Repping et al. Supplementary Methods

Contents

	Page
Assays for potential structural polymorphisms	2
Non-ampliconic structural differences from chimpanzee are not polymorphic	6
Reproducibility of measurements of distal-Yq heterochromatin length	7
Verification of *TSPY* array-size results	8
STSs and probes used to investigate PD339 and YCC038	9
Minimum-mutation histories of structural polymorphisms	10
Total branch length in genealogical tree of 47 Y chromosomes	10
Maximum likelihood analysis of rate of IR3/IR3 inversion events	18
Haplotype of the human Y-chromosome reference sequence	18
Rates of large-scale structural mutations compared to other kinds of mutations	20
Assaying distal-Yq heterochromatin length in WHT3299	21
Assaying the number of gray amplicons in WHT2426	22
Availability of cell lines	23
References	24

Assays for potential structural polymorphisms

Figure SM-1 and Table SM-1 summarize the assays we used to test for potential structural polymorphisms among the 47 chromosomes studied (Supplementary Table 1 and main text Fig. 2).

Figure SM-1 Locations of FISH probes in Table SM-1. The locations of FISH probes are indicated below a diagram of the human Y chromosome annotated with regions of possible structural polymorphism. Main text Figure 4 provides details on additional FISH probes used in *AZFc*.

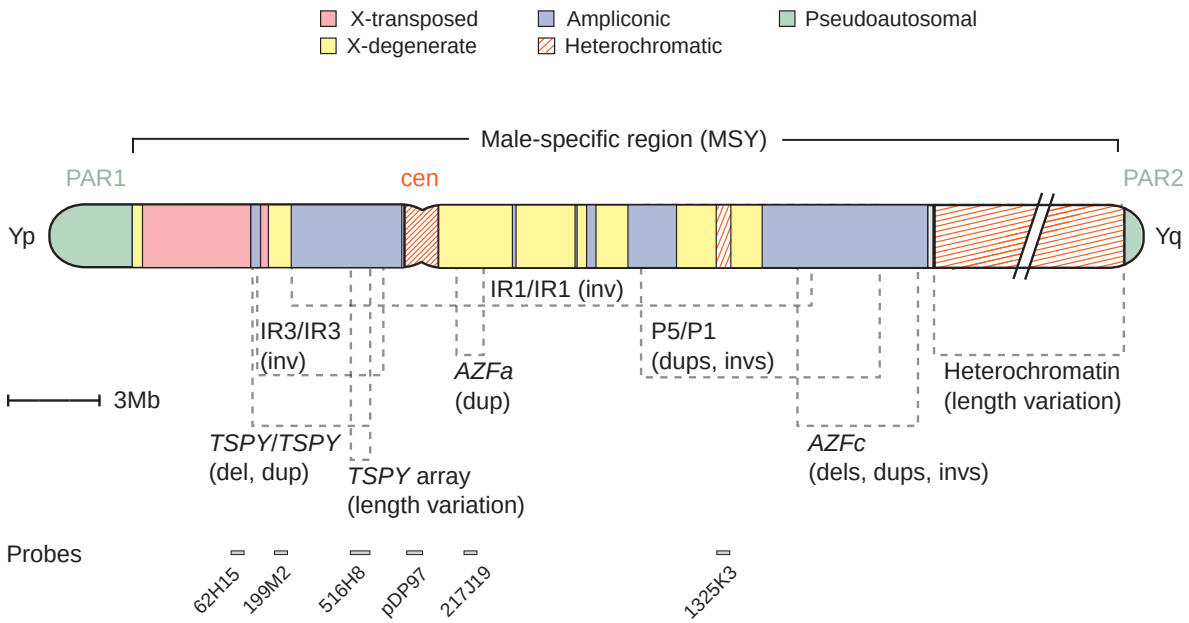


Table SM-1 Assays to detect potential structural polymorphism among human Y chromosomes.

Potential structural polymorphism	Predicted mutational mechanism	Empirical evidence from previous studies	Assays used in present study
IR3/IR3: inversion	Homologous recombination between inverted IR3 repeats[1,2].	Two orientations of this area reported[2-5]. Prevalence, distribution in Y genealogy, and mutational origins largely unexamined.	Interphase FISH, probes as shown (main text Fig. 3c–f).
TSPY array: length	Unequal crossing over between direct repeats in the array.	Length variation reported[5]. Not examined in context of Y genealogy.	*Pme*I pulsed-field gel Southern blots (main text Fig. 3b).
TSPY/*TSPY*: deletion or duplication	Homologous recombination between single *TSPY* gene in IR3 and large *TSPY* array[1,6].	Deletion reported in two unrelated Sri Lankan men[6]. Distribution in Y genealogy unknown. Complementary duplication not reported.	Interphase FISH, signal count of probe 199M2 (Fig. SM-1).
IR1/IR1: pericentromeric inversions	Homologous recombination between IR1 repeat on Yp and complete and partial IR1 copies on Yq.	Cytogenetic reports of pericentromeric inversions[7,8]. Prevalence, distribution in Y genealogy unknown.	Metaphase FISH: signal order of probes 1325K3, pDP97 (Fig. SM-1), and DAPI-stained long-arm heterochromatin.
AZFa: duplication	Ectopic homologous recombination between flanking 10-kb HERV proviruses[9-12].	*AZFa* duplications have been reported. Prevalence unclear[12].	Interphase FISH, signal count of probe 217J19 (Fig. SM-1).
P5/P1: duplications or inversions, and subsequent rearrangements	Homologous recombination between the center of P5 and parts of P1	P5/P1 deletions are rare (~1/12,000 men) and cause spermatogenic failure[13]. Predicted inversions and duplications not reported.	Interphase FISH, signal count of probe 1325K3 (Fig SM-1). Also, signal order and count of probe pairs shown in main text Figure 4.
AZFc: multiple, possibly compound inversions, deletions duplications	Hundreds of organizations could be generated by homologous recombination between *AZFc* amplicons (Supplementary Table 2, Supplementary Fig. 2).	Common polymorphism in this area was first detected in the mid 1980's (ref. [14] and D.C.P.'s unpublished observations). Recently, several possibly common architectural variants were described[15-18].	Interphase FISH with probe pairs as shown in main text Figure 4. Supplementary Figure 2 and Supplementary Table 2 show predicted architectures and assay results.
Distal-Yq heterochromatin: length	Unequal crossing over within tandem arrays.	Length variation reported[19,20]. Not examined in context of Y genealogy.	Metaphase quinacrine staining, measurement relative to total chromosome length (main text Fig. 3a)[19,20].
Other deletion variants	Mechanisms other than homologous recombination between amplicons	Rare Y deletions arose via non-homologous events or recombination between short repeats[17,21].	49 plus/minus STSs (Table SM-2).

Table SM-2 Y-chromosome STSs used in deletion screening

STS	Location	GenBank Accession or Reference
sY14	*SRY*	G38356
sY274	*RPS4Y1*	G38351
sY238	*ZFY* intron 2	G38352
sY1254	*TGIF2LY*	G75612
sY1240	*PCDH11Y*	G75486
sY1256	*TSPY*	G75613
sY276	*AMELY*	G38362
sY1238	*TBL1Y* exon 11	G75611
TranD	*PRKY* 5' end	ref. [22]
sY1319	*PRKY* 3' end	BV210882
sY1250	Proximal boundary of major *TSPY* array	G75495
sY78	Centromere-*DYZ3*	G38359
sY1251	Centromere/Yq boundary	G75496
sY1317	*USP9Y* exon 3	BV210880
sY1316	*USP9Y* exon 26	BV210879
sY1234	*DDX3Y* (*DBY*) exon 9	BV210873
sY1231	*UTY* exon 8	BV210871
sY1230	*TMSB4Y*	BV210870
sY90	*KALY*	G38357
sY1220	*VCY*	BV210869
sY1239	*NLGN4Y* exon 1	BV210876
sY210	*STSP*	G38361
sY1235	*XKRY*	BV210874
sY1260	*CDY2*	BV210877
sY1237	*HSFY* exon 2	BV210875
sY121	Immediately distal to palindrome P4	G38341
sY1322	Between *CYorf15A* and *CYorf15B*	BV210883
SH34Y/SH35Y (sY280)	*SMCY (JARID1D*)	ref. [23]
sY1233	*EIF1AY* exon 1	BV210872
sY627	*RBMY1*-specific	G67175
sY142	Proximal to *AZFc*	G38345
sY1258	Boundary between u1 and b1 in *AZFc*	G75499
sY1161	*PRY*	G66148

