Footprints of X-to-Y Gene Conversion in Recent Human Evolution

Beniamino Trombetta,†,[1] Fulvio Cruciani,†,[1] Peter A. Underhill,[2] Daniele Sellitto,[3] and Rosaria Scozzari*,[1]

[1]Dipartimento di Genetica e Biologia Molecolare, Sapienza Università di Roma, Rome, Italy

[2]Department of Psychiatry and Behavioral Sciences, Stanford University School of Medicine

[3]Istituto di Biologia e Patologia Molecolari, Consiglio Nazionale delle Ricerche, Rome, Italy

†These authors contributed equally to this work.

***Corresponding author:** E-mail: rosaria.scozzari@uniroma1.it.

Associate editor: Yoko Satta

Research article

Abstract

Different X-homologous regions of the male-specific portion of the human Y chromosome (MSY) are characterized by a different content of putative single nucleotide polymorphisms (SNPs), as reported in public databases. The possible role of X-to-Y nonallelic gene conversion in contributing to these differences remains poorly understood. We explored this issue by analyzing sequence variation in three regions of the MSY characterized by a different degree of X–Y similarity and a different density of putative SNPs: the *PCDH11Y* gene in the X-transposed (X–Y identity 99%, high putative SNP content); the *TBL1Y* gene in the X-degenerate (X–Y identity 86–88%, low putative SNP content); and *VCY* genes–containing region in the P8 palindrome (X–Y identity 95%, low putative SNP content). Present findings do not provide any evidence for gene conversion in the *PCDH11Y* and *TBL1Y* genes; they also strongly suggest that most putative SNPs of the *PCDH11Y* gene (and possibly the entire X-transposed region) are most likely X–Y paralogous sequence variants, which have been entered in the databases as SNPs. On the other hand, clear evidence for the *VCY* genes in the P8 palindrome having acted as an acceptor of X-to-Y gene conversion was obtained. A rate of 1.8×10^{-7} X-to-Y conversions/bp/year was estimated for these genes. These findings indicate that in the *VCY* region of the MSY, X-to-Y gene conversion can be highly effective to increase the level of diversity among human Y chromosomes and suggest an additional explanation for the ability of the Y chromosome to retard degradation during evolution. Present data are expected to pave the way for future investigations on the role of nonallelic gene conversion in double-strand break repair and the maintenance of Y chromosome integrity.

Key words: gene-conversion hot spot, MSY, segmental duplications, human sex chromosomes, dbSNP.

Introduction

One remarkable feature of the human genome is the abundance of large duplications characterized by a high degree of sequence identity (International Human Genome Sequencing Consortium 2001). It has been estimated that about 5% of the human genome is composed of sequences that are duplicated either on the same or different chromosomes, with an identity >90% and a length >1 kb (Bailey et al. 2001, 2002; She et al. 2004; Zhang et al. 2005). Once initially formed, these genomic regions, called segmental duplications (SDs), can promote further genomic rearrangements through nonallelic homologous recombination (NAHR) between paralagous sequences, thus playing an important role in genome evolution (Eichler 2001; Samonte and Eichler 2002; Armengol et al. 2003; Stankiewicz et al. 2004; Cheng et al. 2005) and disease (Pentao et al. 1992; Chen et al. 2004; Visser et al. 2005; Bailey and Eichler 2006). NAHR between SDs may also result in interparalog gene conversion, that is, the nonreciprocal transfer of genetic information from a "donor" sequence to a homologous "acceptor" (Bailey and Eichler 2006; Arnheim et al. 2007). Gene conversion between duplicated sequences can have two effects: On the one hand, it can increase overall sequence similarity between both inter and intrachromosomal paralogous sequences; on the other hand, it can generate an excess of genetic diversity among allelic copies, thus increasing the single nucleotide polymorphism (SNP) content of SD.

In the last few years, numerous putative SNPs have been entered in the public databases (dbSNP, http://www.ncbi.nlm.nih.gov/SNP/) in the process of determining the sequence of the human genome. Analyses of public and private human genome sequences revealed that SNPs are overrepresented in SDs (Bailey et al. 2002; Estivill et al. 2002). Putative SNPs in these genomic regions may be nothing more than differences between paralogous sequences (paralogous sequence variants: PSVs), which have been misassembled in the draft of the human-genome sequence (Estivill et al. 2002; Cheung et al. 2003; Bailey and Eichler 2006). Alternatively, the high SNP density observed in SD could be the consequence of gene conversion between duplicated sequences (Giordano et al. 1997; Hurles

2002; Hallast et al. 2005; Chen et al. 2007). Although gene conversion between intrachromosomal SD is well studied (Bussaglia et al. 1995; Reyniers et al. 1995; Roesler et al. 2000; Vázquez et al. 2001; Rozen et al. 2003; Bosch et al. 2004; Hurles et al. 2004; Hallast et al. 2005; Adams et al. 2006; Chen et al. 2007), the evolution and dynamics of interchromosomal gene conversion remain more unclear (Gupta et al. 2005; Chen et al. 2007; Eickbush and Eickbush 2007; Benovoy and Drouin 2009). Sequence-diversity analysis of SDs in different individuals could, in principle, shed light on the dynamics of gene conversion; however, in most of the genome, the phenomenon of diploidy complicates the comparative sequencing strategies because of sequence diversity between alleles. The haploid male-specific region of the human Y chromosome (MSY) has no such a limitation; moreover, it is particularly enriched in both intra and interchromosomal SDs. Therefore, it represents an ideal system in which to study gene conversion between paralogous/gametologous (Garcia-Moreno and Mindell 2000) sequences. Importantly, because the phylogenetic relationships of chromosomes belonging to different Y haplogroups are precisely known (Karafet et al. 2008), the directionality of new single mutational events can be clearly assessed, thus allowing to infer the involvement of gene conversion, if any.

Three classes of duplicated sequences are present in the euchromatic portion of the MSY: X-transposed, X-degenerate, and ampliconic (Skaletsky et al. 2003). The X-transposed regions (99% of sequence identity with their homologous counterpart in Xq21) are interchromosomal SD originated from an X-to-Y transposition 4.7 Ma (Ross et al. 2005). The X-degenerate sequences (X–Y identity 60–96%) are remnants of ancient autosomes from which the modern human sex chromosomes evolved. The ampliconic sequences, composed of intrachromosomal duplications, are largely made up of eight large palindromes, with a total length of 5.7 Mb (one quarter of the euchromatic MSY), and arm-to-arm nucleotide identities >99.9%, mostly due to frequent Y–Y gene-conversion events (Rozen et al. 2003).

The aim of this work was to evaluate the possible role of X-to-Y interchromosomal gene conversion in shaping the extent and distribution of sequence diversity in the human MSY. If gene-conversion events between X and Y chromosomes have contributed to the Y chromosome variation, we expect to find an excess of mutations at X–Y gametologous sequence variant (GSV) sites, where the Y-linked derived allele is the same as the gametologous sequence on the X. Also, if different GSVs are physically close to one another, a single gene-conversion event might eventually create allelic diversity simultaneously at multiple nucleotide sites. Thus, past gene-conversion events could be easily detected in the form of multiple variants, which are both physically neighboring and phylogenetically equivalent mutations.

With this aim, we analyzed by comparative sequencing three MSY regions having different evolutionary histories (Lahn and Page 1999; Skaletsky et al. 2003; Ross et al. 2005) in a number of major Y chromosome haplogroups representing a large coverage of the worldwide MSY diversity. By using this approach, we identified one X-to-Y hot spot for gene conversion in the *VCY* gene. The presence of a second X-to-Y hot spot in the *ARSD* pseudogene region was hypothesized by the analysis of the available information on current MSY variation.

Materials and Methods

DNA Samples

DNA samples came from collections of the authors, and haplogroup information is as reported (Underhill et al. 2000, 2001; Cruciani et al. 2002, 2004, 2006, 2007).The *VCY* region was analyzed in 122 unrelated males belonging to different Y haplogroups (supplementary table 1, Supplementary Material online). Subsets of 15, 11, and 15 Y chromosomes, chosen for their wide coverage of the Y phylogeny, were, respectively, analyzed for the *PCDH11Y*, *TBL1Y*, and P8 palindrome regions (supplementary table 1, Supplementary Material online).

Amplification and Sequencing

Overall, 23.2 kb from the *PCDH11Y* within the X-transposed region, 14.5 kb from the *TBL1Y* gene within the X-degenerate region, and 17.2 kb from the ampliconic P8 palindrome (8.6 kb for each arm) were sequenced (supplementary table 2, Supplementary Material online).

