Toward Male Individualization with Rapidly Mutating Y-Chromosomal Short Tandem Repeats

Kaye N. Ballantyne, 1, 2 Arwin Ralf, 1 Rachid Aboukhalid, 3 Niaz M. Achakzai, 4 Maria J. Anjos, 5 Qasim Ayub, 6 Jože Balažič, 7 Jack Ballantyne, 8, 9 David J. Ballard, 10 Burkhard Berger, 11 Cecilia Bobillo, 12 Mehdi Bouabdelhass, 12 Helen Burri, 13 Tomas Capal, 14 Stefano Caratti, 15 Jorg Cárdenas, 16 François Cartault, 17 Elizeu F. Carvalho, 18 Monica Carvalho, 19 Baowen Cheng, 19 Michael D. Coble, 20 David Comas, 21 Daniel Corach, 12 Maria E. D’Amato, 22 Sean Davison, 23 Peter de Knijff, 23 Maria Corazon A. De Ungria, 25 Ronny Decorte, 25 Tadeusz Dobosz, 25 Berit M. Dupuy, 25 Samir Elmgrgni, 29 Sara C. Gomes, 30 Laurens Grol, 31 Cordula Haas, 32 Erin Hanson, 33 Jürgen Henke, 33 Lotte Henke, 32 Fabiola Herrera-Rodríguez, 32 Carolyn R. Hill, 20 Gunilla Holmlund, 34, 35 Katsuya Honda, 36 Uta-Dorothee Immel, 37 Shota Inokuchi, 38 Patrick Munier, 39 Miriam Sirker, 40 Nederland, 41 E-mail: m.kayser@erasmusmc.nl

Contract grant sponsors: Netherlands Genomics Initiative (NGI)/Netherlands Organization for Scientific Research (NWO) within the framework of the Genomics Research Consortium Netherlands (GFCN); The Wellcome Trust (098853); The Wellcome Trust (087976); Leverhulme Trust (F/00.212/AM); ALPKULTUR and the Austrian Academy of Sciences’ “Alpine Research Programme”; Spanish Government (CDI2010-1494/BOS); UBACYT 2002010010744 and PIP CONICET N122010100914; Argentina; FWO-Vlaanderen; National Natural Science Foundation of China (31100906 and 81271386); Medical Research Council of South Africa and the National Research Foundation of South Africa; Philippine Council for Industry, Energy and Emerging Technology Research and Development (PCIEERD) of the Department of Science and Technology; FINNISH FOUNDATIONS' PROFESSORS POOL, PAULO FOUNDATION.

© 2014 The Authors. "Human Mutation" published by Wiley Periodicals, Inc.

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
ABSTRACT: Relevant for various areas of human genetics, Y-chromosomal short tandem repeats (Y-STRs) are commonly used for testing close paternal relationships among individuals and populations, and for male lineage identification. However, even the widely used 17-loci Yfiler set cannot resolve individuals and populations completely. Here, 52 centers generated quality-controlled data of 13 rapidly mutating (RM) Y-STRs in 14,644 related and unrelated males from 111 worldwide populations. Strikingly, >99% of the 12,272 unrelated males were completely individualized. Haplotype diversity was extremely high (global: 0.9999985, regional: 0.99836–0.9999988). Haplotype sharing between populations was almost absent except for six (0.05%) of the 12,156 haplotypes. Haplotype sharing within populations was generally rare (0.8% nonunique haplotypes), significantly lower in urban (0.9%) than rural (2.1%) and highest in endogamous groups (14.3%). Analysis of molecular variance revealed 99.98% of variation within populations, 0.018% among populations within groups, and 0.002% among groups. Of the 2,372 newly and 156 previously typed male relative pairs, 29% were differentiated including 27% of the 2,378 father–son pairs. Relative to Yfiler, haplotype diversity was increased in 86% of the populations tested and overall male relative differentiation was raised by 23.5%. Our study demonstrates the value of RM Y-STRs in identifying and separating unrelated and related males and provides a reference database.
Introduction

Genetic characterization of male individuals and populations by means of Y-chromosome DNA polymorphisms is relevant in various fundamental and applied areas of human genetics such as in evolutionary genetics and population history, for example, for modeling global and regional human evolution, mapping migration patterns across the globe, and tracking cultural and demographic factors such as patriarchy, extrapair paternity, endogamy, and polygyny. Further, Y-chromosome DNA analysis is important in genetic genealogy and for community genetic purposes such as personal ancestry identification, as well as for the identification of male lineages and inferring paternal genetic ancestry for judicial and investigative purposes [Kaysar et al., 1997; Underhill et al., 2000; Hammer et al., 2001; Oota et al., 2001; Jobling and Tyler-Smith, 2003; Roewer et al., 2005; Shi et al., 2010]. Similarities at Y-chromosome DNA markers are usually interpreted as indicating shared paternal ancestry of individuals and populations, whereas differences are used to conclude the absence of close paternal relationships. Such interpretations, however, depend in part on the underlying mutation rates of the Y-DNA markers used. Slowly evolving Y-chromosomal single-nucleotide polymorphisms (Y-SNPs), with an average mutation rate of about $3 \times 10^{-8}$ per nucleotide per generation [Xue et al., 2009; Poznik et al., 2013], are especially suitable for studying distant relationships between male individuals and populations [Underhill et al., 2000; van Oven et al., 2014]. More quickly evolving Y-chromosomal short tandem repeat (Y-STRs) polymorphisms, also referred to as Y-microsatellites, with an average mutation rate of about $10^{-3}$ per locus per generation [Goedbloed et al., 2009; Ballantyne et al., 2010], have proven useful for testing short to medium timescale paternal relationships [Kaysar et al., 2003; Kaysar et al., 2005; Roewer et al., 2005; Coble et al., 2009; van Oven et al., 2011] such as enabling male lineages to be connected through common paternal ancestry [Coble et al., 2009], for the same lineages be separated and individualized within that shared ancestral lineage [King and Jobling, 2009], and for the origins of entire population groups be elucidated [Parkin et al., 2007; Rebala et al., 2007]. Forensic usage of Y-STRs has largely focused on identifying paternal lineages using a core set of markers, linking suspects, and crime samples for investigation purposes [Kaysar et al., 1997; Roewer, 2009]. Haplotypes generated from conventional Y-STRs, such as the widely used 17 markers included in the commercially available AmpFISTR® Yfiler® PCR Amplification Kit (Life Technologies, San Francisco, CA) (subsequently referred to as Yfiler), suffer from two main limitations: (1) their inability to conclusively resolve some male lineages due to identical haplotypes arising in individuals that are not of common descent because of recurrent mutation, and (2) their inability to differentiate between paternally related males due to the moderately low mutation rate of the loci tested. In general, Y-STRs with much higher mutation rates than those of the conventionally used loci are expected to overcome or at least reduce both limitations. In a previous comprehensive Y-STR mutation rate study where nearly 200 Y-STRs were investigated in almost 2,000 father–son pairs confirmed by autosomal DNA analysis [Ballantyne et al., 2010], 13 Y-STRs with exceptionally high ($>10^{-7}$ per locus per generation) mutation rates were identified and termed rapidly mutating (RM) Y-STRs. Furthermore, in this previous study and a subsequent study [Ballantyne et al., 2012], theoretical and the first empirical evidence were provided to show that this set of 13 RM Y-STRs is able to achieve an order of magnitude higher male relative differentiation than is available with the commonly used Yfiler set, as well as to drastically improve male lineage differentiation over Yfiler.

