Y-chromosome Short Tandem Repeat DYS458.2 Non-consensus Alleles Occur Independently in Both Binary Haplogroups J1-M267 and R1b3-M405

Natalie M. Myres[1], Jayne E. Ekins[1], Alice A. Lin[2], L. Luca Cavalli-Sforza[2], Scott R. Woodward[1], Peter A. Underhill[2]

[1]Sorenson Molecular Genealogy Foundation, Salt Lake City, Utah, USA
[2]Department of Genetics, Stanford University School of Medicine, Stanford, Calif, USA

The first two authors contributed equally to this work.

Aim To determine the human Y-chromosome haplogroup backgrounds of non-consensus DYS458.2 short tandem repeat alleles and evaluate their phylogenetic substructure and frequency in representative samples from the Middle East, Europe, and Pakistan.

Methods Molecular characterization of lineages was achieved using a combination of Y-chromosome haplogroup defining binary polymorphisms and up to 37 short tandem repeat loci, including DYS388 to construct haplotypes. DNA sequencing of the DYS458 locus and median-joining network analyses were used to evaluate Y-chromosome lineages displaying the DYS458.2 motif.

Results We showed that the DYS458.2 allelic innovation arose independently on at least two distinctive binary haplogroup backgrounds and possibly a third as well. The partial allele length pattern was fixed in all haplogroup J1 chromosomes examined, including its known rare sub-haplogroups. Within the alternative R1b3 associated M405 defined sub-haplogroup, both DYS458.0 and DYS458.2 allele classes occurred. A single chromosome also allocated to the R1b3-M269*(xM405) classification. The physical position of the partial insertion/deletion occurrence within the normal tetramer tract differed distinctly in each haplogroup context.

Conclusions While unusual DYS458.2 alleles are informative, additional information for other linked polymorphic loci is required when using such non-conforming alleles to infer haplogroup background and common ancestry.

> **Correspondence to:**
Peter A. Underhill
Department of Genetics
Stanford University School of Medicine
300 Pasteur Drive
Alway Building, Room M302
Stanford, CA 94305-5120, USA
under@stanford.edu

> **Received:** June 9, 2007
> **Accepted:** July 2, 2007

> **Croat Med J. 2007;48:450-9**

Ever since Y-chromosome specific polymorphic short tandem repeat (Y-STR) or microsatellite loci were first developed as practical polymerase chain reaction (PCR)-based amplification assays in the early 1990s (1), these loci have become important reagents in the molecular genetic analysis of the non-recombining portion of this paternally transmitted haploid chromosome. Currently, several hundred potentially polymorphic short tandem repeat loci have been recognized on the human Y-chromosome (2). This abundance, combined with their high variability, makes them useful for studies of population substructure (3), temporality of population dynamics (4), forensic identification (5), and genealogical investigations (6), although homoplasy (ie, recurrent mutation) phenomena may alter true genetic distance, thereby creating potentially false evidence of actual affinity (7,8). Their development and extensive use have been facilitated, in part, by the availability of well-calibrated, standardized commercial multiplex PCR kits, compatible with capillary electrophoresis platforms and quantitative fragment sizing algorithms. These validated kits, that contain anywhere from 12-15 different Y-STR loci, are becoming more widely used throughout the genetic analysis community and will generate a substantial amount of population data for the particular loci involved. One such locus contained in the AmpFLSTR Y filer PCR amplification kit (Applied Biosystems, Foster City, CA, USA) is DYS458 (9) that is composed of a polymorphic tetra (GAAA) nucleotide repeat motif. Interestingly, null alleles at this locus have been associated with amelogenin allele negative men, caused by large scale (>1 Mb) deletions (10) that often correspond to specific Y-chromosome binary haplogroup affiliation (11) suggestive of common ancestry. Additionally, unusual partial DYS458.2 insertion/deletion alleles have been reported in Tunisian Berbers (12).

The ability of Y-chromosome STR diversification to help our understanding of population substructure and group membership is fundamentally linked to our knowledge regarding the molecular resolution of the haploid binary marker defined phylogeny. During the past 10 years, significant progress in reconstructing the detailed branching order of the gene tree topology for the non-recombining portion of the Y-chromosome (13,14) has paralleled the developing understanding and application of Y-STR loci. This binary haplogroup defined gene tree is the scaffold upon which all Y-STR data are partitioned (15). As suggested by de Knijff (16), Y-chromosomes identified by STRs are designated haplotypes. Y-chromosomes that are defined only by biallelic markers are called haplogroups or clades, and the combination of biallelic markers and Y-STRs are called lineages.