We designed polymerase chain reaction (PCR) and sequence primers (supplementary table 3, Supplementary Material online) on the basis of the Y-chromosome sequence reported in the March 2006 assembly of the UCSC Genome Browser using Primer3 software (http://genome.ucsc.edu/; http://frodo.wi.mit.edu/primer3/). Sequencing templates were obtained through PCR in a 50-μl reaction containing 50 ng of genomic DNA, 200 μM each deoxyribonucleotide (dNTP), 2.5 mM $MgCl_2$, 1 unit of Taq polymerase, and 10 pmols of each primer. A touchdown PCR program was used with an annealing temperature decreasing from 62 to 55 °C over 14 cycles, followed by 30 cycles with an annealing temperature of 55 °C. A single PCR reaction was performed to coamplify homologous sequences of the proximal and distal arm of the P8 palindrome, whereas a long-range PCR, specific for a portion (chrY: 14602218–14607913) of the proximal arm of the palindrome, was used for chromosomes carrying multiple mutations (haplogroups B-M146 and L-M11*). Long-range PCR was performed in a final volume of 50 μl containing 500 ng of genomic DNA, 400 μM each dNTP, 2.5 mM $MgCl_2$, 5 U of TaKaRa LA Taq, and 1 μM of each primer (for: 5′-CCTCCTGGGTTCCCGCCATT-3′, rev: 5′-CCGCCATGTCCTGATGGTGCT-3′). The reaction was initiated with a denaturation at 94 °C for 1 min, followed by 30 cycles of denaturation at 98 °C for 10 s, and annealing/elongation at 68 °C for 7 min. A final step at 72 °C for 10 min was performed.

Y-specificity of the PCR products was confirmed by using female genomic DNAs as a negative control. Following DNA amplification, PCR products were purified using the

QIAquick PCR purification kit (Qiagen, Hilden, Germany). Cycle sequencing was performed using the BigDye Terminator Cycle Sequencing Kit with Amplitaq DNA polymerase (Applied Biosystems, Foster City, CA) and an internal or PCR primer. Cycle sequencing products were purified by ethanol precipitation and run on an ABI Prism 3730XL DNA sequencer (Applied Biosystems). Chromatograms were aligned and analyzed for mutations using Sequencher 4.8 (Gene Codes Corporation, Ann Arbor, MI).

Data Analysis

Nucleotide diversity (π) and its standard deviation were calculated according to Nei (1987). Tajima's *D* neutrality test (Tajima 1989) was performed using the DNAsp ver. 4.50.1 software (Rozas et al. 2003).

Alignments between X and Y chromosome sequences were performed using the ClustalW2 software (http://www.ebi.ac.uk/Tools/clustalw2/; Larkin et al. 2007) and default parameters. A site was considered a GSV whenever a difference was found between the ancestral Y sequence (as inferred from the present experimental data) and the reference X chromosome sequence (as reported in the March 2006 assembly of the UCSC Genome Browser). It is worth noting that the X-chromosome reference sequence may not contain the ancestral allele, which would result in an underestimate of the GSV sites. The P8 ampliconic sequences analyzed here align to one to four different X chromosome regions (supplementary table 4, Supplementary Material online); in this case, a site was considered a GSV whenever there was a difference between the ancestral Y sequence and at least one of the X reference sequences. By comparing *PCDH11Y*, *TBL1Y*, and the *VCY*-flanking palindromic region with their gametologous counterparts on the X chromosome, we observed not-aligning sequences with lengths up to more than 10 contiguous bases, whereas a maximum of two not-aligning contiguous bases were observed between *VCY* and *VCX* genes. We arbitrarily chose not to consider sequences of more than five not-aligning contiguous bases as an X–Y GSV. For each region, the expected number of SNPs falling in X–Y GSVs was calculated as the product of the total number of SNPs detected in the present study and the proportion of GSVs observed in the same region. To evaluate whether the Y SNPs identified were randomly distributed with respect to X–Y GSV and non-GSV sites on the MSY, we used a Fisher exact test on 2×2 contingency tables.

To estimate the rate of "per" nucleotide X-to-Y gene conversion (c) in the *VCY* region, we used a modified version of the equation of Repping et al. (2006) by using the following notations:

the number of X-to-Y gene conversion driven mutations, N ($N = 9$),
the total time (t_{tot}) spanned by all branches in the tree of 122 chromosomes,
the length of the analyzed region, l ($l = 1{,}616$ bp), and
the average X–Y sequence diversity (d) between each of the 122 Y chromosomes analyzed and the gametologous regions on the X.

The equation for c is

$$c = \frac{N/t_{tot}}{l \times d}.$$

To estimate t_{tot} (899,750 years, about 45,000 generations for a 20-year generation) we used 8 as the total number of SNPs identified in the analyzed regions (considering only SNPs that were not due to X-to-Y gene conversion), 1.05 as the average number of mutations on path from the root, and 118,000 years (Repping et al. 2006) as the time to the most recent common ancestor (TMRCA). The 95% confidence interval for the TMRCA (78,000–169,000 years) reported by Repping et al. (2006) was used to obtain a range for the X-to-Y gene-conversion rate estimate. We have calculated the sequence divergence between the *VCY* regions and each of the four gametologous regions on the X chromosome for all the 122 Y chromosomes analyzed. Averaging all these values, we obtained a value of $d = 0.035$. Note that the quantity $l \times d$ corresponds to the number of X–Y GSVs in the region. Thus, the rate c can be considered as a gene conversion–driven mutation rate per GSV.

Results

Rationale for Selecting MSY Regions to Be Sequenced

An Excess of Putative SNPs. Mining the information contained in the build 129 of the dbSNP, we realized that the X-transposed regions and, to a lesser extent, portions of the ampliconic regions in the euchromatic MSY contain a higher number of putative SNPs than other MSY sequences (fig. 1). This excess is particularly pronounced for the X-transposed regions, where the putative SNPs are overrepresented by a factor ~20 with respect to X-degenerate MSY regions. Whether the observed increased density of SNPs within X-transposed regions arises from alignment of X–Y gametologous (rather than allelic) sequences or reflects a true enhanced nucleotide diversity, possibly due to X-to-Y gene conversion, remains to be elucidated. To shed light on this issue, we determined the sequence of 23.2 kb of the *PCDH11Y* gene and 5′ flanking region within the X-transposed (supplementary table 2, Supplementary Material online) from 15 human Y chromosomes representing a wide coverage of the Y chromosome diversity. As shown in figure 1 and supplementary table 5, Supplementary Material online, the region analyzed is fully representative of the entire X-transposed region with respect to the relative abundance of SNPs as reported in dbSNP.

A Higher X–Y Sequence Similarity with Respect to Flanking Regions. Comparison of X and Y gametologous sequences revealed that an MSY region extending from the 3′ portion of the *KALP* pseudogene through the *VCY* genes is characterized by a higher X–Y sequence similarity than other regions within the same evolutionary group 4 (Skaletsky et al. 2003). This finding was interpreted as due to either different ages of divergences between 3′ *KALP*/*VCY* and other group 4 genes or relatively recent X–Y gene conversion (~7 Ma) (Skaletsky et al. 2003 and supplementary fig. 10 therein).

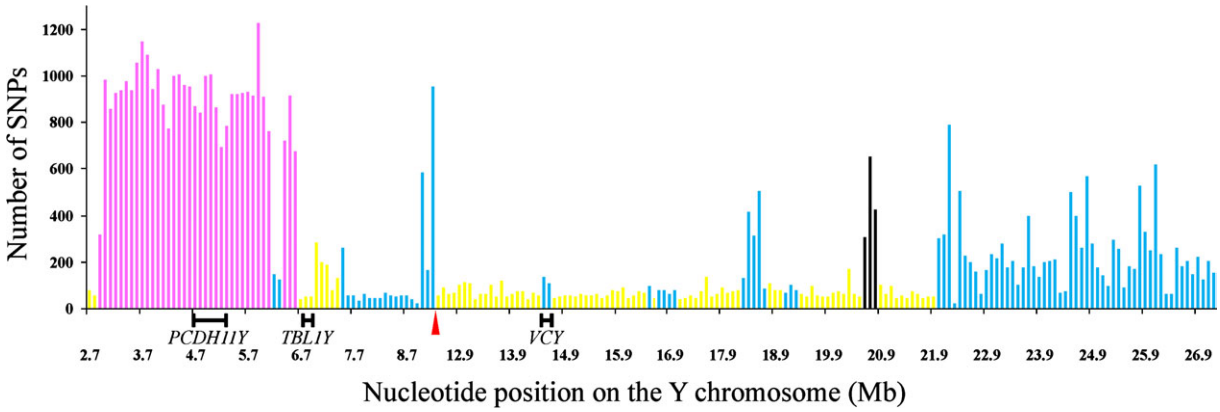

FIG. 1. Distribution of putative SNPs in the euchromatic portion of the human MSY as reported in dbSNP. SNPs as reported in the build 129 of dbSNP. *X*-axis: partitioning of the MSY in 100-kb windows, starting from position 2,700,000 of the sequence reported in the March 2006 (hg18) assembly of the human Y-chromosome. *Y*-axis: number of putative SNPs for each 100-kb window. Different regions of the Y chromosome (Skaletsky et al. 2003) are shown in different colors (pink: X-transposed regions; blue: dupliconic regions; yellow: other sequences, mainly X-degenerate sequences; and black: heterochromatic block). The location of the genes analyzed is also shown. Centromeric and pericentromeric sequences (red arrow) are not shown.