However, to fully explore the potential for the RM Y-STR set in differentiating unrelated as well as related males for various purposes, much more data are needed. Therefore, the International RM Y-STR Study Group, a worldwide collaboration between 52 laboratories, was formed. Group members, chosen based on pre-existing practical experience in Y-STR analysis, genotyped under quality-controlled conditions the 13 RM Y-STRs in 14,644 males including 12,272 unrelated males from 111 worldwide populations and 4,744 closely related males. To compare the RM Y-STR set with conventional Y-STRs, Yfiler data were gathered in a subset of 7,784 unrelated men from 65 worldwide populations as well as in most of the male relatives analyzed for RM Y-STRs. The large collection of RM Y-STR haplotypes presented and explored in the present study enables a better understanding of their value to differentiate related and unrelated male individuals and populations, and additionally provides a suitable reference database for future use of RM Y-STRs in forensic, genealogical, anthropological, and population genetic studies.

Materials and Methods

DNA Samples

With a worldwide coverage in mind, contributing laboratories were identified and recruited based on their previous submission of Y-STR population data to the Y-chromosome Haplotype Reference Database (YHRD; www.yhrd.org), hence having proven experience in Y-STR analysis. Each of the 52 contributing laboratories within the International RM Y-STR Study Group genotyped a selection of their own in-house population sample sets, each consisting of seven to 634 individuals (median 100 individuals per population sample) across 111 defined population samples per laboratory. Population sets were to include only unrelated males. Three population samples were notable exceptions—the Biaka Pygmy group and the Bhutanese Lhokpu and Mönpa language groups. These populations, and particular sample sets used here, are known to contain high numbers of male relatives (Biaka), or were specifically selected as having extremely low resolution with Yfiler (Bhutan). Because these samples were ascertainment differently to the other 108 populations within the study, they were excluded from the continental groups during analysis to avoid bias. The Aboriginal Australian samples used here were also not ascertainment completely randomly, as they were selected based on Y-SNP haplogroups known to be authentic for Aboriginal Australians to avoid admixture effects as described elsewhere [Taylor et al., 2012].

A subset of 7,784 individuals from 65 populations were additionally genotyped with the AmpFISTR Yfiler PCR Amplification Kit (Life Technologies) for the most commonly used panel of 17 Y-STRs. Relative pairs were also analyzed, with 2,339 newly genotyped father–son pairs, 30 brother pairs, and three uncle–nephew pairs, for which relationship had previously been confirmed by autosomal DNA analysis; this dataset of male relatives was supplemented by 156 previously published relative pairs [Ballantyne et al., 2012].

Y-STR genotyping

The organizing laboratory (Department of Forensic Molecular Biology, Erasmus MC University Medical Centre Rotterdam)
provided genotyping protocols, allelic ladders, and tools for allele calling to all participating laboratories, and organized a quality control exercise prior to population data generation (for details see below). The 13 single- or multicycopy RM Y-STR markers (DYF38751, DYF39951, DYF40351, DYF40451, DYS449, DYS518, DYS526, DYS547, DYS570, DYS576, DYS612, DYS626, and DYS627) were amplified in three multiplex PCRs, as described in Supp. Tables S1 and S2. PCR-amplified products were separated and detected using participating laboratories’ standard protocols for analyzing STRs—in either ABI310, ABI3100, ABI3130, ABI3500, or ABI3730 Genetic Analyzers with POP-4, POP-6, or POP-7, and size standards ILS-600, ILS-500, or Genescan 500-LIZ. Allele calling was performed with GeneMapper (Life Technologies) or GeneMarker (SoftGenetics, State College, PA) utilizing custom panel and bin sets. The participating laboratories used their own primers and other consumables (except allelic ladders as provided by the host).

The commercially available AmpFlSTR® Yfiler® PCR Amplification Kit (Life Technologies) targeting 17 conventional Y-STRs was used and allele calling was performed according to the manufacturer’s instructions.