Table SM-2 Y-chromosome STS used in deletion screening, continued

STS	Location	GenBank Accession
sY1197	Internal boundary of palindrome P3	G67168
sY1191	In u3 sequence in *AZFc*	G73809
sY1035	*BPY2*	BV210868
sY1318	*DAZ* exon 11	BV210881
sY254	*DAZ* exon 3	G38349
sY1291	*AZFc* red/gray boundary	G72340
sY1125	Blue/gray boundaries in *AZFc*	G67164
sY1054	Blue/yellow boundaries in *AZFc*	G67163
sY1190	Yellow amplicon in *AZFc*	G67165
sY1263	*CDY1*	BV210878
sY1206	Yellow/green boundaries in *AZFc*	G67171
sY1201	Distal boundary of distal gray amplicon in *AZFc*	G67170
sY1246	Proximal portion of distal-Yq heterochromatin	G75492
sY160	Satellite 3 sequence in distal-Yq heterochromatin (*DYZ1*)	G38343
sY1166	Between distal-Yq heterochromatin and pseudoautosomal region 2	G66149
sY1682	*RPS4Y2* exon 1	BV444811

We observed four of the nine classes of potential structural polymorphism that we assayed for (Table SM-1). None of the four observed classes was originally detected as polymorphic because of enrichment in clinical samples.

1. Distal Yq-heterochromatin: The discovery of specific staining of the distal-Yq heterochromatin with quinacrine (refs. [24-26]) led to many studies of length variation in this region. The lack of obvious enrichment for extremes of length among clinically ascertained samples was immediately evident[19,20].
2. *TSPY* array: At the time of the initial description of variation in the size of the *TSPY* array (1988) it was not known to contain genes, and size variation was clearly not ascertained because of enrichment in clinical samples[5]. The *TSPY* gene family was described five years later[27].
3. IR3/IR3 inversion: The IR3/IR3 inversion was not detected because it was enriched in clinical samples, but was discovered accidentally in the course of mapping the breakpoints of naturally occurring deletions within the chromosome[3-5].
4. *AZFc* variation: Large-scale polymorphism was detected in *AZFc* via Southern-blot mapping with 50f2/*DYS7* in the mid 1980's (ref. [14] and D.C.P.'s unpublished observations in this period; note the low DYS number.) The importance of this region for spermatogenic failure was not demonstrated until ten years later [28].

Similarly, most of the other variants that we did *not* detect in this study were also originally ascertained independently of any enrichment in clinical patients. These include pericentric inversions[7,8] and the *TSPY*/*TSPY* deletion[6] and its hypothesized complementary duplication. Furthermore, the deletions tested for by plus/minus STSs were also selected neutrally. Finally, one would not expect *AZFa* duplications[12] and P5/P1 duplications or inversions to be enriched in clinical samples.

Non-ampliconic structural differences from chimpanzee are not polymorphic

We recently reported a comparison of the X-degenerate sequences of the human and chimpanzee Y chromosomes[29]. In the human lineage, there was a large inversion event involving 1.5 Mb (Figure 1 and Supplementary Figure 3 in ref. [29]). We determined that this inversion likely occurred before the last common ancestor of extant human Y chromosomes. To do this, we confirmed the presence of the two inversion breakpoints in

(i) three chromosomes maximally diverged from the reference chromosome (samples 4566, GM03043, GM06342, all in haplogroup A; Supplementary Fig. 1), and

(ii) a chromosome from a haplogroup indistinguishable from the reference sequence (sample GM02294), serving as a positive control.

We confirmed the presence of the breakpoints by sequencing PCR products amplified from these chromosomes (Table SM-3). In addition, in the human lineage, there were four separate deletions of >20 kb of X-degenerate sequence. We confirmed the presence of the corresponding deletion junctions in these four chromosomes, again by sequencing the PCR-amplified breakpoint junctions (Table SM-3).

Table SM-3 STSs used to check for polymorphism of X-degenerate structural differences between the human and chimpanzee reference Y chromosomes.

STS	Location	GenBank Accession
sY1275	Proximal boundary of P6 (tested because chimpanzee has a different proximal boundary)	G75502
sY1719	Proximal inversion breakpoint (in human Y)	BV679241
sY1720	Distal inversion breakpoint (in human Y)	BV679242
sY1721	Junction of 24-kb deletion in human lineage	BV679243
sY1722	Junction of 33-kb deletion in human lineage	BV679244
sY1723	Junction 21-kb deletion in human lineage	BV679245
sY1724	Junction of 165-kb deletion in human lineage	BV679246

Reproducibility of measurements of distal-Yq heterochromatin length

We assessed the reproducibility of quinacrine measurements of distal-Y heterochromatin length by means of replicates, as summarized in Table SM-4.

Table SM-4 Replication experiments for lengths of the distal-Yq heterochromatin. The most divergent replicate measurements were 43.6% versus 39.5% (WHT3257; experiments 1 and 2, separate cell culture and suspensions, both after 10 minutes of colcemid treatment)

	Results per experiment (% of total chromosome length ± standard error over 25 nuclei within experiment)			Standard deviation between experiments
Sample	Experiment 1	Experiment 2	Experiment 3	
WHT2426 [a]	49.2 (±2.12)	48.4 (±2.15)	—	0.6
WHT2630 [a]	41.7 (±2.04)	40.9 (±2.11)	—	0.5
PD123 [b]	54.3 (±2.56)	53.8 (±1.54)	54.8 (±2.22)	0.3
WHT3257 [b]	43.6 (±2.33)	39.5 (±1.66)	40.9 (±1.71)	2.9
4566 [b]	36.4 (±2.09)	38.9 (±1.89)	40.3 (±2.14)	1.8

[a] The two experiments consisted of two slides created from the same suspension.

[b] The three experiments consisted of two separate cultures of the cell line and one additional suspension (and slide) that had been treated with colcemid for 30 minutes as opposed to the standard 10 minutes.

Verification of *TSPY* array-size results

We used two approaches to verify the *TSPY* array sizes that we obtained from *Pme*I Southern blots.

1. We assayed all 47 samples on pulsed field gel Southern blots prepared with the six-cutter *Xba*I, as previously described[5], except that we used the PCR product of STS sY1256 (GenBank/dbSTS G75613) as the probe. In every sample, the *Xba*I-estimated sizes were similar to or smaller than the *Pme*I-estimated sizes. It was expected that some *Xba*I-estimated sizes would be smaller than *Pme*I-estimated sizes because, as previously reported[5], in many Y chromosomes, small *Xba*I fragments (50-100 kb) arose from polymorphic *Xba*I sites within the array. The number of repeat units accounted for by these small fragments could not be assessed because it was not possible to reliably determine the sizes and number of copies they represented. By contrast, *Pme*I is an eight-cutter, and, in fact, no result suggested more than one *Pme*I fragment originating from the array.
2. In all 47 samples, we confirmed the presence of the *Pme*I site proximal to the *TSPY* array by sequencing the products of STSs sY1725 and sY1726 (GenBank/dbSTS BV679247 and BV679248). We did not sequence the *Pme*I site distal to the array, because loss of that site would increase the size of the *Pme*I fragment by only 16 kb.

STSs and probes used to investigate PD339 and YCC038

Tables SM-2 and SM-5 list STSs and hybridization probes used to investigate PD339 (Supplementary Fig. 3) and YCC0038 (Supplementary Fig. 4).

Table SM-5 Additional STSs and probes used in investigating PD339 and YCC038.

