Copy number variation arising from gene conversion on the human Y chromosome

Wentao Shi^{1,2} · Andrea Massaia^{1,3} · Sandra Louzada¹ · Ruby Banerjee¹ · Pille Hallast^{1,4} · Yuan Chen¹ · Anders Bergström¹ · Yong Gu¹ · Steven Leonard¹ · Michael A. Quail¹ · Qasim Ayub^{1,5} · Fengtang Yang¹ · Chris Tyler-Smith¹ · Yali Xue¹

Received: 26 September 2017 / Accepted: 28 November 2017 / Published online: 5 December 2017
© The Author(s) 2017. This article is an open access publication

Abstract

We describe the variation in copy number of a ~ 10 kb region overlapping the long intergenic noncoding RNA (lincRNA) gene, *TTY22*, within the IR3 inverted repeat on the short arm of the human Y chromosome, leading to individuals with 0–3 copies of this region in the general population. Variation of this CNV is common, with 266 individuals having 0 copies, 943 (including the reference sequence) having 1, 23 having 2 copies, and two having 3 copies, and was validated by break-point PCR, fibre-FISH, and 10× Genomics Chromium linked-read sequencing in subsets of 1234 individuals from the 1000 Genomes Project. Mapping the changes in copy number to the phylogeny of these Y chromosomes previously established by the Project identified at least 20 mutational events, and investigation of flanking paralogous sequence variants showed that the mutations involved flanking sequences in 18 of these, and could extend over > 30 kb of DNA. While either gene conversion or double crossover between misaligned sister chromatids could formally explain the 0–2 copy events, gene conversion is the more likely mechanism, and these events include the longest non-allelic gene conversion reported thus far. Chromosomes with three copies of this CNV have arisen just once in our data set via another mechanism: duplication of 420 kb that places the third copy 230 kb proximal to the existing proximal copy. Our results establish gene conversion as a previously under-appreciated mechanism of generating copy number changes in humans and reveal the exceptionally large size of the conversion events that can occur.

Wentao Shi, Andrea Massaia, and Sandra Louzada contributed equally to this work.

Electronic supplementary material The online version of this article (<https://doi.org/10.1007/s00439-017-1857-9>) contains supplementary material, which is available to authorized users.

✉ Chris Tyler-Smith
cts@sanger.ac.uk

✉ Yali Xue
ylx@sanger.ac.uk

¹ Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridge CB10 1SA, UK

² Department of Genetics, School of Basic Medical Sciences, Tianjin Medical University, Tianjin 30070, China

³ National Heart and Lung Institute, Imperial College London, London SW7 2AZ, UK

⁴ Institute of Molecular and Cell Biology, University of Tartu, Tartu 51010, Estonia

⁵ Present Address: School of Science, Monash University Malaysia, Jalan Lagoon Selatan, Bandar Sunway, 47500 Subang Jaya, Selangor Darul Ehsan, Malaysia

Introduction

Copy number variation (CNV, which also refers to copy number variants) is well-documented in the human genome, affecting more nucleotides than are affected by SNP variation and contributing abundantly to phenotypic diversity (Lupski 2015; Sudmant et al. 2015a, b). CNVs arise by several mechanisms, including non-allelic homologous recombination, non-homologous end joining, microhomology-mediated break-induced replication, and retroelement insertions (Hastings et al. 2009; Carvalho and Lupski 2016). Gene conversion, the non-reciprocal transfer of genetic information from one locus or allele to another, is also well-documented in the human genome as a possible consequence of double-strand breaks and subsequent repair involving homologous regions, including the formation and resolution of a double Holliday junction during meiosis (Szostak et al. 1983). It generally involves short stretches of DNA converting between alleles or nearby paralogs, requires high sequence similarity, and occurs less frequently between

different chromosomes. Allelic conversion tract lengths of up to 22 kb are known (Wang et al. 2012), while lengths for non-allelic events are shorter, although > 9 kb has been reported (Chen et al. 2007; Hallast et al. 2013). These two processes of CNV and gene conversion are generally considered quite distinct, but here, we describe a CNV, whose origin is best explained by gene conversion events, linking the two processes.