Genotyping Quality Control

To ensure genotyping consistency between the laboratories, all participants received allelic ladders prepared by the organizing laboratory, and six blind control DNA samples previously genotyped by the organizing laboratory. Genotyping of population samples was only allowed after a participating laboratory demonstrated the correct genotyping of these six blind control DNA samples at all 13 RM Y-STRs. If a participant reported erroneous genotypes, screenshots were requested and submitted to the organizing laboratory for an evaluation of the possible cause of error(s). Once the cause of error was identified, three additional blind control DNA samples were provided and genotyped by the participant. If these samples were typed correctly, the laboratory was allowed to type their population samples. In cases of unexpected results such as previously unknown alleles or microvariants, screenshots were requested and submitted to the organizing laboratory for inspection. Independent genotyping, and in some cases DNA sequencing, was performed to resolve differences in genotyping of certain samples. Samples with missing data from more than one marker were excluded from data analysis to prevent low quality samples affecting genotype and haplotype distributions. Further, any differences observed between relative pairs of individuals, pairwise FST genetic distances, and analysis of molecular variance (AMOVA) were calculated with Arlecore v3.5.1.3 [Excoffier and Lischer, 2010]. Theta (θ) values were calculated as per the reference Weir and Hill (2002). Molecular relationships between samples were analyzed using Network v4.6.1.1., applying the median-joining method. Weighting was applied as described elsewhere [Qamar et al., 2002], although based on updated mutation rates from Ballantyne et al. (2010) and www.yhrd.org. Only single-copy markers were utilized in network construction, due to the inability to assign multicycopy genotypes to individual loci. Testing of statistical significance (t-tests) was performed in SPSS v17.0. Theoretical estimates for the rates of relative differentiation for each Y-STR panel were calculated using the formula $P(k = 0) = 1 - P(k = 0) = 1 - e^{-Nm}$ [Ballantyne et al., 2010], where $N$ represents the number of markers and $m$ represents the average mutation rate of the set of markers obtained from the sampling from the posterior distribution.

Results and Discussion

Global Diversity of RM Y-STR Haplotypes and Male Lineage Differentiation

The RM Y-STR set applied here generated exceptionally high haplotype diversity in the worldwide sample set analyzed with 12,156 unique (i.e., not matching any other individual in this dataset) haplotypes observed in 12,272 unrelated samples from 111 worldwide populations (Table 1). The global haplotype diversity estimate was 0.9999985 (sampling variance = 2.85 × 10⁻⁸). Of the 102 haplotypes that were nonunique within this dataset (i.e., matched other individuals in this dataset), 90 were shared between two males, 10 between three males, and two between four males. All nonunique haplotypes were shared between males of the same geographic region, and all but six were from the same sampling population. The six nonunique haplotypes shared between different populations involved males from Argentina (Chubut and Rio Negro regions), Greenland (from two separate samples of Inuit), Czech Republic (from two sample sets), Cologne and Warsaw, and Bhutan (the Lhokpu and Mónpa tribes), respectively. As a result, the proportion of haplotypes shared between populations was exceptionally low at 0.05%, and the proportion of haplotypes shared within populations was slightly higher at 0.8%.

Within each regional geographic group (see Table 1), similar high levels of haplotype diversity were obtained ranging from 0.9999988 and 0.999996 in Europeans and Migrant Europeans (i.e., European populations sampled in North America and Australia), respectively, down to 0.99836 in Central Asians (although the sample size of the latter group was much lower than that of all other regional groups studied) (Table 1). The average number of differing loci was highest with 17.5 loci in Admixed Native Americans and South Asians, respectively, as well as with 17.4 loci in Native Americans and Europeans, respectively, and was lowest with 15.26 loci in Middle Easterners (Table 1). As the maximal possible number is 21, these results illustrate how strikingly different the haplotypes were within each of the regional groups.

At the population level (Supp. Table S4), similar high levels of haplotype diversity were observed across all populations tested. Of
the 111 populations, 67 (60%) displayed haplotype diversities of 1.0, meaning that every individual tested per population (sample size ranging from seven to 467 in the various populations) had a different haplotype. The remaining populations had haplotype diversities ranging from 0.972 (Wichi, Salta Province, Argentina) to 0.999994 (Cologne, population 2).

Due to its highly multicopy nature, DYF403S1a-b caused concerns with some study group members regarding genotyping accuracy, so that additional quality control was performed for this marker (as well as for DYF399S1), as described in Materials and Methods section. There are four separate loci at the tetranucleotide DYF403S1a-b, three of which overlap in size. Differences in repeat sequences between the copies lead to the presence of partial (0.1, 0.2, and 0.3) alleles, which can challenge interpretation even in single source samples due to a lack of single base resolution with POP-4 during capillary electrophoresis. To determine the effect of removing this potentially problematic marker from the RM Y-STR panel, the entire dataset was additionally analyzed without DYF403S1a-b. The global haplotype diversity slightly decreased from 0.9999985 with the full 13-loci RM Y-STR set to 0.9999981 with the 12-loci set. This translated to an increase in the number of nonunique haplotypes from 116 to 152 in the 12,272 samples—a decrease of zero haplotypes within each of the 111 populations. Populations were exceptionally low (Table 1; Supp. Table S4), indicating relatively little population substructure detected with the RM Y-STR set. Indeed, AMOVA based on FST values (the inability to accurately assign the alleles of the multicopy markers to specific loci prevented FST values being calculated) demonstrated that 99.98% of haplotype variation was within populations, extremely high for Y-chromosome markers [Willuweit and Roewer, 2007], whereas 0.02% was among populations within the same regional group, and only 0.002% among regional groups. The global FST was as low at 0.00017. Between the regional groups, the average pairwise FST value was only 0.000127, with a maximum value of 0.00058 observed between Aboriginal Australians and Middle Eastern populations. Even when considering individual populations, the highest pairwise FST value observed was only 0.02815 between the Bhutan Lhokpu and Argentinian Wichi from Salta Province (data not shown). Across all population comparisons, an average pairwise FST value of 0.000826 was obtained.

