Brief Communication: New Y-Chromosome Binary Markers Improve Phylogenetic Resolution Within Haplogroup R1a1

Horolma Pamjav,1* Tibor Fehér,1 Endre Németh,1 and Zsolt Pádár2

1Network of Forensic Science Institutes, Institute of Forensic Medicine, DNA Laboratory, Budapest, Hungary
2Genoid Ltd, DNA Laboratory, Budapest, Hungary

KEY WORDS Y chromosome; haplogroup R1a1; Hungarians; Roma; human demographic history

ABSTRACT Haplogroup R1a1-M198 is a major clade of Y chromosomal haplogroups which is distributed all across Eurasia. To this date, many efforts have been made to identify large SNP-based subgroups and migration patterns of this haplogroup. The origin and spread of R1a1 chromosomes in Eurasia has, however, remained unknown due to the lack of downstream SNPs within the R1a1 haplogroup. Since the discovery of R1a1-M458, this is the first scientific attempt to divide haplogroup R1a1-M198 into multiple SNP-based sub-haplogroups. We have genotyped 217 R1a1-M198 samples from seven different population groups at M458, as well as the Z280 and Z93 SNPs recently identified from the “1000 Genomes Project”. The two additional binary markers present an effective tool because now more than 98% of the samples analyzed assign to one of the three sub-haplogroups. R1a1-M458 and R1a1-Z280 were typical for the Hungarian population groups, whereas R1a1-Z93 was typical for Malaysians Indians and the Hungarian Roma. Inner and Central Asia is an overlap zone for the R1a1-Z280 and R1a1-Z93 lineages. This pattern implies that an early differentiation zone of R1a1-M198 conceivably occurred somewhere within the Eurasian Steppes or the Middle East and Caucasus region as they lie between South Asia and Eastern Europe. The detection of the Z93 paternal genetic imprint in the Hungarian Roma gene pool is consistent with South Asian ancestry and amends the view that H1a-M82 is their only discernable paternal lineage of Indian heritage. Am J Phys Anthropol 000:000–000, 2012. © 2012 Wiley Periodicals, Inc.

To this date, many efforts have been made to identify SNP-based subgroups and migration patterns of haplogroup R1a1. A deeper interpretation of prehistoric and historic events of R1a1 chromosomes in Eurasian populations has, however, remained unavailable due to the lack of new downstream SNPs within R1a1 haplogroup.

Haplogroup R1a1-M198 is a major clade of human Y chromosomal haplogroups that is distributed all across Eurasia—from Scandinavia to India, and from the Balkans to Mongolia. The sub-classification and origins of the R1a1 haplogroup have been discussed by different researchers. Haplogroup R1a1-M198 is interpreted to have expanded in the territory of present day Ukraine (Semino et al., 2000) and to have been spread by the Kurgan culture, which migrated into both Europe and the east, resulting in the expansion of Indo-European languages (Gimbutas, 1970). Some other researchers have suggested that the R1a1 haplogroup may have originated in India (Sengupta et al., 2006; Sharma et al., 2009) or in Central Asia (Wells et al., 2001). The reason for these different interpretations was the lack of new Y chromosomal SNP markers with which to provide a higher resolution of the R1a1 haplogroup.

The study by Underhill et al. (2010) was the first successful attempt to identify an R1a1-M198 sub-haplogroup occurring at informative frequencies. The new marker M458, which was reported by the study, was a very important step as a new subgroup was discovered which had a specific geographic distribution focusing in Eastern Europe. Underhill et al. (2010) also found another new SNP - R1a1-M434 - at low frequency in Pakistan. To this date, however, no further subdivision of the R1a1-M198 clade has been described in academic papers.

Recently, the “1000 Genomes Project” (http://www.1000genomes.org/about) has provided many complete Y-DNA genomes in their database, and researchers have started to analyze the data for new SNPs. In cooperation with private researchers, the private genealogy DNA testing company “Family Tree DNA” established the primers for these new SNPs, which received a “Z” label in the nomenclature (those SNPs which were discovered by the company itself, have the label “L”).