The occasional occurrence of an unusually short Y-STR allele that has lost or dramatically reduced its mutagenicity (and hence has properties approaching those of an evolutionary stable binary marker) has lead to reliable clues of the authentic affinity of Y-STR haplotypes. Some examples of such uncommon "short" alleles include DYS390 in Australian haplogroup C chromosomes (17) and DYS388 in a subset of J1-M267 Turkish chromosomes (7). Reduced allelic variability at microsatellite loci also results from nucleotide substitutions within the usual repetitive elements (18). In a similar manner, non-consensus partial insertion/deletion events within the repeat motif of loci like DYS458 have the potential to provide clues to common Y-chromosome lineal ancestries within STR haplotype-based data sets. While Y-STR mutation rates are exceptionally high relative to binary mutations, network analysis (19) provides a useful technique to help disentangle complex multilocus haplotype data and potentially identify chromosomes with distinctively different evolutionary histories.

During the evaluation of potential patterns in Y-STR haplotypes characterized by up to 37 loci in data from 17 646 samples generated by

the Sorenson Molecular Genealogy Foundation (SMGF), a subset of chromosomes characterized by the presence of atypical DYS458.2 alleles was selected for further scrutiny. This allelic designation style follows recommended guidelines (20) in which the 0.2 label denotes the presence of allele repeats of intermediate size, in addition to variation in numbers of typical repeats. Such partial sized alleles arise by insertion/deletion events most probably caused by slipped-strand mis-pairing events within the locus during spermatogenesis (21).

The trinucleotide locus DYS388 that was also included in the SMGF data set played an important role in framing the course of this investigation. This locus is known to deviate from the stepwise mutational process when the allele frequency spectrum is analyzed (22). Subsequently, the larger DYS388 allele size category in the bimodal distribution was shown to be affiliated with samples with geographic ancestry in the Middle East that displayed the derived allele for a binary marker (12f2.1) used to define haplogroup J (23) within the standardized nomenclature (24). While DYS388 short allele representatives are known to occur within a subset of J1-M267 derived chromosomes (7), the majority of haplogroup J1 representatives have DYS388 allele sizes ≥15 repeats (7,25,26). This article reports on the results of our explorations into haplogroup affiliations within DYS458.2 designated chromosomes from various European, Middle Eastern, and Pakistani populations.

Although ambiguities caused by Y-STR homoplasmy can be mitigated by typing large numbers of such loci, especially in genealogical situations, this article illustrates how the intersection of binary haplogroups and Y-STR haplotypes in combination more clearly reveals authentic lineal relationships and underscores the vulnerability of using Y-STR data alone to infer common ancestry, even when unusual allelic variants are involved.

Materials and methods

Samples were collected according to approved informed consent protocols. The DNA samples studied were predominantly those from the inventory of the Sorenson Molecular Genealogy Foundation (SMGF), and were extracted from saliva or blood. These included approximately 17 646 samples with geographic ancestry representing more than 100 countries. Their Y-STR haplotypes were determined at up to 37 loci by means of custom designed amplification panels of multiplexed loci using fluorescent labeled primers, capillary electrophoresis analyzers with internal size standards, and quantitative fragment analysis software. Conversion of absolute fragment size to a number of allele repeats was achieved using the results obtained from sequencing both strands of control samples independently amplified with unlabeled primers. DNA sequencing of the DYS458 locus was conducted in selected samples to determine precise motif structure and circumstance within the amplified fragment.

SMGF samples were initially collected within the context of genealogical studies and, therefore, contained some closely related individuals. To minimize the bias caused by related samples, genealogical records were used to identify a subset of 1965 unrelated SMGF samples for subsequent population, haplogroup, and network analyses. This data set consists of samples with geographic ancestry, primarily from European and Middle Eastern populations. In addition, 53 previously characterized Turkish and Pakistani DNA samples belonging to haplogroup J1-M267 and its rare sub-haplogroups (7,13,27) were analyzed at the DYS458 locus. Further, 76 Turkish and 6 Pakistani M269 derived chromosomes were genotyped for the M405, M467, and DYS458 polymorphisms. Median-joining network analysis was conducted using SMGF sample data from 21 loci-defined Y-STR haplotypes for DYS458.2 category-positive individuals. Allelic frequency

distributions for DYS388 and DYS458 were obtained by the simple direct counting method.

Networks were processed first by the median-joining method and then by the maximum parsimony Steiner method (19,28). Networks were generated without weighting any of the STR loci. Networks were constructed using the Network 4.1.1.1 application (*http://www.fluxus-engineering.com*).

