Forensic Science International: Genetics 7 (2013) 593–600

Contents lists available at SciVerse ScienceDirect

Forensic Science International: Genetics

journal homepage: www.elsevier.com/locate/fsig

Multiple recurrent mutations at four human Y-chromosomal single nucleotide polymorphism sites in a 37 bp sequence tract on the ARSDP1 pseudogene

Harald Niederstätter [a], Burkhard Berger [a], Daniel Erhart [a], Sascha Willuweit [b], Maria Geppert [b], Christoph Gassner [c,d], Harald Schennach [c], Walther Parson [a,e,*], Lutz Roewer [b]

[a] Institute of Legal Medicine, Innsbruck Medical University, Innsbruck, Austria
[b] Department of Forensic Genetics, Institute of Legal Medicine and Forensic Sciences, Charité-Universitätsmedizin, Berlin, Germany
[c] Central Institute for Blood Transfusion & Immunological Department, Innsbruck, Austria
[d] Blood Transfusion Service Zürich, SRC, Schlieren, Switzerland
[e] Penn State Eberly College of Science, University Park, PA, USA

ARTICLE INFO

Article history:
Received 27 February 2013
Received in revised form 22 May 2013
Accepted 23 May 2013

Keywords:
Homoplasy
Gametologues
Gene conversion
Y-SNP
YHRD

ABSTRACT

The male-specific region of the human Y chromosome (MSY) is passed down clonally from father to son and mutation is the single driving force for Y-chromosomal diversification. The geographical distribution of MSY variation is non-random. Therefore, Y-chromosomal single nucleotide polymorphisms (Y-SNPs) are of forensic interest, as they can be utilized, e.g. for deducing the bio-geographical origin of biological evidence. This extra information can complement short tandem repeat data in criminal investigations. For forensic applications, however, any targeted marker has to be unequivocally interpretable.

Here, we report findings for 17 samples from a population study comprising specimens from ~3700 men living in Tyrol (Austria), indicating apparent homoplasic mutations at four Y-SNP loci on haplogroup R-M412/L51/S167, R-U152/S28, and L-M20 Y chromosomes. The affected Y-SNPs P41, P37, L202, and L203 mapped to a 37 bp region on Yq11.21. Observing in multiple phylogenetic contexts up to four homoplasic mutations within such a short sequence tract is unlikely to result from a series of independent parallel mutations. Hence, we rather propose X-to-Y gene conversion as a more likely scenario.

Practical implications arising from markers exhibiting paralogues on the Y chromosome or sites with a high propensity to recurrent mutation for database searches are addressed.

1. Introduction

In 2003 the 23-mega base pair sequence of the euchromatic portions of a human Y chromosome was published by Skaletsky et al. [1]. This physical information resulted in a revised view of the male-specific region of the Y chromosome (MSY) and tremendously improved insights into its organization, function, and evolution. Due to the MSY's clonal mode of intergenerational transmission, mutation is in principle the only driving force behind Y-chromosomal diversification and the slowly mutating binary Y-chromosomal single nucleotide polymorphisms (Y-SNPs) constitute the marker-class with the highest density on the MSY (for review e.g. see [2]). Y-SNPs are usually considered the result of unique single base substitutions (unique event polymorphisms) during human evolution. However, there are exceptions, even when barring "private" mutations. For instance, about 2% of the Y-SNP positions listed in the 2008 update of the Y chromosome consortium (YCC) Y-SNP tree [3] show evidence of recurrent mutations (homoplasies). Further, nearly 3% of the Y-SNPs in the YCC 2008 tree are present in more than just one copy on the MSY.

Sets of uniquely mapped binary Y-SNP markers form stable paternal lineages that can be arranged to robust haplogroups. This enabled the reconstruction of the Y chromosome's evolution by deducing a maximum parsimony haplogroup tree with unified nomenclature from the present-day MSY variation [2–4].

* Corresponding author at: Institute of Legal Medicine, Innsbruck Medical University, Müllerstraße 44, 6020 Innsbruck, Austria. Tel.: +43 512 9003 70651; fax: +43 512 9003 73640.
E-mail address: walther.parson@i-med.ac.at (W. Parson).

1872-4973/$ – see front matter © 2013 Elsevier Ireland Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.fsigen.2013.05.010

The Y-SNPs' low mutation rates make them superb tools to shed light on ancient events in human history and evolution, by analysing modern genetic variation. Furthermore, patterning in present-day MSY diversity is non-random. It can be accounted for by phenomena such as genetic bottlenecks and founder effects, patrilocality, migration, and genetic drift. Because the geographical distribution of MSY variation is non-random, Y-SNPs are also of forensic interest as they can be utilized, e.g. for drawing inferences about the bio-geographical origin or the ethnical background of DNA evidence. When used judiciously, this extra information has the potential to complement short tandem repeat (STR) data in criminal investigations.

For forensic applications special quality requirements apply. Any targeted marker has to be unequivocally interpretable. Rigorous testing is necessary to assure a marker's forensic suitability. In a population study comprising samples from ~3700 men living in Tyrol (Austria), we identified four Y-SNP sites within a 37 bp area of X-degenerate sequence that showed multiple recurrent base substitution events in different phylogenetic contexts. These findings suggest inter-chromosomal X-to-Y gene conversion [5–7] as mutation mechanism rather than a series of independent parallel base substitutions. Practical implications for database searches arising from unreliable markers exhibiting paralogues on the Y chromosome or sites with a high propensity to recurrent mutation will be discussed.