The male-specific region of the Y chromosome offers unique opportunities for investigating both CNV and gene conversion, because (1) it is particularly tolerant of genetic variation (Poznik et al. 2016), so a wide variety of variants persist in the population and (2) the lack of recombination between different Y lineages allows the history of variants to be identified from the phylogeny (Jobling 2008; Jobling and Tyler-Smith 2017; Massaia and Xue 2017; Trombetta and Cruciani 2017). We have previously described the genetic variation in a set of 1244 diverse worldwide Y chromosomes sequenced in phase 3 of the 1000 Genomes Project, including the identification and validation of 110 CNVs (Poznik et al. 2016). The current study presents a detailed analysis of one of these, dbVar ID esv3818053 [chrY: 9,640,466–9,653,590 GRCh37 (hg19); chrY: 9,802,857–9,815,981, GRCh38 (hg38)], here designated by the more descriptive name *TTY22*-CNV, because it overlaps with ~ 85% of this lincRNA gene and changes the number of its functional copies.

Results and discussion

Validation of *TTY22*-CNV

TTY22-CNV lies within the ~ 300 kb long inverted repeat 3 (IR3), which has two copies on the short arm of the Y chromosome, at approximately 6.1–6.4 and 9.4–9.7 Mb (GRCh37) (Skaletsky et al. 2003). In the reference sequence (Skaletsky et al. 2003), derived mostly from haplogroup R1b (Wei et al. 2103), the proximal copy contains an additional ~ 10 kb segment which carries most of *TTY22* (Fig. 1a). Variation in copy number of this additional segment was initially identified using Genome STRiP and validated by array-CGH (Poznik et al. 2016), revealing 0–3 copies, compared with the single copy in the reference sequence. *TTY22*-CNV was further validated here by establishing a PCR assay using one pair of flanking primers which generated a 568 bp product in the absence of the *TTY22*-CNV, and one pair within *TTY22* which generated a 249 bp product in its presence (Table S1). Structures matching the reference sequence with one copy of *TTY22*-CNV generate both products, those with 0 copies generate only the 568 bp band, and those with 2 or 3 copies only the 249 bp band (Fig. 1c). In addition, fibre-FISH experiments using probes

generated from two partially overlapping BAC clones (P1 and P2) spanning *TTY22*-CNV, together with two 5 kb custom PCR-generated probes (P3 and P4, combined; Table S1) lying mainly within *TTY22*-CNV are expected to show differential patterns: (1) hybridization of only the BAC clones plus a small fragment of P4 when the *TTY22*-CNV is absent and (2) hybridization of all four probes when the gene is present. As expected, both patterns were detected in similar proportions (7:9, Fig. 1b, Fig. S1) in samples with structures matching the reference sequence. Conversely, only the former pattern (18:0, Fig. S1) was detected in samples with 0 copies of *TTY22*-CNV and only the latter pattern (0:12, Fig. S1) was observed in samples with 2 copies of *TTY22*-CNV (Fig. 1b). Samples with three copies showed the latter pattern plus a separate, additional fibre-FISH signal, and are described separately below. Finally, 10x Genomics Chromium linked-read data (Zheng et al. 2016) were available for samples with one and two copies; the sample with one copy shows uniform distributions of barcode sharing and read depth across both *TTY22*-CNV locations, while the sample with two copies shows increased barcode sharing and read depth in IR3_{proximal} and decreased sharing and depth in IR3_{distal} (Fig. S2). Thus, this combination of validation approaches confirms both the predicted details of the breakpoints (PCR) and the broader context of the *TTY22*-CNV location within IR3 (fibre-FISH and linked-read sequencing), so we conclude that individuals with 0, 1, or 2 copies of *TTY22*-CNV can be explained by variation within the IR3 repeats.

Population variation and phylogeny of *TTY22*-CNV

In the 1234 Y chromosomes examined for copy number (ten samples were excluded because of Y-chromosomal mosaicism), 266 had 0 copies, 943 had 1, 23 had 2 copies, and two had 3 copies (Table S2). Examination of the Y-chromosomal phylogeny established for these samples using Y-SNP variation (Poznik et al. 2016) showed that, as expected, the different copy numbers of *TTY22*-CNV were often clustered in branches of the phylogeny, so that the number of mutational events inferred was just 20 (Fig. 2, and see below for further details), nevertheless indicating a relatively high mutation rate compared with SNPs. IR3 (including *TTY22*-CNV) is not detectable in the chimpanzee or gorilla genome sequences (the longest BLAST hit is 7.4 kb, in block B, and shares only 87.97% sequence identity with chimpanzee Y:15,046,773–15,054,094) and the *TTY22*-CNV PCR primers do not detect any product in male chimpanzee or gorilla (Fig. 1c), so examination of an outgroup is not informative about the ancestral state of *TTY22*-CNV. Within the human Y-chromosomal phylogeny sampled, all major haplogroups include structures matching the reference with one copy of *TTY22*-CNV (Fig. 2), so we infer that the most recent