To illustrate the magnitude of haplotypic differences between populations and geographic areas, multidimensional scaling (MDS) analysis was performed on Slatkin’s linearized FST values (Fig. 1). The majority of the populations formed a loose cluster, with moderate dispersion across both dimensions. While there were several outlier populations (for example Pakistani Punjabis, Japanese Gunma, and Angola Kimbundos), overall there was no geographic pattern emerging, in line with the AMOVA results. It is notable that populations from the same geographic group did not necessarily cluster together (Fig. 1). The differences in population sample size (Supp. Table S4) did not greatly affect the dispersion of populations—similar distributions were observed when the population sizes were restricted to 100/population (inset, Fig. 1), or 25/population (data not shown). This analysis, together with the AMOVA results, illustrates that no genetic–geographic structuring is detected with this set of RM Y-STR haplotypes, since their high mutation rates have likely removed signals of shared population history, and has driven the high number of unique haplotypes. The low between-population differentiation and lack of substructure could be expected as a result of the high within-population diversities, and the effect of mutation rates on the estimation of θ [Meirmans and Hedrick, 2011]. As a practical consequence, no population substructure correction needs to be applied when using this RM Y-STR set in forensic or other applications, as is usually needed for other DNA marker systems. However, the lack of substructure correction required does not necessarily infer a lack of structure in the distribution of

### Table 1. RM Y-STR Haplotype Characteristics in a Global Set of 12,272 Individuals from 111 Populations Summarized for Regional Groups

<table>
<thead>
<tr>
<th>Group</th>
<th>Individuals</th>
<th>Populations</th>
<th>Haplotypes</th>
<th>Haplotype diversity</th>
<th>θ</th>
<th>Average number of differing loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sub-Saharan Africa</td>
<td>303</td>
<td>5</td>
<td>300</td>
<td>0.99993</td>
<td>0.00006</td>
<td>16.5</td>
</tr>
<tr>
<td>North Africa</td>
<td>452</td>
<td>4</td>
<td>445</td>
<td>0.99992</td>
<td>0.00010</td>
<td>16.2</td>
</tr>
<tr>
<td>Middle East</td>
<td>100</td>
<td>1</td>
<td>100</td>
<td>1.0</td>
<td>0</td>
<td>15.26</td>
</tr>
<tr>
<td>Central Asia</td>
<td>86</td>
<td>1</td>
<td>80</td>
<td>0.99836</td>
<td>0.0012</td>
<td>16.18</td>
</tr>
<tr>
<td>South Asia</td>
<td>661</td>
<td>8</td>
<td>644</td>
<td>0.99992</td>
<td>0.00009</td>
<td>17.5</td>
</tr>
<tr>
<td>East Asia</td>
<td>967</td>
<td>7</td>
<td>964</td>
<td>0.999994</td>
<td>0.000006</td>
<td>17.3</td>
</tr>
<tr>
<td>South East Asia</td>
<td>634</td>
<td>6</td>
<td>630</td>
<td>0.99998</td>
<td>0.00002</td>
<td>16.6</td>
</tr>
<tr>
<td>Aboriginal Australian</td>
<td>100</td>
<td>1</td>
<td>96</td>
<td>0.99919</td>
<td>0.00069</td>
<td>16.8</td>
</tr>
<tr>
<td>Native American</td>
<td>365</td>
<td>10</td>
<td>357</td>
<td>0.99986</td>
<td>0.00015</td>
<td>17.4</td>
</tr>
<tr>
<td>Admixed Native American</td>
<td>764</td>
<td>12</td>
<td>758</td>
<td>0.99998</td>
<td>0.00002</td>
<td>17.5</td>
</tr>
<tr>
<td>European</td>
<td>5,618</td>
<td>38</td>
<td>5599</td>
<td>0.9999988</td>
<td>0.000001</td>
<td>17.4</td>
</tr>
<tr>
<td>Migrant Sub-Saharan Africa</td>
<td>663</td>
<td>5</td>
<td>659</td>
<td>0.99998</td>
<td>0.000004</td>
<td>16.3</td>
</tr>
<tr>
<td>Migrant European</td>
<td>731</td>
<td>3</td>
<td>730</td>
<td>0.999996</td>
<td>0.000004</td>
<td>17.0</td>
</tr>
<tr>
<td>Migrant Asian</td>
<td>649</td>
<td>7</td>
<td>645</td>
<td>0.99998</td>
<td>0.00002</td>
<td>16.7</td>
</tr>
<tr>
<td>Bhutan</td>
<td>78</td>
<td>2</td>
<td>56</td>
<td>0.99434</td>
<td>0.00794</td>
<td>13.9</td>
</tr>
<tr>
<td>Baka Pygmy</td>
<td>101</td>
<td>1</td>
<td>94</td>
<td>0.99822</td>
<td>0.00226</td>
<td>17.5</td>
</tr>
<tr>
<td>Global</td>
<td>12,272</td>
<td>111</td>
<td>12,156</td>
<td>0.9999985</td>
<td>0.0000238</td>
<td>16.5</td>
</tr>
</tbody>
</table>

* Bhutan and Biaka Pygmies were not assigned to a larger geographic group as the individuals sampled were ascertained differently (Bhutan), or are known to have different population characteristics to all other study populations (Biaka, see Materials and Methods and DNA Samples).
Figure 1. Two-dimensional plot of MDS analysis of Slatkin’s linearized $F_{ST}$ values for RM Y-STR haplotypes in a global sample of 12,272 individuals from 111 populations (stress 0.07462). Smaller inset MDS shows the effect of equalized sample size ($N = 100$ individuals or fewer per population, stress = 0.02416).

haplotypes across geographic regions. Nonrandom distributions, caused by shared population histories and common origins, would infer the need to generate and utilize regional or meta-population-specific databases for frequency estimation. As such, the ability for RM Y-STRs to differentiate between male lineage within and between different populations and the relative distribution of haplotypes between regional groups was investigated.