Prompted by the suggestion for X–Y gene conversion, we resequenced a 17.2-kb portion of a palindromic region (the P8 palindrome) encompassing the *VCY* genes (supplementary table 2, Supplementary Material online) in the same sample set as the *PCDH11Y* gene from the X-transposed. Unlike other dupliconic sequences (fig. 1), the portion of the P8 palindrome analyzed here does not show any significant excess of putative SNPs in the dbSNP database (supplementary table 4, Supplementary Material online).

No A Priori Indication for Gene-Conversion Events. A region of 14.5 kb of the *TBL1Y* gene within the X-degenerate portion of the MSY was resequenced (supplementary table 2, Supplementary Material online) in 11 subjects representing a subset of those analyzed for the *PCDH11Y*- and *VCY*-containing regions.

SNP Content and Sequence Diversity in Human MSY

Overall, 54.9 kb of the MSY has been resequenced. No fixed differences between the Y chromosomes analyzed here and the MSY reference sequence were found, confirming the high accuracy of the current human Y chromosome reference sequence (NCBI build 36.1). We found 16 SNPs and 1 3-bp ins/del polymorphism in 23.2 kb of the *PCDH11Y* gene within the X-transposed region by the analysis of 15 Y chromosomes (fig. 2 and table 1). None of these polymorphisms corresponds to any of 193 SNPs as reported in dbSNP for the same region (supplementary table 5, Supplementary Material online). As shown in table 2, the nucleotide diversity was estimated as $1.1 \pm 0.2 \times 10^{-4}$. A similar value of nucleotide diversity, $\pi = 1.5 \pm 0.3 \times 10^{-4}$, was estimated for the 14.5 kb of the *TBL1Y* gene analyzed in 11 Y chromosomes (table 2). In the latter region, we identified seven SNPs (fig. 2 and table 1). One of these (V24) corresponds to SNP rs9786374 present in the dbSNP (five SNPs reported for this region in the dbSNP, see supplementary table 5, Supplementary Material online).

In the 17.2-kb portion of the P8 palindrome (Skaletsky et al. 2003) studied here (8.6 kb for each palindromic arm), 15 SNPs (14 single nucleotide substitutions and one ins/del of 1 nt) were found through the analysis of the same 15 Y chromosomes as the X-transposed region. One of the SNPs (V109) corresponds to one of the eight SNPs reported in the dbSNP for the same region (rs2072420, see supplementary table 5, Supplementary Material online). As shown in figure 2, ten polymorphisms (V120, V121, V118, V123, V122, V111, and V113–V116) only involved one arm of the palindrome, whereas five polymorphisms (V117, V119.1, V109, V110, and V112) were found in both the arms, likely as a consequence of Y−Y gene conversion. The nucleotide diversity for this region was estimated as $\pi = 2.0 \pm 0.4 \times 10^{-4}$ (table 2).

The 15 SNPs identified in the P8 palindromic region did not appear to be equally distributed, 7 of them (V117–V121 and V126–V127) being restricted to a short duplicated DNA segment of 0.8 kb containing the *VCY* genes (chrY: 14606996–14607803 and chrY: 14677475–14678282). In order to evaluate the relative contribution of the *VCY* genes to the diversity value observed for the P8 palindrome, we obtained new estimates of the π value, by considering *VCY* and the rest of the P8 palindrome separately. The P8 palindrome without the *VCY* genes, showed a π-value of $1.2 \pm 0.4 \times 10^{-4}$. On the other hand, a π-value of $10.0 \pm 2.4 \times 10^{-4}$ was obtained for *VCY*, a significantly higher value than that estimated for *PCDH11Y* and *TBL1Y* genes and other MSY genes as reported in the literature (Shen et al. 2000) (table 2 and supplementary fig. 1, Supplementary Material online). It seems unlikely that the extensive Y–Y gene conversion previously reported (Rozen et al. 2003) is the cause of the high nucleotide diversity observed in the *VCY* genes, because the rest of the P8 palindrome, where Y–Y gene conversion also occurs, does not show any statistically significant difference with either the *PCDH11Y* or *TBL1Y* genes (table 2).

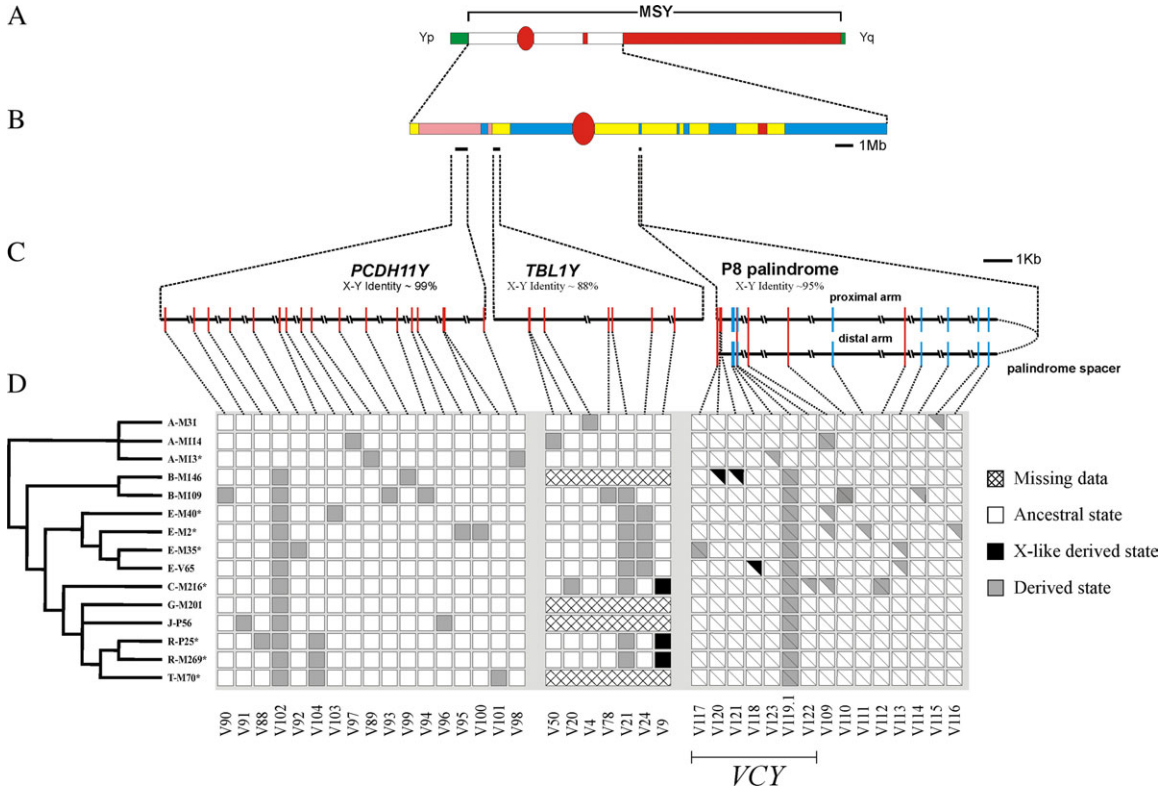

Fig. 2. SNPs identified in the sequenced regions. (*A*) Schematic representation of the Y chromosome. Green, PARs; red, eterochromatic sequences; and white, euchromatic sequences. (*B*) Enlarged view of the euchromatic portion of the MSY, showing the location of three classes of sequences: pink, X-transposed sequences; yellow, X-degenerate sequences; and blue, ampliconic sequences. (*C*) Location of the SNPs identified in the three MSY regions analyzed. SNPs are depicted as vertical bars (red and blue). Blue bars in the P8 palindrome indicate SNPs present on either the proximal or the distal arm of the palindrome. (*D*) To the left, a simplified version of the Y-chromosome tree (Karafet et al. 2008) showing the phylogenetic relationships of the chromosomes analyzed. SNP names are given at the bottom. Square colors represent the allelic state for each SNP: white, ancestral allele; black, the SNP has arisen in an X–Y GSV and the derived state is equal to the gametologous base on the X; and gray, the SNP has arisen in a site identical between X and Y. SNPs V50 and V24 are in regions that do not align with any region of the X chromosome. SNPs within the *VCY* regions are indicated. For the P8 palindrome, each square is divided into two triangles, representing the paralogous sites in the two arms of the palindrome.