Single nucleotide T to G substitution marker M267 (7) was genotyped either by DHPLC (29) or direct sequencing of the PCR amplified fragment. Marker M405 was genotyped by using either a double-labeled fluorogenic probe 5′-3′ nucleolytic single tube reaction PCR-based assay (ie, Taqman) (30) or NcoI RFLP analysis that identifies a C to T transition at position 147 within the 295 bp PCR fragment amplified using the following 5′-3′ primers: forward, CCTCCACTTACCGACCCGCA and reverse, GGAAATGGGTGGCAGATGCA. Marker M405 was determined independently but is identical to the U106 marker shown by phylogenetic analysis (31) to form a sub-clade of haplogroup R1b3-M269 that is common in western Europe (32,33) and Turkey (7). Like M405, nucleotide transition M467 was developed independently but is equal to the U198 marker (31) shown by phylogenetic analysis to form a sub-clade within sub-haplogroup R1b3-M405. Marker M467 was amplified as a 263 bp fragment using 5′-3′ primers: forward, CAAGATAATATTTACCTGCACTCC and reverse, ATCTAAATAATAACTCTCTGTTTGG. M467 was genotyped by Taqman, direct sequencing or by Bsr I RFLP analysis that interrogates a G to A transition at position 148.

Results

While inspecting haplotype data constructed from 37 loci in 1965 samples from the SMGF collection, a total of 202 chromosomes with non-consensus DYS458.2 allele sizes were detected. The allele size frequency distribution normalized to total sample counts for both DYS458.0 and DYS458.2 alleles indicated that DYS458.2 chromosomes were skewed to slightly larger allele sizes (Figure 1).

Median-joining network analysis of 192 DYS458.2 chromosomes, using 21 additional Y-STR loci, revealed a distinctive bimodal constellation (Figure 2), suggesting different evolutionary trajectories for each cluster. A list of the haplotypes used in the network analysis is given as web-extra material.

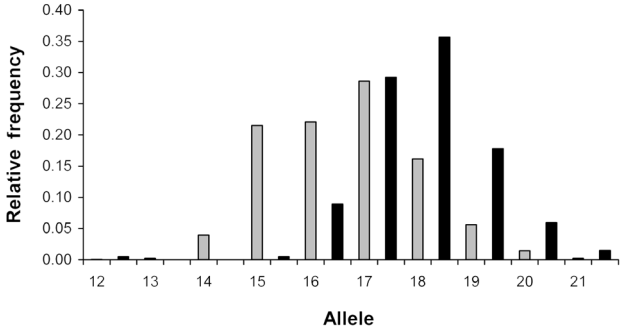

Figure 1. DYS458 allele frequency distribution normalized for total sample counts.

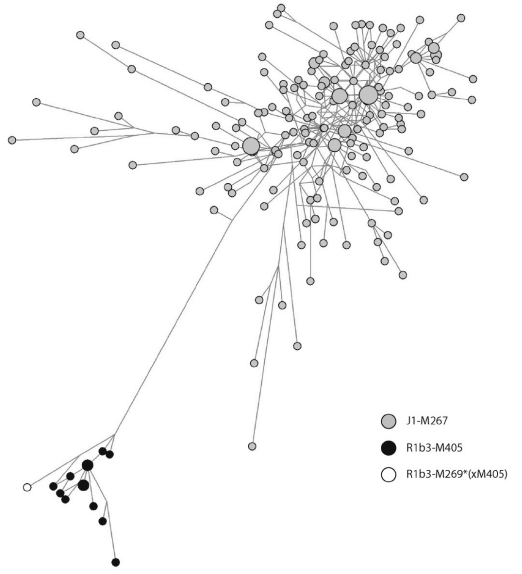

Figure 2. Median-joining network analysis of 192 chromosomes with non-consensus DYS458.2 alleles using loci: DYS388, 389I, 389B, 390, 391, 392, 393, 394, 437, 439, 445, 448, 454, 456, 460, 461, 462, GGAAT1B07, YGATA-A10, YGATA-C4, YGATA-H4.1 (all equally weighted).

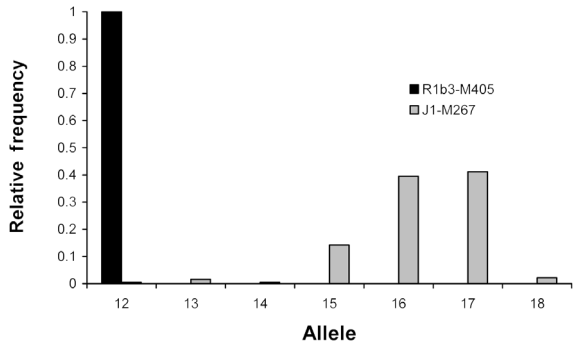

Figure 3. DYS388 allele frequency distribution in DYS458.2 demarcated chromosomes normalized for total sample counts.

