The American Journal of Human Genetics, Volume *103*

Supplemental Data

Selection Has Countered High Mutability to Preserve the Ancestral Copy Number of Y Chromosome Amplicons in Diverse Human Lineages

Levi S. Teitz, Tatyana Pyntikova, Helen Skaletsky, and David C. Page

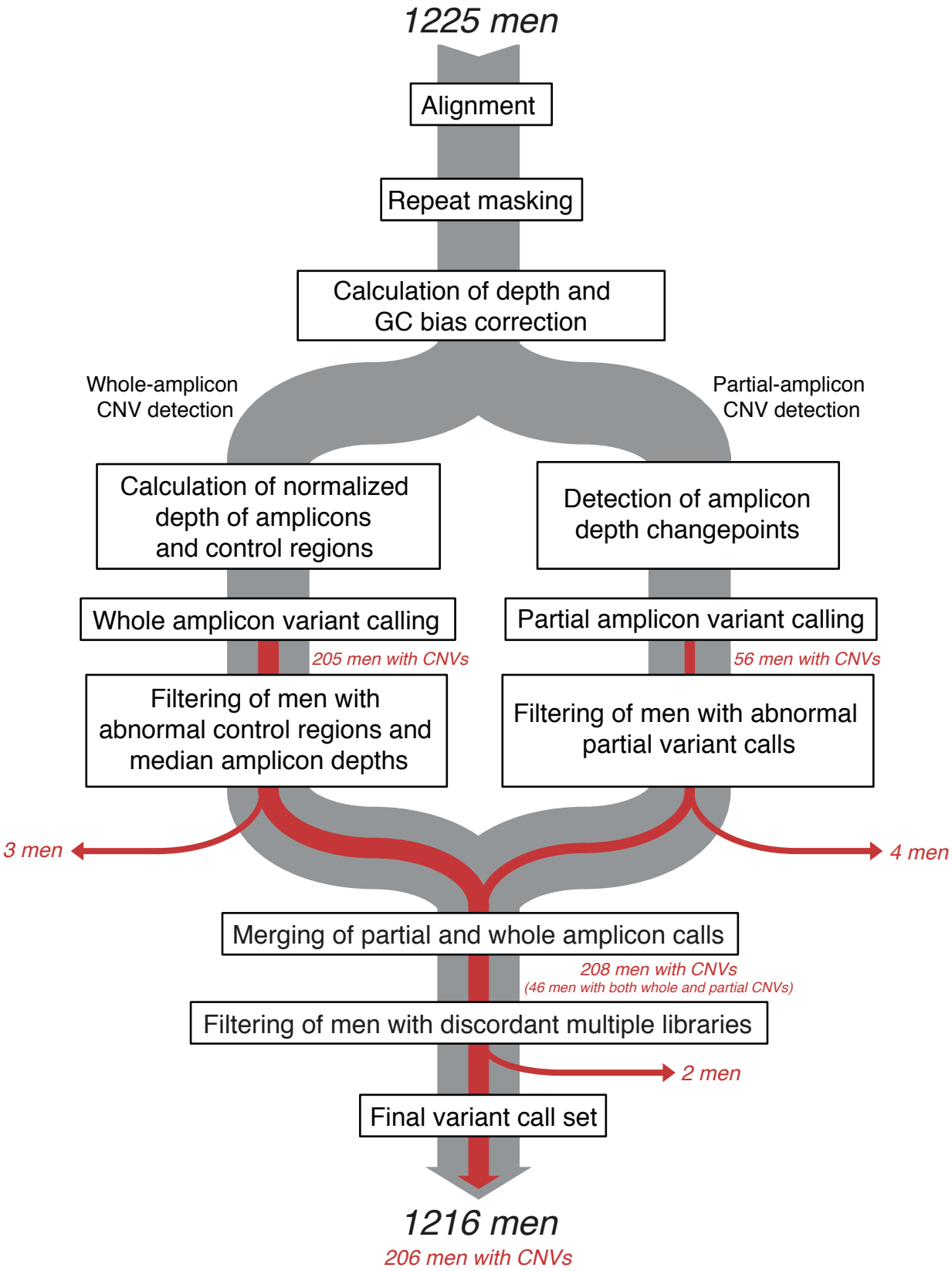

Figure S1. Pipeline for detecting amplicon CNVs from sequencing data.

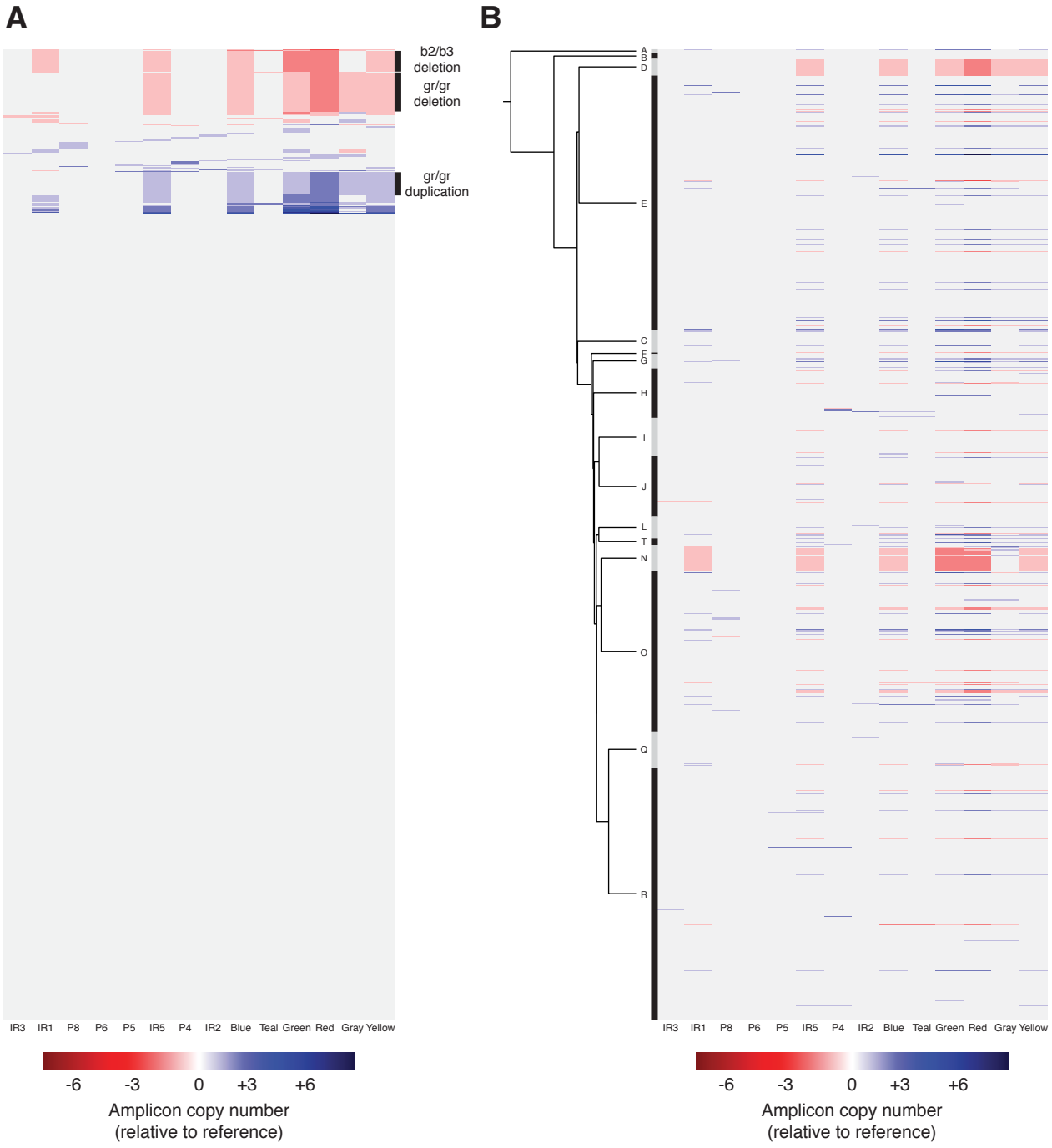

Figure S2. Amplicon copy number in 1000 Genomes Project males. Rows: individual males. Columns: amplicons. (A) Males sorted by variant type. The spans of the three most common CNVs are shown as black bars. (B) Males sorted by phylogenetic relationship. The tree of major haplogroups is drawn to the left. The span of each haplogroup is shown as alternating black and gray bars. Fixed ancestral deletions can be seen in haplogroups D and N.