We have undertaken a survey of R1a1 sub-haplogroup variation in 217 R1a1-M198 samples from Europe and Asia. The aim of the study was to (1) verify previously observed findings for M458 and (2) assess the utility of the Z280 and Z93 SNPs at resolving additional substructure within the R1a1-M198 haplogroup. Furthermore, 10 Y-STR loci were included in phylogenetic analysis to estimate population expansion times and possible migration scenarios for the subgroups M458, Z280, and Z93.

Additional Supporting Information may be found in the online version of this article.

The authors declare that there is no conflict of interest.

*Correspondence to: Horolma Pamjav, Budapest, PO Box 216, H-1536, Hungary. E-mail: phorolma@hotmail.com

Received 23 April 2012; accepted 5 September 2012

DOI 10.1002/ajpa.22167

Published online in Wiley Online Library (wileyonlinelibrary.com).
TABLE 1. Populations analyzed in present study

<table>
<thead>
<tr>
<th>Population</th>
<th>n</th>
<th>M198</th>
<th>M458</th>
<th>Z280</th>
<th>Z93</th>
<th>Source</th>
<th>Color in network</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hungarian (287)</td>
<td>78</td>
<td>4</td>
<td>30</td>
<td>41</td>
<td>3</td>
<td>Present study</td>
<td>Red</td>
</tr>
<tr>
<td>Szekler (88)</td>
<td>15</td>
<td>0</td>
<td>3</td>
<td>11</td>
<td>1</td>
<td>Present study</td>
<td>Yellow</td>
</tr>
<tr>
<td>Csango (106)</td>
<td>22</td>
<td>0</td>
<td>10</td>
<td>11</td>
<td>2</td>
<td>Present study</td>
<td>Blue</td>
</tr>
<tr>
<td>H. Roma (256)</td>
<td>35</td>
<td>0</td>
<td>16</td>
<td>2</td>
<td>17</td>
<td>Present study</td>
<td>Black</td>
</tr>
<tr>
<td>M. Indian (301)</td>
<td>53</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>53</td>
<td>Present study</td>
<td>Grey</td>
</tr>
<tr>
<td>Uzbek (40)</td>
<td>9</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>6</td>
<td>Present study</td>
<td>Green</td>
</tr>
<tr>
<td>Mongolian (222)</td>
<td>5</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>4</td>
<td>Present study</td>
<td>White</td>
</tr>
<tr>
<td>Total (N = 1,300)</td>
<td>217</td>
<td>4</td>
<td>60</td>
<td>68</td>
<td>85</td>
<td>–</td>
<td>–</td>
</tr>
</tbody>
</table>

*N*, all chromosomes investigated (1,300 males); *n*, R1a1 chromosomes tested for the new markers and included in the networks; H. Roma, Hungarian Roma; M. Indian, Malaysian Indian.

**MATERIALS AND METHODS**

**Ethnic composition of the samples**

The three SNPs were genotyped in 217 males comprising the R1a1-M198 component, totaling 1,300 males overall from seven populations. These samples included Hungarians (Hu = 78), Szeklers (Sz = 15), Csangos (Cs = 22), Roma (HuRo = 35), Malaysian Indians (MaIn = 53), Uzbeks (Uz = 9), and Mongolians (Mo = 5). All samples were previously tested and published for Y-STRs and Y-SNPs except for the Szekler, Csango, Uzbek, and Mongolian samples (Volgyi et al., 2009; Pamjav et al., 2011). The Szeklers and Csangos are Hungarian-speaking ethnic groups living in Romania. Malaysian Indians are descendants of South Indian migrants from during the British colonization of Malaysia.

**DNA isolation and PCR amplification**

DNA was isolated from saliva samples using organic extraction methods as described previously by Comey et al. (1994). The samples were quantified using the ABI 7500 Real-time PCR System (Applied Biosystems, Foster City, CA).

**Testing of Y-STR and Y-SNP markers**

DNA was amplified using the PowerPlex Y (Promega, Madison, WI) amplification kit including 12 Y-STR loci according to the manufacturer’s instructions. Fragment sizes and allele designations were determined using a Genetic Analyzer ABI3130 (Applied Biosystems, Foster City, CA) using GeneMapper ID-X v1.2 software.