The allele frequency distribution for associated DYS388 locus data (Figure 3) for the 196 DYS458.2 denoted samples indicated that they contained both short (<15 repeats) and long (≥15 repeats). Since DYS388 alleles sizes ≥15 repeats are known to be common in haplogroup J chromosomes, the haplogroup J defining M304 SNP (7) was genotyped by direct sequencing in representative DYS458.2 chromosomes. The result was that all the ≥13 repeat allele sized chromosomes that were analyzed at M304 distributed to haplogroup J, while all those that were determined to not belong to haplogroup J were DYS388-12 repeats with the exception of a single haplogroup J sample that had 12 repeats at DYS388. The modal nature of the major network cluster suggested that this aggregate were likely to all be haplogroup J representatives. All these chromosomes were subsequently determined to carry the derived allele for the haplogroup J1, defining M267 nucleotide substitution.

As the minor network cluster of non-J chromosomes in the data set were of western European ancestry, they were subsequently genotyped first at the M9 transversion (29) substitution and then the M269 transition. This cluster of DYS458.2 chromosomes apportioned to haplogroup R1b3-M269, demonstrating that the non-conforming length variant has at least two independent origins. The binary haplogroup resolution of the M269 derived chromosomes was subsequently extended to reveal that all but one member of this cluster of DYS458.2 variants belong to the R1b3-M269-related M405(xM467) derived sub-clade. The single exception assigned to haplogroup R1b3-M269*(xM405) variety.

Next, DYS458 amplicons from representative haplogroup J1 and R1b3 chromosomes were directly sequenced on both strands' locus to determine the precise nature of the two nucleotide base length anomaly responsible for causing the abnormal DYS458.2 allelic signal. The sequence context of each of the sequences is given in Table 1. The two nucleotide length differences were determined to be an AA couplet in both J1 and R1b3 contexts. Since this sequence variant occurred within the characteristic GAAA motif, we cannot distinguish if there was a GA deletion or an AA insertion. While both haplogroup classes have the same AA sequence variant, the relative polarity of where it occurs within the tract of GAAA repeats differs. In the haplogroup R1b3-M405* situation, the non-conforming AA motif occurs near the beginning of the GAAA tract, while in haplogroup J1 chromosomes it occurs near the end of the tetra repeat polymer (Table 1). Within the single R1b3-M269*(xM405) sample, the physical position of the occurrence of the non-conforming AA feature within the GAAA tract is unique from the others (Table 1 and Table 2). The phylogenetic position of DYS458.2 variants within the basic known cladistic frameworks of haplogroups J1 and R1b3 are shown in Figures 4A and 4B respectively.

Although additional binary phylogenetic resolution (24) exists that is not shown within haplogroup R1b3-M269 (Figure 4B), our results indicate that such sub-haplogroups do not involve the presence of DYS458.2 differentiated alleles that remain restricted to a portion of the M405 diversity. In addition, binary markers M37 and SRY2627 (24) displayed ancestral alleles in the sole R1b3-M269*(xM405) chromosome. Table 3 presents the allele frequency distributions by

Table 1. Sequence characterization of DYS458.2 insertion/deletion in haplogroup backgrounds R1b3-M405, R1b3-M269*(xM405), and J1-M267

Sample	Country	Haplogroup	Allele	Sequence*	Polarity	Rpt#
control			15	AGCAACAGGAATGAAACTCCAAT... $[GAAA]_{15}$ GGAGGGTGGGCGTGGTGG		—
1	United States	R1b3-M405	15.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{1}$ GAAA**AA*** $[GAAA]_{13}$ GGAGGGTGGGCGTGGTGG	3′	2
2	England	R1b3-M405	16.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{1}$ GAAA**AA** $[GAAA]_{14}$ GGAGGGTGGGCGTGGTGG	3′	2
3	England	R1b3-M405	16.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{1}$ GAAA**AA** $[GAAA]_{14}$ GGAGGGTGGGCGTGGTGG	3′	2
4	United States	R1b3-M405	16.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{1}$ GAAA**AA** $[GAAA]_{14}$ GGAGGGTGGGCGTGGTGG	3′	2
5	United States	R1b3-M405	16.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{1}$ GAAA**AA** $[GAAA]_{14}$ GGAGGGTGGGCGTGGTGG	3′	2
6	Ireland	R1b3-M405	17.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{1}$ GAAA**AA** $[GAAA]_{15}$ GGAGGGTGGGCGTGGTGG	3′	2
7	Ireland	R1b3-M269* (xM405)	16.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{4}$ GAAA**AA** $[GAAA]_{11}$ GGAGGGTGGGCGTGGTGG	3′	5
8	Chile	J1-M267	16.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{13}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
9	England	J1-M267	17.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{14}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
10	Israel	J1-M267	17.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{14}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
11	Bolivia	J1-M267	17.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{14}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
12	Reunion Island	J1-M267	18.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{15}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
13	Russia	J1-M267	18.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{15}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
14	Italy	J1-M267	18.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{15}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
15	Oman	J1-M267	18.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{15}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
16	Oman	J1-M267	18.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{15}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
17	United States	J1-M267	19.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{16}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3
18	England	J1-M267	20.2	AGCAACAGGAATGAAACTCCAAT...$[GAAA]_{17}$ GAAA**AA** $[GAAA]_{2}$ GGAGGGTGGGCGTGGTGG	5′	3

*__AA__ insertion/deletion designated in bold within tandem repeat motif, flanked by primer sequences.

