Gemma Berniell-Lee
Karla Sandoval
Isabel Mendizabal
Elena Bosch
David Comas

Unitat de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Barcelona, Spain

Received February 5, 2007
Revised March 9, 2007
Accepted April 19, 2007

Research Article

SNPlexing the human Y-chromosome: A single-assay system for major haplogroup screening

SNPs are one of the main sources of DNA variation among humans. Their unique properties make them useful polymorphic markers for a wide range of fields, such as medicine, forensics, and population genetics. Although several high-throughput techniques have been (and are being) developed for the vast typing of SNPs in the medical context, population genetic studies involve the typing of few and select SNPs for targeted research. This results in SNPs having to be typed in multiple reactions, consuming large amounts of time and of DNA. In order to improve the current situation in the area of human Y-chromosome diversity studies, we decided to employ a system based on a multiplex oligo ligation assay/PCR (OLA/PCR) followed by CE to create a Y multiplex capable of distinguishing, in a single reaction, all the major haplogroups and as many subhaplogroups on the Y-chromosome phylogeny as possible. Our efforts resulted in the creation of a robust and accurate 35plex (35 SNPs in a single reaction) that when tested on 165 human DNA samples from different geographic areas, proved capable of assigning samples to their corresponding haplogroup.

Keywords:
Haplogroup / Multiplex PCR / SNPlex technology / SNPs / Y-Chromosome phylogeny
DOI 10.1002/elps.200700078

1 Introduction

SNPs are mainly used as genetic markers in biomedical and pharmaceutical research to understand the causes of human diseases and try and identify the relevant genes associated with them. The need for fast and accurate genotyping of hundreds to thousands of SNPs in genome-wide genetic mapping has led to the rapid development of numerous medium-to-high throughput technologies (*e.g.*, MALDI-TOF MS, microarrays, and bead-based assays), which currently allow the typing of over half a million SNPs in a collection of samples. SNPs are also used in population genetic studies to reconstruct human demographic history which generally require the typing of few and select SNPs for targeted research. The genotyping methods used in this field are of low-to-medium throughput (*e.g.*, allele-specific probes and single base primer extension). For this reason, strong efforts are being made to increase the number of SNPs typed in a single reaction.

Due to its exclusively paternal inheritance, the human Y-chromosome has been extensively used in evolutionary and forensic genetics [1]. It is a mosaic of complex and highly repetitive sequences [2] and besides containing large rearrangements, small indels (insertions/deletions), and repetitive motifs such as STRs, it also contains hundreds of well-characterized SNPs. The evolutionary relationships between more than 200 of these SNPs are represented in a robust and well-established Y-chromosome phylogeny [3] which defines 18 major haplogroups (named A–R), in turn divided into subhaplogroups. The determination of what male lineage or haplogroup a sample belongs to requires the successive typing of markers in a hierarchical manner; typing first the markers which define the main haplogroups and then those defining the subhaplogroups within the main haplogroup identified. This process is not only extremely laborious but also highly time- and DNA-consuming. The Y-plexes created until now are mainly used for the refinement of subhaplogroups, and are based on the principle of single base extension (SBE)[4–8]. Although this technique is highly efficient and precise, it only allows a certain degree of multiplexing to take place. Under these observations, we decided

Correspondence: Dr. David Comas, Unitat de Biologia Evolutiva, Departament de Ciències Experimentals i de la Salut, Universitat Pompeu Fabra, Doctor Aiguader 88, E-08003 Barcelona, Spain
E-mail: david.comas@upf.edu
Fax: +34-93-3160901

Abbreviations: ASO, allele-specific oligonucleotide; **LSO,** locus-specific oligonucleotide; **OLA/PCR,** oligo ligation assay/PCR

to create a Y multiplex that both (i) amplified as many SNPs in a single reaction as possible (trying to increase the current maximum of 29 [9]) in order to reduce the DNA used and the time spent during the process, and (ii) defined all the main haplogroups and as many subhaplogroups as possible. In this way, a complete Y-plex would offer the possibility of determining the subhaplogroup (or at least the main haplogroup) to which a sample belongs in a single reaction. Depending on the scrutiny of the study, if subhaplogroup information was not fully covered, further refinement could then be carried out with specific subhaplogroup multiplexes. In order to create our main screening tool we decided to employ an oligo ligation assay (OLA) in a single reaction.