2. Materials and methods

2.1. Ethics statement

This study was approved by the ethics commission of the Innsbruck Medical University (study classification number UN2598, session number 241/4.5). Voluntary blood donors received detailed information regarding the study *ex ante* by regular mail and in a one-to-one interview on-location. All participants gave written consent and personal data (full name, date of birth, and the family's place of residence at the time of the donor's birth). Additionally, they were also asked to provide their father's and paternal grandfather's personal data. A dedicated in-house database (Access 2.0) was used for management of these data and all associated genotyping results.

2.2. Biological samples

Blood samples (~5 ml, $n = 3713$) from healthy West Eurasian men living in Tyrol (Austria) were collected during a series of blood drives organized by the "Central Institute for Blood Transfusion and Immunological Department Innsbruck" in the years 2007 and 2008. This population sample was not filtered for specimens from close paternal relatives. Ninety-two percent ($n = 3416$) of the donors in this study were born in Tyrol.

Total genomic DNA (gDNA) was extracted in the 96-well format from blood-aliquots (200 μl) using the nexttec genomic DNA isolation kit for blood (nexttec Biotechnologie, Leverkusen, Germany) according to the manufacturer's instructions and stored at −80 °C in racked. 2D-barcoded polypropylene tubes (Thermo Fisher Scientific, Hudson, NH, USA) until further use. To minimize the risk of sample mix-up, a Tecan Genesis robot (Tecan Group, Männedorf, Switzerland) was used for automated liquid handling in genotyping experiments.

2.3. Primer design

Previously unpublished primers for polymerase chain reaction (PCR) amplification and Sanger-type sequencing of the phylogenetically informative Y-SNP P37 were designed using the freeware programs Primer-BLAST [8] and Primer3Plus [9]. The sequences of these oligonucleotides can be inferred from Fig. 1. All nucleotide (nt) position numbers (ntps) refer to the February 2009 GRCh37/hg19 human genome reference sequence assembly (http://genome.ucsc.edu).

2.4. SNP genotyping and haplogroup inferences

The phylogeny of all Y-SNPs interrogated in this study is shown in Fig. 2. The inferred haplogroups were designated by the respective single letter label of the basal branch followed by the name of the most derived (i.e. "final") analyzed SNP marker, defining the deepest sub-branch therein.

Asterisks were used to indicate that in the particular haplogroup all analyzed subgroup-defining SNPs were present in the ancestral state.

For haplogroup assignment of the Y chromosomes in this study, in a first round of analysis the 19 phylogenetically informative Y-SNPs M9 (rs3900), M17 (rs3908), M45 (rs2032631), M78, M89 (rs2032652), M96 (rs9306841), M170 (rs2032597), M173 (P241, Page29, rs2032624), M201 (rs2032636), M223, M253 (rs9341296), M269 (rs9786153), M304 (rs13447352), M343 (rs9786184), P15, P37.1/.2, $SRY_{10831.1/.2}$ ($SRY_{1532.1/.2}$, rs2534636), U106 (M405, S21, rs16981293), and U152 (S28, rs1236440) were simultaneously amplified and genotyped by multiplexed PCR and single-nucleotide primer extension (SNPE) as detailed in [10]. Automated Y-SNP haplogroup inferences were achieved by means of a routine implemented in the database that was used for data storage and management.

2.5. Sanger-type sequencing

Further analyses were performed on all samples that yielded with PCR-SNPE results for marker P37 that were inconsistent with the published Y-SNP phylogeny.

2.5.1. Confirmatory sequencing of phylogenetically inconsistent P37 results

On all "affected P37" samples Sanger-type sequencing of the P37 locus and its flanking regions was used to verify the SNPE results. The singleplex PCR cocktails for generating the sequencing templates had a final volume of 20 μl and consisted of 1× KOD buffer (Novagen, Merck, Darmstadt, Germany), 2.5 μg non-acetylated bovine serum albumin (Sigma–Aldrich, St. Louis, MO, USA), 1 mM $MgSO_4$, 200 μM of each dNTP, 0.4 units KOD DNA polymerase (Novagen), and ~4 ng gDNA. PCR primers were used at an assay concentration of 250 nM.

Amplifications employing the primer pair $P37_{XY}F_{seq}$ ↔ ARSDP1-R (455 bp amplicon, Fig. 1) were carried out in conventional thermal cyclers. The cycler protocol comprised an initial denaturation step at 95 °C for 2 min, followed by 32 cycles of denaturation at 95 °C for 15 s, primer-annealing at 70 °C for 30 s, and primer extension at 72 °C for 1 min. The final extension step at 72 °C was extended by 10 min.

Post-PCR treatment of the amplicons, fluorescent dye terminator cycle sequencing of both strands utilizing the amplification primers, product clean-up by size exclusion chromatography, laser induced fluorescence detection capillary electrophoretic separation, and analysis of the raw sequence data were performed as specified in [11]. All sequencing primers were used at an assay-concentration of 160 nM.