Table 2. Full fragment sequence of DYS458.2 insertion/deletion in Haplogroup backgrounds R1b3-M405, R1b3-M269 (xM405) and J1-M267*

Haplogroup	Sequence
R1b3-M405(xM467)	AGCAACAGGAATGAAACTCCAATGAAAGAAAGAAAAGGAAG/GAAA/GAAA**AA**/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GGAGGGTGGGCGTGGTG
R1b3-M269*(xM405)	AGCAACAGGAATGAAACTCCAATGAAAGAAAGAAAAGGAAG/GAAA/GAAA/GAAA/GAAA/GAAA**AA**/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GGAGGGTGGGCGTGGTG
J1-M267	AGCAACAGGAATGAAACTCCAATGAAAGAAAGAAAAGGAAG/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA/GAAA**AA**/GAAA/GAAA/GGAGGGTGGGCGTGGTG

*__AA__ insertion/deletion designated in bold within tandem repeat motif, flanked by underlined primer sequences.

population of DYS458 in haplogroup J1. Table 4 shows the frequency of M405, M467, and associated DYS458 haplogroup varieties in the populations surveyed for this study.

Discussion

Nucleotide point mutations located nearby but physically outside the actual repeating elements are known to occur on the human Y-chromosome (20). Also, the occasional occurrence of unusually short Y-STR alleles that have lost or dramatically reduced mutagenicity (and hence act like a proxy binary marker) has lead to reliable clues of authentic affinity of Y-STR haplotypes. However, currently little information has been published regarding non-conforming STR

Table 3. Frequencies of haplogroup J1 and DYS458.2 variants by population*

Population	Sample size	J1-M267	DYS458.2
Austria	23	0	0
Central and South America†	33	66.7	66.7
Czech Republic	36	0	0
Denmark	116	0.9	0.9
Eastern Europe‡	44	4.5	4.5
England	139	5	5
France	57	3.5	3.5
Germany	333	1.2	1.2
Ireland	105	1.9	1.9
Italy	285	4.2	4.2
Jordan	76	52.6	52.6
Middle East§	43	69.8	69.8
Netherlands	94	0	0
Oceania‖	43	0	0
Oman	29	100	100
Pakistan¶	177	3.4	3.4
Palestine	47	29.8	29.8
Poland	110	0	0
Russia	57	3.5	3.5
Slovenia	105	0	0
Switzerland	91	0	0
Turkey**	523	9	9
Ukraine	32	3.1	3.1
United States	58	25.9	25.9

*Populations with fewer than 20 representatives are not included in frequency calculations.
†Central and South America include Bolivia, Brazil, Chile, Mexico, Peru, and Uruguay.
‡Eastern Europe includes Belarus, Croatia, Greece, Hungary, Lithuania, Romania, Serbia, and Slovakia.
§Middle East includes Egypt, Gaza, India, Iraq, Israel, Kuwait, Lebanon, Saudi Arabia, Syria, and West Bank.
‖Oceania includes American Samoa, Australia, Hawaii, New Zealand, Samoa, Tahiti, Tonga, Tuamotu, and Vanuatu.
¶Samples from Sengupta et al (27).
**Samples from Cinnioglu et al (7).

allele classes for the Y-chromosome, especially with regards to binary haplogroup affiliations. Inspection of known alleles for polymorphic Y-chromosome STR loci indicates that such non-consensus allelic variants occur in numerous loci (*http://www.smgf.org*). While uncertainty about common descent caused by parallel mutation at Y-STR loci can be reduced by typing multiple Y-STR loci (34,35), our results for DYS458.2 illustrate how the combination of binary haplogroups and Y-STR haplotypes used together with non-consensus alleles adds further sophistication to our knowledge of Y-chromosome diversity.

Our data prove that DYS458.2 alleles have at least two different evolutionary origins in J1-M267 and R1b3-M405, with a third possible independent origin in R1b3-M269*(xM405). The evidence includes the fact that such alleles occur on different binary haplogroups that are well separated in the global genealogy and the added detail that the physical position of the occurrence of the non-conforming alleles within the normal tetramer repeat is different and distinctive for each haplogroup. While it is possible that the single DYS458.2 chromosome in R1b3-M269*(xM405) may be the result of a reversion of the M405 derived allele back to the ancestral state, the unique placement of the AA mutation within the GAAA tract (Table 1 and Table 2) suggests a potential third independent origin for this non-conforming DYS458.2 allele. Since the standard quantitative sizing techniques usually used to genotype Y-STR loci are unable to reveal such differential sequence-based features that are associated with the three represented haplogroups, knowledge from other STR and binary polymorphisms are required to more accurately assign genetic affinity.