2 Materials and methods

2.1 Samples

Blood samples were obtained from a total of 165 healthy unrelated males from five different geographic areas; Gabon ($n = 54$), Mexico (10), Gypsies from Spain (12), Cuba (40), and Island of Reunion (25). It was difficult to obtain individual samples representing all the 31 (sub-)haplogroups defined by the multiplex, and whose affiliation had been previously established using other SNP genotyping methods. For this reason we decided to use samples of very different origin, which from the literature [3], we considered would fall into very different clades. In order to check the accuracy of the multiplex, we also used a set of internal controls: 24 samples from the Basque country (Spain) whose haplogroup affiliation had been previously determined using other genotyping methods [10]. DNA extraction was carried out using a standard phenol chloroform method.

2.2 SNP selection

A total of 215 Y-chromosome SNPs defining the main haplogroups and subhaplogroups were chosen [3]. The flanking sequences for each of the SNPs were then investigated for the creation of the SNP-specific ligation probes. Flanking sequences of around 110–140 base pairs (bp) both upstream and downstream of the SNPs were obtained by aligning previously published primers [4, 11] with the human genome using the Basic Local Alignment Search Tool program (BLAST) (in this case nucleotide–nucleotide BLASTn), and verified using the program Ensembl. Flanking sequences that were strictly identical between the two databases and nonambiguous, were then submitted to the SNPlex assay design pipeline (Applied Biosystems, Foster City, CA, USA) in FASTA format for multiplex pool design (these sequences are available on request). The design algorithm behind the SNPlex pooling system tries to minimize the interactions between allele-specific oligonucleotide (ASO) and locus-specific oligonucleotide (LSO) ligation probes, and creates assays of minimally interacting SNPs. Consequently, on many occasions, the SNP pool designed did not contain all the SNPs desired. Several trial-and-error attempts were therefore made at pool designing before the final 48plex was achieved, which is the maximum number of SNPs that can be genotyped with the present technology. This process involved continually replacing SNPs with other evolutionary alternatives (SNPs that define the same clade on the human Y-chromosome phylogeny [3]) when possible, or completely eliminating them and adding others when not possible. Great care was taken when doing this because unlike with other genotyping methods where the efficiency of a multiplex can be tested "*a posteriori*" while it is being created, adding and/or substituting SNPs that did not show good genotyping results until the optimum combination is obtained, with SNPlex technology this is not possible. Once the final pool has been designed, no adding or replacing of SNPs can take place unless a completely new pool design is performed.

The sequences of two monomorphic positions on the Y-chromosome were included in the final pool in order to have positive controls of the technique.

2.3 SNP genotyping and haplogroup classification

A total of 40 ng of genomic DNA was used for genotyping. The OLA/PCR reaction was carried out on a dual 384-well GeneAmp® 9700 thermal cycler (Applied Biosystems) following the guidelines given by Applied Biosytems. The SNPlex system is based on multiplex OLA/PCR followed by CE. Genomic DNA containing the target SNPs is interrogated with highly specific multiplexed sets of ASO and LSO ligation probes. A pair of universal PCR primers then amplify all the ligation products simultaneously. Amplicons containing internal universal cZipCode oligonucleotide sequences (sequences attached at the 5′-end of the genomic equivalent sequence within each ASO probe) are then hybridized to fluorescently labeled universal ZipChute reagents (probes). These probes contain sequences that are complementary to the cZipCode sequences. When they are subsequently eluted by CE they move at a unique rate, allowing their identification and posterior association with the target SNPs with GeneMapper® Analysis software (for specific details on the functioning of this technique please refer to [12]**).** PCR amplicons were run on a 3730x*l* DNA analyzer (Applied Biosystems), and genotypes were determined by means of the GeneMapper® Analysis Software version 3.5 (see Fig. 1). Alleles were called using the SNPlex clustering method, which make calls based on a minimum confidence value for a sample in a particular cluster. As the Y-chromosome is a haploid marker that does not recombine, only two clusters were observed for every SNP analyzed, each cluster representing one of the two alleles of the SNP (see Fig. 2). No heterozygote clusters were found throughout the study. Each sample was assigned to its Y-chromosome haplogroup according to the Y-chromosome phylogeny [3].