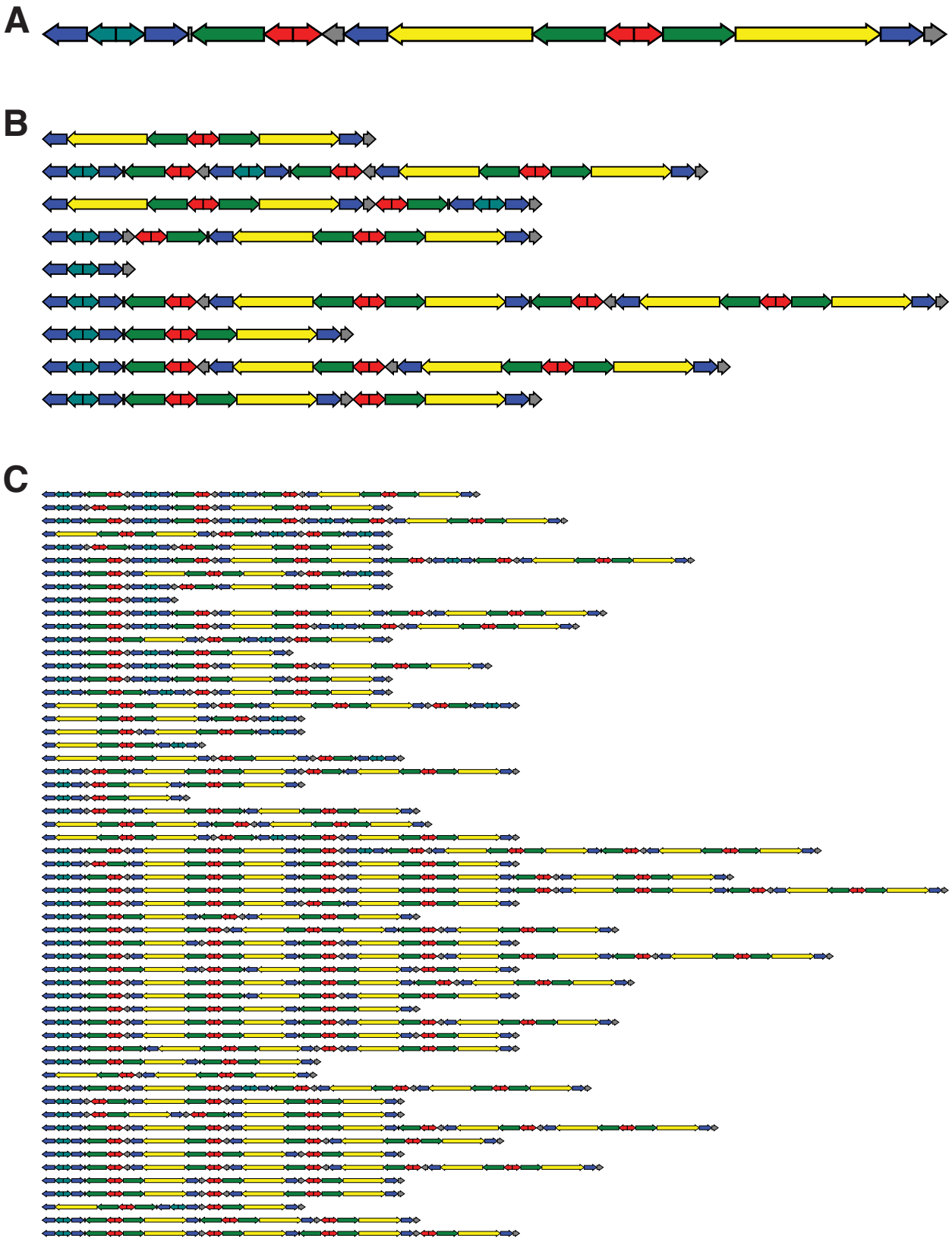

Figure S3 (Page 1).

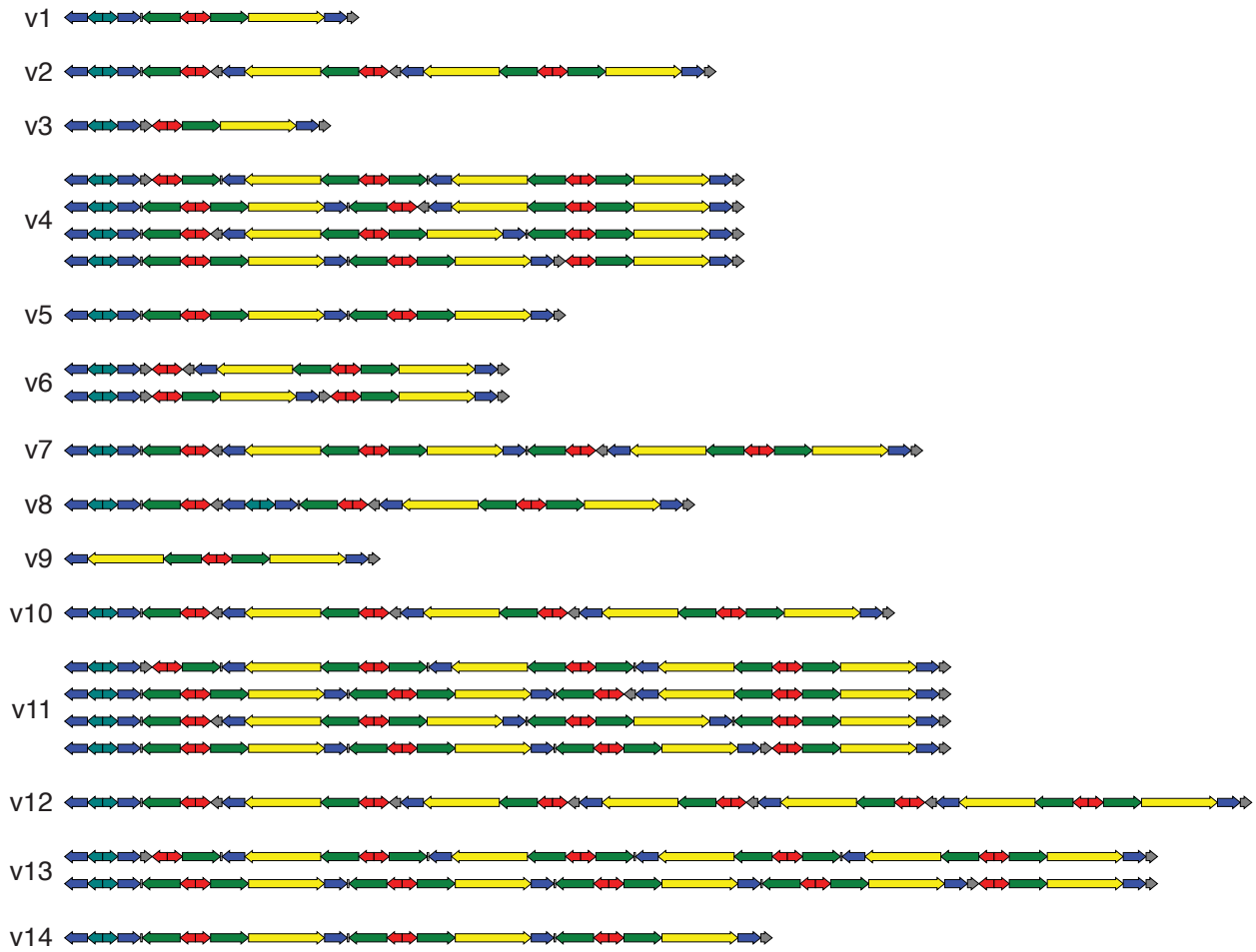

Figure S3 (Page 2). Predicted *AZFc* CNV states arising through NAHR. (A) *AZFc* reference architecture. (B) *AZFc* architectures formed by one NAHR event between amplicon copies. (C) *AZFc* architectures formed by two NAHR events between amplicon copies. The 799 *AZFc* architectures formed by three NAHR events are not shown. (D) *AZFc* architectures corresponding to copy number states found in 1000 Genomes Project males. Some copy number states are concordant with multiple amplicon architectures.