Fig. 1 Location and validation of *TTTY22*-CNV. **a** Top: the Y chromosome, with the red box indicating the location of the ~ 3.7 Mb region carrying the two copies of IR3 (black) and intervening sequence (grey). Middle: the IR3 structure in the reference sequence (GRCh37) showing IR3_{distal} (left) divided into two blocks (B and A), and IR3_{proximal} (right) similarly divided into two blocks (A and B) separated by the *TTTY22*-CNV (small white block). F1, R1 and F2, R2 are pairs of PCR primers used in validation (see part C). The lighter purple block shows the location of the previously-reported IR3 inversion breakpoint interval (Turner et al. 2006). Coordinates are given for GRCh37 in this study to be consistent with previous work on this CNV in the same samples (Poznik et al. 2016). Lower: location of probes P1, P2, P3, and P4 used in fibre-FISH validation (see part B) and schematic representation of gene conversion events which would change the single copy of *TTTY22*-CNV in the reference sequence to 0 copies (left) or 2 copies (right). **b** Fibre-FISH valida-

tion: HG00096, with 1 copy of *TTTY22*-CNV matching the reference sequence, shows both fibre-FISH patterns that include P3 (entirely within *TTTY22*-CNV, upper) and patterns that lack P3 (lower). NA19146, with 0 copies of *TTTY22*-CNV, shows only the pattern that lacks P3. NA18953, with 2 copies of *TTTY22*-CNV, shows only the pattern that includes *TTTY22*-CNV. **c** PCR validation: primers F1 and R1 amplify across *TTTY22*-CNV (see part A) and produce a 568 bp product in the absence of *TTTY22*-CNV, while primers F2 and R2 amplify a fragment within *TTTY22*-CNV and produce a 249 bp product in the presence of *TTTY22*-CNV. HG00096, with 1 copy of *TTTY22*-CNV matching the reference sequence, shows both products; NA19146, with 0 copies of *TTTY22*-CNV, shows only the 568 bp product; NA18953, with 2 copies, shows only the 249 bp product; NA19661, with 3 copies, shows only the 249 bp product: the same pattern as the 2-copy sample; neither human female, male chimpanzee nor male gorilla produce any product

common ancestor of the Y chromosomes examined probably also carried one copy.

A gene conversion mechanism for most *TTTY22*-CNV variation

Since the copy number of the IR3 repeat as a whole does not vary in the samples examined (see below for a partial

Fig. 2 Phylogenetic distribution of *TTY22*-CNV copy number variation. Left: simplified Y-chromosomal phylogeny showing the branching pattern of the lineages considered in this study; branch lengths do not correspond to the number of SNPs. Subsequent columns show the numbers of samples in our data set of 1234 chromosomes with 1, 0, 2, or 3 copies of *TTY22*-CNV, the number of deletion (Dels) and duplication (Dups) events inferred from the full phylogeny (total of 20) (Poznik et al. 2016), and for comparison, the IR3 orientation reported for these haplogroups, where available (Repping et al. 2006). The simplified haplogroups common between the two data sets were named according to the International Society of Genetic Genealogy (2013): Y-DNA Haplogroup Tree 2013, Version: (8.89), Date: (31 December 2013), <http://isogg.org/tree/2013/index13.html>

exception), the most likely mechanisms to generate 0, 1, or 2 copies of *TTY22*-CNV variation would be gene conversion or double crossover in the IR3 sequences flanking *TTY22*-CNV, either replacing the proximal copy with the distal to generate 0 copies of *TTY22*-CNV, or replacing the distal copy with the proximal to generate 2 copies (Fig. 1a, lower section). These flanking sequences are 99.45% identical (99.65% for Block A; 99.13% for Block B), but show occasional variants in the reference sequence differentiating the proximal and distal IR3 copies, known as Paralogous Sequence Variants (PSVs) or Sequence Family Variants (SFVs), which are expected to be present in the population as well. The low-coverage sequences available from the 1000 Genomes Project prevented reliable de novo detection of PSVs, because first, an erroneous PSV allele call might be made in a single read and misinterpreted as a true variant, and second, PSV state would not be called at all if there was zero coverage of a particular position. In contrast, the genotypes of PSVs known in the reference sequence can be extracted reliably, since the probability of a single-nucleotide variant that matches a known PSV allele being genuine is much higher than for a novel variant, even for low-coverage sequences. By aligning the reference sequences of the distal and proximal IR3 copies, we identified 165 PSVs and counted the number of reads carrying each PSV allele in each sample, usually detecting both alleles (Table S3). In this way, sizes of mutations including *TTY22*-CNV and all such *TTY22*-CNV mutations throughout IR3 were estimated (Figs. 3, 4, respectively; Table S4). Missing data increase the uncertainty of these estimates and thus decrease the resolution of the mapping; low sample depth increases the probability of data being missing and thus decreases the resolution. The resolution of the size estimates was entirely determined by the location of the PSVs: we can only detect a mutation when a PSV is affected. Therefore, for every mutation event, we give two numbers: the maximum and minimum lengths given the information from the PSVs.