2.5.2. Sequencing for refined haplogroup assignments

To attain a more detailed picture of the haplogroup memberships of the "affected P37" samples in haplogroups R-M269*(xU106, U152) and R-U152, Sanger-type sequencing of

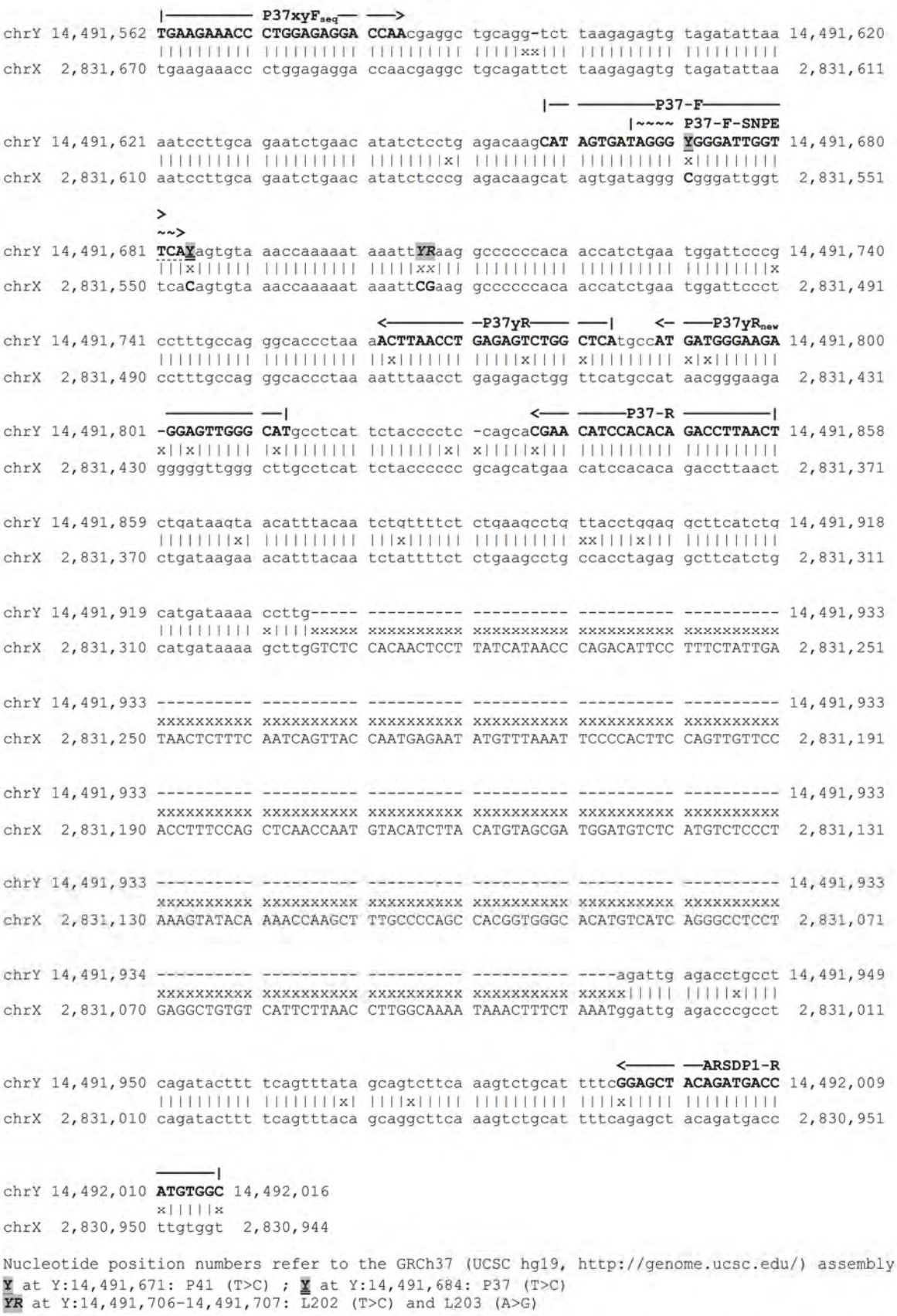

Fig. 1. Sequence alignment. Alignment of the 455 bp *ARSDP1* segment analyzed by confirmatory Sanger-type sequencing and its X-chromosomal gametologue depicting GSVs, positions of known Y-SNPs and primer (binding) sequences.