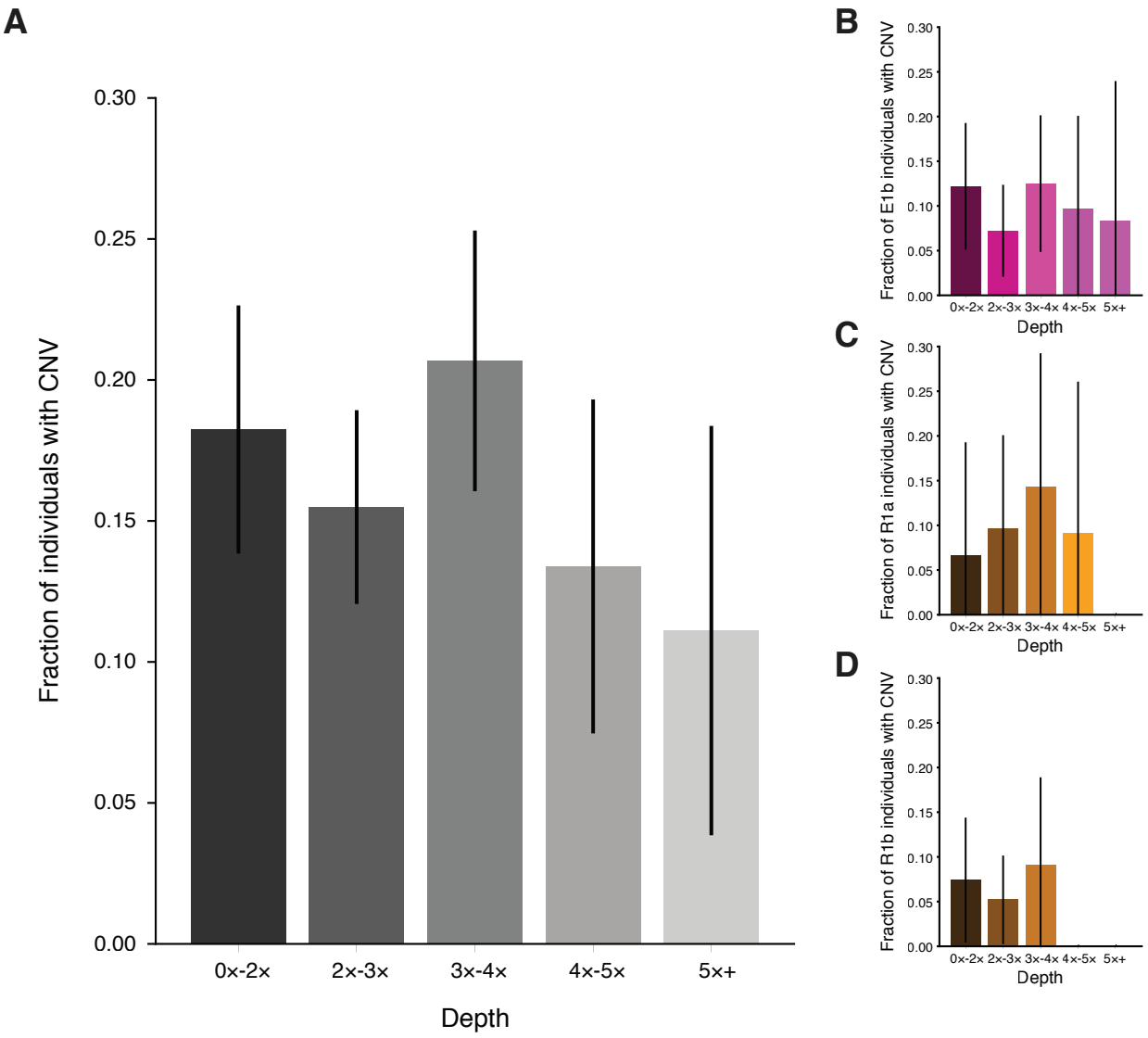

Figure S4. Sequencing depth is not correlated with CNV calls. (A) Fraction of males with CNV calls in different ranges of sequencing depth. Error bars represent binomial 95% confidence intervals. (B-D) Fraction of males with CNV calls in different ranges of sequencing depth in well-represented sub-haplogroups (B) E1b (n=294), (C) R1a (n=81), and (D) R1b (n=206). This controls for the possibility of the whole-dataset results being affected by, for example, a haplogroup with a high fraction of males with CNVs that was sequenced more deeply than other haplogroups. Error bars represent binomial 95% confidence intervals.

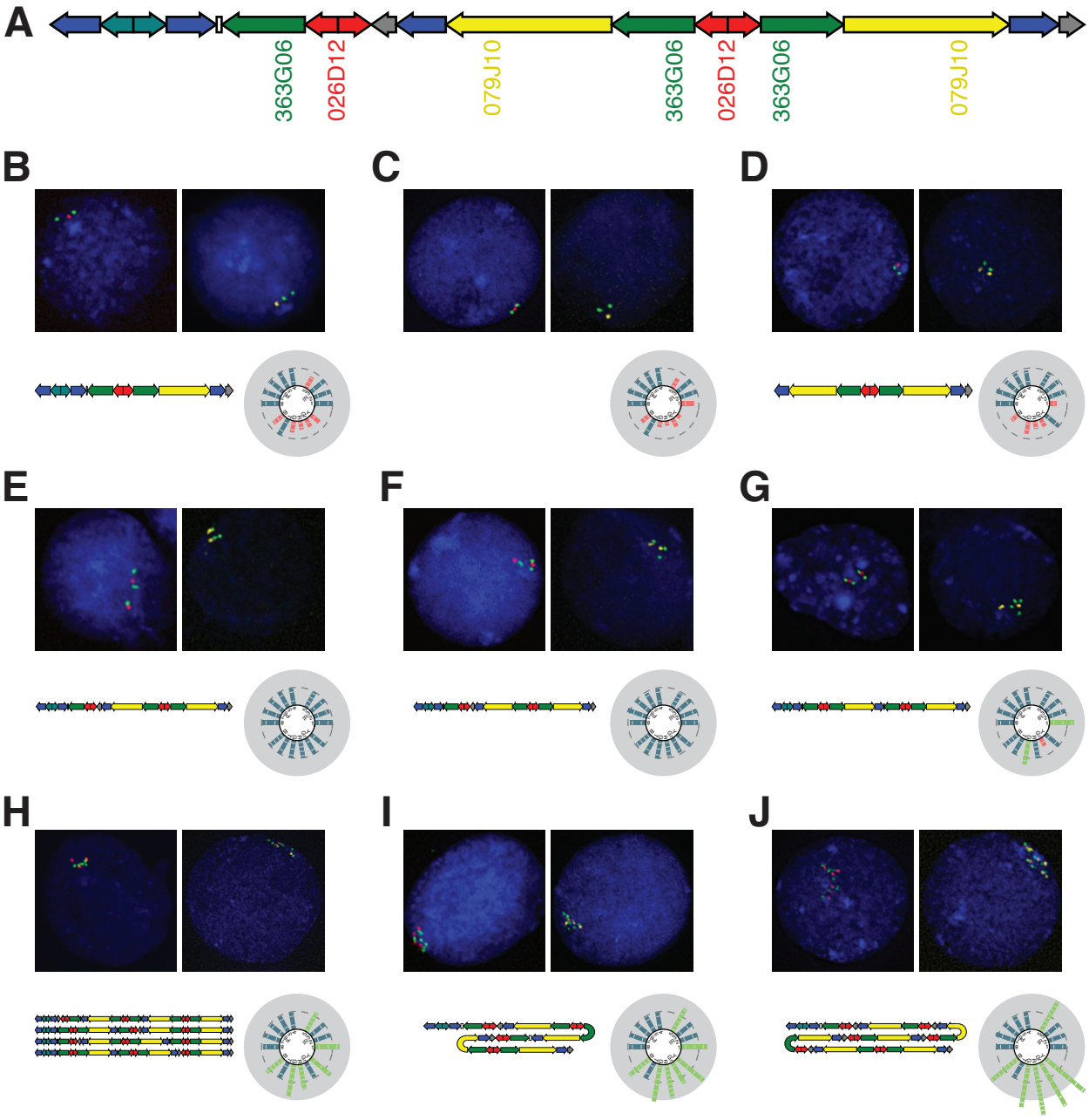

Figure S5. Results of two-color FISH analysis. (A) Hybridization locations of FISH probes. (B-J) Selected FISH images and copy number plots of the remaining 9 males on whom FISH was performed. *AZFc* architectures are shown for males whose computational CNV calls matched a predicted architecture. (B-D) Males with deletions. FISH of the male in (C) detected an error in the computational CNV call. (E,F) Males with the reference copy number. (G-J) Males with duplications. At high copy numbers, FISH underestimates the number of amplicon copies.