While the DYS458.2 allele class is fixed in all the haplogroup J1-M267 chromosomes inves-

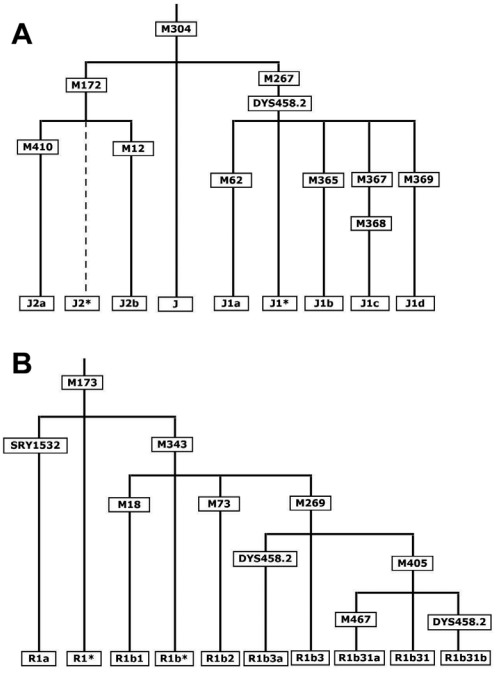

Figure 4. Phylogenetic relationships of DYS458.2 chromosomes in panel A – Haplogroup J1 and panel B – Haplogroup R1b3.

Table 4. Frequencies of haplogroup R1b3, its sub-clades, and DYS458.2 variants by population*

Population	Sample size	R1b3 Clade	M269(xM405)	M405(xM467)	M467	DYS458.2†	
						M269(xM405)	M405(xM467)
Austria	22	27.3	4.5	22.7	0	0	0
Central and South America‡	33	0	0	0	0	0	0
Czech Republic	36	27.8	13.9	13.9	0	0	0
Denmark	113	34.5	16.8	16.8	0.9	0	0
Eastern Europe§	44	4.5	4.5	0	0	0	0
England	138	57.2	35.5	20.3	1.4	0	4.3
France	56	51.8	44.6	7.1	0	0	0
Germany	332	43.1	22.6	18.7	1.8	0	0.6
Ireland	102	80.4	74.5	5.9	0	1	1
Italy	284	37.3	33.8	3.5	0	0	0
Jordan	76	0	0	0	0	0	0
Middle East‖	43	0	0	0	0	0	0
Netherlands	94	54.3	17	35.1	2.1	0	2.1
Oceania¶	43	0	0	0	0	0	0
Oman	29	0	0	0	0	0	0
Pakistan**	177	3.4	3.4	0	0	0	0
Palestine	47	0	0	0	0	0	0
Poland	110	22.7	14.5	8.2	0	0	0
Russia	56	21.4	14.3	5.4	1.8	0	0
Slovenia	105	17.1	13.3	3.8	0	0	0
Switzerland	90	57.8	44.4	13.3	0	0	0
Turkey††	523	14.5	14.1	0.4	0	0	0
Ukraine	32	25	15.6	9.4	0	0	0
United States	58	5.2	0	5.2	0	0	5.2

*13 samples with undetermined R1b3-M269 sub-clade status are not included in frequency calculations. Populations with fewer than 20 representatives are not included in frequency calculations.
†DYS458 status is undetermined for 24 samples.
‡Central and South America include Bolivia, Brazil, Chile, Mexico, Peru, Uruguay.
§Eastern Europe includes Belarus, Croatia, Greece, Hungary, Lithuania, Romania, Serbia, and Slovakia.
‖Middle East includes Egypt, Gaza, India, Iraq, Israel, Kuwait, Lebanon, Saudi Arabia, Syria, and West Bank.
¶Oceania includes American Samoa, Australia, Hawaii, New Zealand, Samoa, Tahiti, Tonga, Tuamotu, and Vanuatu.
**Samples from Sengupta et al (27).
††Samples from Cinnioglu et al (7).