STS	Location	GenBank Accession or Primers
sY1196	Internal boundary of palindrome P3 (opposite side from sY1197)	G67167
sY1616	STS used to localize deletion breakpoint in PD339	BV210884
sY1617	STS used to localize deletion breakpoint in PD339	BV210885
sY1618	STS used to localize deletion breakpoint in PD339	BV210886
sY1192	In u3 sequence in *AZFc*	G67166
sY1315	Proximal tip of proximal copy IR2	G75515
sY1302	Proximal inner boundary of IR2	G75513
sY1294	Distal inner boundary of IR2	G75512
sY1259	Distal tip of distal copy of IR2	BV444812
sY132	Southern probe for 50f2/E and 50f2/C	G12023
15467/8	FISH probe for center of palindrome P3 (10680 bp, RP11-477B5 used as template) [a]	TCTGAAAGCCGTTTGGCAACATTTAAGA CAGTGAGGCAGTCAGGATTTGGAGAAAG
15469/70	FISH probe immediately proximal to 50f2/E (10446 bp, RP11-209I11 used as template) [a]	CTTTTCCTGCCATTGCTTTTGGTGTTTT CAAGGGAGCCTTGATCAGCACTTTTCTT
17916/7	FISH probe for NORF sequence in palindrome P2 (9674 bp, RP11-95B23 used as template) [a]	AACCCCATCCAAACCTTACCAGATTGTG TTGGATGTCTTCACGTGTTTGTGGCTTA
17918/9	FISH probe for "end" of green amplicon in *AZFc* (near red) in YCC038 (10119 bp, RP11-290O3 used as template) [a]	AATTCACACTGGTGAGAAACCCCACAAA CTGATTTGGCCCTTGTGTCATGGAATTA
17920/1	FISH probe for "end" of green amplicon in *AZFc* (near red) in YCC038 (10194 bp, RP11-290O3 used as template) [a]	GCCTTTACCCGCTCCTCAACCCTTATTA GGCCTCGGAGCTGAACTCTTTGTTTCTA
17926/7	Southern probe for YCC038 (373 bp, RP11-450B24 used as template) [b]	TGGGGTGTGGATAATACCGT GACTACCCCTTGAGCATCCA
17930/1	Southern probe for YCC038 (509 bp, RP11-65G9 used as template) [b]	TGGGTTATGTTCAGGGAAGG GGCACCAAGGTTGTCAGTTT
18114/5	Southern probe for YCC038 (432 bp, RP11-470K20 used as template) [b]	CATGCCTGTCTGCCACATAC CTGACATGCCCCAACTTTCT
18901/2	Southern probe for YCC038 (539 bp, RP11-178M5 used as template) [b]	TTTTGGGTTGGAGAGAGGTG GCATAGCTGCTTCTTCCCAC

[a] Long range PCR using Advantage 2 Taq polymerase (BD Biosciences) according to the manufacturer's instructions. Each primer at 1 uM final concentration. For a 100 ul reaction, template DNA was either 20 ul of a 1/10 dilution of an overnight BAC inoculant or 50 ng of extracted BAC DNA. Cycling: 95°C for 1 min; 30 X (95°C for 30 s; 68°C for *n* min), where *n* = product size in kb; 68°C for *n* min.

[b] Amplified from 10 ul of 1/10 dilution of overnight BAC inoculant as template in a 40 ul reaction. Each primer at 1 uM final concentration. Taq polymerase at 0.05 units/ul Cycling: 94°C 3 min; 35 X (94°C 1 min, 61°C 1 min, 72°C 1 min), 72°C 5 min.

Minimum-mutation histories of structural polymorphisms

For IR3/IR3, ≥12 inversion events must have occurred to account for the observed orientations (Fig. SM-2). This number was determined by inspection and confirmed by an implementation of Sankoff's algorithm[30] (code available on request). For *AZFc*, ≥20 rearrangement events must have occurred to account for the observed *AZFc* variants (Fig. SM-3). For the distal-Yq heterochromatin and the *TSPY* array, it was necessary to accommodate experimental variance in the measured sizes. For the distal-Yq heterochromatin we allowed Sankoff's algorithm to adjust the postulated length upward or downward of the measured length, L, by 4.1% (i.e. within the range [L – 4.1%, L + 4.1%]) in order to minimize the number of mutations in a postulated history. The rationale for choosing 4.1% was that this was the largest difference observed between replicate experiments (Table SM-4). Figure SM-4 shows an example of a minimum-mutation (most parsimonious) history of large changes in heterochromatin length. For the *TSPY* array, we allowed the algorithm to adjust the measured array size upward or downward by 30 kb in order to minimize the number of mutations in a postulated history. This value (30 kb) was slightly larger than 2x the average standard deviation of the *Pme*I-based size estimates for 14 samples for which we made replicate measurements (excluding WHT3883, whose very large array of ~64 repeat units was outside the optimal size range for measurement). Figure SM-5 shows an example minimum-mutation history of changes in *TSPY* array size.

Total branch length in genealogical tree of 47 Y chromosomes

To estimate the total time spanned by all branches in the tree of 47 chromosomes, we used:

the total number of single nucleotide mutations in the tree, S_{tot},

the average number of single nucleotide mutations on paths from the root to the leaves, i.e. the average single nucleotide mutation height of the tree, S_h,

the time, t_h, to the last common ancestor of extant human Y chromosomes, and

the equation for total time, $t_{tot} = t_h \cdot S_{tot}/S_h$.

Use of previously reported single-nucleotide mutations for this purpose could have led to bias, for example, if some parts of the Y-chromosome genealogical tree were subject to greater intensity of SNP discovery than others. Therefore we resequenced ~80 kb in the 47 chromosomes, thereby identifying 94 SNPs in an unbiased way. Figure SM-6 shows the number of single nucleotide substitutions on each path from the root to a tip in the Y chromosome genealogy. The average, S_h, is 8.617. The total number of nucleotide substitutions in the tree, S_{tot}, is 95, a number that includes a reversion of one of the SNPs. We based the age of the last common ancestor of extant human Y chromosomes, t_h, on

Figure SM-2 At least 12 IR3/IR3 inversion events are needed to explain the observed orientations of the intervening sequence. This figure shows one of several possible minimum-mutation (most-parsimonious) assignments of inversion events to the Y-chromosomal genealogy. This assignment postulates branching orders in the genealogy that are currently unknown but that are consistent with the available data. This genealogy includes the SNP RPS4Y2+771, which is described in this report and which affects the analysis of IR3/IR3 inversions in branch J.

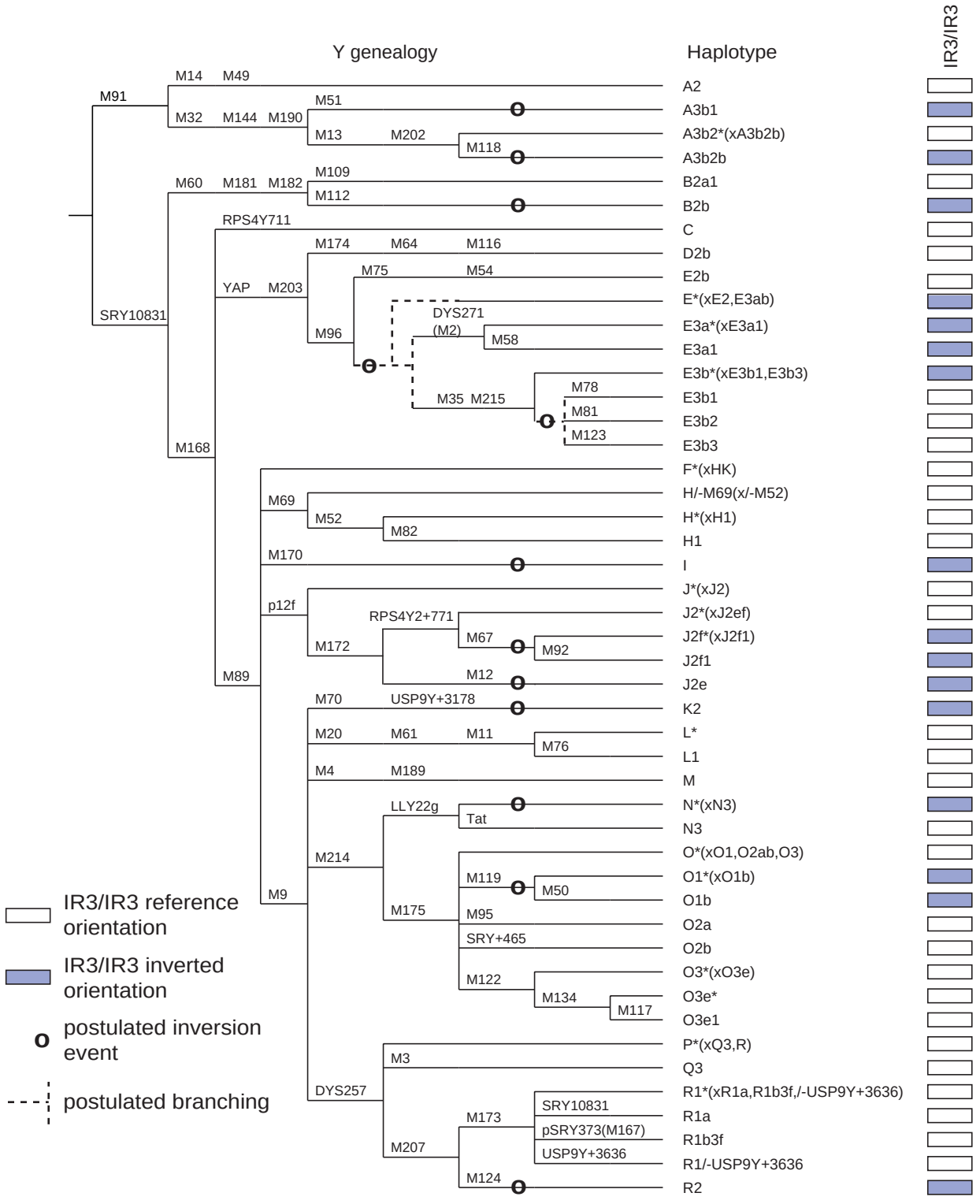


Figure SM-3 Minimum-mutation (most parsimonious) history of *AZFc* architectures among the 47 chromosomes studied. *See legend on next page.*

