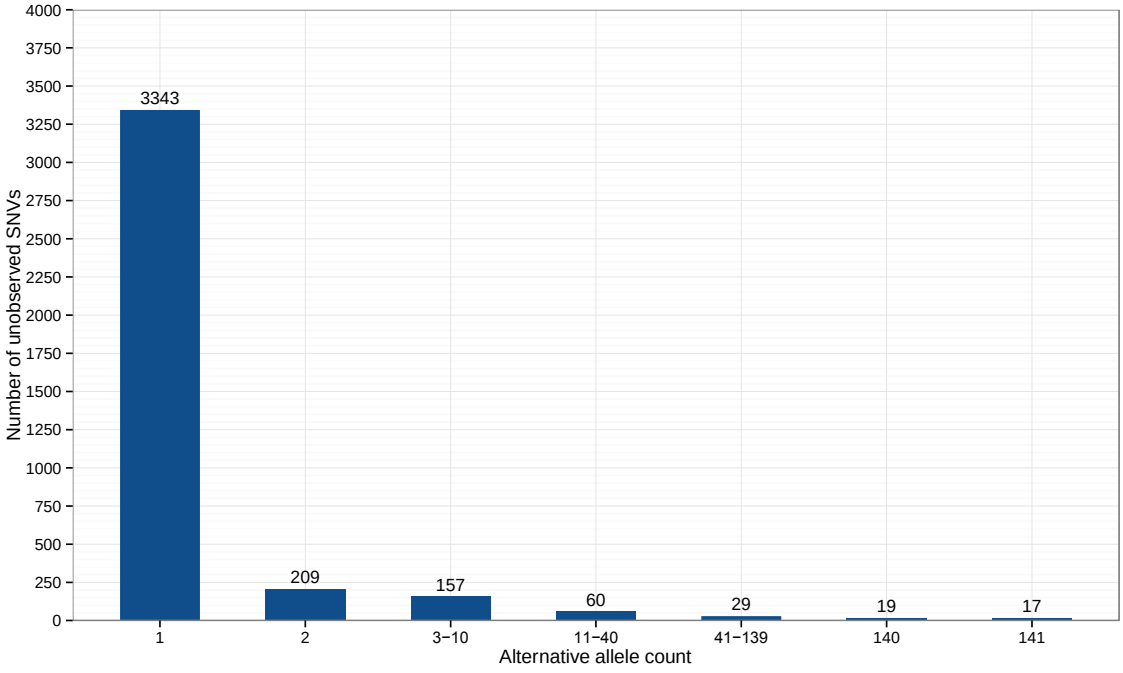
SUPPLEMENTARY INFORMATION

Punctuated bursts in human male demography inferred from 1,244 worldwide Y-chromosome sequences

G. David Poznik[1,2,25], Yali Xue[3,25], Fernando L. Mendez[2], Thomas F. Willems[4,5], Andrea Massaia[3], Melissa A. Wilson Sayres[6,7], Qasim Ayub[3], Shane A. McCarthy[3], Apurva Narechania[8], Seva Kashin[9], Yuan Chen[3], Ruby Banerjee[3], Juan L. Rodriguez-Flores[10], Maria Cerezo[3], Haojing Shao[11], Melissa Gymrek[5,12], Ankit Malhotra[13], Sandra Louzada[3], Rob Desalle[8], Graham R. S. Ritchie[3,17], Eliza Cerveira[13], Tomas W. Fitzgerald[3], Erik Garrison[3], Anthony Marcketta[14], David Mittelman[15,16], Mallory Romanovitch[13], Chengsheng Zhang[13], Xiangqun Zheng-Bradley[17], Goncalo R. Abecasis[18], Steven A. McCarroll[19], Paul Flicek[17], Peter A. Underhill[2], Lachlan Coin[11], Daniel R. Zerbino[17], Fengtang Yang[3], Charles Lee[13,20], Laura Clarke[17], Adam Auton[14], Yaniv Erlich[5,21,22], Robert E. Handsaker[9,19], The 1000 Genomes Project Consortium[23], Carlos D. Bustamante[2,24] & Chris Tyler-Smith[3]