In principle, a higher mutation rate and/or X-to-Y gene conversion could generate a higher diversity among allelic sequences. If gene conversion is operating between X–Y SDs, we should expect 1) that the number of Y SNPs found in X–Y GSV sites is higher than expected by chance and 2) that the Y-linked derived allele (as inferred from the Y phylogeny) is the same as the gametologous site on the X.

Comparing Diversity in X–Y GSV and Non-GSV Sites

As shown in table 1, none of the 17 polymorphisms identified in the X-transposed *PCDH11Y* gene (99% X–Y identity) are SNPs at GSVs. Only one (V9) of the seven SNPs identified in the *TBL1Y* gene (86–88% X–Y identity) was found at a GSV site (expected number = 0.6; $P = 0.35$ for one-tailed Fisher's exact test). In the P8 palindrome (95% X–Y identity), we found three SNPs at X–Y GSV sites (V120, V121, and V118, fig. 3) (expected number = 1.2; $P = 0.12$). Based on the phylogenetic information, it was not possible to establish the direction of the mutation for the highly homoplasic SNP V109 unambiguously; therefore, this SNP was not further considered. Interestingly, all three SNPs at X–Y GSV sites were found within the short 0.8-kb duplicated segment containing the *VCY* genes, that is, the region showing the highest π value among those analyzed. The excess of SNPs in correspondence of X–Y GSVs in the *VCY* genes was statistically significant (expected number = 0.4; $P = 0.005$). The Y-linked derived allele for each of these SNPs was the same as the gametologous site on the X, strongly suggesting the involvement of X-to-Y gene-conversion events in generating the observed diversity in this region. This hypothesis was further reinforced by the finding of two closely spaced SNPs in contiguous X–Y GSVs (V120 and V121) occurring in the same phylogenetic context (Y haplogroup B-M146).

To further explore whether X-to-Y gene conversion has been the cause of the high diversity in the *VCY* genomic region, we resequenced the duplicated 0.8-kb segment containing the *VCY* genes in additional 107 globally diverse Y chromosomes. Nine additional variant sites were found, for a total of 16 SNPs (fig. 3 and table 1), only one of which involved a C-to-T mutation at a CpG site. The sequences showed a greatly elevated value of π ($\pi = 7.8 \pm 0.8 \times 10^{-4}$, table 2 and supplementary fig. 1, Supplementary Material

Table 1. List of the Polymorphisms Identified in the Present Study.

SNP	Y-Position[a]	Mutation[b]	ChrX[c]	Haplogroup[d]	Region
V90	4900743	C to G	C	B-M109 (B2a1a)	
V91	4922349	C to T	C	J-P56 (J1d)	
V88	4922861	C to T	C	R-P25* (R1b1*)	
V102	4932610	G to C	G	BT-SRY$_{10831.1}$ (BT)	
V92	4933798	C to T	C	E-M35* (E1b1b1*)	
V104	4934629	T to C	T	K-M9 (KT)	
V103	4934864	C to T	C	E-M40* (E*)	
V97	4961948–4961950	del CTT	ins	A-M114 (A2a)	
V89	4962306	A to C	A	A-M13* (A3b2*)	
V93	4980405	C to G	C	B-M109 (B2a1a)	
V99	4986299	G to A	G	B-M146 (B1a)	
V94	5029475	T to A	T	B-M109 (B2a1a)	
V96	5334763	G to A	G	J-P56 (J1d)	
V95	5334968	C to G	C	E-M2* (E1b1a*)	
V100	5395763	G to A	G	E-M2* (E1b1a*)	
V101	5395800	A to T	A	T-M70 (T)	
V98	5427907	A to G	A	A-M13* (A3b2*)	*PCDH11Y*
V50	6905936	T to C	NA	A-M114 (A2a)	
V20	6905955	G to A	G	C-M216* (C*)	
V4	6906482	G to A	G	A-M31 (A1a)	
V78	6919558	C to G	C	B-M109 (B2a1a)	
V21	6919691	C to T	C	BT-SRY$_{10831.1}$ (BT)	
V24	6921075	G to A	NA	E-M40 (E)	
V9	6924267	A to G	G	CT-P143 (CT)	*TBL1Y*
V131	14607001 and 14678277	C/C to T/T	T	K-M9* (K*)	
V117	14607005 and/or 14678273	T/T to (C/T or T/C) to C/C	T	E-M215 (E1b1b) multiple branches	
V129	14607093 or 14678185	C/C to (G/C or C/G)	C	J-M92 (J2a2a)	
V120	14607120	A to G	G	B-M146 (B1a)	
V121	14607130	A to G	G	B-M146 (B1a)	
V124	14607470 or 14677808	G/G to (A/G or G/A)	G	A-M171 (A3b2a)	
V132	14607559 or 14677719	T/T to (C/T or T/C)	T	E-V19 (E1b1b1a3b)	
V118	14607569 and/or 14677709	A/A to (G/A or A/G) to G/G	G	E-P147 (E1) multiple branches	
V123	14607604 or 14677674	G/G to (C/G or G/C)	G	A-M13* (A3b2*)	
V128	14607626 or 14677652	C/C to (T/C or C/T)	T	O-M122* (O3*)	
V125	14677612	T to C	C	L-M11* (L*)	
V126	14677607	T to G	G	L-M11* (L*)	
V119.1	14607717 and 14677561	T/T to G/G	T	BT-SRY$_{10831.1}$ (BT)	
V119.2	14677561	G to T	T	L-M11* (L*)	
V122	14607734 or 14677544	C/C to (A/C or C/A)	C	C-M216* (C*)	
V127	14677533	A to G	G	L-M11* (L*)	
V130	14607799 and/or 14677479	C/C to (T/C or C/T) to T/T	C	Q-M3* (Q1a3a*), Q-M194 (Q1a3a2)	*VCY*
V109	14608058 and/or 14677220	G/G - (A/G or G/A) - A/A	A	Y multiple branches	
V110	14610085 and 14675193	A/A to C/C	A	B-M109 (B2a1a)	
V111	14613053 or 14672225	A/A to (G/A or A/G)	A	E-M2* (E1b1a*)	
V112	14614957 and 14670321	A/A to G/G	A	C-M216* (C*)	
V113	14615459 or 14669819	G/G to (A/G or G/A)	G	E-M35* (E1b1b1*), E-V65 (E1b1b1a4)	
V114	14617261 or 14668017	A/A to (G/A or A/G)	A	B-M109 (B2a1a)	
V115	14618433 or 14666845	T/T to (del/T or T/del)	T	A-M31 (A1a)	
V116	14618798 or 14666480	A/A to (G/A or A/G)	A	E-M2* (E1b1a*)	P8 palindrome

[a] Position according to the March 2006 human Y-chromosome reference sequence (NCBI build 36.1).
[b] As to the P8 palindromic region, the allelic state was assigned by considering the 5′–3′ and the 3′–5′ direction for the proximal and the distal arm, respectively. The ancestral state of the mutation V109 could not be unambiguously determined (see text).
[c] Gametologous base on the X chromosome. NA = no X–Y alignment.
[d] Haplogroup nomenclature by lineage in parentheses. Nomenclature according to Karafet et al. (2008).

online), which is more than 5-fold greater than the *PCDH11Y* and the *TBL1Y* genes and the average nucleotide diversity of the MSY ($\pi = 1.52 \times 10^{-4}$) (The International SNP Map Working Group 2001). Nine of the 16 SNPs were found at positions in which the MSY ancestral sequence differed from the X-linked gametologous sequence, resulting in a highly significant enrichment of SNPs at these sites (expected number: 0.9; $P < 10^{-6}$). All these SNPs resulted in the homogenization of preexisting X–Y sequence differences, with the derived allele equal to the gametologous base on the X. Under the hypothesis of neutral mutation and different relative mutation rates among the 4 nt (Li et al. 1984), this result is expected to occur with a probability of 5.0×10^{-4}. Three of the mutations that occurred at GSV sites determine nonsynonymous substitutions, and four mutations are located in the UTRs regions of the gene. In theory, natural selection could account for the observed relative enrichment of SNPs at GSV sites. Positive selection

Table 2. SNP Content and Genetic Diversity of the Analyzed Regions in the Human MSY.