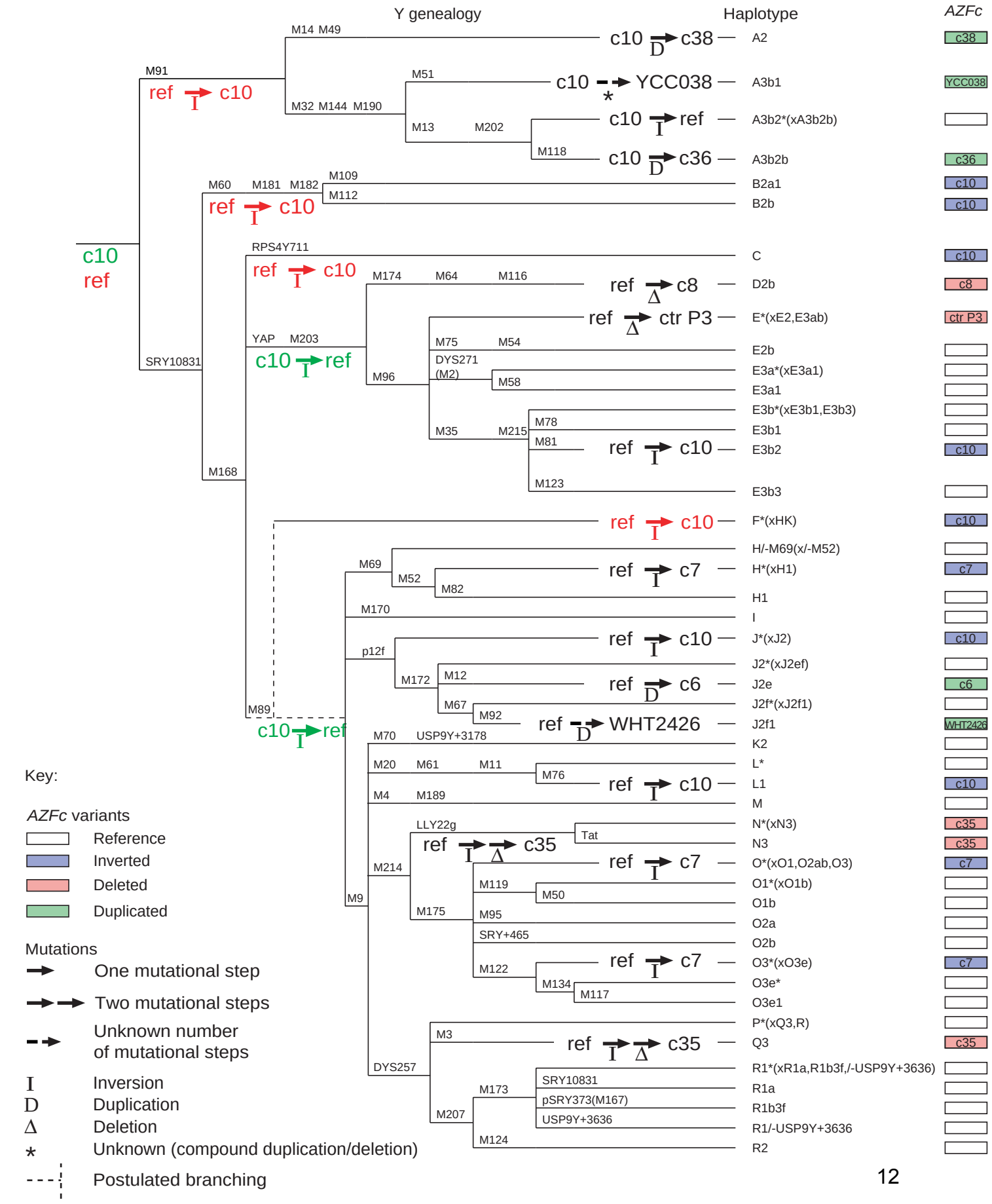


Figure SM-3, Legend Minimum-mutation (most parsimonious) history of *AZFc* architectures among the 47 chromosomes studied.

Shown is a minimum-mutation assignment of *AZFc* recombination events to branches of the Y-chromosome genealogy to account for the observed *AZFc* architectures. Recombination events are indicated by arrows linking different *AZFc* architectures. For example, "ref → c10" indicates a mutation from the reference sequence to architecture c10. Double arrows indicate two recombination events. Arrows with a gap indicate a mutation produced by an unknown recombination event or events. For example, the organization observed in YCC038 might have required several sequential mutations, or might be the product of single event that produced a complex rearrangement.

Mutations indicated in green indicate a model in which the ancestral Y chromosome had architecture c10, and mutations indicated in red indicate a model in which the ancestral Y chromosome had the reference *AZFc* architecture. The former model is more parsimonious. An equally parsimonious model can be obtained by positing a branching order in which the YAP and M89 subtrees were sister clades with branch C (defined by RPS4Y711) as an outgroup. In this model or the one shown, and assuming that single mutations account for the architectures of YCC038 and WHT2426, 20 mutations are needed to account for the observed distribution of *AZFc* architectures.

Symbols below the arrows (I, D, or Δ) indicate an inversion, duplication, or deletion event, respectively. The asterisk at the mutation in YCC038 indicates that the nature of the mutation is unknown, as discussed above. There are a total of 11 inversion events, 4 duplication events, 4 deletion events, and one unknown event in this assignment of *AZFc* mutations to the Y genealogy as shown in the figure. These results are inconsistent with a null hypothesis in which inversion, duplication, and deletion events are equally probable, at $P<.038$ for one-sided binomial test against the null hypothesis that the proportion of inversion events (11/20) is 1/3. (This analysis counts the mutation at YCC038 as a single non-inversion event and the mutation at WHT2426 as a single duplication.) As discussed in the main text, a predominance of inversion events could be caused by: (i) more frequent inversion events than deletion or duplication events, or (ii) natural selection against deletions and duplications but not inversions.

Figure SM-4 A minimum-mutation history of distal-Yq heterochromatin length on bifurcating tree derived from human Y genealogy (main text Fig. 1). Double green bars indicate large changes in heterochromatin length in this postulated history. There are 12 such changes in this history. Blue numbers indicate inferred heterochromatin length as percent of the length of the metaphase Y chromosome. Tips are also labeled by abbreviations of their haplotype designation (see Supplementary Fig. 1)