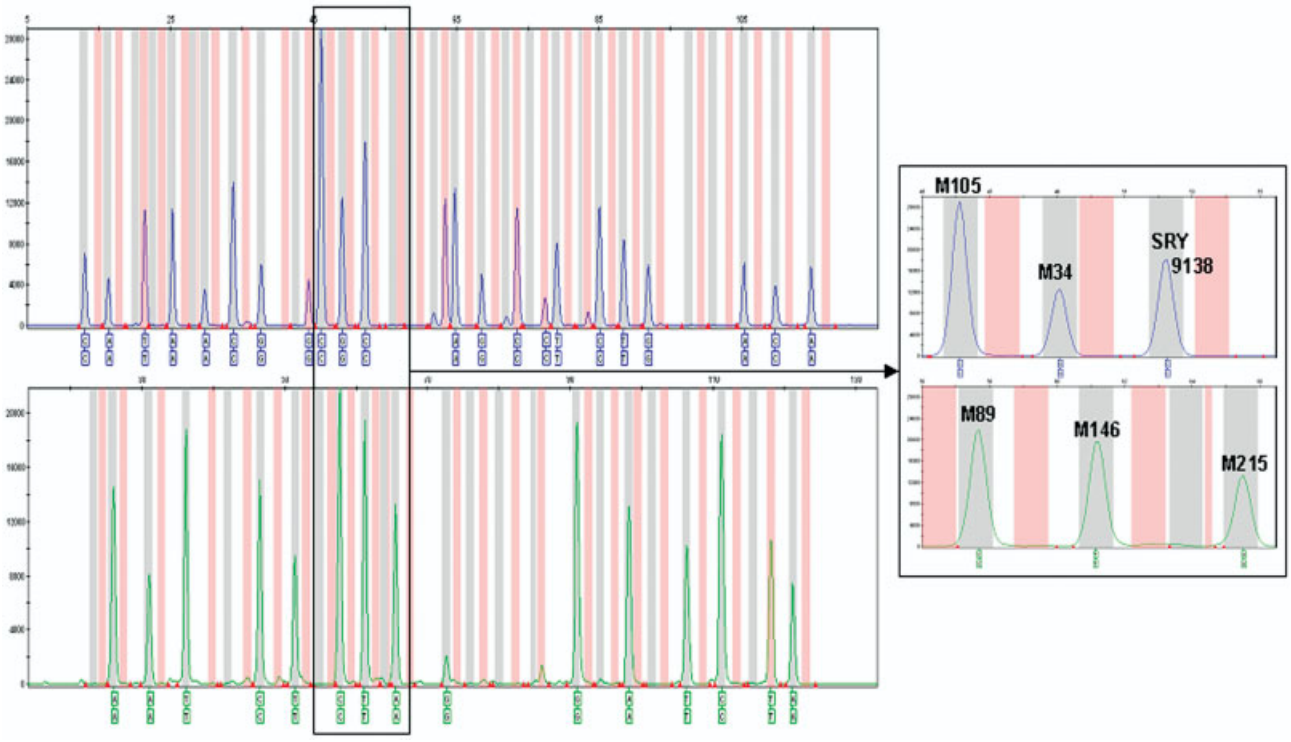

Figure 1. (a) Electropherogram of the 37 Y-markers typed in the multiplex. Each labeled peak represents a SNP, and two of them are the monomorphic positions used as positive controls. (b) Zoomed image of an electropherogram showing six Y-SNPs typed. Each labeled peak represents a SNP. In both (a) and (b) the gray and pink shading represents the bins for each SNP in the GeneMapper® software.

3 Results

Of the 48 genetic markers in the pool, 37 were successfully typed, being 77% the pass rate for the whole multiplex. Two of the 37 successful markers were the positive controls, which as expected, were monomorphic throughout the study. The number of successful genotypes for every SNP in the pool (call rate) was high (average call rate >97%), with the exception of M180 for which the call rate was 40% (Table 1). The final pool of 35 SNPs was able to distinguish a total of 31 different haplogroups and subhaplogroups, 20 of which are present in the populations typed, and 10 of which are shared by two or more populations (Fig. 3). The unfortunate failure of 11 SNPs, left some main haplogroups uncharacterized. Within the main haplogroups defined, most of the subhaplogroups are well characterized, with some main haplogroups being divided into as many as 3–7 subhaplogroups. Haplogroup composition varied greatly across the different populations sampled as expected from the known phylogeography of the samples used [13].