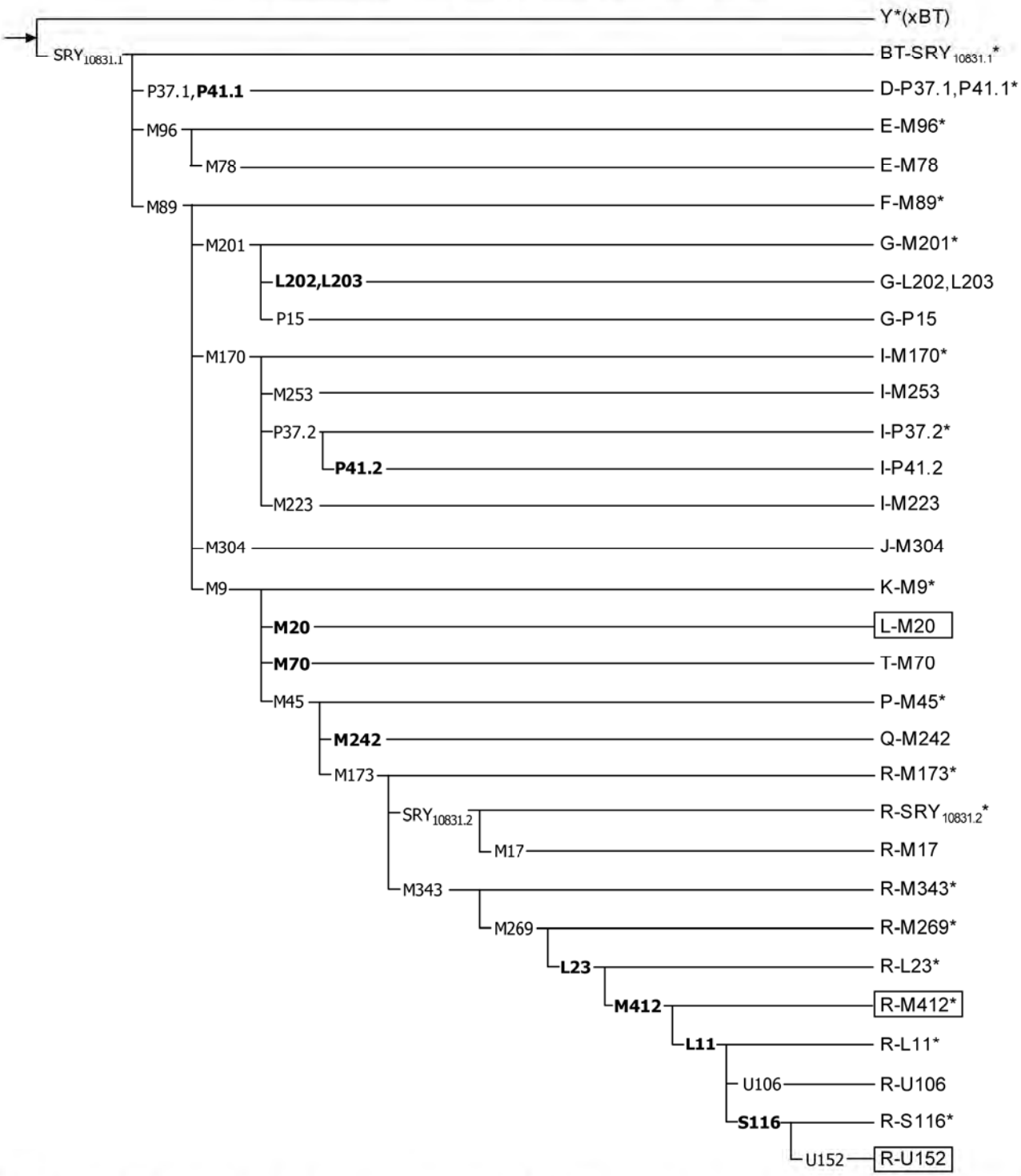

Fig. 2. Y-SNP haplogroup tree. Phylogenetic tree depicting the relationships among the binary Y-chromosomal haplogroups defined by the 29 SNP sites addressed in this study. Markers that were typed by Sanger-type sequencing instead of a multiplexed PCR/single nucleotide primer extension assay are indicated by bold typeface SNP designations. Bounding boxes indicate those haplogroups for which Y-chromosomes with phylogenetically mismatched P37 C allele calls were observed.

PCR products was used to interrogate the Y-SNPs L11 (S127, rs9786076), L23 (S141, rs9785971), M412 (L51, S167, rs9786140), and S116 (P312, rs34276300). Accordingly, for Y chromosomes placed by 19-plex PCR-SNPE in haplogroup K-M9*, the Y-SNPs M20 (rs3911), M70 (rs2032672), and M242 (rs8179021) were analyzed (Fig. 2). For experimental details see [10].

2.6. Y-STR haplotyping

Seventeen-locus Y-STR haplotypes were determined from all "affected P37" samples in this study with the AmpFlSTR Yfiler PCR amplification kit (Life Technologies, Paisley, UK) as described in [12].

Table 1
Genotyping results. Genotypes obtained by confirmatory Sanger-type sequencing, Y-SNP haplogroup assignments, and 17-locus Yfiler haplotypes of the 17 samples showing P37 allelic states discordant with the published Y-SNP phylogeny. Samples originating from a particular Tyrolean region feature identical upper case letter region designations. Matching surnames of the donors are indicated by identical numbers. Y-SNP haplogroup assignments were based either on the results of PCR-SNPE (19 SNP loci) or PCR-SNPE plus Sanger-type sequencing (29 SNP loci).

Sample ID	Paternal relatedness	Tyrolean region	Surname	Sequence variation: nucleotide position on chrY (GRCh37/hg19)										
				14,491,596.1	14,491,597	14,491,649 P53.1, P53.2	14,491,671 P41.1, P41.2	14,491,684 P37.1, P37.2	14,491,706 L202	14,491,707 L203	14,491,740	14,491,763	14,491,776	14,491, 781
chrY	n/a	n/a	n/a	–	G	T>C	T>C	T>C	T>C	A>G	G	C	T	C
chrX	n/a	n/a	n/a	A	T	C	C	C	C	G	T	T	A	T
1	None known	A	1	–	G	T	T	C	T	A	G	C	T	C
2	None known	B	2	–	G	T	T	C	T	A	G	C	T	C
3, 4	Brothers	B	3	–	G	T	T	C	T	A	G	C	T	C
5, 6	Brothers	B	3	–	G	T	T	C	T	A	G	C	T	C
7, 8	Father–son	B	3	–	G	T	T	C	T	A	G	C	T	C
9, 10	Cousins	B	3	–	G	T	T	C	T	A	G	C	T	C
11	None known	B	3	–	G	T	T	C	T	A	G	C	T	C
12	None known	B	3	–	G	T	T	C	T	A	G	C	T	C
13	None known	B	3	–	G	T	T	C	T	A	G	C	T	C
14	None known	C	4	–	G	T	T	C	C	G	G	C	T	C
15	None known	D	5	–	G	T	T	C	C	G	G	C	T	C
16	None known	E	6	–	G	T	T	C	C	G	G	C	T	C
17	None known	F	7	–	G	T	C	C	C	G	G	C	T	C
Control male 1	n/a	n/a	n/a	–	G	T	T	T	T	A	G	C	T	C
Control male 2	n/a	n/a	n/a	–	G	T	T	T	T	A	G	C	T	C
Control male 3	n/a	n/a	n/a	–	G	T	T	T	T	A	G	C	T	C
Control female 1	n/a	n/a	n/a	noSeq	noSeq	noSeq	noSeq	noSeq	noSeq	noSeq	noSeq	noSeq	noSeq	noSeq