When testing Y-SNP markers, amplifications of 3–5 ng genomic DNA were performed in an ABI 7500 and in GeneAmp 9700 thermal cyclers with TaqMan probes using the programs designed by the manufacturer (Applied Biosystems, Foster City, CA). The relative fluorescence of the PCR products was evaluated on an ABI 7500 Real-time PCR System using its SDS software as described in the manufacturer’s manual (Applied Biosystems, Foster City, CA) and in former publications (Biró et al., 2009; Volgyi et al., 2009). Based on the Blat search, two copies of PCR products may be amplified by primers and probes used. Thus, we always used controls for testing (negative and positive). Negative controls are two male DNA from different haplogroups and positive controls are two Z280 carriers.

A complete list of primers and Taqman probes for binary markers (SNPs) is included in Supporting Information Table 1. The nomenclature of haplogroups followed the results of the “1000 Genomes Project”, as reflected on the 2012 ISOGG tree (http://isogg.org/tree/ISO-GG_HapgrpR.html). To avoid later misunderstandings in nomenclature, a simpler nomination (“R1a1-SNP”) was used in the article.

**Phylogenetic study**

To examine the STR variation within the sub-haplogroups, networks were constructed using the Network 4.5.1.0 program (Bandelt et al., 1999). Repeats of the locus DYS389II were subtracted from the locus DYS389I and, as is common practice, the locus DYS385 was excluded from the network. The rho statistic within the network program was used to estimate the time to the most recent common ancestor (TMRCA) of haplotypes within the haplogroups R1a1-M458, R1a1-Z93, and R1a1-Z280. Evolutionary time estimates were calculated according to Zhivotovsky et al. (2004) and STR mutation rate was assumed to be 6.9 × 10⁻⁴/locus/25 years.

**RESULTS**

**Distribution of the R1a1-haplogroup in the samples**

Distribution of the R1a1 sub-haplogroups is presented in Table 1. The R1a1*-M198 parahaplogroup was observed only in the Hungarian samples. R1a1-M458 was the second most frequent haplogroup for both Hungarians and Roma. It was absent from the Malaysian and Mongolian samples, whereas only one occurrence was observed among the three R1a1 Uzbek samples.

As shown in Table 1, haplogroup R1a1-Z280 was present in more than 50% of R1a1 Hungarian males, whereas it was absent from the Indian population group. This haplogroup was detected at low frequencies in Hungarians and Roma, whereas R1a1-M458 to R1a1-Z280 occurrences was relatively similar in the Hungarian and Csango population groups, whereas Z280 chromosomes were present in most of the Szekler samples. Despite the possibility of duplicated copies, the received results were always relevant.

Haplogroup R1a1-Z93 was frequent in the Roma, Uzbek, and Mongolian samples, whereas the Malaysian Indians belonged exclusively to this haplogroup. It occurred with low frequency in the Hungarian samples (Table 1).

**Phylogenetic analysis**

Based on 10 Y-STR loci (DYS385 excluded), networks were constructed within each of the sub-haplogroups detected. Median (or modal) haplotypes were used to
The main aim of this study was to establish the phylogeographic subdivisions within the R1a1-M198 haplogroup by analyzing a few Eurasian population samples in the context of geographical specificity and population origin. The most important observation of the study is that the new binary markers present an effective tool for subtyping the R1a1-M198 haplogroup because more than 98% of the R1a1 samples analyzed belonged to one of the three sub-haplogroups. The highest frequency of the R1a1-M458 haplogroup is confined to Eastern Europe (Hungary and Hungarian ethnic groups in Romania) and is virtually absent in Asia, which coincides with the findings of Underhill et al. (2010). Despite the limited data available for Z280 and Z93, some general inferences can be drawn from the geographic distributions of these two haplogroups. The R1a1-Z280 subclade is a strong candidate for covering the R1a1a* (xM458) in Eastern Europe, which was found in high frequency by Underhill et al. (2010). The tested set of 53 Malaysian Indian samples presented 100% frequency of the R1a1-Z280 subclade, without co-existence Z280 or Z93 sub-haplogroups. Inner and Central Asia seem to be the overlap zones for the R1a1-Z280 and R1a1-Z93 chromosomes as both forms were observed at low frequencies. This is again consistent with the observations described for R1a1a* spread in Central Asia and in the Altai region by Underhill et al. (2010). This pattern suggests that the origin of R1a1-M198 arguably occurred somewhere between South Asia and Eastern Europe. Potential candidates could be the Eurasian Steppes (Ukraine – Southern Russia – Kazakhstan – Caucasus) or the Middle East. European populations showed higher
M458 and Z280, whereas Asian populations presented higher Z93 frequencies, indicating that the new markers can be effectively used to distinguish between the European and Asian branches of the haplogroup R1a1-M198.