No phylogenetic incompatibilities were found between the alleles of any of the different SNPs. In other words, if a sample showed to be derived for a SNP that lead to a certain branch of the tree, it was also derived for all the SNPs leading to that branch, but ancestral for all the rest of the SNPs leading to other branches. For example, if a sample was derived for M201 (allele T), it was also derived for M89 (allele T), M168 (allele T), M42 (allele T), and $SRY_{10831.1}$ (allele G), but was ancestral for all the rest of the SNPs in the pool (see Table 1 and Fig. 2). Although some of the haplogroups (*i.e.*, haplogroups C, D, and M) were not present in the populations sampled, all the samples typed showed the ancestral allele for the SNPs defining these branches.

The results obtained for the control samples (Basque samples) were exactly the same as those previously obtained with other genotyping systems [10], giving a concordance rate of 100%.

4 Discussion

The objective of this study was to create a Y-SNP multiplex that typed as many SNPs in a single reaction as possible, and that could be used as a major haplogroup screening tool. The use of this multiplex would therefore reduce both the amount of time spent on typing and the amount of DNA necessary. Typing a large number of SNPs rather than only typing a subset also has the additional advantage of allowing us to detect possible inconsistencies (if they exist) on the Y-chromosome phylogeny. The 35plex created has proved suc-