Region	Length (bp)	Number of Y Chromosomes	SNPs	Nucleotide Diversity ($\pi \pm$ SD)
PCDH11Y	23,198	15	17[a]	$1.1 \pm 0.2 \times 10^{-4}$
TBL1Y	14,502	11	7	$1.5 \pm 0.3 \times 10^{-4}$
P8 palindrome	17,194	15	15	$2.0 \pm 0.4 \times 10^{-4}$
Total	54,894		39	
P8 palindrome[b]	15,578	15	8	$1.2 \pm 0.4 \times 10^{-4}$
VCY	1,616	15	7	$10.0 \pm 2.4 \times 10^{-4}$
VCY	1,616	122	16	$7.8 \pm 0.8 \times 10^{-4}$

NOTE.—SD, Standard deviation

[a] The 3-bp ins/del polymorphism in the *PCDH11Y* region has been also included.

[b] P8 palindromic sequence without *VCY* genes.

may favor mutations at GSV sites resulting in a higher *VCX–VCY* similarity, whereas negative selection may erase mutations occurring at gametologous X–Y sites. The latter possibility is unlikely, due to the relatively high nucleotide diversity value ($\pi = 4.0 \pm 0.8 \times 10^{-4}$) estimated for the *VCY* gene, when GSV sites were not taken into account. As to the possibility of positive selection, we observed a slightly significant negative Tajima *D* value ($D = -1.92$; $P < 0.05$). It is worthwhile to note, however, that, as previously pointed out (Hammer et al. 2003), a biased excess of low-frequency variants and negative Tajima *D* values may well be the result of global sampling strategies, as used in this study. Gene conversion therefore remains the most likely scenario. An additional support to the hypothesis that X-to-Y gene conversion has been active in these regions is given by the finding of a further example of multiple, contiguous and phylogenetically equivalent SNPs in the global sample of 122 Y chromosomes (four SNPs spanning 80 bp, see fig. 3).

Number, Tract Length, and Rate of X-to-Y Gene Conversion

Under the hypothesis that gene conversion generated the observed pattern of sequence diversity at the *VCY* genes, we counted the number of X-to-Y gene-conversion events by considering the Y chromosome phylogeny and making the following assumptions: 1) a mutation at a single GSV resulting in the same base as the gametologous sequence on the X chromosome was considered as due to X-to-Y gene conversion; 2) when the same mutation/s was/were found in different chromosomes within a single clade, we considered it/them as the product of a single gene-conversion event that occurred at the root of that clade; 3) multiple and contiguous converted GSVs on the same branch of the phylogeny were also considered as due to a single event of gene conversion; and 4) for the seemingly homoplasic mutation V118 in the *VCY* gene, we made the conservative hypothesis that a single X-to-Y gene conversion occurred at the root of haplogroup E-P147 (fig. 3) and that its recurrence on the E1 portion of the Y tree is due to Y–Y gene-conversion events. Following these criteria, we counted a minimum of five putative independent gene-conversion events, two of which involved multiple SNPs. It was not possible to identify the donor X-copy for any of the X-to-Y gene-conversion events, as the MSY converted bases were present in either all the four *VCX* genes (four cases) or two of them (one case).

The observed minimum gene-conversion tract, measured as the nucleotide segment including the outermost converted GSVs, ranged from 1 to 80 bp. The maximum gene-conversion tract, measured as the distance between the two nearest nonconverted GSVs flanking the converted site/s, ranged from 19 to 169 bp.

By using a slightly modified version (see Materials and Methods) of the equation by Repping et al. (2006), we estimated a rate of 1.8×10^{-7} X-to-Y conversions/bp/year (range 1.2×10^{-7}–2.7×10^{-7}) or of 3.6×10^{-6} conversions/bp/generation (range 2.5×10^{-6}–5.4×10^{-6}, with a 20-year generation) for the *VCY* genes. This is clearly an underestimate of the true value because 1) gene conversion involving sites that are identical between X and Y would be undetectable and 2) Y-to-Y gene-conversion events (Rozen et al. 2003) could lead to back mutation to the ancestral Y allele. The estimated rate values are about 60-fold lower than that estimated for Y–Y gene conversion by Rozen et al. (2003) and two orders of magnitude higher than the estimated MSY mutation rate of 1.6×10^{-9} events/bp/year (Rozen et al. 2003). Thus, X-to-Y gene conversion can be highly effective to increase the level of diversity among human Y chromosomes, but this effect is restricted to the X–Y GSV sites. As a consequence, a highly increased nucleotide diversity is expected to be observed at GSV sites with respect to non-GSV sites in regions actively engaged in ectopic gene conversion, as compared with other regions. Accordingly, we estimated a much higher π value at GSV sites with respect to non-GSV sites in *VCY* genes ($29.9 \pm 9.2 \times 10^{-4}$ vs. $4.0 \pm 0.8 \times 10^{-4}$) as compared with *PCDH11Y*, *TBL1Y* and P8 palindrome without *VCY* (0.0 vs. $1.1 \pm 0.2 \times 10^{-4}$, $3.4 \pm 1.0 \times 10^{-4}$ vs. $1.3 \pm 0.2 \times 10^{-4}$, and 0.0 vs. $1.2 \pm 0.3 \times 10^{-4}$, respectively).

Discussion

Mammalian sex chromosomes evolved from a pair of autosomes through the suppression of X–Y recombination over progressively larger regions. It has long been accepted that recombination between X and Y chromosomes in humans is limited to short telomeric portions known as pseudoautosomal regions (PARs). On the Y chromosome, PARs delineate the borders of a male-specific region comprising 95% of the entire chromosome that came to be known as NRY (nonrecombining region on the Y). The view that homologous recombination in the human Y chromosome only occurs in the PARs has been recently dismissed by the discovery that duplicated sequences are present in the human MSY, within which multiple Y–Y gene-conversion events take place per generation (Rozen et al. 2003; Skaletsky et al. 2003). In this study, we report on the identification of a further mode of "productive" recombination involving the human MSY in form of X-to-Y gene conversion.

We exploited the haploid nature of the Y chromosome and the availability of a stable and reliable human Y

