members of haplogroup D descend from Cameroon (3), Ivory Coast (1), Philippines (1), and Hawaii (1); those of haplogroup Q-M242 originate from Brazil (1), Mexico (1), and Peru (1); and those of haplogroup R are of predominantly European populations.

Both DYS447.2 and DYS449.2 are present in haplogroups R1a1-M198, R1b1b2-M269, and non-R lineages of haplogroup K. Additionally, the variant DYS447.2 may comprise an informative sublineage of haplogroup E1b1b1c -M123. The two samples included in this group descend from Kuwait and Palestine, indicating the M123 associated DY447.2 lineage may localize within Middle Eastern populations.

DYS392.2 appears to delineate new phylogenetic substructure within haplogroup I1*-M253. All seven samples carrying this variant have the derived allele for M253. Network analysis (Figure 2) of the DYS392.2 samples with ninety additional I1-M253 chromosomes shows six of the variants clustering tightly together, indicating they are likely of common descent. The seventh variant is located 2 steps from other variants, and all variants are located within two steps of the most frequent I1-M253 haplotype. Overall, the network suggests a common phylogenetic connection among the DYS392.2 variants. Additionally, one 0.2 variant sample was sequenced at markers M21, M72, and M227, which define the known subclades of I1-M253, and it was found to carry the ancestral allele of all of these subclade markers. Given the low frequencies generally displayed by I1-subhaplogroups, it is unlikely the remaining DYS392.2 variants would apportion to any of them, although further testing is needed to be certain. Consistent with previously reported I1-M253 distributions (Underhill 2007), DYS392.2 samples descend from Germany (4), Denmark (2), and Wales (1).

Population	Sample Size	DYS385.2	DYS392.2	DYS441.1	DYS447.2	DYS449.2	DYS464.1	DYS464.3
Africa								
Cameroon	1169	0.28	0	0	0	0.93	0	0
Ghana	106	0	0	0	0	0	0	0.06
Mali	450	0	0	0	0	0	0	0.18
Nigeria	209	0	0	0	0	0.47	0	0
Togo	259	0	0	0	0	0	0	0
Asia								
China	374	0	0	0.25	0	0	0.02	0.09
India	100	0	0	0	0	0.98	0	0
Japan	105	0	0	0	0	0.92	0	0.12
Kyrgyzstan	205	0	0	0.47	0	0	0	0.15
Mongolia	434	0	0	0	0	0	0	2.2
Central and South America								
Bolivia	77	0.63	0	0	0	0	0	0.08
Brazil	605	0.12	0	0	0	0.33	0.04	0.16
Chile	258	0.19	0	0.39	0	0	0.08	0.15
Colombia	65	0	0	0	0	0	0	0.19
Mexico	972	0.13	0	0.3	0	0	0.07	0.19
Peru	623	0.08	0	0	0	0	0.04	0.13
Uruguay	51	0.49	0	0	0	0	0.13	0.26
Europe								
Austria	70	0	0	0	0	0	0.29	0.38
Denmark	475	0.05	0.41	0	0	0	0	0.36
England	2707	0.07	0	0	0.08	0	0.04	0.23
Finland	134	0	0	0	0	0	0	6.63
France	134	0	0	0	0	0	0.05	0.35
Germany	1318	0	0.3	0	0	0.15	0.04	0.38
Czech Republic	75	0	0	0	0	0	0	1.06
Ireland	874	0.06	0	0	0	0	0	0.06
Hungary	207	0	0	0	0	0	0.03	0.59
Netherlands	143	0	0	0	0	0	0.09	0.23
Norway	220	0	0	0	0	0	0	0.36
Poland	208	0	0	0	0	0	0.06	0.55
Portugal	79	0	0	0	0	0	0.32	0
Russia (Western)	306	0	0	0	0	0	0.14	0.96
Scotland	552	0.09	0	0.18	0	0	0	0.21
Slovenia	99	0	0	0	0	0	0.19	0.25
Spain	103	0	0	0	0	0	0	0.18
Sweden	407	0	0	0	0	0.24	0.02	0.58
Switzerland	151	0	0	0	0	0	0.31	0.27
Wales	219	0.11	0	0	0	0.45	0	0.2
Middle East								
Jordan	120	0	0	0	0	0	0	0.52
Pakistan	144	0	0	0	0	0	0	0
Palestine	63	0	0	0	0	0	0	0.91
Oceania								
Australia	66	0	0	0	0	0	0	0.19

Table 2:
YSTR microvariant frequency by country. Values are in %.
* Multi-copy locus frequencies normalized according to copy number.