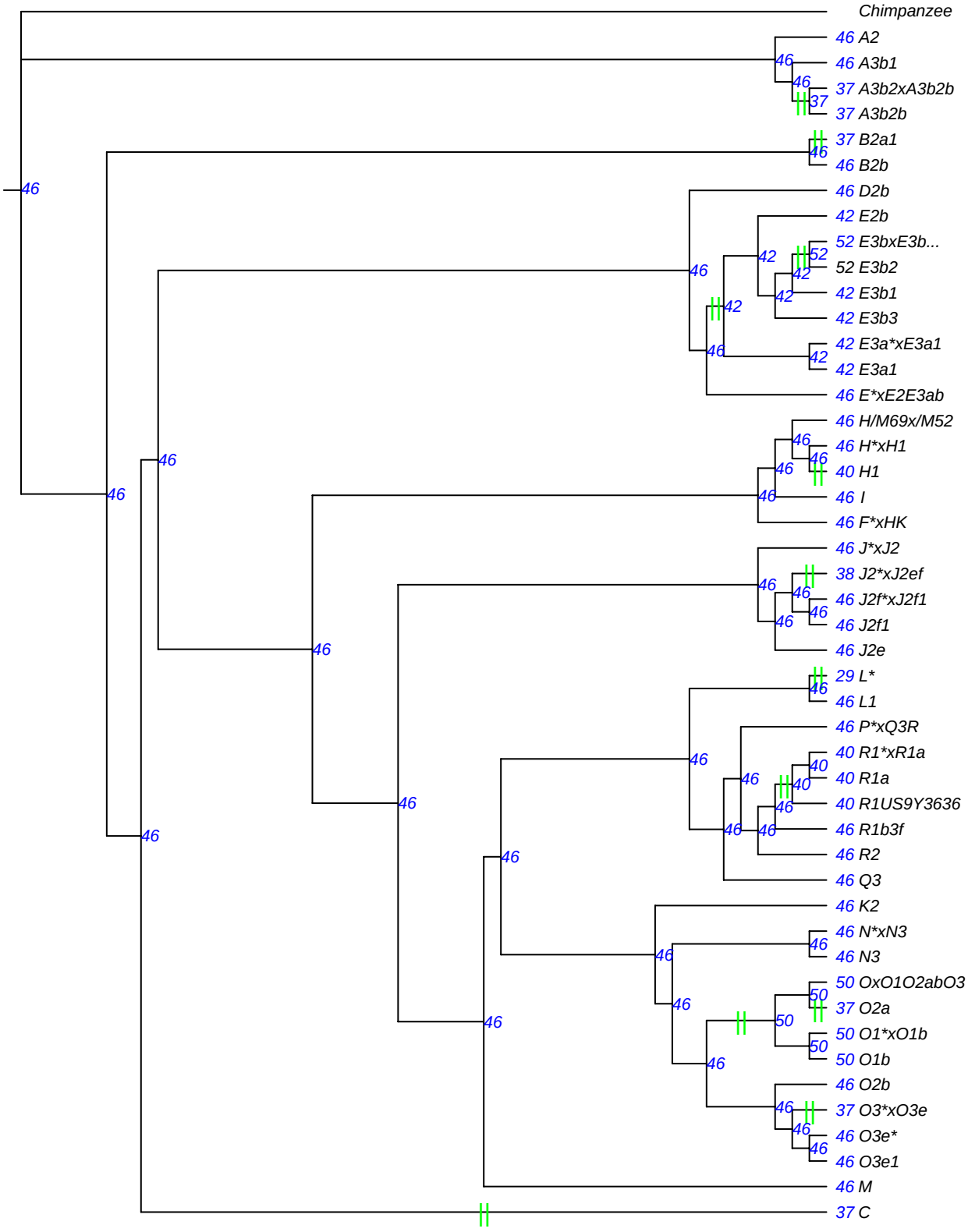


Figure SM-5 A minimum-mutation history of *TSPY* array length on bifurcating tree derived from human Y genealogy (main text Fig. 1). Double green bars indicate changes in array length in this postulated history. There are 23 of these changes in this history. Blue numbers indicate inferred number of repeat units in the array. Tips are also labeled by abbreviations of their haplotype designation (see Supplementary Fig. 1)

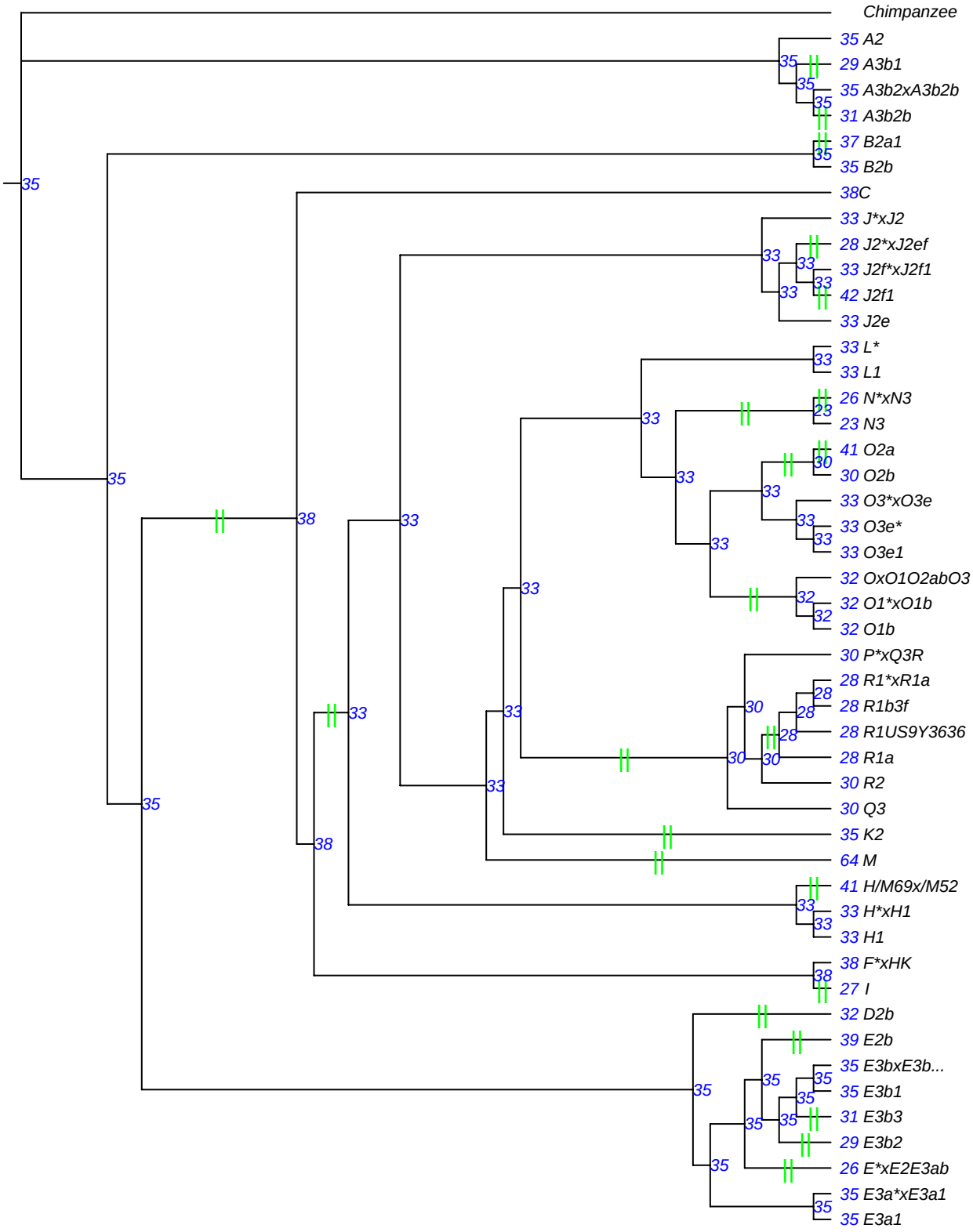
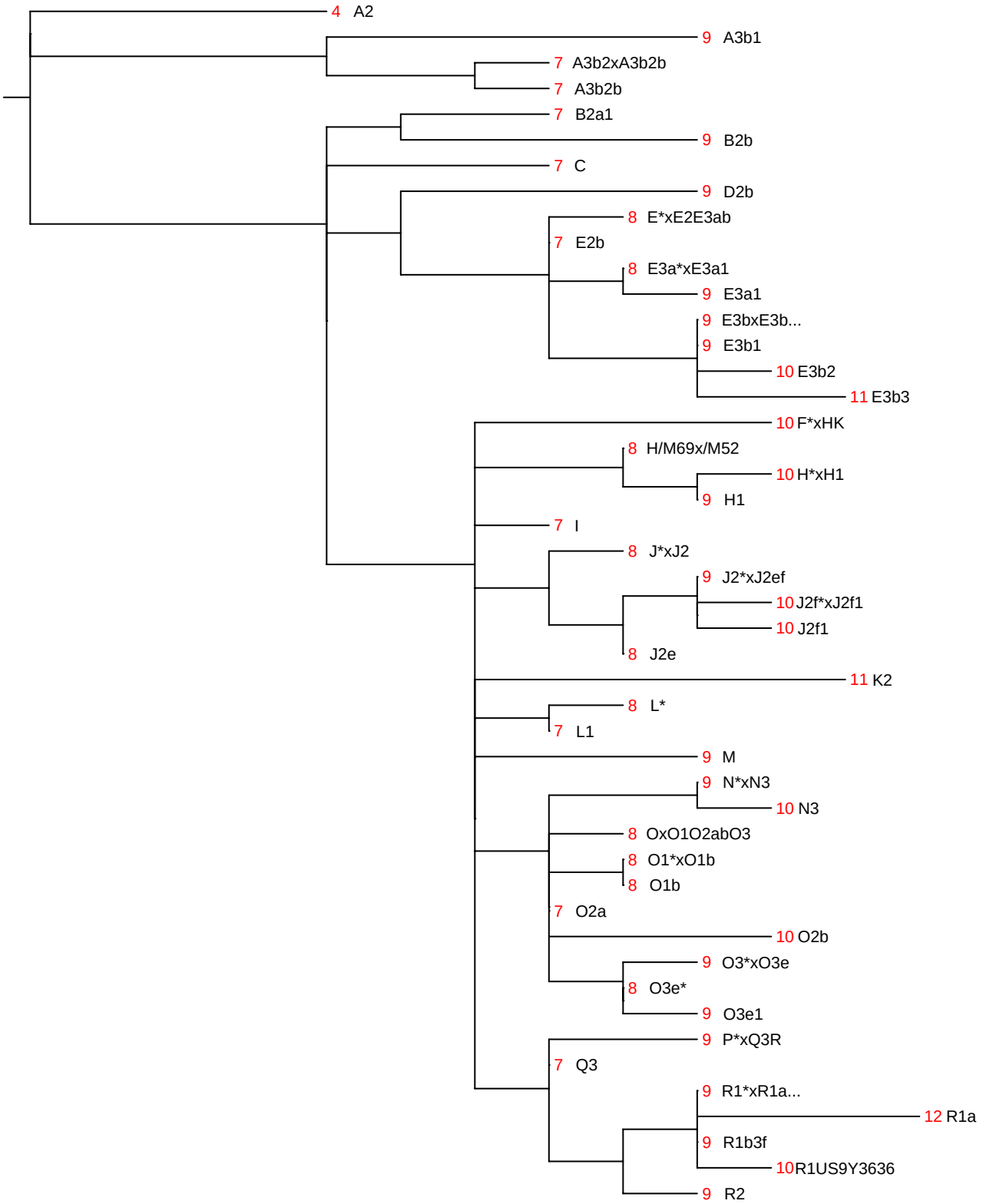