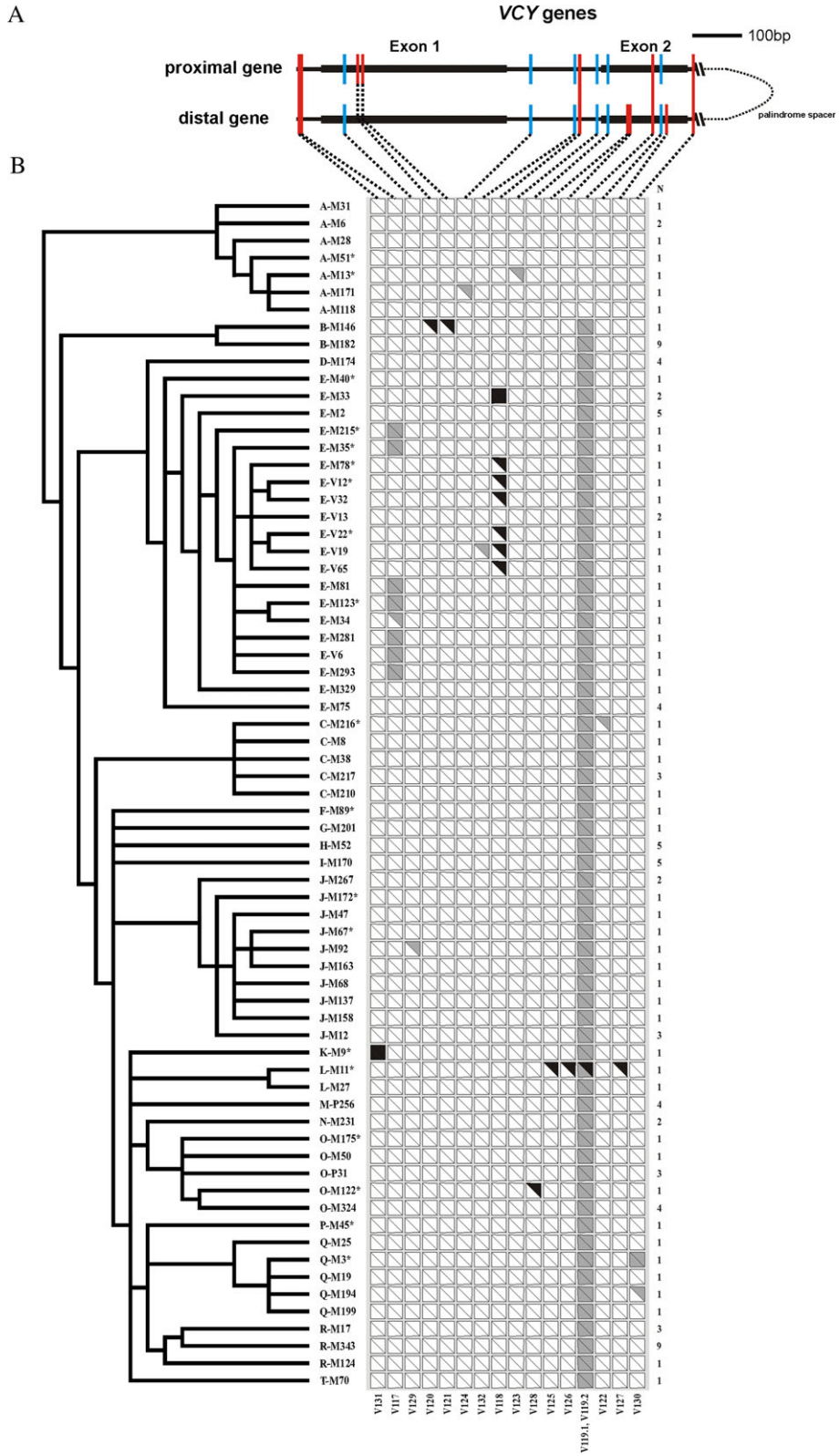

FIG. 3. SNPs identified in the *VCY* region (chrY: 14606996–14607803 and chrY: 14677475–14678282) by sequencing 122 Y chromosomes. (*A*) Location of the SNPs identified in the *VCY* genes. SNPs are depicted as vertical bars (red and blue). Blue bars indicate SNPs present on either the proximal or the distal copy of the gene. (*B*) To the left, a simplified version of the Y chromosome tree (Karafet et al. 2008) showing the phylogenetic relationships of the chromosomes analyzed. SNP names are given at the bottom. Square colors represent the allelic state for each SNP: white, ancestral allele; black, the SNP has arisen in an X–Y GSV and the derived state is equal to the gametologous base on the X; and gray, the SNP has arisen in a site identical between X and Y. *N* represents the number of Y chromosomes sequenced for each of the haplogroups shown to the left; when $N > 1$, chromosomes belonging to different subhaplogroups were chosen. Each square is divided into two triangles, representing the paralogous sites in the two arms of the palindrome.

chromosome phylogeny, to provide insights into the dynamics of interchromosomal gene conversion. The X-transposed sequences were analyzed because of the high density of SNPs reported for this region in the dbSNP (fig. 1), a finding that, in principle, might be due to nonallelic gene conversion (Hurles 2002). Also, because gene conversion is more common among sequences exhibiting high sequence similarity, one might have expected any such process to be more active between recently duplicated regions, as the X-transposed (99% X–Y identity), than other regions. Our resequencing analysis of the X-transposed *PCDH11Y* gene in chromosomes representative of a wide range of the human MSY diversity did not show any evidence for X-to-Y gene conversion being responsible for the relatively high level of putative SNP content. Indeed, our findings did not even corroborate previous reports on the relative enrichment of putative SNPs in this region as compared with other regions. There are 193 SNPs (8.3/kb) in the *PCDH11Y* region analyzed here, as reported in dbSNP, 92.7% of which are in X–Y GSV sites (supplementary table 5, Supplementary Material online). We found only 17 SNPs (0.7/kb) by analyzing 15 Y chromosomes for the same region. None of these SNPs is reported in dbSNP or is in an X–Y GSV site. In principle, differences in the populations where the samples are coming from could account for the differences in SNP content between the dbSNP and present findings. However, the sampling strategy adopted here makes this as an unlikely possibility, unless most SNPs reported in dbSNP for this region are internal to a particular branch of the Y phylogeny not included in our analysis. This suggests that the excess of SNPs of the *PCDH11Y* gene (and possibly the entire X-transposed region), as reported in dbSNP, is artifactual and represents differences between gametologous sequences rather than allelic variants. Unlike *PCDH11Y*, clear evidence was provided indicating that an X-to-Y gene conversion hot spot has been active in the *VCY* gene–containing region. In principle, mutational hot spots could explain the high sequence diversity we found in the *VCY* genes. Also, transient germline hypermutability could have generated closely spaced multiple mutations such as those we observed (Chen et al. 2009). However, some important features, such as the significant excess frequency of mutations at X–Y GSV sites, and the resulting homogenization of the X and Y sequences for all these mutations, point to gene conversion as the underlying mutational mechanism. Altogether, five X-to-Y gene-conversion events were identified, two of which involved multiple substitutions, with a minimum observed tract of 1–80 bp and a maximum tract length of 19–169 bp. Although these results could be also explained by double crossover, gene conversion was assumed to have occurred, due to the short length of the observed sequence changes (Chen et al. 2007).

In mammals, ancient episodes of recombination between highly similar X–Y homologous sequences have been previously only detected in Felidae (Slattery et al. 2000) and in chimpanzees (Bhowmick et al. 2007) by an interspecific phylogenetic approach. Interestingly, signals of gene conversion were found in the latter species by comparing available *VCX* and *VCY* primate sequences; however, no such a signal was found in humans for the same region (Bhowmick et al. 2007). By comparing X and Y human genomic sequences, Skaletsky et al. (2003) observed a region of high X–Y similarity that extended from the 3′ end of the *KALP–KAL1* gene pair through the *VCY–VCX* gene pair. This feature led the authors to suggest either a recent divergence or extensive gene conversion (about 7 Ma) between these X–Y paralogs (see legend to supplementary fig. 10 in Skaletsky et al. 2003). *VCY* genes align with four *VCX* genes, showing a sequence identity 96.2–96.8% (supplementary table 4, Supplementary Material online). By using the present approach, based on an intraspecific phylogenetic study, we demonstrate for the first time that X-to-Y gene conversion has been ongoing in recent human evolution and may have contributed to the high *VCY–VCX* sequence similarity. Outside *VCY* genes, the P8 palindromic sequences analyzed here show a similarly high X–Y sequence identity with one of the X gametologous regions (supplementary table 4, Supplementary Material online), despite the absence of apparent X-to-Y conversion events. Very ancient episodes of gene conversion, undetectable by the approach used here, could account for this discrepancy.

The present study has focused on X-to-Y gene conversion, and no data are available on whether conversion from *VCY* to *VCX* has been also active in the human lineage. Interestingly, five SNPs among those identified here (V120, V121, V126, V127, and V128) are shared between *VCY* and one-to-four gametologous *VCX* genes (rs1058248, rs3198862, rs5934423, rs1728763, and rs6639945, respectively), indicating possible bidirectional gene-conversion events. Comparative sequence diversity analyses of *VCX* genes may help in elucidating this issue.

Combined data from the report of Bhowmick et al. (2007) and this study suggest that the *VCX–VCY* hot spot has been evolutionarily persistent, and its existence possibly predates the divergence of chimpanzee and human species, as previously suggested on different grounds (Skaletsky et al. 2003). This finding is at odds with the observation that allelic homologous recombination (AHR) hot spots minimally overlap between chimpanzee and human species (Ptak et al. 2004; Winckler et al. 2005), whereas it correlates with previous reports of human–chimp conserved NAHR hot spots in the Y chromosome (Rozen et al. 2003; Bosch et al. 2004; Hurles et al. 2004). Further comparative sequencing is required to investigate whether there is a difference in the evolutionary persistence between AHR and NAHR hot spots and/or if the correspondence across species in the locations of NAHR hotspots, if any, is specific to the Y chromosome.

In an attempt to explain the preservation of *VCY* on the human Y chromosome, an adaptive model (the team-work model) has been proposed (Lahn and Page 2000), according to which members of the VCX–VCY protein family can work together by complementing each other in functions involved in spermatogenesis (Lahn and Page 2000; Van Esch et al. 2005). Later, it has been suggested that genes in palindromes might have been able to resist the evolutionary

decay of the Y chromosome thanks to Y–Y gene conversion (Rozen et al. 2003). The revelation of X-to-Y gene conversion in the *VCX–VCY* region suggests an additional explanation for the preservation of this important class of genes, for which a possible role in cognitive development has also been hypothesized (Van Esch et al. 2005).