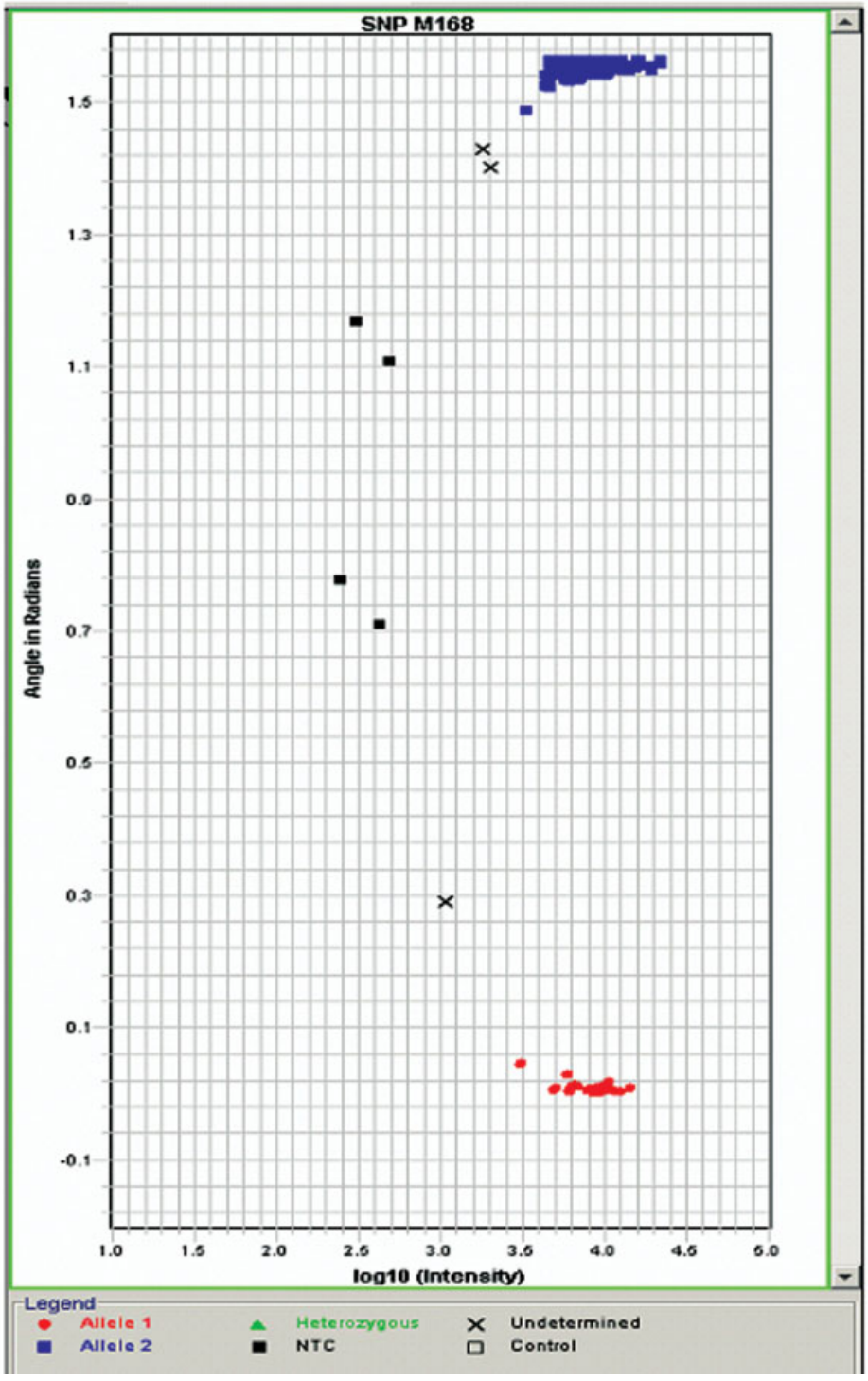

Figure 2. An example of a cartesian cluster plot (marker M168). Samples are plotted according to peak height (intesity), where peak height for the ancestral allele is plotted along the *x*-axis, the red dots, and peak height for the derived allele is plotted along the *y*-axis, the blue dots. Crosses represent samples for which allele determination was not possible due to low intensity. Black squares represent negative controls, data points where no DNA has been used (nontemplate controls NTC).

cessful in both of its objectives. It is capable of assigning samples of different ethnic origin to their corresponding haplogroup and, furthermore, depending on which main haplogroup a sample belongs to, the 35plex also allows the discrimination of subhaplogroups to a considerably detailed level (*e.g.*, haplogroup E). This degree of detail may be enough for certain studies and further typing may no longer be required.

The pass rate (77% successfully genotyped SNPs) and the average call rate (>97% successful genotypes) are slightly lower than previous values reported for 521 SNPs from human chromosome 21 using the same technology [12].

Table 1. Genetic markers in the original multiplex, the Y-haplogroup they define, their respective ancestral and derived alleles, and their call rate (successful genotype calls)

Polymorphic marker	Ancestral/ derived	Y-Haplogroup	Call rate (%)
M5	C/T	M	100
M9	C/G	K-R	100
M11	A/G	L	Failed
M14	T/C	A2	Failed
M15	C/A	D1	100
M31	G/C	A1	Failed
M32	T/C	A3	Failed
M33	A/C	E1	99
M34	G/T	E3b3a	100
M35	G/C	E3b	Failed
M38	A/C	C2	100
M42	A/T	B-R	100
M52	A/C	H	100
M55	T/C	D2	100
M61	C/T	L	Failed
M70	A/C	K2	99
M75	G/A	E2	100
M78	C/T	E3b1	99
M81	C/T	E3b2	100
M89	C/T	F-R	100
M96	C/G	E	100
M105	C/T	C1	100
M106	A/G	M	100
M112	G/A	B2b	Failed
M119	T/G	O1	100
M122	A/G	O3	98
M130	G/A	C	Failed
M132	G/T	E1	100
M146	T/G	B1	99
M150	C/T	B2a	Failed
M168	C/T	C-R	100
M170	A/C	I	100
M173	A/C	R1	100
M180	T/C	E3a	40
M181	T/C	B	100
M201	G/T	G	100
M214	T/C	N-O	100
M215	A/G	E3b	100
M216	C/T	C	100
M217	A/C	C3	99
P1	C/T	E3a	Failed
P27	G/A	P-R	Failed
P31	T/C	O2	99
SRY10831[a)1, 2]	A/G[1], G/A[2]	B-R[1], R1a[2]	99
SRY9138	C/T	K1	100
50f2(P)	C/G	B2b	98

Markers in bold are those present in the final 35plex.

a) This SNP has mutated twice in its history (recurrent mutation). The first mutation ([1]) was from (A) to (G) defining the major branching of haplogroup B through to haplogroup R. The second mutation ([2]) reverted back from (G) to (A) defining the subhaplogroup R1a.