Figure SM-6 Numbers of single nucleotide mutations (substitutions) on paths from root to tips are show in red at tips. These mutations correspond to the 94 SNPs ascertained by resequencing 80 kb in each of the 47 chromosomes. The total number of single nucleotide mutations in tree = 95 (including one reversion). The average number of single nucleotide mutations in paths from root to tip = 8.617. Tips are also labeled by abbreviations of their haplotype designation (see Supplementary Fig. 1).



analysis by Tang and colleagues[31], which, unlike other analyses[32], did not depend on estimating the history of the effective population size of Y chromosomes. However, Tang and colleagues' estimate was based on a human-chimpanzee divergence time of 5 million years ago, and more recently an estimate of 6-7 million years ago has become widely accepted[33,34]. Therefore, we scaled Tang's estimate of 91,000 years (95% CI 60,000 to 130,000 years) by 6.5/5, to yield t_h = 118,000 years (95% CI 78,000 to 169,000 years). Thus,

$$t_{tot} = t_h \cdot S_{tot}/S_h. = \frac{1.18\text{x}10^5 \text{ years} \cdot 95 \text{ substitutions / tree}}{8.617 \text{ substitutions}} = 1.3\text{x}10^6 \text{ years / tree.}$$

To convert t_{tot} to generations we used a male generation time of 25 years. This time is shorter than some results for recent populations would indicate[35,36], but is conservative for obtaining a lower bound on mutation rates per generation, in the sense that it will lead to lower rates of structural mutation per generation than a larger male generation time. With a 25-year generation, a t_{tot} of 1.3x10^6 years corresponds to 52,000 generations. Table SM-6 summarizes the conclusions about the rates of mutations generating structural polymorphism based on the total time represented in the tree.

Table SM-6 Rates of structural mutations per generation. Lower bounds on the numbers of mutations were obtained from minimum-mutation histories (maximum parsimony) as discussed. These lower bounds were divided by the total branch length in the Y genealogical tree to obtain lower bounds on rates. For IR3/IR3 inversions, we also obtained rate estimates by maximum likelihood analyses as discussed below.

Type of variation	Total number of mutations in tree	Mutations per father-to-son Y transmission (x10^{-4})
Distal-Yq heterochromatin length (minimum-mutation-based lower bound)	12	2.3
TSPY array length (minimum-mutation-based lower bound)	23	4.4
IR3/IR3 orientation (minimum-mutation-based lower bound)	12	2.3
IR3/IR3 orientation (likelihood-based lower bound)	19	3.7
IR3/IR3 orientation (maximum likelihood)	48	9.2
AZFc architecture (minimum-mutation-based lower bound)	20	3.8

For assignment of single nucleotide mutations to branches of the tree in Figure SM-6 and to verify consistency of the newly detected SNPs with previously reported Y-chromosome genealogies[37,38], we used the PHYLIP "pars" program (v 3.6.3, Felsenstein, http://evolution.genetics.washington.edu/phylip.html).

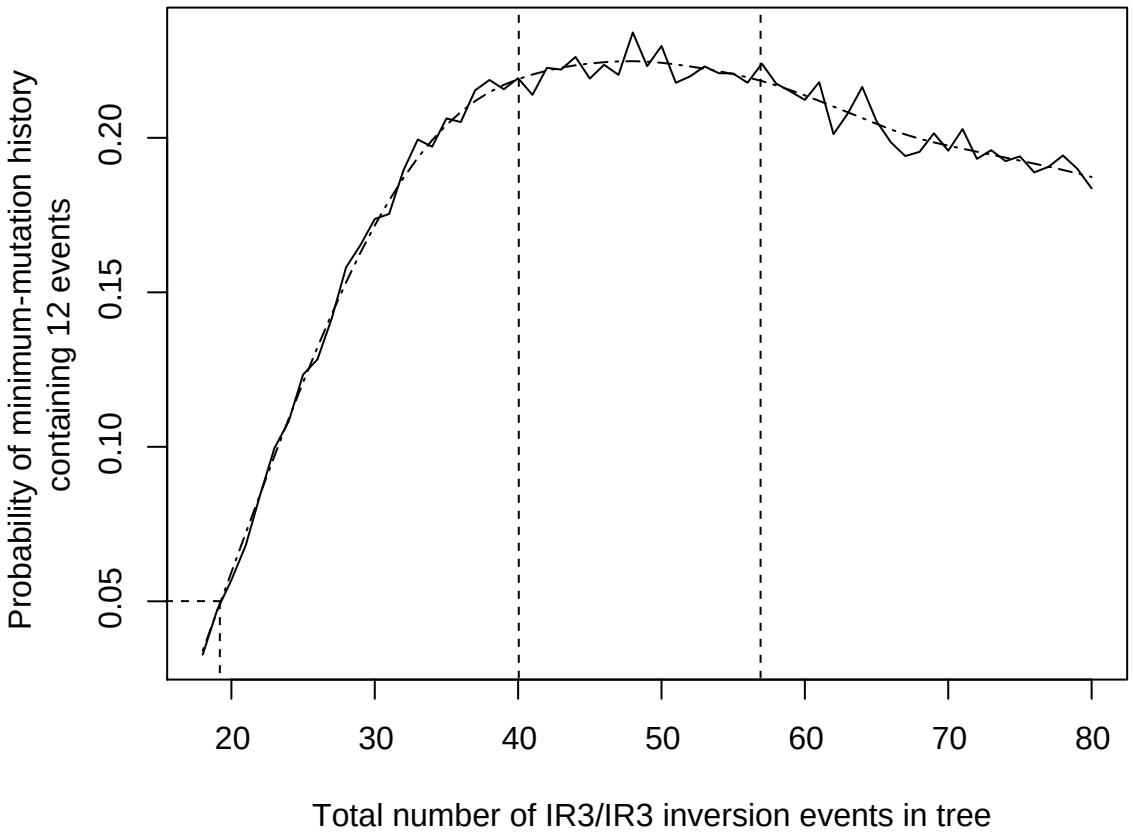
Maximum likelihood analysis of rate of IR3/IR3 inversion events