Only one short region was found to be involved in X-to-Y gene conversion by sequencing more than 50 kb of the MSY. Whether this process is also active in other regions of the MSY remains to be elucidated. A reanalysis of the molecular data reported in Karafet et al. (2008) could help to shed light on this issue. We found that among about 600 phylogenetically characterized Y SNPs reported in Karafet et al. (2008), 202 SNPs are within regions of X–Y identity >80%, and 31 of these 202 SNPs are compatible with X-to-Y gene conversion–driven mutations, having the derived allele equal to the X gametologous base. Interestingly, among these 31 mutations, three pairs of recurrent SNPs (P37.1, P37.2, P41.1, P41.2, P53.1, and P53.2) were found to be located in a narrow DNA region of 36 bp within the *ARSDP* pseudogene. Although parallel de novo mutations cannot be ruled out as the source for these recurrent SNPs, gene conversion between X and Y is a more likely scenario, because two of these mutations (P37.1 and P41.1) are also phylogenetically equivalent (D-M55 haplogroup). As a support to the X-to-Y gene-conversion scenario, in a population genetics survey of the P37 mutation (Scozzari R, unpublished data), we identified five additional SNPs within a 380-bp region containing the P37/P41/P53 markers (supplementary table 6, Supplementary Material online). Three of these SNPs involve contiguous X–Y GSV sites, are phylogenetically equivalent, and have the derived allele equal to the gametologous base on the X chromosome (supplementary fig. 2, Supplementary Material online). These findings, along with the significant excess of MSY SNPs at X–Y GSV sites ($P < 10^{-5}$) in the 380-bp *ARSDP* sequence, strongly suggest this region as being a second X-to-Y gene-conversion hot spot, although the rate and amount of this event still remain to be defined.

Our current findings raise relevant questions for studies of 1) the recent human evolution and 2) the evolutionary partitioning of human sex chromosomes. 1) Based on low mutation rate, most SNPs are used as stable markers in the construction of the Y phylogeny, as being interpreted as of monophyletic origin. We estimated an X-to-Y gene conversion rate in the *VCY* region that is two orders of magnitude higher than the average Y mutation rate. Thus, similarly to previous considerations on the use of Y–Y PSVs (Adams et al. 2006), we suggest caution in using Y SNPs in regions characterized by a high X–Y identity, with the derived allele equal to the X gametologous base. 2) The extent of sequence similarity between X and Y gametologous regions has been recently used to infer the structuring of the X chromosome in different evolutionary strata, these reflecting a stepwise suppression of X–Y recombination (Lahn and Page 1999; Ross et al. 2005). Lahn and Page (1999) observed that genes on the X chromosome having a homologous counterpart on the Y are arranged in four discrete groups (evolutionary strata 1–4), with an orderly location with respect to the degree of X–Y identity (as measured by *Ks*). Subsequently, Ross et al. (2005) reevaluated the degree of identity between X–Y genomic sequences located in the two youngest evolutionary strata (strata 3 and 4). Based on the presence of two contiguous regions characterized by different degrees of identity within stratum 4, they defined a new fifth evolutionary stratum, characterized by markedly great X–Y sequence similarity. However, the most proximal location of the *VCX*-containing region in the stratum 4 (characterized by an overall X–Y identity below 90%) was not consistent with the increased similarity of *VCX* and *VCY* genes (about 95%). According to the authors (Ross et al. 2005), a recent emergence of this gene family in the simian lineage could account for this discrepancy. The revelation of a hot spot of X–Y gene conversion in the *VCY* region offers an alternative/additional explanation for the unusual features of the *VCX*-containing portion of the stratum 4 and corroborates previous suggestions (Marais and Galtier 2003; Lemaitre et al. 2009) that X-to-Y gene conversion might be an important confounding factor in similarity-based evolutionary studies.

A better understanding of gene conversion should provide a clearer picture of the role of this process in double-strand break repair (DSB) and the maintenance of genome integrity. Gene-conversion events are initiated by DSBs and can be defined as the copying of one intact stretch of DNA into another that contains the DSB. Gene conversion can occur between sister chromatids, homologous chromosomes, or paralogous sequences on either the same or different chromosomes. It could be argued that before the DNA-replication event, no homologous sequence can be used by the haploid MSY as a substrate to repair DSB. The degree to which the transfer of genetic information from the X to the Y chromosome, as that shown in this study, can make up for this insufficiency, remains to be determined. Since this manuscript was submitted, Rosser et al. (2009) have obtained evidence for X-to-Y gene conversion at a human translocation hot spot.

Supplementary Material

Supplementary tables 1–6 and supplementary figures 1 and 2 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

This work was supported by the Sapienza Università di Roma, Grandi Progetti di Ateneo, and the Italian Ministry of the University, Progetti di Ricerca di Interesse Nazionale 2007 (both to R.S.). B.T. was supported by an "Ateneo della Scienza e della Tecnica" postdoctoral fellowship. The Sorenson Molecular Genealogy Foundation provided support for P.A.U.

References

Adams SM, King TE, Bosch E, Jobling MA. 2006. The case of the unreliable SNP: recurrent back-mutation of Y-chromosomal