[1]Program in Biomedical Informatics, Stanford University, Stanford, California, USA.
[2]Department of Genetics, Stanford University, Stanford, California, USA.
[3]The Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, Cambridgeshire, UK.
[4]Computational and Systems Biology Program, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[5]New York Genome Center, New York, New York, USA.
[6]School of Life Sciences, Arizona State University, Tempe, Arizona, USA.
[7]Center for Evolution and Medicine, The Biodesign Institute, Arizona State University, Tempe, Arizona, USA.
[8]Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, New York, USA.
[9]Program in Medical and Population Genetics, Broad Institute of MIT and Harvard, Cambridge, Massachusetts, USA.
[10]Department of Genetic Medicine, Weill Cornell Medical College, New York, New York, USA.
[11]Institute for Molecular Bioscience, University of Queensland, St Lucia, Australia.
[12]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA.
[13]The Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA.
[14]Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, USA.
[15]Virginia Bioinformatics Institute, Virginia Tech, Virginia, USA.
[16]Department of Biological Sciences, Virginia Tech, Virginia, USA.
[17]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK.
[18]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, MI, USA.
[19]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA.
[20]Department of Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea.
[21]Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, New York, USA.
[22]Center for Computational Biology and Bioinformatics, Columbia University, New York, New York, USA.
[23]A list of members and affiliations appears in the Supplementary Note.
[24]Department of Biomedical Data Science, Stanford University, Stanford, California, USA.
[25]These authors contributed equally to this work.
Correspondence should be addressed to C.D.B. (cdbustam@stanford.edu) or C.T.-S. (cts@sanger.ac.uk).

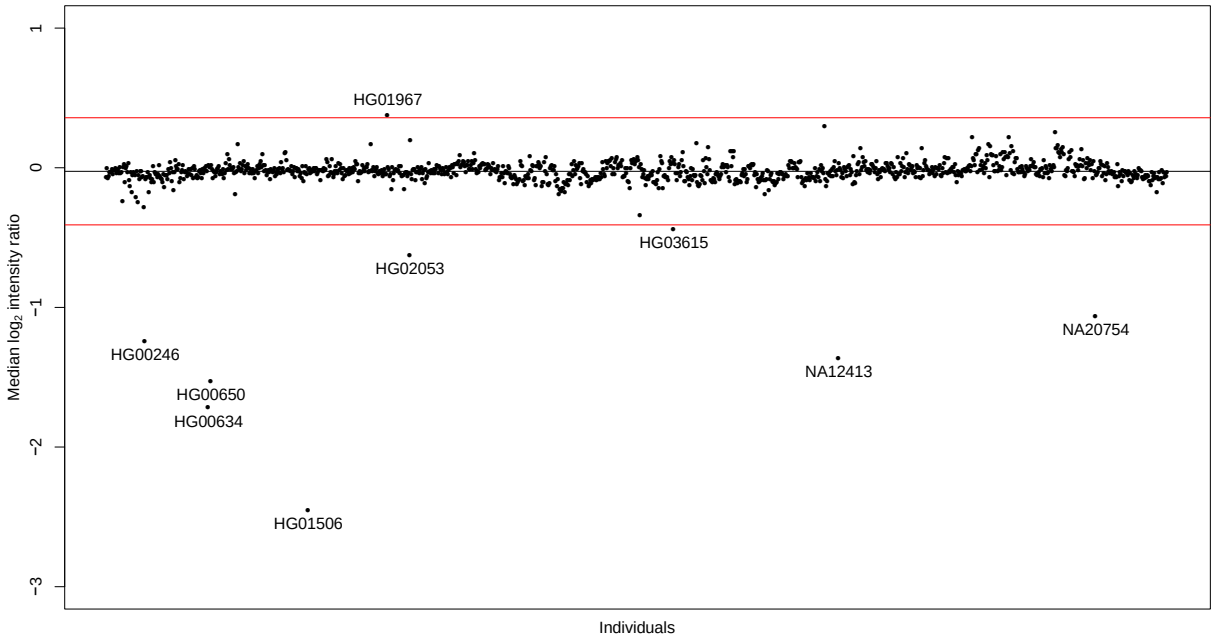
Supplementary Figures



Supplementary Figure 1

Unobserved single-nucleotide variants (SNVs).