Our approach was to determine the total number of inversions in the tree that had the maximum likelihood of creating a mutation history for which reconstructed minimum-mutation histories (maximum parsimony) yielded 12 inversions. For n = 18…80, over 11,000 replicates each, we generated n mutations randomly on the branches of the tree, and determined the proportion of the replicates in which reconstructed minimum mutation histories had 12 inversion events. Figure SM-7 shows this likelihood as a function of n. As shown, at n=19 (a total of 19 inversion events in the tree), in 5% of the replicates, minimum-mutation histories had 12 inversions. The likelihood peak is broad, but seems to be located between n=40 and n=57. Likelihood declines only slowly as n increases beyond 57. We hypothesize that this is because the data contain little information to bound the inversion rate from above, and that the distribution of IR3/IR3 orientations is almost consistent with a random distribution (main text Fig 2). To make as much information as possible available to this analyses, we used the 94 SNPs that we ascertained by re-sequencing ~80 kb in each of the 47 samples augmented with 54 additional, publicly available biallelic markers (a total of 148 markers). We used as branch lengths the number of biallelic mutations on each branch. As a check on this maximum likelihood analysis, we also used the "discrete" program (http://www.ams.rdg.ac.uk/zoology/pagel)[39] to estimate the likelihood of the specific distribution of IR3/IR3 orientations at the tips of the branches given different rates of IR3/IR3 inversion events (this rate being the parameter over which to maximize likelihood). The lower bound of the 95% confidence interval was at 18 to 19 inversion events in the tree.

Haplotype of the human Y-chromosome reference sequence

We determined the haplotype of the reference sequence as follows. Almost of all of the reference sequence was generated from BACs from the RPCI-11 library[1,40]. However, M173 and USP9Y+3636 are in *AZFa*, a 800-kb region of the Y chromosome that was sequenced earlier than the rest of the Y chromosome, using BACs from an individual different from the RPCI-11 donor[41]. To genotype the RPCI-11 donor at M173 and USP9Y+3636 (ref. [41], also known as M222; ref. [42]), we genotyped the BACs RP11-460B21 and RP11-576E9 by sequencing PCR products containing these polymorphisms. Furthermore, electronic analysis of Y-chromosome sequence derived from RPCI-11 BACs showed that it has the derived allele at M269, which would imply the derived allele at M173, thereby confirming the experimentally determined M173 genotype. Additional electronic analysis of Y chromosome sequence from the RPCI-11 donor indicates ancestral alleles for M37, M65, M126, M153, and M160 (ref. [37]).

Figure SM-7 Likelihood analysis of rates of IR3/IR3 inversion events. The maximum appears to be in the region [40,57], indicated by vertical dashed lines.



Rates of large-scale structural mutations compared to other kinds of mutations

Rates of single nucleotide substitutions in the human Y chromosome are ~2.3×10^{-8} mutations / nucleotide-generation, based on a human-chimpanzee divergence of 1.23% (ref [29]), and, for consistency with our conservative calculation of the branch length of the genealogical tree, a 25-year male generation time. (A 33-year generation time would yield a substitution rate of 3.1×10^{-8} mutations / nucleotide-generation.) Minisatellite mutation rates are extremely variable, depending on both the particular minisatellite and the allele length. They range from $<5 \times 10^{-5}$ to $>10^{-1}$ (refs. [43-46]). The mutation rates of microsatellites are also variable, ranging from $<4 \times 10^{-4}$ to 7×10^{-3} (refs. [47,48]).

Assaying distal-Yq heterochromatin length in WHT3299

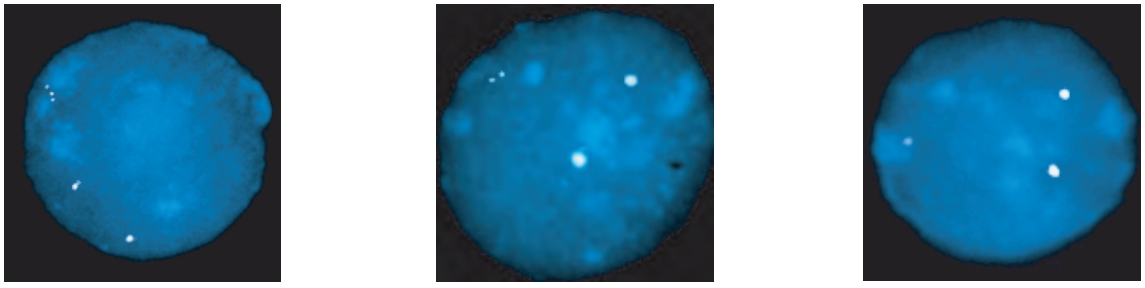
We measured the length of the heterochromatin in WHT3299 by FISH with probe RP11-242E13 (Fig. SM-8). We calibrated this measurement to the quinacrine and FISH measurements of the heterochromatin in YCC038. Assayed with quinacrine, the proportion of heterochromatin in YCC038 was 47.01%. Assayed with RP11-242E13, it was 78.22%. Assayed with RP11-242E13, the proportion of heterochromatin in WHT3299 was 48.64%, which, scaled by 47.01/78.22, yields 29.23%

Figure SM-8 Assaying the length of the distal-Yq heterochromatin in WHT3299.

Assaying the number of gray amplicons in WHT2426

As we did not observe a consistent order of "green" (RP11-363G6) and "yellow" (RP11-79J10) probes for WHT2426, we sought to investigate its *AZFc* organization further. We developed a FISH assay for the gray amplicons in *AZFc*[21] using BAC RP11-366C6 as a probe. This BAC originated from chromosome 1, and was predicted to detect both chromosome 1 and the chromosome-1-homologous gray amplicons in *AZFc* (main text Fig. 4a). We selected a chromosome-1 BAC because Y-chromosome BACs containing the gray amplicon also contained segments of neighboring amplicons, and thus would cross hybridize not only to chromosome 1 but to other parts of *AZFc* in addition. In hybridizations to nuclei with known *AZFc* organizations, we observed (i) consistent spatial clustering of signals emanating from the Y chromosome and (ii) brighter chromosome 1 signals (Fig. SM-9). Thus, in practice it was straightforward to identify the Y-chromosome signals.

Figure SM-9 Assaying the number of gray amplicons in WHT2426.



WHT2426	GM02294 control with reference *AZFc* architecture (2 gray amplicons)	WHT3453 control with an *AZFc* architecture containing 1 gray amplicon (b1/b3-deleted; *AZFc* architecture c2)

Availability of cell lines

Table SM-7 Availability of cell lines representing large-scale structural variants of the human Y chromosome.

Y-chromosome architecture	Sample ID	Source
Reference sequence	GM02294	Coriell, cell line GM02294
*AZF*c architecture c6 or similar duplication	PD388	Coriell, cell line GM15191
b2/b3 inversion; *AZFc* architecture c7	PD073	Coriell, cell line GM15050
gr/gr deletion; *AZFc* architecture c8	PD178	Coriell, cell line GM15093
gr/rg inversion; *AZFc* architecture c10	PD061	Coriell, cell line GM15357
b2/b3 deletion; *AZFc* architecture c35	PD024	Coriell, cell line GM15594
AZFc architecture c36 and IR3/IR3 inversion	GM06342	Coriell, cell line GM06342
AZFc architecture c38	GM03043	Coriell, cell line GM03043
Duplication and deletion in proximal *AZFc* and extending ~1 Mb proximally, and IR3/IR3 inversion	YCC038	The Y Chromosome Consortium, http://ycc.biosci.arizona.edu/
Deletion in central P3; found in other men in haplotype E*(xE2,E3ab), and IR3/IR3 inversion	PD339	Coriell, cell line GM15420
Other AZFc architecture (see main text Figure 4f).	WHT2426	Coriell, cell line GM20118