Examination of the PSV profiles adjacent to the *TTY22*-CNV thus provided the minimum and maximum size estimates for the genomic region accompanying the CNV change, as shown in Fig. 3. We then mapped the CNV

changes onto the phylogeny constructed using the SNPs in the same sample set (Poznik et al. 2016). We identified 12 different mutation sizes, which could be resolved into 20 mutation events using the phylogeny, since events with indistinguishable sizes sometimes occurred on independent branches of the phylogeny. The largest mutation had a minimum size of > 32 kb.

Since gene conversion and double crossover produce indistinguishable structures (Chen et al. 2007), additional factors have to be considered to distinguish between the two possibilities. It is generally reasoned that crossovers are rare, so the chance of two occurring in close proximity is low, and therefore, structures resulting from exchange of information over lengths of 10 kb or less, which have been abundantly documented on the Y chromosome (Rozen et al. 2003; Bosch et al. 2004; Hurles et al. 2004; Hallast et al. 2013), have been interpreted as resulting from gene conversion. It is notable that gene conversion events of up to 9 kb have been reported on the Y chromosome (Hallast et al. 2013), exceeding the maximum size of ~ 4 kb on other chromosomes (Dumont and Eichler 2013; Trombetta and Cruciani 2017). Nevertheless, the large size of some of the events has been reported (Williams 1998; Repping et al. 2006). Although recurrent, only 12 changes in orientation for this inversion have so far been identified here suggests that reconsideration of the possible involvement of double crossovers is merited. We know of no reported measurement of the frequencies of double crossovers on the human Y chromosome. However, some information about the frequencies of single crossovers is available. Single crossovers between sister chromatids would result in isodicentric or acentric chromosomes, which are both very rare (and evolutionarily lethal), but intra-chromatid single crossovers result in inversions which are unlikely to affect the phenotype and have in fact been inferred between IR3_{proximal} and IR3_{distal} within the Y-chromosomal phylogeny (Repping et al. 2006). Repping et al. counted 12 inversions in the phylogeny they sampled and inferred a mutation rate of $\geq 2.3 \times 10^{-4}$ per generation. This phylogeny consisted of a slightly different set of haplogroups from the current study, so we constructed and examined a consensus phylogeny based on the subset of haplogroups in common between the two studies. On this consensus phylogeny, there were eight inversion events and 13 *TTY22*-CNV copy number changes (Fig. 2). Thus, *TTY22*-CNV mutations are more common than single intra-chromatid crossovers. Double cross-over events close together should be considerably more rare than single crossovers. The known breakpoints of single crossovers are also located in a different region of IR3 (Turner et al. 2006, 2008), suggesting that the known crossovers in IR3 are both structurally independent from, and less frequent than, changes to *TTY22*-CNV copy number. We, therefore, conclude that

Fig. 3 Changes to PSVs flanking *TTTY22*-CNV. Left, Y-chromosomal phylogeny of the haplogroups carrying non-reference copy numbers of *TTTY22*-CNV. In haplogroups R2a and O2b, there are two independent copy number mutations, and we added (1) and (2) after the haplogroup names to distinguish them. The next columns show the sample sizes, and minimum (Min) and maximum (Max) lengths of gene conversion events surrounding each CNV deduced from changes to the PSV patterns. Right, PSV patterns,

colored according to the reference sequence (top) as indicated in Fig. 1a, proximal and distal positions of PSVs (A6, A5 ... B8, B9) are shown in Table S3. The black box indicates the minimum length of gene conversion; the maximum length could extend to the next PSV on either side. The relative paucity of events affecting left-hand-side PSVs reflects the larger physical distances of these PSVs from *TTTY22*-CNV. The PSVs patterns of 934 individuals with one copy of *TTTY22*-CNV are quite diverse and are summarised in Table S6

Fig. 4 Minimum length distributions of gene conversion events inferred from changes to PSV patterns in IR3 as a whole (distal to proximal, blue; proximal to distal, red) and for events involving *TTTY22*-CNV (green), which are on average larger. The events here

are minimum numbers inferred from PSV patterns, and the same pattern will sometimes have arisen independently on different Y-chromosomal lineages

double crossovers do not provide a plausible mechanism for *TTTY22*-CNV copy number change, while gene conversion does, as noted for other Y loci (Hallast et al. 2013).