Sample ID	Y-SNP haplogroup		Y-STR genotype (YHRD order)
	19 loci	29 loci	
chrY	n/a	n/a	n/a
chrX	n/a	n/a	n/a
1	R-M269*	R-M412*	14 13 28 24 10 13 13 11,14 12 12 15 19 16 17 23 12
2	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 15 18 23 12
3, 4	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 15 18 23 12
5, 6	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 15 18 23 12
7, 8	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 15 18 23 12
9, 10	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 15 18 23 12
11	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 15 18 23 12
12	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 12 15 19 16 18 23 12
13	R-M269*	R-M412*	14 13 28 24 11 13 13 11,14 12 11 15 19 15 18 23 12
14	R-U152	R-U152	15 13 29 24 11 13 13 11,14 12 12 14 19 16 16 25 12
15	R-U152	R-U152	15 13 29 24 11 13 13 11,15 12 12 14 19 17 15 25 12
16	R-U152	R-U152	15 13 29 24 10 13 13 11,14 12 11 14 19 16 16 25 12
17	K-M9*	L-M20	14 14 30 23 10 13 12 17,17 10 11 16 19 15 16 20 11
Control male 1	R-U152	n/a	14 13 29 24 11 13 13 11,14 12 11 15 19 15 18 23 12
Control male 2	E-M78	n/a	13 13 30 25 10 11 13 16,18 08 12 14 20 17 15 23 13
Control male 3	I-M253	n/a	14 12 28 23 10 11 13 13,13 10 10 16 20 14 15 23 10
Control female 1	n/a	n/a	n/a

n/a, not analyzed, not applicable; noSeq, no analyzable sequencing trace was obtained.

3. Results and discussion

Using a 19-plex PCR-SNPE approach we obtained Y-SNP haplogroup assignments for 3713 men living in Tyrol (Table S1, part of this dataset has been published previously [10,13]). Notably, in 17 instances we observed the derived P37 allele although the base states at the other 18 phylogenetically informative markers in the assay convincingly indicated haplogroup inferences other than D-P37.1 or I-P37.2 (Fig. 2).

One of these Y chromosomes belonged to haplogroup K-M9*, three were found in R-U152, and the remaining 13 samples featured M269 as the most derived Y-SNP included in the PCR-SNPE assay.

A possible explanation for these inconsistencies with the published Y-SNP phylogeny [3] might arise from the observation that the Y-chromosomal SNP P37 (T > C) lies on a region of the Y chromosome that has a gametologue (i.e. a similar sequence on the opposite sex chromosome [14]) on the minus-strand of the X chromosome. The gametologous sequence variant (GSV; X:2,831,547) of the P37 site (Y:14,491,684) resembles the derived Y-chromosomal SNP allele (Fig. 1). This can cause problems as became evident in initial PCR-SNPE experiments.

When using primers P37-F [10] and P37-R (Fig. 1) for amplification, PCR-SNPE analyses of samples known to carry the ancestral P37 T-allele resulted in base calls for both the ancestral Y-chromosomal and the apparently derived – yet actually X-chromosomal – P37 allelic states. Evidently, the central T:G mismatch of primer P37-F with its X-chromosomal binding sequence (Fig. 1) did not exert a large enough negative effect on the primer hybridization kinetics to prevent PCR success. Furthermore, the single 3′ terminal G:T mismatch of primer P37-R with the binding site on its X-chromosomal gametologue was not likely to significantly impair primer extensibility, as it was subject to the 3′–5′ proofreading activity of the DNA polymerase used for amplification.

Substituting primer P37-R by $P37_{Y}R$ [10], featuring three internal mismatches to its gametologous binding site on the X-chromosome (Fig. 1), however, completely prevented the detection of pseudo-P37 C-alleles with PCR-SNPE.

Likewise, even in the case of a deletion of the P37 target sequence, detectable product formation arising from the X-chromosomal P37 gametologue appeared improbable in the highly competitive multiplex PCR situation with fully matching primers for the other 18 Y-SNP targets.

However, as we could not fully exclude the faint possibility of having genotyped by mistake the X-chromosomal GSV instead of the targeted Y-SNP P37, we set out to confirm all 17 phylogenetically inconsistent P37 base calls by Sanger-type sequencing.

Primers $P37_{XY}F_{seq}$ and ARSDP1-R PCR amplified 455 bp template molecules for Sanger-type sequencing with a 408 bp analysis window (Y:14,491,586–14,491,993). This facilitated the analysis of P37 and its sequence neighborhood including the binding sites of the oligonucleotides used in PCR-SNPE experiments (Fig. 1). For all but one sample primer binding site mutations affecting PCR-SNPE genotyping could be excluded. Sample 17 showed a single mismatch with primer P37-F (Table 1 and Fig. 1). Furthermore, the observed sequence patterns confirmed that those 17 samples indeed displayed P37 results which were discordant with the Y-SNP phylogeny and beyond that, additional base variation was detected in the sequencing analysis window (Fig. 1).