Value of RM Y-STRs in Improving Male Lineage Differentiation

When considering the utility of the RM Y-STR set for future applications in forensic, genealogical or anthropological genetic studies, it is informative to compare its properties to those of conventional Y-STRs, such as the 17 loci included in the Yfiler kit. Hence, RM YSTR and Yfiler haplotypes were compiled for 7,784 individuals across 65 populations as a subset of the global dataset presented here. In every diversity measure applied, the 13 RM YSTR set provided enhanced estimates relative to the 17-loci Yfiler set. Global haplotype diversity was increased from 0.99995 with Yfiler to 0.999997 with RM Y-STRs; the total number of haplotypes detected was increased from 6,975 to 7,714, and the number of singletons from 6,469 to 7,647. All regional groups showed more haplotypes and higher diversity estimates with RM Y-STRs relative to Yfiler, that is, haplotype diversity increased on average by 0.00226 and the number of singletons on average by 77 (largest increase of 1.56-fold in Aboriginal Australians) (Table 2). At the population level, there were increases in the number of haplotypes and in the haplotype diversity estimates for 56 (86.2%) of the 65 populations with the RM Y-STR set compared with Yfiler, whereas for six (9.2%) populations the same number of haplotypes were obtained because all males were already fully individualized with Yfiler (often in small sample-sized populations), and for three populations (4.6%), one haplotype fewer was detected with the RM Y-STRs than with Yfiler (Supp. Table S5). There was a significant difference between the RM Y-STR and the Yfiler sets in the level of haplotype sharing between individuals within regional groups, between populations, and within populations (Fig. 2). At the global level, 506 Yfiler haplotypes were shared between 1,318 individuals, compared with only 66 RM
Y-STR haplotypes shared between 70 individuals. For Yfiler at the population level, only three (4.6%) of the 65 populations did not display any matching individuals either within or outside populations; 58 (89.2%) and 52 (80.0%), respectively, showed within and outside population haplotype matches. Within- and outside-population haplotype matches for Yfiler were as high as 62%, for the Wichis—a Native American group from Argentina (although sample size was N = 13). Notably, the level of outside-population matches compared with within-population matches was markedly higher for most populations for Yfiler haplotypes, whereas strikingly reduced for RM Y-STRs (Fig. 2). Of the 65 populations, 35 (53.8%) did not show any haplotype matches within and outside populations. Only 29 (44.6%) of the 65 populations displayed haplotype matches within populations, with the highest proportion (11%) observed in Argentinian Wichis from the Salta province. For 49 of the 58 populations that showed within-population Yfiler haplotype matches, no RM Y-STR haplotype matches were observed. The reduction in haplotype matches for RM Y-STRs relative to Yfiler is even more striking when considering outside-population matches. Only four (6.2%) populations displayed outside-population RM Y-STR matches, namely, two Czech population samples (one haplotype), and Cologne and Warsaw (one haplotype). The finding that 61 populations (93.8%) showed no outside population haplotype matches for RM Y-STRs contrasts strongly with those found with Yfiler, for which the corresponding number is 13 (20%).

Figure 3 provides an overview of the population pairs that showed between population haplotype matches for Yfiler (Fig. 3A, blue lines) and RM Y-STRs (Fig. 3B, blue lines). The significant reduction in between-population haplotype sharing (τwu = 8.091, $P = 2.23 \times 10^{-11}$) demonstrates the power of the RM Y-STR panel in male lineage differentiation. For RM Y-STRs, no haplotype sharing was observed between different populations sampled from the same countries, such as the five Indian, two Japanese, three Korean, three Austrian, two Belgian, three German, two Italian, two Polish, three US American, and two Hungarian populations, except for the two Czech populations. Further, almost no RM Y-STR haplotype sharing was observed between populations from the same geographic region such as the densely sampled continent of Europe except for Cologne and Warsaw. In contrast, many Yfiler haplotypes were shared between populations within countries, and between populations within regions such as Europe (see Fig. 3, insets). The lack of RM Y-STR haplotype sharing between populations is not limited to the full 13-loci haplotype. Network analysis of 1,000 individuals, selected at random in the same population proportions as the full set of 7,784 samples and based on single-copy Y-STRs, displayed little geographic clustering with RM Y-STR haplotypes (Fig. 4B, perhaps with the exception of North Africans), whereas with Yfiler haplotypes for the same individuals (Fig. 4A), geographic clustering is seen with several groups. For RM Y-STRs, the network was constructed based on nine single-copy Y-STRs (excluding DYS387S1, DYS399S1, DYF403S1, and DYF404S1), and for Yfiler, based on 15 Y-STRs (excluding DYS385a/b). Unlike the Yfiler network, the RM Y-STR network provides almost no clustering of haplotypes according to geographic regions of sampling (except some of the North African haplotypes), which underlines the lack of population structure as also seen in the MDS and AMOVA of the complete data. Hence, for RM Y-STRs, the need for regional (metapopulation) reference databases for haplotype frequency estimation in forensic and other applications is strongly reduced compared with conventional Y-STRs such as those in the Yfiler kit. Combining the RM Y-STR set and the Yfiler set to generate 30-marker Y-STR haplotypes resulted in the individualization of 25 additional men of the 137 not already individualized with the RM Y-STR set alone. While this does represent a slight improvement on the global scale, from 7,714 to 7,737 haplotypes, it is clear that the vast majority of differentiation was achieved by the RM Y-STR set alone.

### Ability of RM Y-STRs to Detect Hidden Paternal Relationships

The observation of RM Y-STR haplotype sharing within, but (almost) not between, populations might indicate that RM Y-STRs allow the detection of unknown (i.e., hidden) paternal relationships in populations. If true, more shared RM Y-STR haplotypes would be expected between putatively unrelated individuals sampled from rural areas, where male relatives often stay in the region (and thus can be sampled), and especially from endogamous populations, than among putatively unrelated individuals sampled from urban areas, where male relatives tend to not to stay (and thus cannot be sampled). To test this hypothesis, all populations were designated