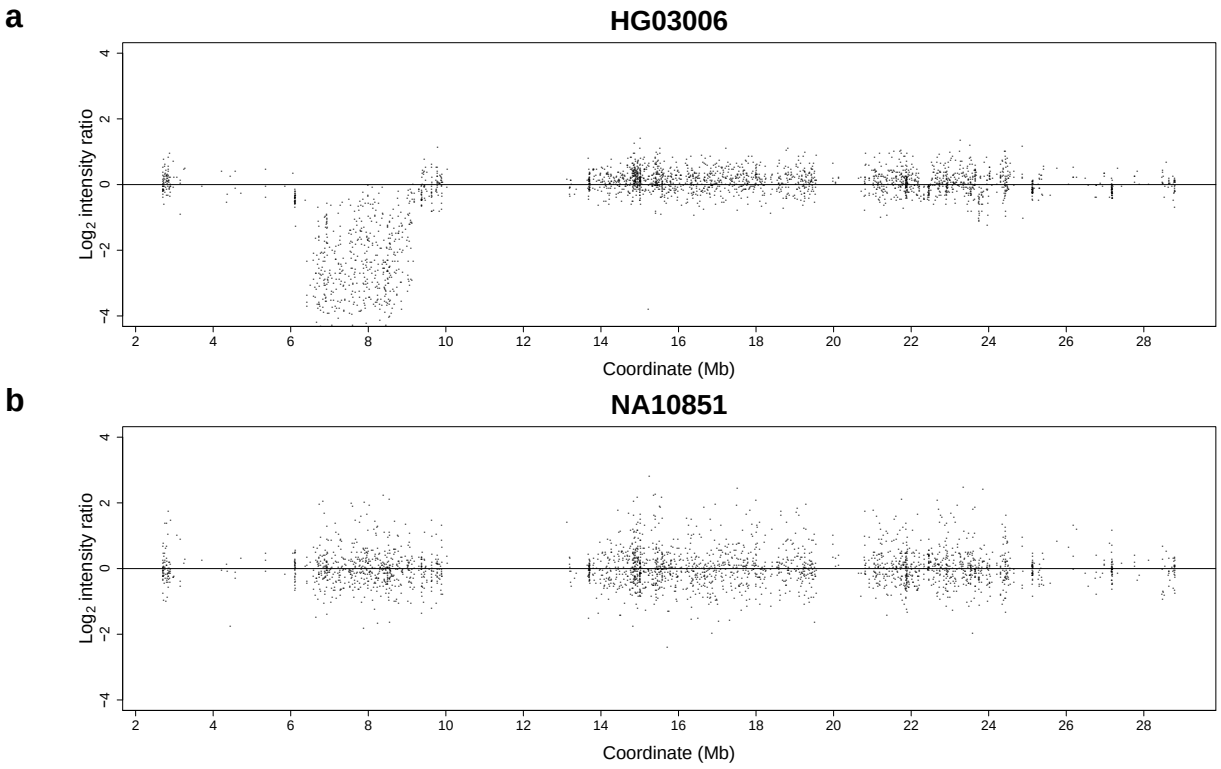
Histogram indicating, for a range of alternative allele frequency bins, the number of SNVs observed in 143 high-coverage Complete Genomics sequences but unobserved in the corresponding low-coverage data (**Supplementary Note 1.3.2**).



Supplementary Figure 2

Sample-wise median aCGH intensity ratios.

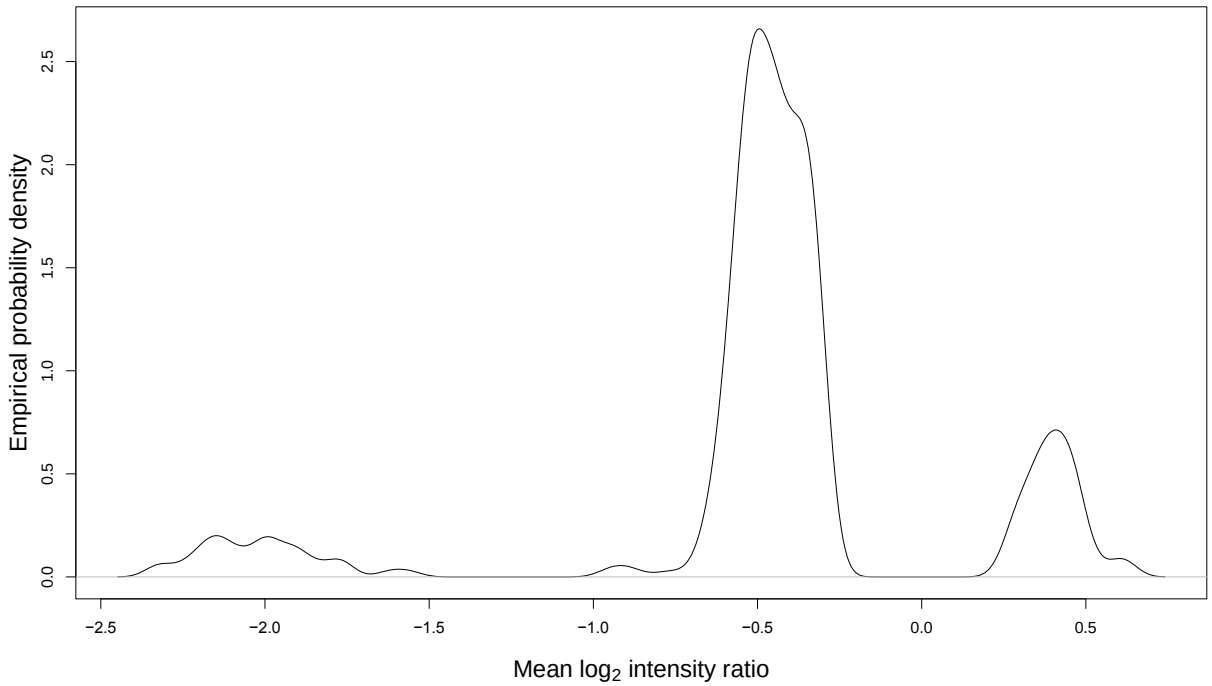
For each sample, median $\log_2$ intensity ratio (prior to normalization) across the 2,714 probes used for CNV calling. The black line represents the mean value across samples, and the red lines indicate an interval of 6 standard deviations, centered on the mean. Samples with values outside this interval are indicated with their IDs (**Supplementary Note 2.1.3**).