A distinct and more standard mechanism generating three copies of *TTTY22*-CNV

The two individuals with three copies of *TTTY22*-CNV are clustered in the phylogeny and share a common origin (Fig. 5a, b) which, in contrast to all the other *TTTY22*-CNV variants, cannot be accounted for by a simple gene conversion event between IR3 copies. Fibre-FISH shows the presence of a third copy of *TTTY22*-CNV and > 100 kb of flanking sequence (the maximum tested by this method) located ~ 230 kb away from one of the reference copies (Fig. 5d). Measurement of read depth in the two individuals carrying three copies identified ~ 420 kb extending from within IR3_{proximal} into the proximal flanking region with increased read depth, and analysis of the PSVs within this segment showed over-representation of proximal PSVs. For the part of IR3 shown by the fibre-FISH and read depth to be present in three copies, the proximal:distal PSV read depth ratio was 1.7 or 1.9 in the two individuals carrying three copies of *TTTY22*-CNV, while for the part present in two copies, the ratios were 1.3 and 1.0 (Table S5). Together, these data suggest that copy 3 originated by tandem duplication of a 420 kb region including *TTTY22*-CNV in IR3_{proximal}, from a chromosome with two copies (Fig. 5c).

Implications of *TTTY22*-CNV variation

Men carrying non-reference copy numbers of *TTTY22*-CNV are likely to have the corresponding numbers of copies of the functional *TTTY22* gene, because ~ 15% of this gene that lies outside the CNV is provided by the flanking sequence. Their high frequency, and the large numbers who sometimes share the same mutational event, most marked in an E1b sub-lineage which carries 0 copies and is very common in sub-Saharan Africa, represented 260 times in the current data set (Poznik et al. 2016), demonstrate that copy numbers 0–3 allow male lineage expansion, and are unlikely to be detrimental. Given the long-term persistence of lineages with 0–3 copies of *TTTY22*-CNV in the population, and the extreme drift experienced by the Y chromosome, this variation is compatible with evolutionary neutrality. Nevertheless, the location of *TTTY22* in the CNV suggests that the abundance of this transcript may vary between men. The gene is transcribed primarily in testis, with a lower level in brain (The GTEx Consortium 2017) and further work is needed to investigate whether or not subtle phenotypic consequences of variation in transcript levels can be detected. Men with three copies of *TTTY22*-CNV carry an additional large duplication of proximal Yp sequences and have in addition three copies of *TTTY23*, compared with two copies (one in each IR3 repeat) in other men. This duplication has also spread in the population, so also seems unlikely to be detrimental.

Overall, our findings highlight gene conversion as an additional mechanism for generating CNV in the human

Fig. 5 Characterization of chromosomes with three copies of *TTY22*-CNV. **a** Read depth in 1 kb bins between IR3_{proximal} and the Y chromosome centromere for the two samples with three copies of *TTY22*-CNV (NA19685 and NA19661, central panels) and two phylogenetically-nearby samples with one copy (NA07357 and NA12144, top and bottom panels). The pink shading highlights a 420 kb region with increased read depth, resulting from increased copy number. The phylogenetic relationships of these four chromosomes are indicated to the left; the number of SNPs on each branch is shown (Poznik et al. 2016). **b** Validation of the copy number differences using published array-CGH data (Poznik et al. 2016), where the same shaded region shows increased log2 ratio. **c** Schematic representation of the origin of the chromosome with three copies of *TTY22*-CNV, starting with a gene conversion event to generate two copies followed by tandem duplication of a 420 kb region to generate the third copy. **d** Fibre-FISH using the same probes as described in Fig. 1 revealing the location of the 420 kb insertion. An expanded view of the fibre-FISH is included in Fig. S1

genome, especially on the Y chromosome, that has previously received little attention. The link between these two processes should be further explored and considered when either is investigated.

Methods

Data sets

We used the existing Y-chromosomal sequence data and CNV calls from phase 3 of the 1000 Genomes Project, where an initial CNV call set had been made using Genome STRiP and calls validated using array-CGH (Poznik et al. 2016). We also generated new 10× Genomics Chromium libraries from two of these samples (HG01097 and NA18953) following the manufacturer's instructions (<https://www.10xgenomics.com/genome/>) followed by sequencing on the Illumina HiSeq X platform with 150 bp paired-end reads to a depth of ~ 30×. The sequence data were processed using the LongRanger 1.0 software using the reference sequence GRCh37 and viewed using the Loupe 2.1.0 software from 10× Genomics. All coordinates in this paper are based on GRCh37 to make them compatible with the initial CNV calls (Poznik et al. 2016).