tigated in this study, the possibility remains that some J1-M267 chromosomes with DYS458.0 allele sizes may exist in other populations or different sample sets. However, the recognition of marker M405, which demonstrates its highest frequency in the Netherlands, and describes a considerable fraction of haplogroup R1b3-M269 chromosomes that predominate in Western Europe, plus the discovery that DYS458.2 alleles occur in a subset of M405 derived chromosomes provides further phylogenetic resolution within this sub-clade. Our results indicate that previously characterized haplogroup R1b3-M269* chromosomes, which are most common in Western Europe can be subdivided into the informative M405 DYS458.0 and M405 DYS458.2 sub-clades. This study confirmed that Y-chromosome non-conforming STR alleles, once integrated into the binary haplogroup Y chromosome gene tree could expose further levels of sub-structure, some of which define sub-categories of chromosomes with more restricted geographic distribution. Understanding the binary haplogroup affiliation of Y-STR non-conforming alleles improves the value of such alleles in a phylogenetic context. By leveraging knowledge concerning the phylogenetic and spatial frequency distribution patterns of such non-conforming STR allele classes for other loci, it will be possible to better understand diversity within the Y-chromosome gene pool. In the future, the binary haplogroup affiliations of other Y-STR loci with non-consensus allele classes that occur at informative frequencies should also be evaluated.

Acknowledgments

We thank all men who donated DNA samples used in this study. We acknowledge the SMGF genealogy staff for their review of the genealogical records of study participants. We also thank Enass Tinah for her sample collection efforts and technical support.

References

1 Roewer L, Arnemann J, Spurr NK, Grzeschik KH, Epplen JT. Simple repeat sequences on the human Y chromosome are equally polymorphic as their autosomal counterparts. Hum Genet. 1992;89:389-94. Medline:1618488

2 Kayser M, Kittler R, Erler A, Hedman M, Lee AC, Mohyuddin A, et al. A comprehensive survey of human Y-chromosomal microsatellites. Am J Hum Genet. 2004;74:1183-97. Medline:15195656

3 Roewer L, Croucher PJ, Willuweit S, Lu TT, Kayser M, Lessig R, et al. Signature of recent historical events in the European Y-chromosomal STR haplotype distribution. Hum Genet. 2005;116:279-91. Medline:15660227

4 Zhivotovsky LA, Underhill PA, Cinnioglu C, Kayser M, Morar B, Kivisild T, et al. The effective mutation rate at Y chromosome short tandem repeats, with application to human population-divergence time. Am J Hum Genet. 2004;74:50-61. Medline:14691732

5 Kwak KD, Jin HJ, Shin DJ, Kim JM, Roewer L, Krawczak M, et al. Y-chromosomal STR haplotypes and their applications to forensic and population studies in east Asia. Int J Legal Med. 2005;119:195-201. Medline:15856270

6 Moore LT, McEvoy B, Cape E, Simms K, Bradley DG. A y-chromosome signature of hegemony in gaelic ireland. Am J Hum Genet. 2006;78:334-8. Medline:16358217

7 Cinnioglu C, King R, Kivisild T, Kalfoglu E, Atasoy S, Cavalleri GL, et al. Excavating Y-chromosome haplotype strata in Anatolia. Hum Genet. 2004;114:127-48. Medline:14586639

8 Shen P, Lavi T, Kivisild T, Chou V, Sengun D, Gefel D, et al. Reconstruction of patrilineages and matrilineages of Samaritans and other Israeli populations from Y-chromosome and mitochondrial DNA sequence variation. Hum Mutat. 2004;24:248-60. Medline:15300852

9 Redd AJ, Agellon AB, Kearney VA, Contreras VA, Karafet T, Park H, et al. Forensic value of 14 novel STRs on the human Y chromosome. Forensic Sci Int. 2002;130:97-111. Medline:12477629

10 Chang YM, Perumal R, Keat PY, Yong RY, Kuehn DL, Burgoyne L. A distinct Y-STR haplotype for Amelogenin negative males characterized by a large Y(p)11.2 (DYS458-MSY1-AMEL-Y) deletion. Forensic Sci Int. 2007;166:115-20. Medline:16765004

11 Cadenas AM, Regueiro M, Gayden T, Singh N, Zhivotovsky LA, Underhill PA, et al. Male amelogenin dropouts: phylogenetic context, origin and implications. Forensic Sci Int. 2007;166:155-63. Medline:16781100

12 Frigi S, Pereira F, Pereira L, Yacoubi B, Gusmão L, Alves C, et al. Data for Y-chromosome haplotypes defined by 17 STRs (AmpFLSTR Yfiler) in two Tunisian Berber communities. Forensic Sci Int. 2006;160:80-3. Medline:16005592

13 Underhill PA, Shen P, Lin AA, Jin L, Passarino G, Yang WH, et al. Y chromosome sequence variation and the history of human populations. Nat Genet. 2000;26:358-61. Medline:11062480

14 Hammer MF, Karafet TM, Redd AJ, Jarjanazi H, Santachiara-Benerecetti S, Soodyall H, et al. Hierarchical patterns of global human Y-chromosome diversity. Mol Biol Evol. 2001;18:1189-203. Medline:11420360