<table>
<thead>
<tr>
<th>Group</th>
<th>Individuals</th>
<th>Populations</th>
<th>Yfiler haplotypes</th>
<th>RM Y-STR haplotypes</th>
<th>Yfiler haplotype diversity</th>
<th>RM Y-STR haplotype diversity</th>
<th>Yfiler average percentage of differing loci</th>
<th>RM Y-STR average percentage of differing loci</th>
</tr>
</thead>
<tbody>
<tr>
<td>North Africa</td>
<td>193</td>
<td>2</td>
<td>173</td>
<td>189</td>
<td>0.99792</td>
<td>0.99973</td>
<td>59%</td>
<td>73%</td>
</tr>
<tr>
<td>Central Asia</td>
<td>83</td>
<td>1</td>
<td>67</td>
<td>77</td>
<td>0.99860</td>
<td>0.99824</td>
<td>63%</td>
<td>77%</td>
</tr>
<tr>
<td>South Asia</td>
<td>497</td>
<td>6</td>
<td>450</td>
<td>483</td>
<td>0.99951</td>
<td>0.99988</td>
<td>68%</td>
<td>83%</td>
</tr>
<tr>
<td>East Asia</td>
<td>633</td>
<td>5</td>
<td>580</td>
<td>630</td>
<td>0.99952</td>
<td>0.99999</td>
<td>64%</td>
<td>82%</td>
</tr>
<tr>
<td>South East Asia</td>
<td>200</td>
<td>2</td>
<td>175</td>
<td>198</td>
<td>0.99759</td>
<td>0.99900</td>
<td>66%</td>
<td>89%</td>
</tr>
<tr>
<td>Aboriginal Australian</td>
<td>100</td>
<td>1</td>
<td>74</td>
<td>96</td>
<td>0.99152</td>
<td>0.99919</td>
<td>68%</td>
<td>89%</td>
</tr>
<tr>
<td>Native American</td>
<td>279</td>
<td>8</td>
<td>233</td>
<td>275</td>
<td>0.99761</td>
<td>0.99988</td>
<td>61%</td>
<td>82%</td>
</tr>
<tr>
<td>Admixed Native American</td>
<td>458</td>
<td>6</td>
<td>444</td>
<td>454</td>
<td>0.99886</td>
<td>0.99996</td>
<td>67%</td>
<td>81%</td>
</tr>
<tr>
<td>European</td>
<td>4,041</td>
<td>25</td>
<td>3,696</td>
<td>4,025</td>
<td>0.99991</td>
<td>0.999998</td>
<td>66%</td>
<td>82%</td>
</tr>
<tr>
<td>Migrant Sub-Saharan Africa</td>
<td>442</td>
<td>3</td>
<td>407</td>
<td>439</td>
<td>0.99909</td>
<td>0.99997</td>
<td>63%</td>
<td>81%</td>
</tr>
<tr>
<td>Migrant European</td>
<td>552</td>
<td>3</td>
<td>541</td>
<td>551</td>
<td>0.99993</td>
<td>0.99999</td>
<td>63%</td>
<td>81%</td>
</tr>
<tr>
<td>Migrant Asian</td>
<td>205</td>
<td>2</td>
<td>176</td>
<td>203</td>
<td>0.99737</td>
<td>0.99990</td>
<td>66%</td>
<td>83%</td>
</tr>
<tr>
<td>Blaka Pygmies</td>
<td>101</td>
<td>1</td>
<td>83</td>
<td>94</td>
<td>0.99505</td>
<td>0.99822</td>
<td>65%</td>
<td>83%</td>
</tr>
<tr>
<td>Global</td>
<td>7,784</td>
<td>65</td>
<td>6,975</td>
<td>7,714</td>
<td>0.99995</td>
<td>0.999997</td>
<td>65%</td>
<td>81%</td>
</tr>
</tbody>
</table>
Figure 2. Proportion of individuals with haplotypes shared within populations (right) and between populations (left) for Yfiler (light blue bars) and the RM Y-STR set (dark red bars).
according to their urbanization status—each set of males was classified as either urban (genetically moderately outbred), rural (more likely to be genetically inbred), mixed (males from both rural and urban populations, in unspecified proportions), or endogamous (populations subscribing to the cultural practice of endogamy—restricting marriage outside of an individual’s social class, ethnic group, or tribe). In support of the RM Y-STR set’s ability to detect hidden paternal relationships, we indeed see significantly lower haplotype diversity and lower proportions of unique haplotypes in the endogamous group than in the rural group than in the urban group (diversity $t_2 = 7,960, P = 1.6 \times 10^{-8}$, shared haplotypes $t_2 = 1.347, P = 0.310$) (Table 3). Although the sample sizes differ considerably between the three categories, this does not affect the conclusions as simulations performed with equalized sample numbers ($n = 558$ per category) showed that the RM Y-STRs still result in decreased diversity in endogamous and rural populations compared with urban populations (data not shown; diversity $t_2 = 8936, P = 1.3 \times 10^{-8}$, shared haplotypes $t_2 = 1.117, P = 0.380$).

On the other hand, these data also demonstrate that even in populations with considerable hidden paternal relationships, such as in endogamous populations, the RM Y-STR set still allows the differentiation of a large number of men. A comparison between Yfiler and the RM Y-STR set for the available populations (Table 4) illustrates the improved value of RM Y-STRs relative to Yfiler in differentiating males when grouping populations into urban, rural, and endogamous. Most tellingly, even in the endogamous group, the haplotype diversity increased and the proportion of shared (i.e., nonunique) haplotypes decreased from 0.99947 and 0.101 with Yfiler to 0.99978 and 0.045 with the RM Y-STR set, respectively, in the rural group from 0.99916 and 0.157 to 0.99997 and 0.008, and even in the urban from 0.99994 and 0.08 to 0.999998 and 0.005.

**Value of RM Y-STRs for Male Relative Differentiation**

In addition, we tested the value of RM Y-STRs for differentiating male relatives. On a theoretical level, the rate of relative differentiation per meiosis can be calculated using the mutation rates of each locus within a genotyping panel (see Materials and Methods section). For the Yfiler set, it is estimated that a mutation at one or more of the 17 Y-STR loci will occur with a probability of 0.047 (95% confidence interval 0.038–0.057) per meiosis, which for the RM Y-STR set is more than fourfold higher at 0.195 (95% CI 0.177–0.21). Notably, for the recently released PowerPlex Y-23 kit (Promega), which targets all 17 Yfiler markers together with two of the 13 RM Y-STR

---

**Figure 3.** Geographic representation of pairwise between-population haplotype sharing. Blue lines connect population pairs showing shared haplotypes for A: Yfiler, and B: the RM Y-STRs set. Smaller insets show Europe enlarged.