Although the 35plex can be used for a wide range of populations, some of the main branches still require further refinement (*i.e.*, branches F, L, N, P, Q). Some of these main branches were defined by SNPs in the initial 48plex but failed during the OLA/PCR reaction. Haplogroup L, for example, was initially defined by mutations M11 and M61. These two mutations could be replaced by their evolutionary alternatives M20 or M22 to improve the strength of the multiplex. The same could be applied to haplogroup P, which was originally defined by P27. This SNP could be substituted for any of its three evolutionary equivalents 92R7, M45, or M74. Unfortunately, it was not possible to include an SNP defining haplogroup Q (such as P36 or MEH2) in the original 48plex due to incompatibilities with other SNPs. Further refinement of haplogroups N and R could be done by adding LLY22g and M207, respectively. Also, because of the low success rate of M180, defining haplogroup E3a, it would be advantageous to substitute this marker for another of its evolutionary alternatives such as M2 or P1.

SNPlex technology has proved a successful and robust method for the creation of Y-chromosome SNP multiplexes. However, the success rate seems to strongly depend on the compatibility and efficiency of the SNPs chosen. The Y-assay created here has proved sensitive, easy to use, and highly automated. Not only can hundreds of samples be typed at the same time, but also, the SNPlex allele-calling clustering algorithm largely facilitates the interpretation of the genotyping results and thus enables the fast and accurate classification of samples into their corresponding haplogroups.

Although there is room for improvement, this has been the first successful attempt, to our knowledge, at using this technology for Y-chromosome multiplexes, the positive outcome being the creation of a robust Y-35plex that can be widely used in human Y-chromosome population studies.

We would like to thank Mònica Vallés and Anna Maria Lluís Cano (UPF) and the staff of the Barcelona Node of the CEGEN (Centro Nacional de Genotipado) for their technical help. SNP genotyping services were provided by the Spanish "Centro Nacional de Genotipado" (CEGEN; www.cegen.org). The present study was supported by the Dirección General de Investigación, Ministerio de Educación y Ciencia, Spain (HF2004-0199 and BFU2006-01235/BMC) and Direcció General de Recerca, Generalitat de Catalunya (2005SGR/00608). G. B. L. received a FI fellowship from the Generalitat de Catalunya, and K. S. received a fellowship from the Consejo Nacional de Ciencia y Tecnología (CONACYT), México.