The detection of the Z93 paternal imprint in the Hungarian Roma gene pool is consistent with South Asian ancestry and amends the view that H1a-M82 is their only discernible paternal lineage of Indian heritage (Gusmão et al., 2005; Pamjav et al., 2011). The presence of the lineage in Hungary is unlikely to be entirely a consequence of just the Roma migrations. It is possible that other gene flows from Central Asia and the Middle East could account for some of the Z93 detected in Europe. The relatively high presence of M458 lineage in the Hungarian Roma population group is due to host population admixture and genetic drift. The network analysis of the M458 haplogroup supports this possibility (Fig. 1A).

The coalescent time calculated by us for R1a1-M458 carriers is consistent with the age calculated by Underhill et al. (2010) in Europe yielding 7.3 KYA versus 7.9 KYA (thousands of years ago). Underhill et al. (2010) also noted the potential association of R1a1-M458 with the Linear Pottery Neolithic culture in the territory of present-day Hungary—this observation is supported by our data. The TMRCA calculated for R1a1-Z280 diversification (10.3 KYA) is approximately in agreement with the estimation of Underhill et al. (2010) for R1a1a*(xM458) chromosomes in Eastern Europe (~11 KYA). However, the coalescent age of 10.3 KYA for R1a1-Z93 chromosomes in this study is lower than that of populations of the Indus Valley (14 KYA) for the STR associated diversity of R1a1a*(xM458) chromosomes calculated by Underhill et al. (2010). Previous publications have pointed out that regions of highest haplogroup frequencies do not always indicate the territory of origin (Cinnioglu et al., 2004) and high STR diversity may not be exclusively an indicator of in-situ diversification but could also be the consequence of repeated gene flow from different sources (Zerjal et al., 2002; Sharma et al., 2009).

It is an open question as to whether the Andronovo and Tarim Basin R1a1-M198 peoples (Keyser et al., 2009; Li et al., 2010) belonged to the R1a1-Z280 group, the R1a1-Z93 group, or both. By discovering the relationship between R1a1-Z280, R1a1-Z93 and the Kurgan expansions, many questions could be answered regarding the history of the Eurasian Steppe. Furthermore, it would be important to check the occurrence of R1a1-Z280 in Finno-Ugric populations because R1a1-M198 is the relatively most frequent haplogroup among Mordvins, Maris, Komis, Estonians (Tambets et al., 2004), and Hungarians (Volgyi et al., 2009).

Comprehensive surveys of the Indo-European, Uralic, Altaic, and Paleo-Siberian populations in Eurasia should be carried out for the R1a1-M198 haplogroup using the M458, Z280, and Z93 SNPs to be able to draw more robust conclusions on the origins, spread, and ethno-linguistic affiliations of the specific R1a1-M198 subclades.

ACKNOWLEDGMENTS

This study could not have been written without the work of G. Magoon, who identified the Z280 and Z93 SNPs among the raw data of the “1000 Genomes Project”. We thank Dr. Eva Susa (General Director of the Network of Forensic Science Institutes) for her support. We also thank Ati Rosselet for the English editing and all sample donors and laboratory assistants.

LITERATURE CITED


Underhill PA, Myres NM, Rootsi S, Metspalu M, Zhivotovsky LA, King RD, Lin AA, Chow CE, Semino O, Battaglia V,