marker P25 through gene conversion. *Forensic Sci Int*. 159: 14–20.
Armengol L, Pujana MA, Cheung J, Scherer SW, Estivill X. 2003. Enrichment of segmental duplications in regions of breaks of synteny between the human and mouse genomes suggest their involvement in evolutionary rearrangements. *Hum Mol Genet*. 12:2201–2208.
Arnheim N, Calabrese P, Tiemann-Boege I. 2007. Mammalian meiotic recombination hot spots. *Annu Rev Genet*. 41:369–399.
Bailey JA, Eichler EE. 2006. Primate segmental duplications: crucibles of evolution, diversity and disease. *Nat Rev Genet*. 7:552–564.
Bailey JA, Gu Z, Clark RA, Reinert K, Samonte RV, Schwartz S, Adams MD, Myers EW, Li PW, Eichler EE. 2002. Recent segmental duplications in the human genome. *Science* 297: 1003–1007.
Bailey JA, Yavor AM, Massa HF, Trask BJ, Eichler EE. 2001. Segmental duplications: organization and impact within the current human genome project assembly. *Genome Res*. 11:1005–1017.
Benovoy D, Drouin G. 2009. Ectopic gene conversions in the human genome. *Genomics* 93:27–32.
Bhowmick BK, Satta Y, Takahata N. 2007. The origin and evolution of human ampliconic gene families and ampliconic structure. *Genome Res*. 17:441–450.
Bosch E, Hurles ME, Navarro A, Jobling MA. 2004. Dynamics of a human interparalog gene conversion hotspot. *Genome Res*. 14:835–844.
Bussaglia E, Clermont O, Tizzano E, et al. (11 co-authors). 1995. A frame-shift deletion in the survival motor neuron gene in Spanish spinal muscular atrophy patients. *Nat Genet*. 11:335–337.
Chen DC, Saarela J, Clark RA, Miettinen T, Chi A, Eichler EE, Peltonen L, Palotie A. 2004. Segmental duplications flank the multiple sclerosis locus on chromosome 17q. *Genome Res*. 14:1483–1492.
Chen J-M, Cooper DN, Chuzhanova N, Férec C, Patrinos GP. 2007. Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet*. 8:762–775.
Chen J-M, Férec C, Cooper DN. 2009. Closely spaced multiple mutations as potential signatures of transient hypermutability in human genes. *Hum Mutat*. 30:1435–1448.
Cheng Z, Ventura M, She X, et al. (12 co-authors). 2005. A genome-wide comparison of recent chimpanzee and human segmental duplications. *Nature* 437:88–93.
Cheung J, Estivill X, Khaja R, MacDonald JR, Lau K, Tsui L-C, Scherer SW. 2003. Genome-wide detection of segmental duplications and potential assembly errors in the human genome sequence. *Genome Biol*. 4:R25.
Cruciani F, La Fratta R, Santolamazza P, et al. (19 co-authors). 2004. Phylogeographic analysis of haplogroup E3b (E-M215) Y chromosomes reveals multiple migratory events within and out of Africa. *Am J Hum Genet*. 74:1014–1022.
Cruciani F, La Fratta R, Torroni A, Underhill PA, Scozzari R. 2006. Molecular dissection of the Y chromosome haplogroup E-M78 (E3b1a): a posteriori evaluation of a microsatellite-network-based approach through six new biallelic markers. *Hum Mutat*. 27:831–832.
Cruciani F, La Fratta R, Trombetta B, et al. (24 co-authors). 2007. Tracing past human male movements in northern/eastern Africa and western Eurasia: new clues from Y-chromosomal haplogroups E-M78 and J-M12. *Mol Biol Evol*. 24:1300–1311.
Cruciani F, Santolamazza P, Shen P, et al. (16 co-authors). 2002. A back migration from Asia to sub-Saharan Africa is supported by high-resolution analysis of human Y-chromosome haplotypes. *Am J Hum Genet*. 70:1197–1214.
Eichler EE. 2001. Recent duplication, domain accretion and the dynamic mutation of the human genome. *Trends Genet*. 17: 661–669.
Eickbush TH, Eickbush DG. 2007. Finely orchestrated movements: evolution of the ribosomal RNA genes. *Genetics* 175:477–485.
Estivill X, Cheung J, Pujana MA, Nakabayashi K, Scherer SW, Tsui L-C. 2002. Chromosomal regions containing high-density and ambiguously mapped putative single nucleotide polymorphisms (SNPs) correlate with segmental duplications in the human genome. *Hum Mol Genet*. 11:1987–1995.
Garcia-Moreno J, Mindell DP. 2000. Rooting a phylogeny with homologous genes on opposite sex chromosomes (gametologs): a case study using avian CHD. *Mol Biol Evol*. 17:1826–1832.
Giordano M, Marchetti C, Chiorboli E, Bona G, Momigliano Richiardi P. 1997. Evidence for gene conversion in the generation of extensive polymorphism in the promoter of the growth hormone gene. *Hum Genet*. 100:249–255.
Gupta PK, Adamtziki E, Budde U, et al. (12 co-authors). 2005. Gene conversions are a common cause of von Willebrand disease. *Br J Haematol*. 130:752–758.
Hallast P, Nagirnaja L, Margus T, Laan M. 2005. Segmental duplications and gene conversion: human luteinizing hormone/chorionic gonadotropin β gene cluster. *Genome Res*. 15: 1535–1546.
Hammer MF, Blackmer F, Garrigan D, Nachman MW, Wilder JA. 2003. Human population structure and its effects on sampling Y chromosome sequence variation. *Genetics* 164:1495–1509.
Hurles M. 2002. Are 100,000 "SNPs" useless? *Science* 298:1509a.
Hurles ME, Willey D, Matthews L, Hussain SS. 2004. Origin of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. *Genome Biol*. 5:R55.
International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* 409: 860–921.
International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* 409:928–933.
Karafet TM, Mendez FL, Meilerman MB, Underhill PA, Zegura SL, Hammer MF. 2008. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. *Genome Res*. 18:830–838.
Lahn BT, Page DC. 1999. Four evolutionary strata on the human X chromosome. *Science* 286:964–967.
Lahn BT, Page DC. 2000. A human sex-chromosomal gene family expressed in male germ cells and encoding variably charged proteins. *Hum Mol Genet*. 9:311–319.
Larkin MA, Blackshields G, Brown NP, et al. (13 co-authors). 2007. ClustalW and ClustalX version 2.0. *Bioinformatics* 23:2947–2948.
Lemaitre C, Braga MDV, Gautier C, Sagot M-F, Tannier E, Marais GAB. 2009. Footprints of inversions at present and past pseudoautosomal boundaries in human sex chromosomes. *Genome Biol Evol*. 1:56–66.
Li W-H, Wu C-I, Luo C-C. 1984. Nonrandomness of point mutation as reflected in nucleotide substitutions in pseudogenes and its evolutionary implications. *J Mol Evol*. 21:58–71.
Marais G, Galtier N. 2003. Sex chromosomes: how X–Y recombination stops. *Curr Biol*. 13:R641–R643.
Nei M. 1987. Molecular evolutionary genetics. New York: Columbia University Press.
Pentao L, Wise CA, Chinault AC, Patel PI, Lupski JR. 1992. Charcot-Marie-Tooth type 1A duplication appears to arise from recombination at repeat sequences flanking the 1.5 Mb monomer unit. *Nat Genet*. 2:292–300.
Ptak SE, Roeder AD, Stephens M, Gilad Y, Pääbo S, Przeworski M. 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. *PLoS Biol*. 2:849–855.
Repping S, van Daalen SKM, Brown LG, et al. (11 co-authors). 2006. High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet*. 38:463–467.

Reyniers E, Van Thienen M-N, Meire F, De Boulle K, Devries K, Kestelijn P, Willems PJ. 1995. Gene conversion between red and defective green opsin gene in blue cone monochromacy. *Genomics* 29:323–328.

Roesler J, Curnutte JT, Rae J, Barrett D, Patino P, Chanock SJ, Goerlach A. 2000. Recombination events between the p47-phox gene and its highly homologous pseudogenes are the main cause of autosomal recessive chronic granulomatous disease. *Blood* 95:2150–2156.

Ross MT, Grafham DV, Coffey AJ, et al. (282 co-authors). 2005. The DNA sequence of the human X chromosome. *Nature* 434:325–337.

Rosser ZH, Balaresque P, Jobling MA. 2009. Gene conversion between the X chromosome and the male-specific region of the Y chromosome at a translocation hotspot. *Am J Hum Genet*. 85:130–134.

Rozas J, Sánchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Rozen S, Skaletsky H, Marszalek JD, Minx PJ, Cordum HS, Waterston RH, Wilson RK, Page DC. 2003. Abundant gene conversion between arms of palindromes in human and ape Y chromosomes. *Nature* 423:873–876.

Samonte RV, Eichler EE. 2002. Segmental duplications and the evolution of the primate genome. *Nat Rev Genet*. 3:65–72.

She X, Jiang Z, Clark RA, Liu G, Cheng Z, Tuzun E, Church DM, Sutton G, Halpern AL, Eichler EE. 2004. Shotgun sequence assembly and recent segmental duplications within the human genome. *Nature* 431:927–930.

Shen P, Wang F, Underhill PA, et al. (13 co-authors). 2000. Population genetic implications from sequence variation in four Y chromosome genes. *Proc Natl Acad Sci USA*. 97:7354–7359.

Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, et al. (40 co-authors). 2003. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423:825–837.

Slattery JP, Sanner-Wachter L, O'Brien SJ. 2000. Novel gene conversion between X–Y homologues located in the non-recombining region of the Y chromosome in Felidae (Mammalia). *Proc Natl Acad Sci USA*. 97:5307–5312.

Stankiewicz P, Shaw CJ, Withers M, Inoue K, Lupski JR. 2004. Serial segmental duplications during primate evolution result in complex human genome architecture. *Genome Res*. 14:2209–2220.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123: 585–595.

Underhill PA, Passarino G, Lin AA, Shen P, Mirazón Lahr M, Foley RA, Oefner PJ, Cavalli-Sforza LL. 2001. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann Hum Genet*. 65:43–62.

Underhill PA, Shen P, Lin AA, et al. (21 co-authors). 2000. Y chromosome sequence variation and the history of human populations. *Nat Genet*. 26:358–361.

Van Esch H, Hollanders K, Badisco L, Melotte C, Van Hummelen P, Vermeesch JR, Devriendt K, Fryns J-P, Marynen P, Froyen G. 2005. Deletion of *VCX-A* due to NAHR plays a major role in the occurrence of mental retardation in patients with X-linked ichthyosis. *Hum Mol Genet*. 14:1795–1803.

Vázquez N, Lehrnbecher T, Chen R, Christensen BL, Gallin JI, Malech H, Holland S, Zhu S, Chanock SJ. 2001. Mutational analysis of patients with p47-phox-deficient chronic granulomatous disease: the significance of recombination events between the *p47–phox* gene (*NCF1*) and its highly homologous pseudogenes. *Exp Hematol*. 29:234–243.

Visser R, Shimokawa O, Harada N, Kinoshita A, Ohta T, Niikawa N, Matsumoto N. 2005. Identification of a 3.0-Kb major recombination hotspot in patients with Sotos syndrome who carry a common 1.9-Mb microdelection. *Am J Hum Genet*. 76:52–67.

Winckler W, Myers SR, Richter DJ, et al. (11 co-authors). 2005. Comparison of fine-scale recombination rates in humans and chimpanzees. *Science* 308:107–111.

Zhang L, Lu HHS, Chung W-y, Yang J, Li W-H. 2005. Patterns of segmental duplication in the human genome. *Mol Biol Evol*. 22:135–141.