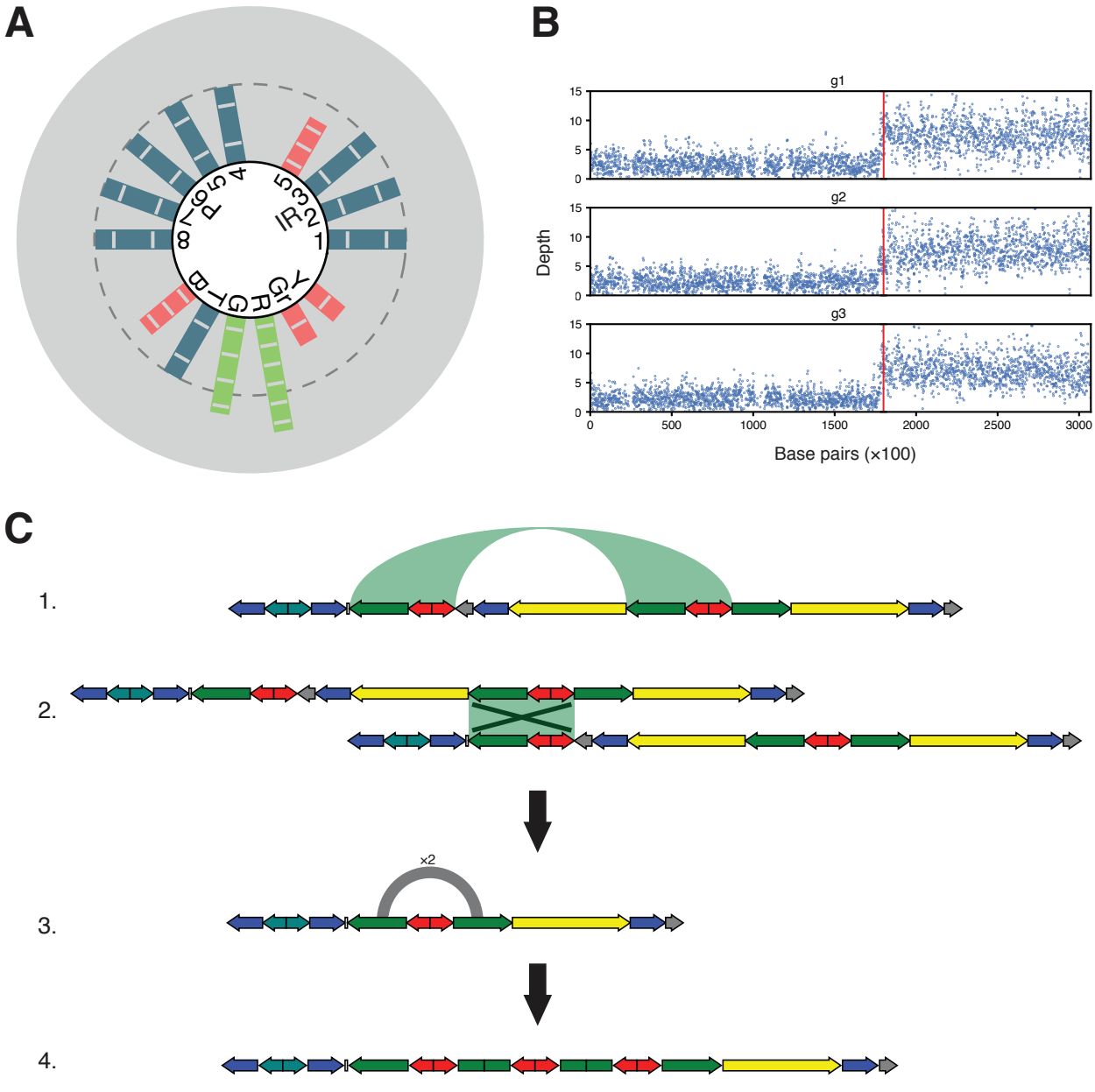

Figure S6. Mechanism of formation of a complex *AZFc* CNV. (A) Copy number calls of affected male. The copy number calls do not match any predicted *AZFc* architecture. (B) Evidence of a partial amplicon mutation event in this male. Blue dots: depth of 100-bp windows. Red lines: predicted change points. (C) Predicted multi-step mechanism of formation for this CNV. 1. Reference *AZFc* architecture. The green arc shows the targets of NAHR on a single copy of the *AZFc* region. 2. Crossing over occurs between two sister chromatids of the

Y chromosome, causing a deletion. An alternative mechanism, in which a single chromatid forms a loop and undergoes NAHR with itself, is not shown. 3. Intermediate deletion stage after NAHR. The gray arc shows breakpoints of the subsequent non-NAHR duplication event. This duplication event occurred twice. 4. Final *AZFc* architecture. Note that the final architecture matches the called copy numbers in (A). The copy number call for the green amplicon results from part of that amplicon being present in two copies and part of it being present in six copies.

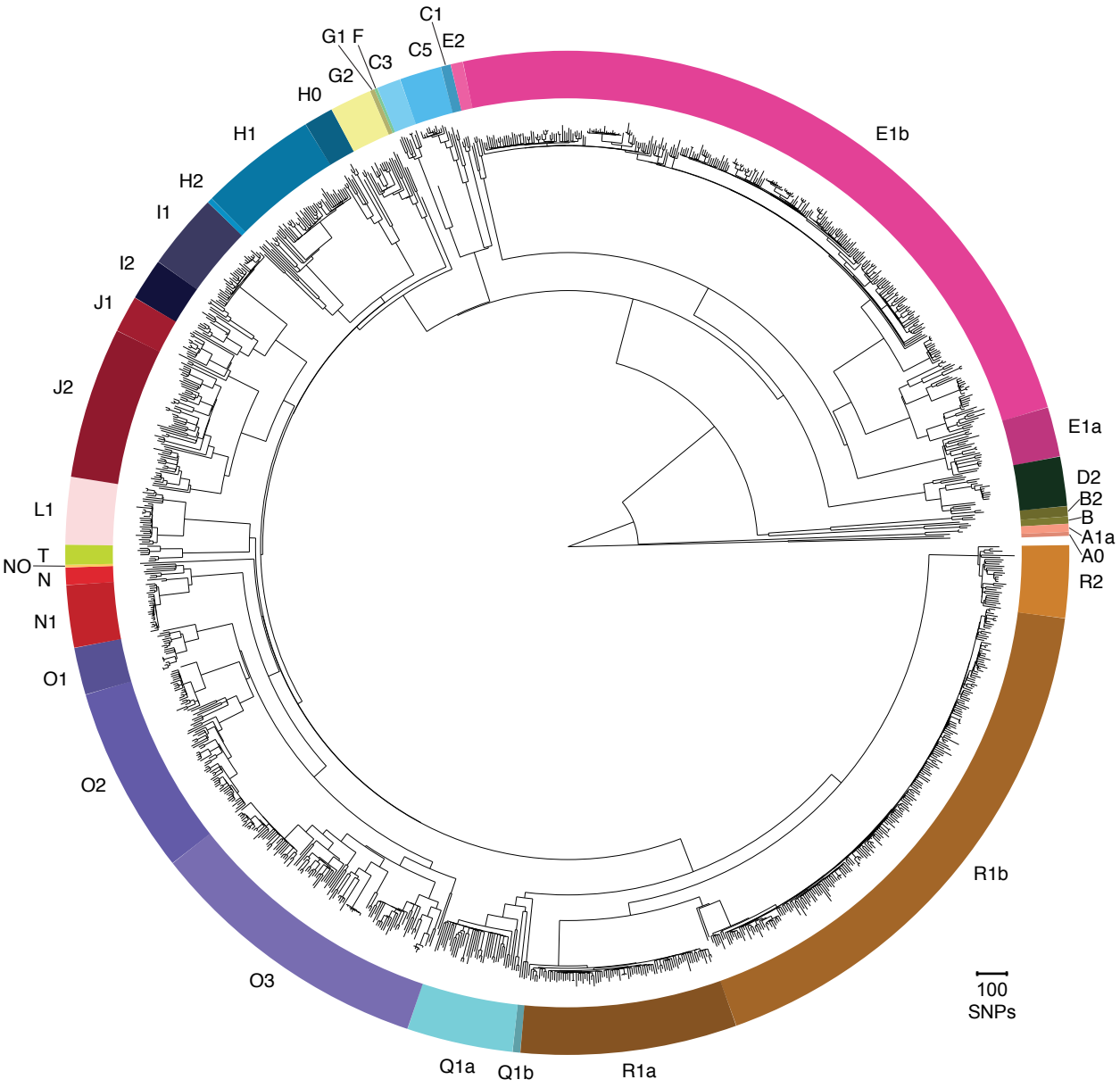

Figure S7. Detailed phylogenetic tree of 1000 Genomes Project Y chromosomes. Haplogroup names are shown around the tree. Branch lengths are measured in SNPs.