TTY22-CNV validation

We designed two sets of primers to validate the presence or absence of the 10 kb insert of *TTY22*-CNV (dbVar description: <https://www.ncbi.nlm.nih.gov/dbvar/variants/esv3818053/#VariantGenome>), one set spanning the breakpoint of the empty site and the other lying within the 10 kb insert region. Primers, PCR conditions, and predicted product sizes are described in Table S1.

Molecular combing fibre-FISH experiments were carried out as described previously (Poznik et al. 2016).

Probes consisted of two BAC clones (RP11-117N22 and RP11-453C1, obtained from the clone archive resource of Wellcome Trust Sanger Institute) used to identify the genomic region of interest, as well as two custom PCR probes of ~ 5 kb each lying mainly within the insert of *TTY22*-CNV (details also in Table S1), which were combined to produce a single probe to distinguish the presence of the insert from its absence. We validated the CNV calls using both PCR and fibre-FISH for the same four samples: NA19146 with 0, HG00096 with 1, NA18953 with 2, and NA19661 with 3 copies. In addition, we applied 10× Genomics Chromium to two samples: HG01097 with 1 and NA18953 with 2 copies of *TTY22*-CNV.

Sequence and phylogenetic analyses

We defined blocks A and B within IR3 (<https://genome.ucsc.edu/>) as being separated by *TTY22*-CNV, and measured their similarity between the two copies of IR3 as 99.65% for the ~ 170 kb of block A and 99.13% for the ~ 101 kb of block B.

We placed the *TTY22*-CNV copy number of each samples onto the full Y-chromosomal phylogeny based on SNPs (Poznik et al. 2016) to infer the number of mutational events (deletion or duplication). To understand the relation between the *TTY22*-CNV and the IR3 inversion events (Repping et al. 2006), we placed both *TTY22*-CNV copy number changes and IR3 inversion events onto a simplified phylogeny as described in the main text and compared their phylogenetic locations.

To investigate gene conversion events, we first identified PSVs (Table S3) between the two copies of IR3 from the reference sequence after aligning them using Basic Local Alignment Search Tool (<https://blast.ncbi.nlm.nih.gov/Blast.cgi>), and then counted the number of reads covering each allele of each PSV from the BAM files for each individual in the 1000 Genomes Project phase 3 data using a custom Perl script. To allow for the low sequence coverage, we established the following protocol. When six or more reads covering the same PSV were found in a sample, the chance of both IR3 copies being represented was ~ 97%, and so a ≥ 6:0 ratio of PSV alleles was taken to indicate that both IR3 copies carried the same allele, implying that a gene conversion (or double crossover) event had occurred. The length of such mutational events was inferred by extending the analysis to PSVs adjacent to the ≥ 6:0 seed, accepting ≥ 1:0 counts of the same allele as resulting from the same event.

Data access

All data on Y-chromosomal variation from the 1000 Genomes Project phase 3 are freely available:<http://>

<ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/>. PSV read coverage is reported in Table S3.

Acknowledgements We thank Ed Hollox for comments. Our work was supported by The Wellcome Trust (098051); W.S. was also supported by the State Scholarship fund (No. 201606940004) of the China Scholarship Council, and a National Natural Science Foundation of China grant (No. 31201029); and P.H. was supported by Estonian Research Council Grant PUT1036.

Compliance with ethical standards

Ethical approval Cell lines and DNA samples were obtained from the Coriell Institute for Medical Research (Camden, New Jersey, USA) following their ethical procedures (<https://catalog.coriell.org/1/Support/FirstOrder>).

Conflict of interest The authors declare that they have no conflict of interest.