Thirteen of these affected samples, all attributed to haplogroup R-M269* by PCR-SNPE, perfectly matched the Y chromosome reference sequence except for carrying the derived C-allele at the P37 site (Table 1). Phylogenetic analysis targeting the Y-SNPs L23, M412, L11, and S116 by Sanger-type sequencing located all of these Y chromosomes in haplogroup R-M412* (Fig. 2). Twelve of these 13 samples showed a clear spatial clustering; they all came from donors born in two neighboring Tyrolean villages (road distance: ~10 km). Furthermore, 11 of these 12 donors had the same surname (Table 1). Another three samples in our study came from donors also bearing this particular name, but their Y-STR profiles and major-haplogroup affiliations differed from those obtained for their 11 namesakes (data not shown). Only one of the three "unaffected" probands with this particular surname had a paternal background in Tyrol. Based on the personal data regarding provenance, surname, and date of birth of the probands, their fathers and paternal grandfathers as well as the Y-STR profiling results, Y chromosomes 2–13 were considered (nearly) identical by descent (Table 1). Sample 1, however, was not a representative of this paternal lineage. The patrilineal background of this individual was in a region outside Tyrol and shared neither the surname nor a Y-STR profile with the other twelve R-M412* + P37 chromosomes.

Within haplogroup R1b (R-M343) P37 had evidently experienced another homoplasic mutation. By sequencing, the derived P37 allele was established for three additional samples that came from non-related donors attributed to haplogroup R-U152, a subbranch of R-M412. The theoretical possibility that these three samples actually belonged to R-M412* + P37 and a homoplasic mutation occurred at the U152 locus could be excluded by the derived allelic states of Y-SNPs L11 and S116 (Fig. 2).

Here, P37 was not the only marker in the sequenced segment showing homoplasy. Y-SNPs L202 and L203, both being mapped to the major haplogroup G in the "ISOGG 2013 tree" (http://www.isogg.org/tree), exhibited derived alleles in all three R-U152 + P37 chromosomes (Table 1). The sequence string containing P37, L202 and L203 comprises only 24 nucleotides. The remaining base calls within the sequencing analysis window matched perfectly with the Y-chromosomal reference sequence (Fig. 1).

The same variation pattern at sites P37, L202 and L203 was also observed for sample 17. Yet, here an additional T > C transition at Y-SNP P41 (Y:14,491,671, Fig. 1 and Table 1) was found. The four sequence variants fell into a 37 bp tract. This particular Y chromosome belonged to haplogroup L-M20, indicating even more homoplasic events.

Notably, in all instances the derived alleles on all four affected Y-SNP sites matched with their GSVs (Fig. 1).

The unique event assumption usually applied to SNPs clearly is violated for markers P37, P41, L202, and L203. Together with the published P37.1 and P37.2 homoplasies, the number of Y-SNP haplogroups featuring the derived P37 allele has risen to five. For markers P41, L202, and L203 this count amounts to three each.

None of these four homoplasic Y-SNPs was in its ancestral state part of a CpG dinucleotide sequence motif (Fig. 1). Deamination driven hypermutation at 5-methylcytosines in CpG dinucleotides could be, therefore, excluded as a possible explanation for our findings.

Thus, repeatedly observing up to four homoplasic mutations within a sequence tract as short as 37 bp appears unlikely to be the result of a series of parallel mutations in different phylogenetic contexts. This view is further strengthened by the observation that Y-SNPs P37 and P41 are phylogenetically *de facto* equivalent. The same holds true for markers L202 and L203.

Unidirectional transfer of a short stretch of X-chromosomal sequence to its Y-gametologue [6,7], a process that would allow for the simultaneous switching of multiple Y-SNP base states at GSV sites, is a more likely scenario.

Trombetta et al. have previously postulated the same region on the MSY analyzed herein to be a hotspot for X-to-Y gene conversion [7] and our experimental data convey support for this hypothesis.

The sequenced segment lies in the arylsulfatase D pseudogene 1 (*ARSDP1*, Y:14,474,762–14,499,204, +strand, NC_000024.9) on Yq11.21, a member of a cluster of X degenerate sulfatase pseudogenes [15]. The functional arylsulfatase D gene (*ARSD*, MIM: 300002, X:2,822,011–2,847,416, −strand, NC_000023.10), resides on Xp22.3. The X–Y reference sequence identity for a 348 bp sequencing analysis window (Y:14,491,586–14,491,933) was 93% (Fig. 1).

For simultaneously changing the base states at the P37, P37/L202/L203, or P41/P37/L202/L203 SNP sites, the minimum number of X-to-Y transferred nucleotides would have had to be 1 (Y:14,491,684), 24 (Y:14,491,684–14,491,707), and 37 (Y:14,491,671–14,491,707), respectively. The corresponding maximum conversion tracts would have been 34 (Y:14,491,672–14,491,705), 68 (Y:14,491,672–14,491,739), and 90 (Y:14,491,650–14,491,739) nt in length (Fig. 1). These values are in good agreement with published data. Trombetta et al. [7] determined for the gametologous *VCX* (MIM 300229) and *VCY* genes (MIM 400012) minimum (maximum) lengths for X-to-Y gene conversion tracts of 1–80 nt (19–169 nt) and in a study on *HSA*, a translocation hotspot adjacent to the gametologous gene pair *PRKX* (MIM 30083) and *PRKY* (MIM 400008), Rosser et al. [6] estimated the average inter-chromosomal conversion tract-length at ~100 nucleotides.