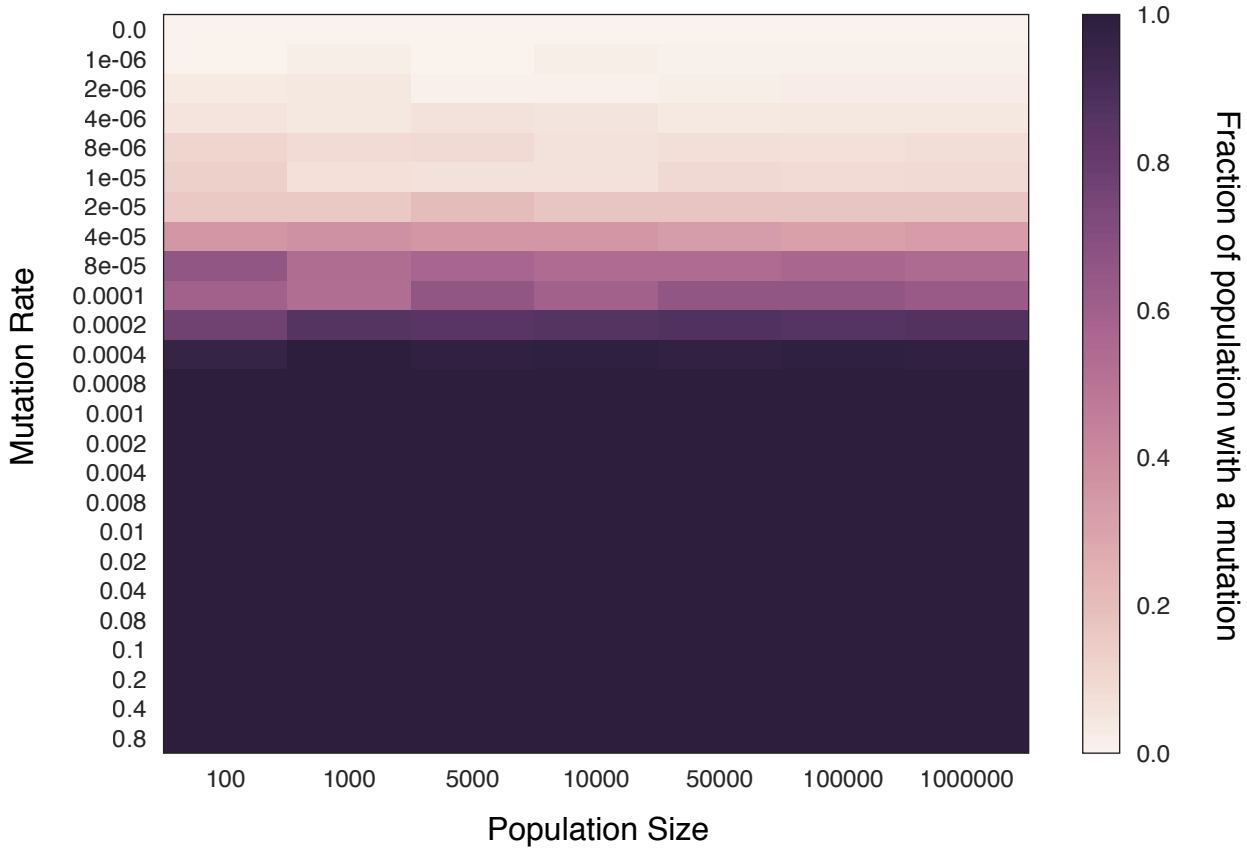

Figure S8. Simulation of A00 mutation and drift. Each cell is the average of 1,000 simulations. The lower bound of the true mutation rate, as calculated from the 1000 Genomes Project data, is 3.83×10^{-4} mutations per father-to-son Y transmission. CNVs are present in the large majority of the population in all simulations at or above the predicted mutation rate.

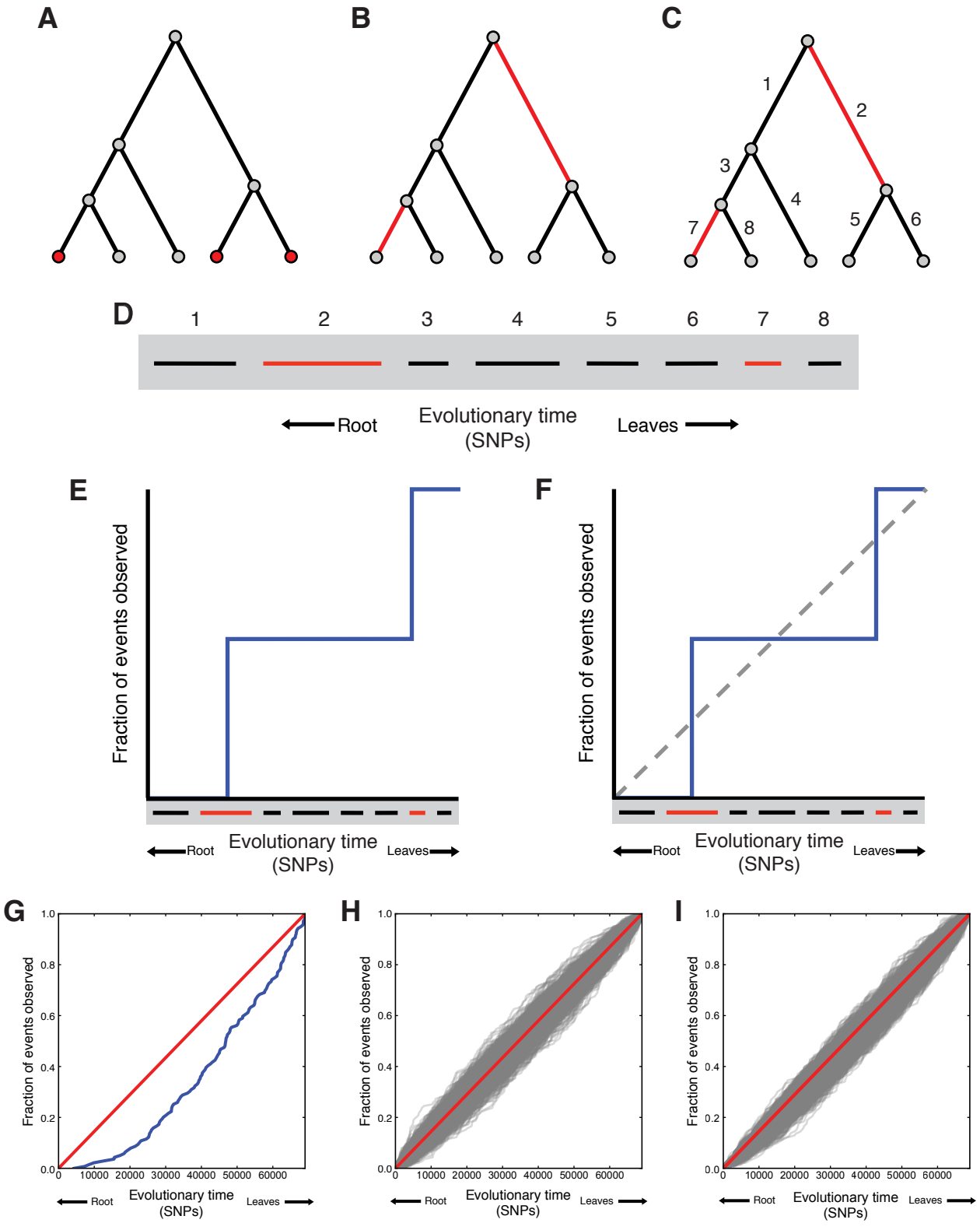

Figure S9. Methodology for calculating CNV distribution over the phylogenetic tree.

(A) Sample phylogenetic tree. Red leaves: individuals with a CNV. (B) Step 1: Find the

edges of the tree in which mutation events occurred by maximum parsimony, shown in red. (C) Step 2: Annotate edges by age. Edge 1 is the oldest branch. Edge 8 is the youngest branch. (D) Step 3: Arrange the edges in a single line and sort edges by age. After sorting, edges closer to the root of the tree will be further to the left, and edges closer to the leaves of the tree will be further to the right. The length of this line is the sum of the edge lengths, which is equal to the total evolutionary time traversed by the tree. Evolutionary time is measured in SNPs, as phylogenetic trees are built using single-nucleotide changes as a molecular clock. (E) Step 4: Plot the cumulative fraction of mutation events observed from the beginning of the line to the end of the line. In this case, there are two edges with mutation events, so 50% of events are observed at branch 2, and 100% of events are observed at branch 7. (F) Step 5: Using the Kolmogorov-Smirnov test, compare the distribution of mutation events to the null distribution (dotted gray line), which represents a constant rate of mutation over time. (G) Distribution of real mutation events over phylogenetic tree. Blue curve: branches of the phylogenetic tree sorted by branch age. Red diagonal line: expected distribution if CNVs were selectively neutral. $p = 1.01 \times 10^{-7}$, KS test. (H) Distribution of shuffled real mutation events over phylogenetic tree. Gray lines: branches of the phylogenetic tree shuffled at random. 1,000 shuffles were performed. Red diagonal line: expected distribution. Minimum p-value of shuffles = 2.00×10^{-3}, KS test. (I) Distribution of simulated mutation events over phylogenetic tree. Gray lines: branches of the phylogenetic tree with simulated mutations sorted by branch age. 1,000 simulations were performed. Red diagonal line: expected distribution. Minimum p-value of simulations = 4.99×10^{-4}, KS test.