**Figure 4.** Weighted median-joining networks from single-copy Y-STRs for 1,000 individuals randomly selected from the total dataset, with regional geographic groups indicated by colors (see inset legend). A: Based on 15-loci Yfiler haplotypes (excluding DYS385a/b). B: Based on 10-loci RM Y-STR haplotypes (excluding DYF399S1, DYF403S1, and DYF404S1).
Table 3. RM Y-STR Haplotype Characteristics in a Global Set of 12,272 Individuals from 111 Populations Summarized for Groups Based on Urbanization Status

<table>
<thead>
<tr>
<th>Population classification</th>
<th>Populations</th>
<th>Individuals</th>
<th>Haplotypes</th>
<th>Singletons</th>
<th>Haplotype diversity</th>
<th>Proportion of nonunique haplotypes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endogamous</td>
<td>7</td>
<td>558</td>
<td>512</td>
<td>478</td>
<td>0.999614</td>
<td>0.143</td>
</tr>
<tr>
<td>Rural</td>
<td>20</td>
<td>1,369</td>
<td>1,354</td>
<td>1,340</td>
<td>0.999982</td>
<td>0.021</td>
</tr>
<tr>
<td>Urban</td>
<td>51</td>
<td>7,198</td>
<td>7,164</td>
<td>7,131</td>
<td>0.999999</td>
<td>0.009</td>
</tr>
<tr>
<td>Mixed</td>
<td>26</td>
<td>2,379</td>
<td>2,361</td>
<td>2,342</td>
<td>0.999994</td>
<td>0.015</td>
</tr>
<tr>
<td>Unknown</td>
<td>7</td>
<td>770</td>
<td>767</td>
<td>764</td>
<td>0.999990</td>
<td>0.008</td>
</tr>
</tbody>
</table>

Table 4. Comparison of RM Y-STRs and Yfiler Haplotype Characteristics in a Global Set of 7,784 Individuals from 65 Populations Summarized for Groups Based on Urbanization Status

<table>
<thead>
<tr>
<th>Population classification</th>
<th>N</th>
<th>Yfiler ht</th>
<th>RM Y-STR ht</th>
<th>Yfiler singletons</th>
<th>RM Y-STR singletons</th>
<th>Yfiler haplotype diversity</th>
<th>RM Y-STR haplotype diversity</th>
<th>Yfiler proportion of nonunique haplotypes</th>
<th>RM Y-STR proportion of nonunique haplotypes</th>
<th>Yfiler percentage average allelic difference</th>
<th>RM Y-STR percentage average allelic difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Endogamous</td>
<td>466</td>
<td>419</td>
<td>445</td>
<td>380</td>
<td>427</td>
<td>0.99947</td>
<td>0.99978</td>
<td>0.101</td>
<td>0.101</td>
<td>68%</td>
<td>83%</td>
</tr>
<tr>
<td>Rural</td>
<td>714</td>
<td>602</td>
<td>708</td>
<td>537</td>
<td>703</td>
<td>0.99916</td>
<td>0.99997</td>
<td>0.157</td>
<td>0.157</td>
<td>70%</td>
<td>85%</td>
</tr>
<tr>
<td>Urban</td>
<td>4,974</td>
<td>4,577</td>
<td>4,951</td>
<td>4,318</td>
<td>4,929</td>
<td>0.99994</td>
<td>0.99998</td>
<td>0.080</td>
<td>0.080</td>
<td>68%</td>
<td>83%</td>
</tr>
<tr>
<td>Mixed</td>
<td>1,422</td>
<td>1,332</td>
<td>1,407</td>
<td>1,267</td>
<td>1,407</td>
<td>0.99988</td>
<td>0.99999</td>
<td>0.063</td>
<td>0.063</td>
<td>67%</td>
<td>84%</td>
</tr>
<tr>
<td>Unknown</td>
<td>208</td>
<td>181</td>
<td>205</td>
<td>169</td>
<td>202</td>
<td>0.99684</td>
<td>0.99986</td>
<td>0.014</td>
<td>0.014</td>
<td>64%</td>
<td>83%</td>
</tr>
</tbody>
</table>

Figure 5. Empirical male relative differentiation using 2,528 paternal relative pairs separated by one to 20 meioses for Yfiler (light blue bars) and the RM Y-STR set (dark red bars). The data combine the new 2,372 relative pairs from the current study with the previously used 156 relative pairs [Ballantyne et al., 2012]. Values above the bars indicate the absolute number of relative pairs the estimated differentiation rate is based upon. Error bars represent binomial confidence intervals.

markers and four additional Y-STRs, this value is 0.092 (95% CI 0.077–0.107), nearly twofold higher than for Yfiler, but more than twofold lower than for the RM Y-STR set studied here.

To compare the theoretical expectation for the RM Y-STR and the Yfiler sets with empirical data, we genotyped 2,372 pairs of male relatives previously confirmed by autosomal DNA analysis (2,339 father–son pairs, 30 brother pairs, and three uncle–nephew pairs), confirmed the observed allelic differences by independent genotyping (and some by sequencing), and combined these new data with those from the 156 relative pairs separated by one to 20 paternal genetic transfers (or meioses) described previously [Ballantyne et al., 2012]. In this combined dataset, the RM Y-STR set allows differentiation by at least one allelic difference (i.e., mutation) in at least one locus in 742 (29%) of the total 2,528 pairs related by one to 20 generations, whereas Yfiler only allowed the separation of 118 (5.5%) of a subset of 2,161 of these relative pairs (Fig. 5). In particular, the RM Y-STR set differentiated fathers from their sons in 26.9% of the cases versus 4.5% with Yfiler, and brothers from each other in 56.3% of the cases versus 10.0% with Yfiler.