Using data published by [7], Cruciani et al. [16] determined for the *VCY* gene conversion hotspot lower and upper bound X-to-Y conversion rate estimates of 1.6×10^{-6} and 8.2×10^{-6} conversions per nt per generation (25 years). For the *PRKY HSA* gene conversion hotspot Rosser et al. [17] calculated rate estimates of 3.8×10^{-8} to 1.7×10^{-6} conversions per nt per generation (25 years) and Cruciani et al. [16] reported similar values. These X-to-Y conversion rates are considerably lower than that quoted for Y–Y intra-palindrome conversions (2.8×10^{-4} conversions per duplicated nt per generation (25 years) [18]), but up to two orders of magnitude higher than the estimated average MSY single base substitution rate (direct approach: 3.0×10^{-8} substitutions per nucleotide per generation (95% CI: 8.9×10^{-9} to 7.0×10^{-8}) [19]; human–chimpanzee comparison: 2.3×10^{-8} substitutions per nt per generation (25 years) [20]).

Thus, at gametologous regions of high sequence similarity inter-chromosomal gene conversion has to be considered a serious source of Y chromosome diversity. Such Y-SNPs can cause problems in a forensic context. Aside from producing genotyping results that are incompatible with the published Y-SNP phylogeny, affected markers yield uncertain haplogroup assignments when defining a terminal sub-branch. Like Y-SNPs exhibiting recurrent mutations due to intra-chromosomal gene conversion between paralogues on the MSY's ampliconic regions (e.g. P25 and 92R7) [21,22], SNPs at GSV sites need to be considered potentially unreliable. Such markers should be either avoided or need to be supplemented with equivalent or downstream SNPs.

4. Practical aspects and consequences

The increased awareness of the practical forensic use of Y-SNPs is accompanied by a lack of standardization of phylogenetic Y chromosome analyses. The seminal publication of the YCC [4] and the updates of the "YCC tree" in 2003 [2] and 2008 [3] are still the basis of most technical protocols. However, a plethora of new markers, branch names, and corrections have been published in the academic literature since then. At the time of writing an updated Y-SNP tree for forensic genetics has been published [23].

Forensic analyses crucially depend on the scientific acceptance, standardization and reproducibility of the applied methodology. Hence, the YHRD (Y chromosome haplotype reference database) [24,25] has been developed as a general source of information concerning lab methods, population frequencies and interpretation tools. Starting with release 31 from 16th November 2009 the YHRD now includes Y chromosomes typed for phylogenetically informative SNPs and submissions are on the increase. Currently 11,438 samples in the database (release 43) have a haplogroup designation based on high-resolution Y-SNP typing. The haplogroup designation is automatically generated from the standard submission form, which includes a list of the ancestral/derived states of the analyzed binary polymorphisms (http://www.yhrd.org/Contribute). Details (branch name according to the most actual version of the YCC tree, mutation type, publication, analysis protocol including primer sequences) are available in the YHRD (http://www.yhrd.org/Research/YSNPs).

We take the supposed inter-gametologue X-to-Y conversions in *ARSDP1* to propose necessary steps for the forensic validation of Y-SNPs:

(a) Discussion and publication of guidelines on mandatory validation requirements for new Y-SNP markers published after 2008.
(b) Set-up of a central scientifically curated repository of validity states and haplogroup designations of Y-SNP markers.
(c) Proposal for a revised, unequivocal nomenclature which accounts for the current Y haplogroup phylogeny with synonymous markers and large numbers of branches (e.g. R-U152 instead of R1b1b2a1a2d where R is the basal branch and U152 the defining marker for the sub-branch).

As a first direct consequence and conditioned by the growing number of Y-SNP studies submitted to the YHRD for validation, we have introduced a "black list" of ambiguous Y-SNP markers, which is available online (http://yhrd.org/BYSNP). This list also includes alternative SNPs which can be used to define the haplogroup unambiguously. Furthermore, we have contacted all those YHRD contributors of population samples who have determined a haplogroup with the ambiguous final marker P37.2. It was decided to assign all those haplotypes to the next analyzed upstream marker (e.g. I2-P215).

Conflicts of interest

The authors declare that they do not have any known conflict of interest.

Acknowledgments

First and foremost the authors would like to thank all voluntary blood donors. The staff of the "Central Institute for Blood Transfusion and Immunological Department" (Landeskrankenhaus Innsbruck) is acknowledged for both organizing the blood drive campaigns and their hospitality. We are also obliged to Stefan Troger (Institute of Legal Medicine, Innsbruck Medical University) for programming and maintaining the in-house database. Hannelore Volderauer (Institute of Legal Medicine, Innsbruck Medical University) is greatly acknowledged for sample management and administration. This work was financially supported by the grants "ALPKULTUR – cultural historical documentation of toponyms in the alpine area" and "High-Resolution Analysis of Y-chromosome Variability in Selected Regions of the Tyrolean Alps (Austria): Influences of Topology and Demographic History on the Genetic Differentiation" within the "Alpine Research Programme" of the Austrian Academy of Sciences. This work has also received funding from the European Union Seventh Framework Programme (FP7/2007–2013) under grant agreement no. 285487 (EUROFORGEN-NoE). D.E. received a scholarship from the Leopold Franzens University of Innsbruck.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.fsigen.2013.05.010.

References

[1] H. Skaletsky, T. Kuroda-Kawaguchi, P.J. Minx, et al., The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes, Nature 423 (2003) 825–837.

[2] M.A. Jobling, C. Tyler-Smith, The human Y chromosome: an evolutionary marker comes of age, Nat. Rev. Genet. 4 (2003) 598–612.

[3] T.M. Karafet, F.L. Mendez, M.B. Meilerman, P.A. Underhill, S.L. Zegura, M.F. Hammer, New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree, Genome Res. 18 (2008) 830–838.