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

- Bosch E, Hurles ME, Navarro A, Jobling MA (2004) Dynamics of a human interparalog gene conversion hotspot. *Genome Res* 14:835–844. <https://doi.org/10.1101/gr.2177404>
- Carvalho CM, Lupski JR (2016) Mechanisms underlying structural variant formation in genomic disorders. *Nat Rev Genet* 17:224–238. <https://doi.org/10.1038/nrg.2015.25>
- Chen JM, Cooper DN, Chuzhanova N, Ferec C, Patrinos GP (2007) Gene conversion: mechanisms, evolution and human disease. *Nat Rev Genet* 8:762–775. <https://doi.org/10.1038/nrg2193>
- Dumont BL, Eichler EE (2013) Signals of historical interlocus gene conversion in human segmental duplications. *PLoS One* 8:e75949. <https://doi.org/10.1371/journal.pone.0075949>
- Hallast P, Balaesque P, Bowden GR, Ballereau S, Jobling MA (2013) Recombination dynamics of a human Y-chromosomal palindrome: rapid GC-biased gene conversion, multi-kilobase conversion tracts, and rare inversions. *PLoS Genet* 9:e1003666. <https://doi.org/10.1371/journal.pgen.1003666>
- Hastings PJ, Lupski JR, Rosenberg SM, Ira G (2009) Mechanisms of change in gene copy number. *Nat Rev Genet* 10:551–564. <https://doi.org/10.1038/nrg2593>
- Hurles ME, Willey D, Matthews L, Hussain SS (2004) Origins of chromosomal rearrangement hotspots in the human genome: evidence from the AZFa deletion hotspots. *Genome Biol* 5:R55. <https://doi.org/10.1186/gb-2004-5-8-r55>
- Jobling MA (2008) Copy number variation on the human Y chromosome. *Cytogenet Genome Res* 123:253–262. <https://doi.org/10.1159/000184715>
- Jobling MA, Tyler-Smith C (2017) Human Y-chromosome variation in the genome-sequencing era. *Nat Rev Genet Online*. <https://doi.org/10.1038/nrg.2017.36>
- Lupski JR (2015) Structural variation mutagenesis of the human genome: impact on disease and evolution. *Environ Mol Mutagen* 56:419–436. <https://doi.org/10.1002/em.21943>
- Massaia A, Xue Y (2017) Human Y chromosome copy number variation in the next generation sequencing era and beyond. *Hum Genet* 136:591–603
- Poznik GD, Xue Y, Mendez FL, Willems TF, Massaia A, Wilson Sayres MA, Ayub Q, McCarthy SA, Narechania A, Kashin S, Chen Y, Banerjee R, Rodriguez-Flores JL, Cerezo M, Shao H, Gymrek M, Malhotra A, Louzada S, Desalle R, Ritchie GR, Cerveira E, Fitzgerald TW, Garrison E, Marcketta A, Mittelman D, Romanovitch M, Zhang C, Zheng-Bradley X, Abecasis GR, McCarroll SA, Flicek P, Underhill PA, Coin L, Zerbino DR, Yang F, Lee C, Clarke L, Auton A, Erlich Y, Handsaker RE, Genomes Project C, Bustamante CD, Tyler-Smith C (2016) Punctuated bursts in human male demography inferred from 1244 worldwide Y-chromosome sequences. *Nat Genet* 48:593–599. <https://doi.org/10.1038/ng.3559>
- Repping S, van Daalen SK, Brown LG, Korver CM, Lange J, Marszałek JD, Pyntikova T, van der Veen F, Skaletsky H, Page DC, Rozen S (2006) High mutation rates have driven extensive structural polymorphism among human Y chromosomes. *Nat Genet* 38:463–467. <https://doi.org/10.1038/ng1754>
- Rozen S, Skaletsky H, Marszałek 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. <https://doi.org/10.1038/nature01723>
- Skaletsky H, Kuroda-Kawaguchi T, Minx PJ, Cordum HS, Hillier L, Brown LG, Repping S, Pyntikova T, Ali J, Bieri T, Chinwalla A, Delehaunty A, Delehaunty K, Du H, Fewell G, Fulton L, Fulton R, Graves T, Hou SF, Latrielle P, Leonard S, Mardis E, Maupin R, McPherson J, Miner T, Nash W, Nguyen C, Ozersky P, Pepin K, Rock S, Rohlfing T, Scott K, Schultz B, Strong C, Tin-Wollam A, Yang SP, Waterston RH, Wilson RK, Rozen S, Page DC (2003) The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* 423:825–837. <https://doi.org/10.1038/nature01722>
- Sudmant PH, Rausch T, Gardner EJ, Handsaker RE, Abyzov A, Huddleston J, Zhang Y, Ye K, Jun G, Hsi-Yang Fritz M, Konkel MK, Malhotra A, Stutz AM, Shi X, Paolo Casale F, Chen J, Hormozdiari F, Dayama G, Chen K, Malig M, Chaisson MJ, Walter K, Meiers S, Kashin S, Garrison E, Auton A, Lam HY, Jasmine Mu X, Alkan C, Antaki D, Bae T, Cerveira E, Chines P, Chong Z, Clarke L, Dal E, Ding L, Emery S, Fan X, Gujral M, Kahveci F, Kidd JM, Kong Y, Lameijer EW, McCarthy S, Flicek P, Gibbs RA, Marth G, Mason CE, Menelaou A, Muzny DM, Nelson BJ, Noor A, Parrish NF, Pendleton M, Quitadamo A, Raeder B, Schadt EE, Romanovitch M, Schlattl A, Sebra R, Shabalin AA, Untergrasser A, Walker JA, Wang M, Yu F, Zhang C, Zhang J, Zheng-Bradley X, Zhou W, Zichner T, Sebat J, Batzer MA, McCarroll SA, Genomes Project C, Mills RE, Gerstein MB, Bashir A, Stegle O, Devine SE, Lee C, Eichler EE, Korbel JO (2015) An integrated map of structural variation in 2504 human genomes. *Nature* 526:75–81. <https://doi.org/10.1038/nature15394>
- Sudmant PH, Mallick S, Nelson BJ, Hormozdiari F, Krumm N, Huddleston J, Coe BP, Baker C, Nordenfelt S, Bamshad M, Jorde LB, Posukh OL, Sahakyan H, Watkins WS, Yepiskoposyan L, Abdullah MS, Bravi CM, Capelli C, Hervig T, Wee JT, Tyler-Smith C, van Driem G, Romero IG, Jha AR, Karachanak-Yankova S, Toncheva D, Comas D, Henn B, Kivisild T, Ruiz-Linares A, Sajantila A, Metspalu E, Parik J, Villesms R, Starikovskaya EB, Ayodo G, Beall CM, Di Rienzo A, Hammer MF, Khusainova R, Khusnutdinova E, Klitz W, Winkler C, Labuda D, Metspalu M, Tishkoff SA, Dryomov S, Sukernik R, Patterson N, Reich D, Eichler EE (2015) Global diversity, population stratification, and