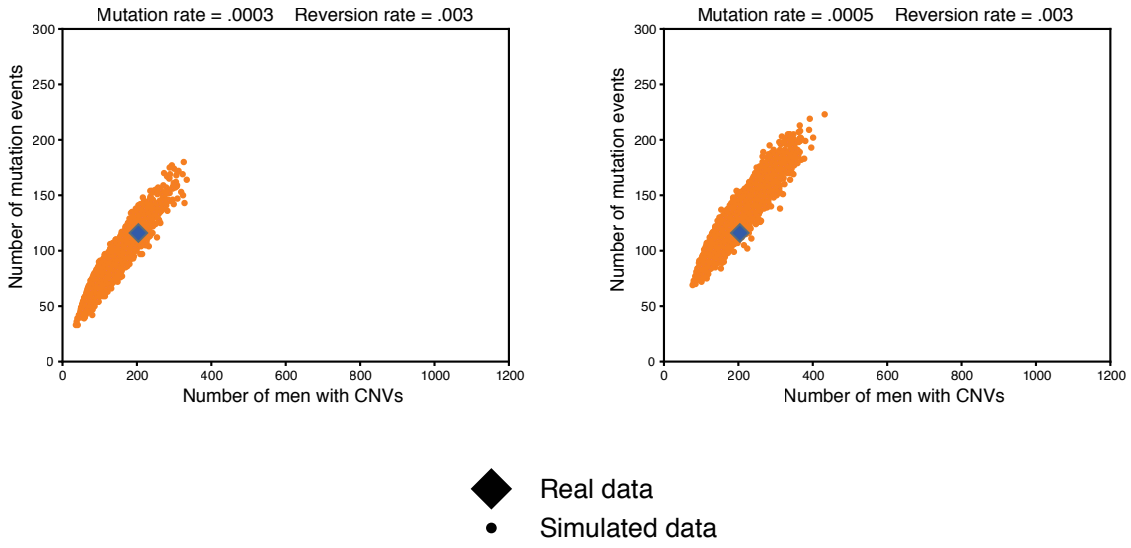

Figure S10. Simulations of neutral evolution with reversion. Mutation events vs. number of males with mutants. Each point represents one simulation over the phylogenetic tree of males in our dataset.

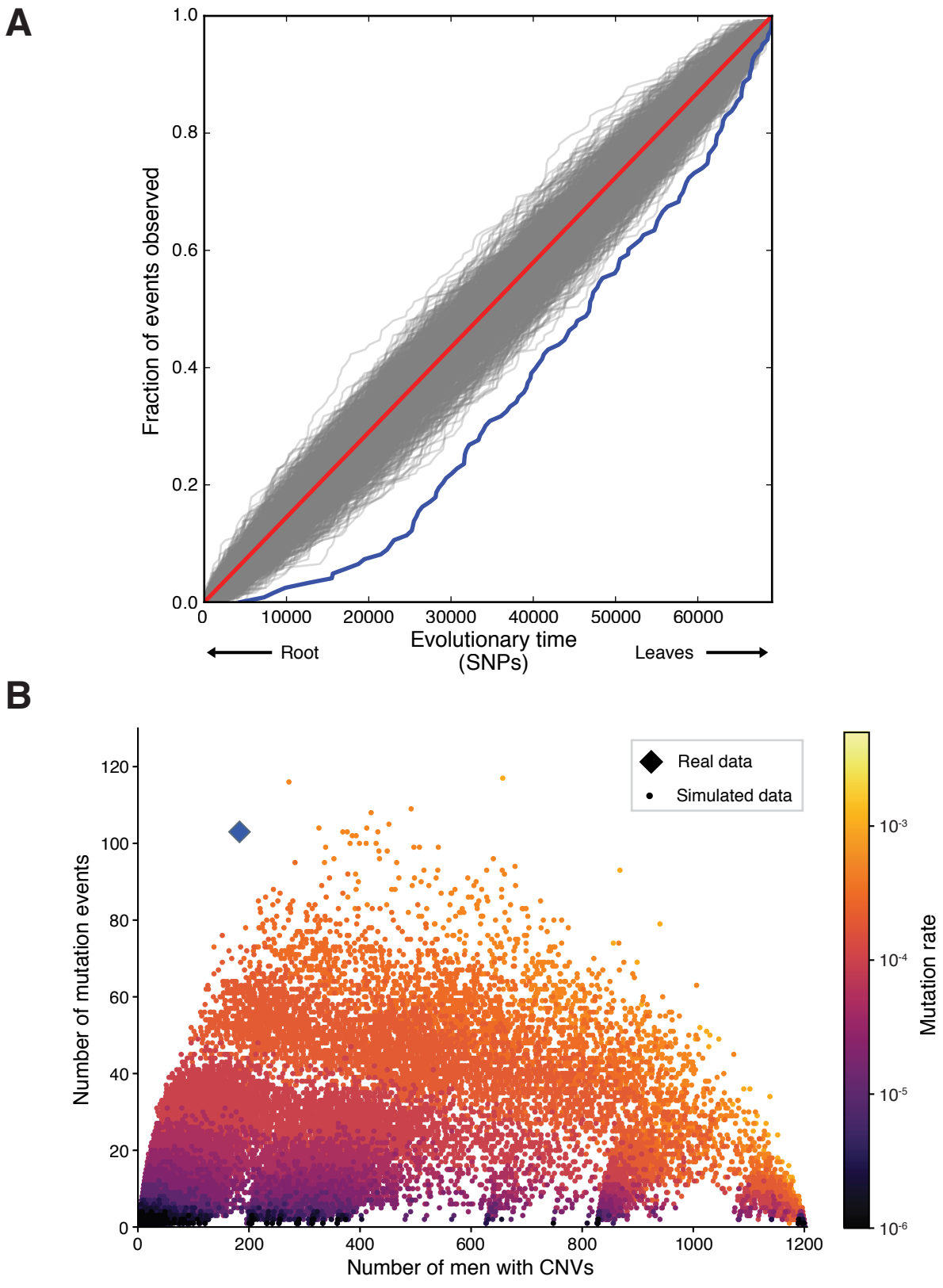

Figure S11. Evolutionary analysis of high-confidence CNV calls. (A) Distribution of CNV

mutation events over the evolutionary tree. Blue curve: branches of the phylogenetic tree of males in our dataset sorted by branch age. Red diagonal line: expected distribution if CNVs were selectively neutral. Gray lines: branches of the phylogenetic tree shuffled at random. 1,000 shuffles were performed. $p = 6.57 \times 10^{-7}$, KS test. (B) Mutation events vs. number of males with CNVs. Each point represents one simulation over the phylogenetic tree of males in our dataset.