[4] Y Chromosome Consortium, A nomenclature system for the tree of human Y-chromosomal binary haplogroups, Genome Res. 12 (2002) 339–348.

[5] J.M. Chen, D.N. Cooper, N. Chuzhanova, C. Férec, G.P. Patrinos, Gene conversion: mechanisms, evolution and human disease, Nat. Rev. Genet. 8 (2007) 762–775.

[6] Z.H. Rosser, P. Balaresque, M.A. Jobling, Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot, Am. J. Hum. Genet. 85 (2009) 130–134.

[7] B. Trombetta, F. Cruciani, P.A. Underhill, D. Sellitto, R. Scozzari, Footprints of X-to-Y gene conversion in recent human evolution, Mol. Biol. Evol. 27 (2010) 714–725.

[8] S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers, Methods Mol. Biol. 132 (2000) 365–386.

[9] A. Untergasser, H. Nijveen, X. Rao, T. Bisseling, R. Geurts, J.A. Leunissen, Primer3-Plus, an enhanced web interface to Primer3, Nucleic Acids Res. 35 (2007) W71–W74. , http://dx.doi.org/10.1093/nar/gkm306.

[10] H. Niederstätter, G. Rampl, D. Erhart, F. Pitterl, H. Oberacher, F. Neuhuber, I. Hausner, C. Gassner, H. Schennach, B. Berger, W. Parson, Pasture names with Romance and Slavic roots facilitate dissection of Y chromosome variation in an exclusively German-speaking Alpine region, PLoS ONE 7 (2012) e41885.

[11] A. Brandstätter, H. Niederstätter, M. Pavlic, P. Grubwieser, W. Parson, Generating population data for the EMPOP database—an overview of the mtDNA sequencing and data evaluation process considering 273 Austrian control region sequences as example, Forensic Sci. Int. 166 (2007) 164–175.

[12] B. Berger, A. Lindinger, H. Niederstätter, P. Grubwieser, W. Parson, Y-STR typing of an Austrian population sample using a 17-loci multiplex PCR assay, Int. J. Legal Med. 119 (2005) 241–246 (Erratum in: Int. J. Legal Med. 120 (2006) 255).

[13] D. Erhart, B. Berger, H. Niederstätter, C. Gassner, H. Schennach, W. Parson, Frequency data for 17 Y-chromosomal STRs and 19 Y-chromosomal SNPs in the Tyrolean district of Reutte, Austria, Int. J. Legal Med. 126 (2012) 977–978.

[14] J. García-Moreno, D.P. Mindell, Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD, Mol. Biol. Evol. 17 (2000) 1826–1832.

[15] G. Meroni, B. Franco, N. Archidiacono, S. Messali, G. Andolfi, M. Rocchi, A. Ballabio, Characterization of a cluster of sulfatase genes on Xp22.3 suggests gene duplications in an ancestral pseudoautosomal region, Hum. Mol. Genet. 5 (1996) 423–431.

[16] F. Cruciani, B. Trombetta, V. Macaulay, R. Scozzari, About the X-to-Y gene conversion rate, Am. J. Hum. Genet. 86 (2010) 495–497.

[17] Z.H. Rosser, P. Balaresque, M.A. Jobling, Response to Cruciani et al., Am. J. Hum. Genet. 86 (2010) 497–498.

[18] S. Rozen, H. Skaletsky, J.D. Marszalek, P.J. Minx, H.S. Cordum, R.H. Waterston, R.K. Wilson, D.C. Page, Abundant gene conversion between arms of palindromes in human and ape Y chromosomes, Nature 423 (2003) 873–876.

[19] Y. Xue, Q. Wang, Q. Long, B.L. Ng, H. Swerdlow, J. Burton, C. Skuce, R. Taylor, Z. Abdellah, Y. Zhao, Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree, Curr. Biol. 19 (2009) 1453–1457.

[20] S. Repping, S.K. van Daalen, L.G. Brown, C.M. Korver, J. Lange, J.D. Marszalek, T. Pyntikova, F. van der Veen, H. Skaletsky, D.C. Page, S. Rozen, High mutation rates have driven extensive structural polymorphism among human Y chromosomes, Nat. Genet. 38 (2006) 463–467.

[21] J.J. Sanchez, M. Brión, W. Parson, A.J. Blanco-Verea, C. Børsting, M. Lareu, H. Niederstätter, H. Oberacher, N. Morling, A. Carracedo, Duplications of the Y-chromosome specific loci P25 and 92R7 and forensic implications, Forensic Sci. Int. 140 (2004) 241–250.

[22] S.M. Adams, T.E. King, E. Bosch, M.A. Jobling, The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal marker P25 through gene conversion, Forensic Sci. Int. 159 (2006) 14–20.

[23] A. Van Geystelen, R. Decorte, M.H.D. Larmuseau, Updating the Y-chromosomal phylogenetic tree for forensic applications based on whole genome SNPs, Forensic Sci. Int. Genet. (2013), http://dx.doi.org/10.1016/j.fsigen.2013.03.010.

[24] L. Roewer, M. Krawczak, S. Willuweit, et al., Online reference database of European Y-chromosomal short tandem repeat (STR) haplotypes, Forensic Sci. Int. 118 (2001) 106–113.

[25] S. Willuweit, L. Roewer, International Forensic Y Chromosome User Group, Y chromosome haplotype reference database (YHRD): update, Forensic Sci. Int. Genet. 1 (2007) 83–87.