References

1. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423**, 825-837 (2003).
2. Tilford, C. *et al.* A physical map of the human Y chromosome. *Nature* **409**, 943-945 (2001).
3. Affara, N.A. *et al.* Variable transfer of Y-specific sequences in XX males. *Nucleic Acids Res.* **14**, 5375-5387 (1986).
4. Page, D.C. Sex reversal: deletion mapping the male-determining function of the human Y chromosome. *Cold Spring Harbor Symposium on Quantitative Biology* **1**, 229-235 (1986).
5. Tyler-Smith, C., Taylor, L. & Muller, U. Structure of a hypervariable tandemly repeated DNA sequence on the short arm of the human Y chromosome. *J. Mol. Biol.* **203**, 837-848 (1988).
6. Santos, F.R., Pandya, A. & Tyler-Smith, C. Reliability of DNA-based sex tests. *Nature Genet.* **18**, 103 (1998).
7. Grace, H.J., Ally, F.E. & Paruk, M.A. 46,Xinv(Yp+q-) in four generations of an Indian family. *J. Med. Genet.* **9**, 293-297 (1972).
8. Bernstein, R., Wadee, A., Rosendorff, J., Wessels, A. & Jenkins, T. Inverted Y chromosome polymorphism in the Gujerati Muslim Indian population of South Africa. *Hum. Genet.* **74**, 223-229 (1986).
9. Sun, C. *et al.* Deletion of azoospermia factor a (*AZFa*) region of human Y chromosome caused by recombination between HERV15 proviruses. *Hum. Mol. Genet.* **9**, 2291-2296 (2000).
10. Blanco, P. *et al.* Divergent outcomes of intrachromosomal recombination on the human Y chromosome: male infertility and recurrent polymorphism. *J. Med. Genet.* **37**, 752-758 (2000).
11. Kamp, C., Hirschmann, P., Voss, H., Huellen, K. & Vogt, P.H. Two long homologous retroviral sequence blocks in proximal Yq11 cause *AZFa* microdeletions as a result of intrachromosomal recombination events. *Hum. Mol. Genet.* **9**, 2563-2572 (2000).
12. Bosch, E. & Jobling, M.A. Duplications of the *AZFa* region of the human Y chromosome are mediated by homologous recombination between HERVs and are compatible with male fertility. *Hum. Mol. Genet.* **12**, 341-347 (2003).
13. Repping, S. *et al.* Recombination between palindromes P5 and P1 on the human Y chromosome causes massive deletions and spermatogenic failure. *Am. J. Hum. Genet.* **71**, 906-922 (2002).
14. Disteche, C.M. *et al.* Small deletions of the short arm of the Y chromosome in 46,XY females. *Proc. Natl. Acad. Sci. U. S. A.* **83**, 7841-7844 (1986).
15. Repping, S. *et al.* Polymorphism for a 1.6-Mb deletion of the human Y chromosome persists through balance between recurrent mutation and haploid selection. *Nature Genet.* **35**, 247-251 (2003).

16. Fernandes, S. *et al.* A large *AZFc* deletion removes *DAZ3*/*DAZ4* and nearby genes from men in Y haplogroup N. *Am. J. Hum. Genet.* **74**, 180-187 (2004).
17. Repping, S. *et al.* A family of human Y chromosomes has dispersed throughout northern Eurasia despite a 1.8-Mb deletion in the azoospermia factor c region. *Genomics* **83**, 1046-1052 (2004).
18. Lin, Y.W. *et al.* Polymorphisms associated with the *DAZ* genes on the human Y chromosome. *Genomics* **86**, 431-438 (2005).
19. Bobrow, M., Pearson, P.L., Pike, M.C. & el-Alfi, O.S. Length variation in the quinacrine-binding segment of human Y chromosomes of different sizes. *Cytogenetics* **10**, 190-198 (1971).
20. Schnedl, W. Flurescenzuntersuchungen ueber die Langenvariabilitaet des Y-Chromosoms beim Menschen. *Humangenetik* **12**, 188-194 (1971).
21. Kuroda-Kawaguchi, T. *et al.* The *AZFc* region of the Y chromosome features massive palindromes and uniform recurrent deletions in infertile men. *Nature Genet.* **29**, 279-286 (2001).
22. Schiebel, K. *et al.* Abnormal XY interchange between a novel isolated protein kinase gene, *PRKY*, and its homologue, *PRKX*, accounts for one third of all (Y+)XX males and (Y-)XY females. *Hum. Mol. Genet.* **6**, 1985-1989 (1997).
23. Agulnik, A.I. *et al.* A novel X gene with a widely transcribed Y-linked homologue escapes X-inactivation in mouse and human. *Hum. Mol. Genet.* **3**, 879-884 (1994).
24. Caspersson, T., Zech, L., Johansson, C. & Modest, E.J. Identification of human chromosomes by DNA-binding fluorescent agents. *Chromosoma* **30**, 215-227 (1970).
25. George, K.P. Cytochemical differentiation along human chromosomes. *Nature* **226**, 80-81 (1970).
26. Pearson, P.L., Bobrow, M. & Vosa, C.G. Technique for identifying Y chromosomes in human interphase nuclei. *Nature* **226**, 78-80 (1970).
27. Manz, E., Schnieders, F., Muller Brechlin, A. & Schmidtke, J. *TSPY*-related sequences represent a microheterogeneous gene family organized as constitutive elements in *DYZ5* tandem repeat units on the human Y chromosome. *Genomics* **17**, 726-731 (1993).
28. Reijo, R. *et al.* Diverse spermatogenic defects in humans caused by Y chromosome deletions encompassing a novel RNA-binding protein gene. *Nature Genet.* **10**, 383-393 (1995).
29. Hughes, J.F. *et al.* Conservation of Y-linked genes during human evolution revealed by comparative sequencing in chimpanzee. *Nature* **437**, 100-103 (2005).
30. Sankoff, D. Minimal mutation trees of sequences. *SIAM J App Math* **28**, 35-42 (1975).
31. Tang, H., Siegmund, D.O., Shen, P., Oefner, P.J. & Feldman, M.W. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. *Genetics* **161**, 447-159 (2002).

32. Thomson, R., Pritchard, J., Shen, P., Oefner, P. & Feldman, M. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. *Proc. Natl. Acad. Sci. U. S. A.* **97**, 6927-6929 (2001).
33. Vignaud, P. *et al.* Geology and palaeontology of the Upper Miocene Toros-Menalla hominid locality, Chad. *Nature* **418**, 152-155 (2002).
34. The Chimpanzee Sequencing and Analysis Consortium. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* **437**, 69-87 (2005).
35. Tremblay, M. & Vezina, H. New estimates of intergenerational time intervals for the calculation of age and origins of mutations. *Am. J. Hum. Genet.* **66**, 651-658 (2000).
36. Helgason, A., Hrafnkelsson, B., Gulcher, J., Ward, R. & Stefansson, K. A populationwide coalescent analysis of Icelandic matrilineal and patrilineal genealogies: evidence for a faster evolutionary rate of mtDNA lineages than Y chromosomes. *Am. J. Hum. Genet.* **72**, 1370-1388 (2003).
37. Underhill, P.A. *et al.* Y chromosome sequence variation and the history of human populations. *Nature Genet.* **26**, 358-361 (2000).
38. The Y Chromosome Consortium. A nomenclature system for the tree of human Y-chromosomal binary haplogroups. *Genome Res.* **12**, 339-348 (2002).
39. Pagel, M. Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. Lond. B. Biol. Sci.* **255**, 37-45 (1994).
40. Osoegawa, K. *et al.* A bacterial artificial chromosome library for sequencing the complete human genome. *Genome Res.* **11**, 483-496 (2001).
41. Sun, C. *et al.* An azoospermic man with a *de novo* point mutation in the Y-chromosomal gene *USP9Y*. *Nature Genet.* **23**, 429-432 (1999).
42. Underhill, P.A. *et al.* The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. *Ann. Hum. Genet.* **65**, 43-62 (2001).
43. Jeffreys, A.J., Royle, N.J., Wilson, V. & Wong, Z. Spontaneous mutation rates to new length alleles at tandem-repetitive hypervariable loci in human DNA. *Nature* **332**, 278-281 (1988).
44. Buard, J., Bourdet, A., Yardley, J., Dubrova, Y. & Jeffreys, A.J. Influences of array size and homogeneity on minisatellite mutation. *EMBO J.* **17**, 3495-3502 (1998).
45. Jobling, M.A., Bouzekri, N. & Taylor, P.G. Hypervariable digital DNA codes for human paternal lineages: MVR-PCR at the Y-specific minisatellite, MSY1 (*DYF155S1*). *Hum. Mol. Genet.* **7**, 643-653 (1998).
46. Bois, P. & Jeffreys, A.J. Minisatellite instability and germline mutation. *Cell. Mol. Life Sci.* **55**, 1636-1648 (1999).
47. Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J. & Rolf, B. Mutation rate in human microsatellites: influence of the structure and length of the tandem repeat. *Am. J. Hum. Genet.* **62**, 1408-1415 (1998).

48. Dupuy, B.M., Stenersen, M., Egeland, T. & Olaisen, B. Y-chromosomal microsatellite mutation rates: differences in mutation rate between and within loci. *Hum. Mutat.* **23**, 117-124 (2004).