DYS464 is a multi-copy locus usually consisting of four allelic copies. Its microvariants DYS464.1 and DYS464.3 occur at substantial overall frequencies of .03% and 0.3% respectively. With the exception of one Guatemalan sample which belongs to haplogroup Q-M242, all DYS464.1 variant chromosomes are members of haplogroup J2-M172. These chromosomes also carry a value of 6 at locus DYS445 which further partitions this variant into haplogroup J2a1k. Since other J2a1k members within the SMGF dataset do not display the 0.1 variant, DYS464.1 exposes further subdivision within the J2a1k clade. DYS464.1 Samples are of predominantly European origin.

The DYS464.3 variant was the second most frequent partial repeat allele observed in our dataset (Table 1). A subset of 265 DYS464.3 chromosomes, representing various geographic locations, was genotyped at the haplogroup K-defining marker M9, of which 160 allocated into the ancestral state and 105 into the derived state. Those derived for K-M9 were not studied further. The remaining samples divide into haplogroups E1b1b1-M35(xV22) and E1b1b1a3-V22. Additional SNP testing is needed to determine whether DYS464.3 variants within M35(xV22) are of common descent or whether they represent various independent lineages within M35(xV22) subclades. Conversely, DYS464.3 variants belonging to haplogroup V22 are likely an informative subset of lineages that define a distinct phylogenetic branch within the V22 clade. Representative samples descend mainly from Europe and the Middle East.

Lastly, locus DYS441 displayed 0.1, 0.2, and 0.3 variants, each appearing rarely or as singletons across several haplogroup backgrounds including those of haplogroups G, I, J, K, and R, suggesting multiple independent origins. The high number of variant types and their sparse yet broad spread across the binary Y-chromosome phylogeny may indicate a universally high mutation rate at locus DYS441 that is relatively independent of the unique demographic histories that have differentially influenced individual haplogroup evolutions.

Haplogroup	Marker(s)	DYS385.2	DYS392.2	DYS447.2	DYS449.2	DYS464.1	DYS464.3
D	M145(xM35)	6					
E1b1b1	M35(xV22)						136
E1b1b1a3	V22						24
E1b1b1c	M123			2			
I1	M253		7				
J2a1k	M172,DYS445=6					52	
K	K				1		105
Q	M242	3				1	
NO	M214			1	1		
R1a1	M198			1	2		
R1b*	M343(xM269)	1					
R1b1b2	M269	13		4	5		
A,BR(xDExK)	M9C(xM145)				8		

Table 3:
Haplogroup membership of micorvariant alleles.

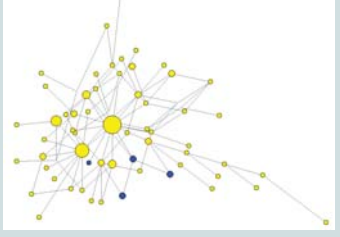

Figure 2:
Median joining network of haplogroup I1-M253 haplotypes. The algorithm was applied to the 10-locus STR haplotypes of 97 haplogroup I1-M253 chromosomes, of which seven haplotypes carry the DYS392.2 microvariant and are indicated in blue. Nodes are proportional to haplotype frequency with the smallest node representing one haplotype.

Conclusion

Partial insertion/deletion polymorphisms within YSTR loci, while relatively rare, typically occur on multiple SNP-defined haplogroup backgrounds. This distribution indicates that micorvariant alleles arise at a rate intermediate to those of binary markers and STR loci. As a result, microvariants can reveal phylogenetic substructure within the Y-chromosome binary tree that is not currently detectable by either Y-SNP or YSTR testing alone, but only when such variants are evaluated within a haplogroup context.

References

Cadenas AM, Regueiro M, Gayden T, et al. (2007) Male amelogenin dropouts: phylogenetic context, origin and implications. Forensic Sci. Int. 166: 155-163.

Cinnioglu C, King R, Kivisild T, et al. (2004) Excavating Y-chromosome haplotype strata in Anatolia. Mol. Biol. Evol. 114: 127-148

Myres NM, Jayne EE, Lin AA, et al. (2007) Y-chromosome short tandem repeat DYS458.2 non-consensus alleles occur independently in both binary haplogroups J1-M267 and R1b3-M405.

Underhill PA, Passarino G, Lin AA, et al. (2001) The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann. Hum. Genet. 65: 43-62

Underhill PA, Myres NM, Rootsi S, et al. New phylogenetic relationships for Y-chromosome haplogroup I: reappraising its phylogeography and prehistory; in Mellars P, Boyle K, Bar-Yosef O, Stringer C (eds): Rethinking the human revolution. Cambridge, UK, 2007, McDonald Institute Monographs, pp 33-42.

SORENSON MOLECULAR GENEALOGY FOUNDATION

www.smgf.org