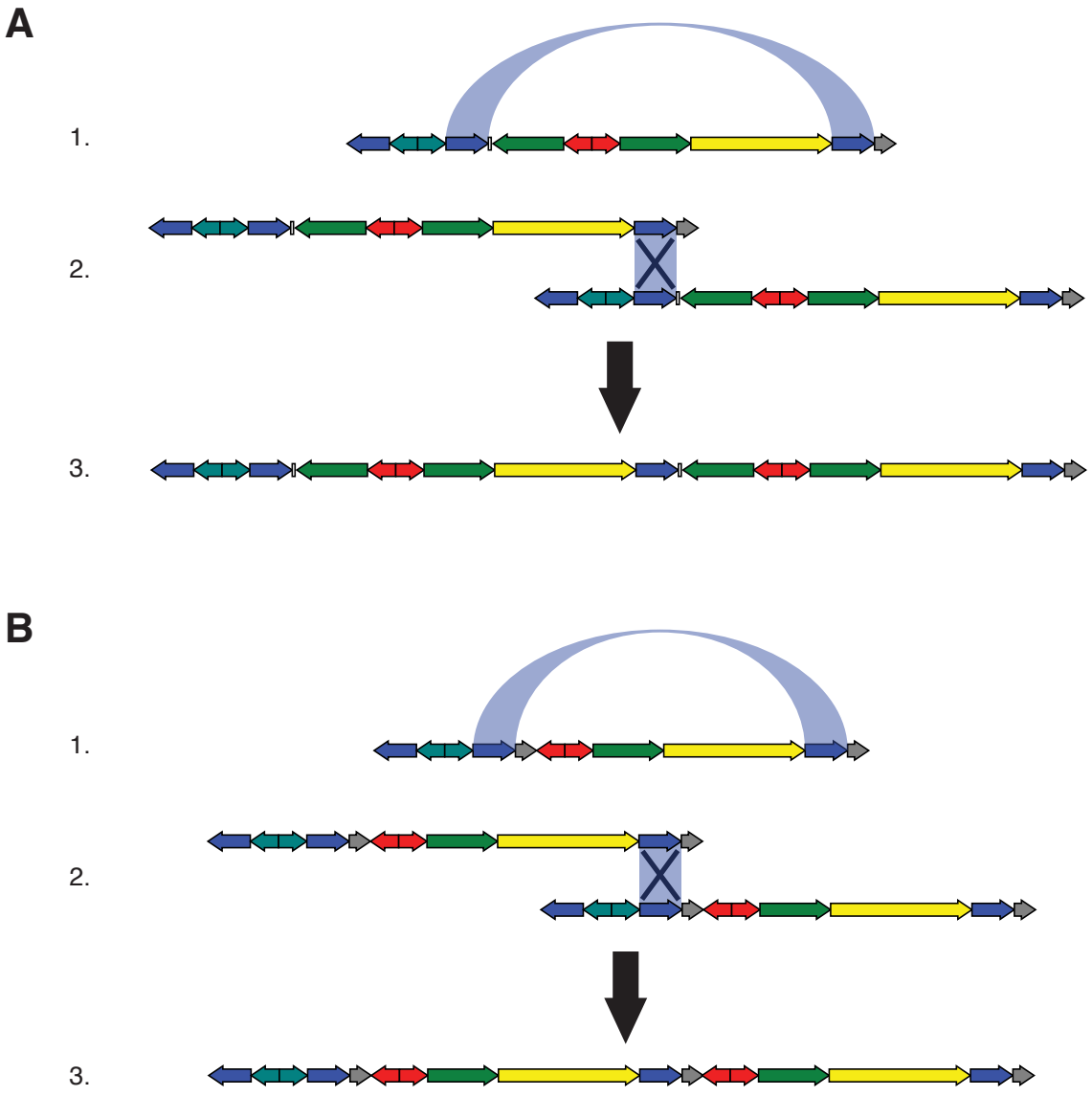

Figure S12. Mechanism of amplicon rescue. (AB) 1. Architecture of the *AZFc* region with a gr/gr deletion (A) and a b2/b3 deletion (B). The blue arc shows the targets of NAHR on a single copy of the *AZFc* region. 2. Crossing over occurs between two sister chromatids of the Y chromosome, causing a duplication. 3. The resulting architecture after NAHR.

CNV[a]	Type	Name	Phenotype	Mechanism	Partial CNVs[b]	NAHR structure[c]	# of males	# of events
+1 IR5, +1 B, +1 G, +2 R, +1 Gr, +1 Y	Duplication	gr/gr duplication		NAHR (1-step)	+Y (1)	v2	28	25
+1 IR1, +1 IR5, +1 B, +2 G, +2 R, +1 Y	Duplication	b2/b3 duplication		NAHR (2-step)		v4	8	7
+1 IR1, +2 B, +2 T, +1 G, +2 R, +1 Gr	Duplication	b1/b3 duplication		NAHR (1-step)		v8	3	3
+1 IR1, +2 IR5, +2 B, +3 G, +4 R, +1 Gr, +2 Y	Duplication	b2/b4 duplication		NAHR (1-step)		v7	3	2
+2 IR1, +2 IR5, +2 B, +4 G, +4 R, +2 Y	Duplication			NAHR (2-step)		v11	1	1
+3 IR1, +3 IR5, +3 B, +6 G, +6 R, +3 Y	Duplication			NAHR (3-step)		v13	1	1
+4 IR5, +4 B, +4 G, +8 R, +4 Gr, +4 Y	Duplication			NAHR (3-step)		v12	1	1
+2 IR5, +2 B, +2 G, +4 R, +2 Gr, +2 Y	Duplication			NAHR (3-step)		v10	1	1
-1 IR5, -1 B, -1 G, -2 R, -1 Gr, -1 Y	Deletion	gr/gr deletion	spermatogenic failure, testis cancer	NAHR (1-step)	+Y (8)	v1	49	25
-1 IR1, -1 IR5, -1 B, -2 G, -2 R, -1 Y	Deletion	b2/b3 deletion	spermatogenic failure[d]	NAHR (2-step)		v3	26	3
-1 IR1, -2 B, -2 T, -1 G, -2 R, -1 Gr	Deletion	b1/b3 deletion	spermatogenic failure	NAHR (1-step)		v9	1	1
+1 IR1, +1 G, -1 Gr	Complex	gr/gr rescue		NAHR (2-step)		v5	5	5
-1 IR1, -1 G, +1 Gr	Complex	b2/b3 rescue		NAHR (3-step)		v6	4	3
+2 IR1, +1 IR5, +1 B, +3 G, +2 R, -1 Gr, +1 Y	Complex			NAHR (3-step)		v14	1	1
+1 IR1, +2 IR5, +2 B, +4 G, +5 R, +1 Gr, +2 Y	Duplication			Both	+G, +R		1	1
+2 IR1, +2 IR5, +2 B, +4 G, +3 R, +2 Y	Duplication			Both			1	1
+1 IR1, +1 IR5, +1 B, +3 G, +2 R, +1 Y	Duplication			Both			1	1
-1 IR5, -1 G, -2 R, -1 Gr, -1 Y	Deletion			Both	+B		1	1
-1 IR5, -1 B, -1 G, -1 R, -1 Y	Deletion			Both	+R		1	1
-1 IR1, -1 IR5, -1 B, -2 G, -1 R, +1 Gr, -1 Y	Complex			Both	+B (2)		3	1
-1 IR1, +1 IR5, +1 B, +1 R, +2 Gr, +1 Y	Complex			Both			1	1
-1 IR5, -1 B, +1 G, +2 R, -1 Gr, -1 Y	Complex			Both	+G		1	1
+1 G	Duplication			Non-NAHR	+G (2)		6	4
+1 R, +1 Gr	Duplication			Non-NAHR	+B (1)		3	2
+1 IR5	Duplication			Non-NAHR	+Y (1)		2	2
+1 G, +1 R	Duplication			Non-NAHR	+G (2)		2	1
+1 Y	Duplication			Non-NAHR	+Y (2)		2	2
+1 B	Duplication			Non-NAHR	+B (1)		2	1
+2 G, +2 R	Duplication			Non-NAHR	+Y (2)		2	2

+1 B, +1 T	Duplication	Non-NAHR	+T	1	1
+1 B, +1 Gr	Duplication	Non-NAHR	+Y	1	1
+1 IR1	Duplication	Non-NAHR	+G	1	1
+1 IR5, +1 B, +2 G, +3 R, +1 Gr, +1 Y	Duplication	Non-NAHR	+G, +R	1	1
+1 IR5, +1 B, +1 Y	Duplication	Non-NAHR	+Y	1	1
+1 IR5, +1 B, +2 G, +2 R, +1 Gr, +1 Y	Duplication	Non-NAHR		1	1
+1 R	Duplication	Non-NAHR	+G, +R	1	1
-1 IR1, -1 B, -1 T, -1 G, -2 R, -1 Gr	Deletion	Non-NAHR	-T	1	1
-1 B, -1 T	Deletion	Non-NAHR	-T	1	1
-1 IR1, -1 IR5, -1 B, -2 G, -3 R, -1 Gr, -1 Y	Deletion	Non-NAHR		1	1

[a]B = blue, T = teal, G = green, R = red, Gr = gray, Y = yellow
[b]Number of males with evidence of each partial CNV shown in parentheses
[c]Structures shown in Figure S3D
[d]Phenotypic impact of the b2/b3 deletion is unclear; see Lu et al. (2009) and Rozen et al. (2012)

Table S5. CNVs located inside the *AZFc* region. Sorted by mechanism, type, and # of males. CNV classes (listed in column "CNV") count only amplicon copy number changes detected by the whole-amplicon CNV pipeline. Amplicons listed in column "Partial CNVs" show evidence of a CNV breakpoint in the middle of that amplicon, and therefore may or may not be listed in the "CNV" column. (Because we expect the partial-amplicon CNV pipeline to have a high false negative rate, splitting CNV classes based on a subset of members having evidence of partial CNVs would artificially inflate the number of CNV classes and mutation events.) The exceptions are six males with evidence of a partial CNV that had no CNV in any amplicon detected by the whole-amplicon CNV pipeline; the "CNV" column for these males corresponds to the detected partial CNVs.