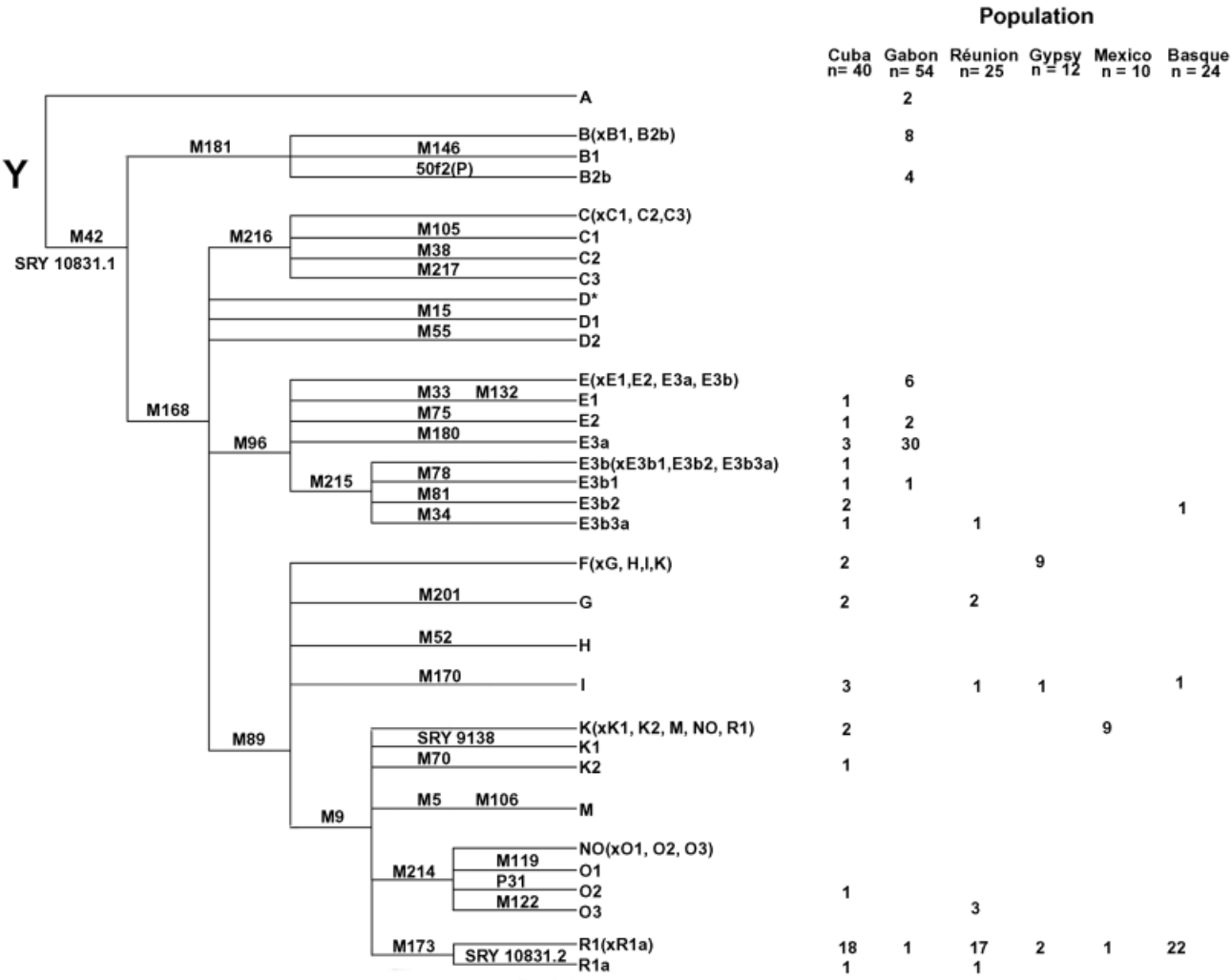

Figure 3. Phylogenetic tree showing the 35 Y-SNPs typed by the multiplex. SNP names and haplogroup names are given above the lines and at the end of the lines, respectively, following the nomenclature proposed by the Y-chromosome consortium [3]. Haplogroup absolute frequencies and the number of samples typed for each population are given on the right-hand side of the tree.

5 References

[1] Jobling, M. A., Pandya, A., Tyler-Smith, C., *Int. J. Legal Med.* 1997, *110*, 118–124.

[2] Skaletsky, H., Kuroda-Kawaguchi, T., Minx, P. J., Cordum, H. S. *et al.*, *Nature* 2003, *423*, 825–837.

[3] Y Chromosome Consortium, *Genome Res.* 2002, *12*, 339–348.

[4] Paracchini, S., Arredi, B., Chalk, R., Tyler-Smith, C., *Nucleic Acids Res.* 2002, *30*, e27.

[5] Lessig, R., Zoledziewska, M., Fahr, K., Edelmann, J. *et al.*, *Forensic Sci. Int.* 2005, *154*, 128–136.

[6] Mengel-Jorgensen, J., Sanchez, J. J., Borsting, C., Kirpekar, F., Morling, N., *Anal. Chem.* 2004, *76*, 6039–6045.

[7] Brion, M., Sobrino, B., Blanco-Verea, A., Lareu, M. V., Carracedo, A., *Int. J. Legal Med.* 2005, *119*, 10–15.

[8] Sanchez, J. J., Borsting, C., Hallenberg, C., Buchard, A. *et al.*, *Forensic Sci. Int.* 2003, *137*, 74–84.

[9] Brion, M., Sanchez, J. J., Balogh, K., Thacker, C. *et al.*, *Electrophoresis* 2005, *26*, 4411–4420.

[10] Bosch, E., Calafell, F., Comas, D., Oefner, P. J. *et al.*, *Am. J. Hum. Genet.* 2001, *68*, 1019–1029.

[11] Hammer, M. F., Blackmer, F., Garrigan, D., Nachman, M. W., Wilder, J. A., *Genetics* 2003, *164*, 1495–1509.

[12] Tobler, A. R., Short, S., Andersen, M. R., Paner, T. M. *et al.*, *J. Biomol. Tech.* 2005, *16*, 398–406.

[13] Underhill, P. A., Passarino, G., Lin, A. A., Shen, P. *et al.*, *Ann. Hum. Genet.* 2001, *65*, 43–62.