- selection of human copy-number variation. *Science* 349:aab3761. <https://doi.org/10.1126/science.aab3761>
- Szostak JW, Orr-Weaver TL, Rothstein RJ, Stahl FW (1983) The double-strand-break repair model for recombination. *Cell* 33:25–35
- The GTEx Consortium (2017) Genetic effects on gene expression across human tissues. *Nature* 550:201–213. <https://doi.org/10.1038/nature24277>
- Trombetta B, Cruciani F (2017) Y chromosome palindromes and gene conversion. *Hum Genet* 136:605–619. <https://doi.org/10.1007/s00439-017-1777-8>
- Turner DJ, Shendure J, Porreca G, Church G, Green P, Tyler-Smith C, Hurles ME (2006) Assaying chromosomal inversions by single-molecule haplotyping. *Nat Methods* 3:439–445. <https://doi.org/10.1038/nmeth881>
- Turner DJ, Tyler-Smith C, Hurles ME (2008) Long-range, high-throughput haplotype determination via haplotype-fusion PCR and ligation haplotyping. *Nucleic Acids Res* 36:e82. <https://doi.org/10.1093/nar/gkn373>
- Wang J, Fan HC, Behr B, Quake SR (2012) Genome-wide single-cell analysis of recombination activity and de novo mutation rates in human sperm. *Cell* 150:402–412
- Williams G (1998) Mapping studies of the centromeric region of the human Y chromosome. D.Phil. Thesis, Oxford University
- Zheng GX, Lau BT, Schnall-Levin M, Jarosz M, Bell JM, Hindson CM, Kyriazopoulou-Panagiotopoulou S, Masquelier DA, Merrill L, Terry JM, Mudivarti PA, Wyatt PW, Bharadwaj R, Makarewicz AJ, Li Y, Belgrader P, Price AD, Lowe AJ, Marks P, Vurens GM, Hardenbol P, Montesclaros L, Luo M, Greenfield L, Wong A, Birch DE, Short SW, Bjornson KP, Patel P, Hopmans ES, Wood C, Kaur S, Lockwood GK, Stafford D, Delaney JP, Wu I, Ordonez HS, Grimes SM, Greer S, Lee JY, Belhocine K, Giorda KM, Heaton WH, McDermott GP, Bent ZW, Meschi F, Kondov NO, Wilson R, Bernate JA, Gauby S, Kindwall A, Bermejo C, Fehr AN, Chan A, Saxonov S, Ness KD, Hindson BJ, Ji HP (2016) Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat Biotechnol* 34:303–311. <https://doi.org/10.1038/nbt.3432>