CNV[a]	Type	Name	Location	Mechanism	Partial CNVs[b]	# of males	# of events
+2 IR2, +1 B, +1 T	Duplication		Both	Non-NAHR		1	1
+1 IR2, +1 G	Duplication		Both	Non-NAHR		1	1
+1 IR5, +1 P4, +1 Y	Duplication		Both	Non-NAHR	+P4 +Y	1	1
-1 IR3, -1 IR1, -1 R	Deletion		Both	Non-NAHR	-IR3 -G -R	1	1
+1 P8	Duplication		Non-*AZFc*			8	5
+2 P4	Duplication		Non-*AZFc*		+P4 (3), +P4 +P5 (1)	4	2
+1 P4	Duplication		Non-*AZFc*		+P4 (1), +P4 +P5 (1)	3	3
+1 IR2	Duplication		Non-*AZFc*		+B (1)	3	3
+1 IR3	Duplication		Non-*AZFc*		+IR3 (2)	2	1
+1 P5, +1 IR5	Duplication		Non-*AZFc*			1	1
+1 P5, +1 P4	Duplication		Non-*AZFc*		+P4	1	1
+2 P8	Duplication		Non-*AZFc*			1	1
+1 P5	Duplication		Non-*AZFc*		+P5	1	1
+2 P5, +2 IR5, +2 P4	Duplication		Non-*AZFc*			1	1
-1 IR3, -1 IR1	Deletion	*AMELY* deletion	Non-*AZFc*	NAHR (between *TSPY* copies)	-IR3 (3)	3	2
-1 P8	Deletion		Non-*AZFc*		-P8 (1)	2	2
-1 P4	Deletion		Non-*AZFc*		-P4	1	1

[a]B = blue, T = teal, G = green, R = red, Gr = gray, Y = yellow
[b]Number of males with evidence of each partial CNV shown in parentheses

Table S6. CNVs located both within and outside the *AZFc* region ("both") or completely outside the *AZFc* region ("Non-*AZFc*"). Sorted by location, type, and # of males. CNV classes (listed in column "CNV") count only amplicon copy number changes detected by the whole-amplicon CNV pipeline. See Table S5 legend for explanation of "Partial CNVs" column.

Supplemental Material and Methods

GC bias correction

The GC content of DNA affects read depth in high-throughput sequencing.[36] This bias can drastically differ between sequencing libraries and is primarily driven by the GC content of the entire DNA fragment, rather than just the sequenced read.[73] To correct for this effect, we built a GC bias curve for each sequencing library and corrected sequencing depth based on those curves. To build a GC bias curve, we began by selecting 10,000,000 positions on the autosomes, excluding repetitive regions as annotated by RepeatMasker (http://www.repeatmasker.org). In order to reduce the possibility of any systematic bias due to unanticipated factors in specific regions of the genome, these locations were different for each curve we built, but were always chosen so that regions with very high and very low GC content—which are relatively rare—were well-represented. Then, using the mapping locations of paired reads in the library, we built an empirical distribution of DNA fragment sizes present in the library. For each of the 10,000,000 chosen locations in the genomes, we randomly selected from the empirical fragment size distribution and calculated the GC content of a window of the selected size starting at that location. We sorted each calculated GC percentage into bins of 0.5%. Then, we calculated the GC content of each fragment from the library that began at one of the chosen locations, and again sorted each calculated percentage in bins of 0.5%. For each bin, we divided the number of real fragments by the number of locations and normalized by total sequencing depth of the library. Finally, we smoothed the resulting GC curve with the LOWESS method, using the Statistics module of Biopython.[74] The value of each bin in the final

curve equals the over- or underrepresentation of observed fragments (fragments with that bin's GC content in the sequencing library) relative to expected fragments (the prevalence of regions with that GC content in the genome).

After calculating GC curves, we calculated corrected sequencing depth. Sequencing depth for a location in the genome is normally calculated by adding 1 for each read that overlaps that location. For corrected depth, instead of adding 1 for each read, we add 1 divided by the value of the GC bias curve for the fragment's GC content. If this equals a number > 3, we add 3 instead. This occurs most often for fragments with extremely high or low GC content, which tend to have very low GC curve values. Capping the depth value of a read at 3 prevents rare instances in which, by chance, a region of such fragments has a high number of reads, leading to its depth being exaggerated to extremely high levels in the absence of such a cap.

Branch-sorting analysis

The branch-sorting test generates an analytical p-value of observing a distribution of amplicon CNVs over the detailed phylogenetic tree under selectively neutral conditions. (See Figure S8 and Material and Methods for a description of this test.) We make the assumption that, under selectively neutral conditions, mutation events will be distributed uniformly over the total evolutionary time covered by the tree. This assumption holds true even if Y chromosomes underwent bursts of population expansion throughout history. The phylogenetic tree of Y chromosomes contains within itself the information about such population dynamics; because this analysis calculates the distribution of CNV

events over the total evolutionary time traversed by the tree, the greater number of males in which mutation can occur after a population expansion is reflected in the greater evolutionary time covered by such males within the tree.

For this analysis, we annotate each mutation event, including events that happen on a Y chromosome that has already undergone a previous amplicon CNV mutation event. In contrast, our other, simulation-based analysis did not count such events. We made this distinction to make the simulation-based method more tractable, at some cost of verisimilitude. Most branches in which a mutation event occurred can be annotated by Fitch's algorithm.[44] For 25 branches in which Fitch's algorithm gave an inconclusive result, we manually annotated mutation events based on the most likely mechanism of mutation. For example, when two different variants are child nodes of the same parent node, Fitch's algorithm is inconclusive. If one of the variants could result from a mutation event occurring on a chromosome with the other variant, we annotated the branch of the parent node and the branch of the former variant as having mutation events. If those two variants could not occur from an event occurring to the other variant, we annotated both child node branches as having mutation events.

For a p-value to be valid, its values when testing data that conforms to the null hypothesis must be uniformly distributed between 0 and 1. Therefore, to test the validity of this analysis, we shuffled the order of the branches within the tree, maintaining the presence or absence of a mutation event in each branch, and calculated a p-value in the same way we calculated the p-value of the sorted branches. We performed this process 1,000 times

and calculated the distribution of resulting p-values. We found that the p-values generated by shuffling the branches were indeed uniformly distributed, demonstrating that the test does perform well in this case.

However, two further tests demonstrate the limitations of our method. First, we took the empirical tree structure of the 1000 Genomes Project males and randomly assigned mutation events to branches with various mutation frequencies, ranging from 5×10^{-1} to 5×10^{-7} mutations per father-to-son Y transmission. This generated a number of trees, one per mutation rate, and each with a different total number of mutation events. We then performed 1,000 shuffles of each of these trees. In the trees with a high number of mutation events, the p-value distribution of the resulting shuffled trees was skewed towards low p-values. Second, we simulated amplicon mutation over the tree structure 1,000 times using a mutation rate of 3.83×10^{-4} mutations per father-to-son Y transmission, which is the lower bound calculated from the real data. Unlike our other simulations, branches in which a mutation had occurred in an ancestral branch were allowed to mutate a second time; this allowance of re-mutation is necessary to match our assumption that mutation events should be uniformly distributed over the evolutionary time within the tree. For each simulation, we calculated p-values as described. In this case too, the distribution was skewed towards low p-values (Figure S8I). Additionally, the simulated trees tended to curve below the line representing neutral evolution (i.e. they had more mutation evens in the recent past and fewer in the ancient past).

These results occur for two reasons: first, the KS test is designed to test continuous distributions. Here, the distribution of mutation events is discrete, as we place the mutation event at then center of the branch in which it occurred. Second, our model only allows a single mutation event per branch.

When mutations are rare (as is the case with the real data), these factors make little difference. However, when mutations are more common, the fact that all mutation events are in the center of each branch, combined with the fact that branches not all the same length, creates enough deviation from the continuous null uniform distribution to skew the p-values toward lower values. Further, the fact that longer branches tend to be in the more ancient parts of the tree means that it is more likely that two mutation events (either true or simulated) would occur in a single branch in the more ancient parts of the tree. Those events are only counted as a single event by our method, reducing the number of events counted in the ancient branches of the tree.

Allowing multiple mutations to occur in each branch and distributing them randomly within the branch, rather than in the center, ameliorated these issues. For analysis of our real data, we chose not to do this, to keep the method as simple as possible. We note that the true data had a more extreme KS statistic than all 1,000 simulations; further, the minimum p-value of the 1,000 simulations was 4.99×10^{-4}, compared to $p = 1.01 \times 10^{-7}$ for the real data. Therefore, although our test exaggerates the significance of the p-value, the deviation of the real data from neutral expectation is nevertheless extremely significant. However, our method must be modified for trees that are more densely

populated with mutation events and for trees in which the signature of selection is less extreme.

Supplemental References

36. Dohm, J.C., Lottaz, C., Borodina, T., and Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. Nucleic Acids Res. *36*, e105.

44. Fitch, W.M. (1971). Toward defining the course of evolution: minimum change for a specified tree topology. Systematic Zoology 20, 406-416.

73. Benjamini, Y., and Speed, T.P. (2012). Summarizing and correcting the GC content bias in high-throughput sequencing. Nucleic Acids Res. *40*, e72.

74. Cock, P.J., Antao, T., Chang, J.T., Chapman, B.A., Cox, C.J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., et al. (2009). Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics *25*, 1422-1423.