It should be noted, however, that only for the 2,378 father–son pairs investigated here, was the sample size large enough to allow reliable conclusions, whereas for more distantly related males, future studies need to deliver more data. However, as the chance of mutations increases with each meiosis, it is expected that the more distantly related men are, the higher the chance that they will have different RM Y-STR haplotype; a trend that can be seen in our data (Fig. 5). The importance of sufficient sample size is illustrated in the following for father–son pairs. Our initial estimate of
father–son differentiation was 70% based on only 20 pairs [Ballantyne et al., 2010], which was subsequently revised to 49% based on 39 pairs, including the initial 20 pairs [Ballantyne et al., 2012]. In the present study, we managed to increase the number of father–son pairs drastically to 2,378 pairs, now achieving a differentiation rate of 26.9% (95% confidence interval 25.1%–28.7%). Notably, the empirical rate achieved, now based on a large sample size, corresponds more closely to the theoretical estimate of 19.5%.

Removing DYF403S1a+b from the RM Y-STR panel resulted in the nondifferentiation of 84 relative pairs that could be differentiated when DYF403S1a+b was included; the overall relative differentiation rate dropped to 26%. However, the difference only occurred in pairs with fewer than five meioses separating the individuals—in all other pairs, there were sufficient mutations at other RM Y-STR loci to allow differentiation (data not shown).

As pointed out previously [Ballantyne et al., 2012], RM Y-STRs are not particularly useful for paternity testing or familial searching because of their high mutation rates and the resultant power of differentiation of male relatives. If, however, they are used for such purposes, correction factors would be needed to compensate for the likely observed mutations in the probability calculations. It has been suggested before that the presence of at least three mutations at any Y-STR loci (most commonly with the 17 Yfiler Y-STRs) is sufficient to exclude paternity [Kayser and Sajantila 2001; Gjertson et al., 2007]. In the current dataset however, 3.8% of the DNA-confirmed father–son pairs displayed three or four RM Y-STR mutations; as such, the number of mutations constituting exclusion should be increased when using the RM Y-STRs. As outlined elsewhere (Ballantyne et al., 2012), instead of an ad hoc cut-off, a dynamic threshold shall be used, estimated from the number of Y-STR loci genotyped and their locus-specific mutation rates. Because three mutations have also been observed occasionally among 17 Yfiler loci in DNA-confirmed father–son pairs [Goedbloed et al. 2009], this notion also applies to conventional Y-STR sets such as Yfiler, which due to their moderate mutation rates in principle are more appropriate for paternity and kinship testing than RM Y-STRs.

Conclusions

The large worldwide dataset compiled, presented, and analyzed here demonstrates the exceptional value of the RM Y-STR panel for differentiating male lineages on a global and regional scale. Many of the current limitations of Y-chromosome applications are reduced with this RM Y-STR set, providing increased utility and effectiveness to the genetic analysis of male populations and lineages. In particular, this 13 RM Y-STR set provides near-complete paternal lineage differentiation in general populations as well as in populations with otherwise reduced Y-chromosome diversity, due to peculiarities in population history or cultural practices. The effects of this near-complete male individualization will be of great benefit to numerous fields using the Y-chromosome genetics to investigate male lineages, such as in genealogical studies (e.g., to detect extrapair paternity or adventitious haplotype matches within different surnames/lineages), population genetic and genetic history studies (e.g., to assist in differentiating between lineages shared between populations with common history), and in forensic applications (e.g., to reduce the inclusion of innocent individuals in investigations due to adventitious haplotype matches). Moreover, the immense value of this RM Y-STR set to differentiate between both close and distant male relatives will have beneficial effects on these same fields, especially in forensic genetics, providing increased confidence that haplotype matches between unknown samples such as those from crime scenes and reference samples such as those from suspects are due to individual identity, rather than relatedness.

On the other hand, however, the extreme degree of RM Y-STR haplotype diversity highlights the limitations of the current approach for placing a statistical weight on Y-STR haplotype matches by using empirically derived haplotype frequency estimates obtained from Y-STR haplotype reference databases. Although already noted with the currently used Y-STR sets such as Yfiler, the problem that an observed haplotype is not present in the large and growing reference database, and how to deal with rare haplotypes in obtaining frequency estimates and estimating match probabilities for which no consensus has been reached yet [Krawczak, 2001; Brenner, 2010; Buckleton et al., 2011; Willuweit et al., 2011], will become drastically increased when using RM Y-STRs, as can be seen from the data presented here. Therefore, new statistical solutions shall be developed to estimate the weight of a RM Y-STR haplotype match with particular relevance for forensic applications. To assist future studies utilizing RM Y-STRs, complete dataset of RM Y-STR and Yfiler haplotypes obtained from the 12,272 individuals analyzed here is made available via Supp. Table S6, with allele frequencies for individual populations, regional groups, and the complete dataset summarized in Supp. Table S7. Relative pair haplotypes for RM Y-STRs and Yfiler are available in Supp. Table S8.

Acknowledgments

The authors thank the following colleagues for their expert technical assistance: Linda S. H. Gan (Singapore) for the Singapore samples, Minttu Hedman (Helsinki) for the Finnish samples, Helena Nilsson (Linköping) for the Swedish samples, Kevin Cloote (Cape Town) for the South African samples, Nancy Vanderheyden (Leuven) for the Belgium samples, Baowen Cheng (Kunming) and Yanan Sun (Kunming) for the Kunming Chinese samples, Juliana Januzzi (Rio de Janeiro) for the Rio Brazil samples, Maria Lourdes D. Ronrado (Quito City) for the Italian samples, and Harald Niederstätter (Innsbruck) for the Innsbruck Austrian samples as well as for his comments on the manuscript. Oscar Lao (Rotterdam) is acknowledged for statistical advice and help in some calculations. Certain commercial equipment, instruments, and materials are identified to specify experimental procedures as completely as possible. In no case does such identification imply a recommendation or endorsement by the National Institute of Standards and Technology (USA), nor does it imply that any of the materials, instruments, or equipment identified are necessarily the best available for the purpose.

Disclosure statement: KNB and MKay, together with two colleagues from Life Technologies, are coinventors on a registered patent about the RM Y-STRs used here and their forensic utility; however, no personal or institutional benefits were received. All other authors declare no conflict of interest.

References