15 Underhill PA. Inferring human history: clues from Y-chromosome haplotypes. Cold Spring Harb Symp Quant Biol. 2003;68:487-93. Medline:15338652

16 de Knijff P. Messages through bottlenecks: on the combined use of slow and fast evolving polymorphic markers on the human Y chromosome. Am J Hum Genet. 2000;67:1055-61. Medline:11023811

17 Kayser M, Brauer S, Weiss G, Schiefenhovel W, Underhill PA, Stoneking M. Independent histories of human Y chromosomes from Melanesia and Australia. Am J Hum Genet. 2001;68:173-90. Medline:11115381

18 Jin L, Macaubas C, Hallmayer J, Kimura A, Mignot E. Mutation rate varies among alleles at a microsatellite locus: phylogenetic evidence. Proc Natl Acad Sci U S A. 1996;93:15285-8. Medline:8986803

19 Bandelt HJ, Forster P, Röhl A. Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol. 1999;16:37-48. Medline:10331250

20 Gusmao L, Butler JM, Carracedo A, Gill P, Kayser M, Mayr WR, et al. DNA Commission of the International Society of Forensic Genetics (ISFG): an update of the recommendations on the use of Y-STRs in forensic analysis. Forensic Sci Int. 2006;157:187-97. Medline:15913936

21 Levinson G, Gutman GA. Slipped-strand mispairing: a major mechanism for DNA sequence evolution. Mol Biol Evol. 1987;4:203-21. Medline:3328815

22 Nebel A, Filon D, Hohoff C, Faerman M, Brinkmann B, Oppenheim A. Haplogroup-specific deviation from the stepwise mutation model at the microsatellite loci DYS388 and DYS392. Eur J Hum Genet. 2001;9:22-6. Medline:11175295

23 Nebel A, Filon D, Brinkmann B, Majumder PP, Faerman M, Oppenheim A. The Y chromosome pool of Jews as part of the genetic landscape of the Middle East. Am J Hum Genet. 2001;69:1095-112. Medline:11573163

24 Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. Genome Res. 2002;12:339-48. Medline:11827954

25 Di Giacomo F, Luca F, Popa LO, Akar N, Anagnou N, Banyko J, et al. Y chromosomal haplogroup J as a signature of the post-neolithic colonization of Europe. Hum Genet. 2004;115:357-71. Medline:15322918

26 Nebel A, Landau-Tasseron E, Filon D, Oppenheim A, Faerman M. Genetic evidence for the expansion of Arabian tribes into the Southern Levant and North Africa. Am J Hum Genet. 2002;70:1594-6. Medline:11992266

27 Sengupta S, Zhivotovsky LA, King R, Mehdi SQ, Edmonds CA, Chow CE, et al. Polarity and temporality of high-resolution y-chromosome distributions in India identify both indigenous and exogenous expansions and reveal minor genetic influence of central asian pastoralists. Am J Hum Genet. 2006;78:202-21. Medline:16400607

28 Polzin T, Daneschmand SV. On Steiner trees and minimum spanning trees in hypergraphs. Operations Research Letters. 2003;31:12-20.

29 Underhill PA, Jin L, Lin AA, Mehdi SQ, Jenkins T, Vollrath D, et al. Detection of numerous Y chromosome biallelic polymorphisms by denaturing high-performance liquid chromatography. Genome Res. 1997;7:996-1005. Medline:9331370

30 Lee LG, Connell CR, Bloch W. Allelic discrimination by nick-translation PCR with fluorogenic probes. Nucleic Acids Res. 1993;21:3761-6. Medline:8367293

31 Sims LM, Garvey D, Ballantyne J. Sub-populations within the major European and African derived haplogroups R1b3 and E3a are differentiated by previously phylogenetically

undefined Y-SNPs. Hum Mutat. 2007;28:97. Medline:17154278

32 Cruciani F, Santolamazza P, Shen P, Macaulay V, Moral P, Olckers A, et al. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. Am J Hum Genet. 2002;70:1197-214. Medline:11910562

33 Flores C, Maca-Meyer N, González AM, Oefner PJ, Shen P, Pérez JA, et al. Reduced genetic structure of the Iberian peninsula revealed by Y-chromosome analysis: implications for population demography. Eur J Hum Genet. 2004;12:855-63. Medline:15280900

34 Zerjal T, Xue Y, Bertorelle G, Wells RS, Bao W, Zhu S, et al. The genetic legacy of the Mongols. Am J Hum Genet. 2003;72:717-21. Medline:12592608

35 Xue Y, Zerjal T, Bao W, Zhu S, Lim SK, Shu Q, et al. Recent spread of a Y-chromosomal lineage in northern China and Mongolia. Am J Hum Genet. 2005;77:1112-6. Medline:16380921