Supplementary Figure 3

***AMELY* deletion.**

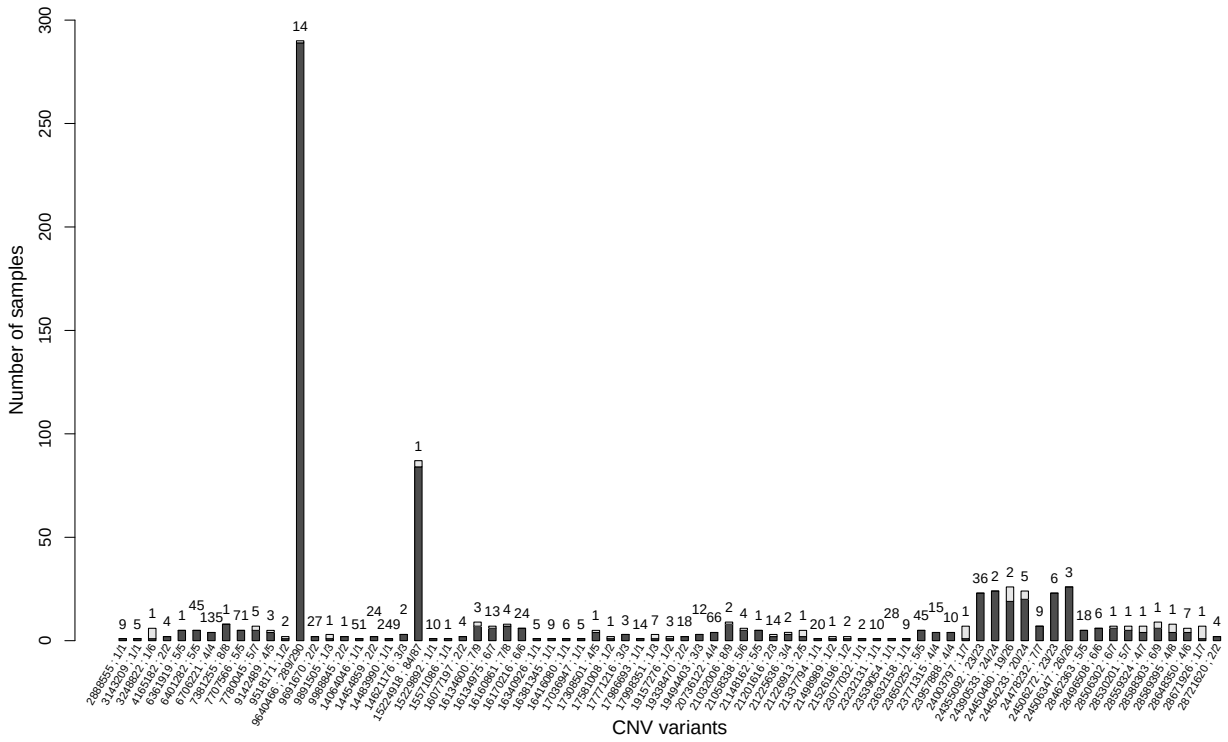
Array CGH $\log_2$ intensity ratios versus genomic coordinate (Mb) for (**a**) sample HG03006 and for (**b**) the reference sample, NA10851. Each point represents a single probe. Lower intensities are clearly visible for HG03006 in Y:6,103,728–9,397,666, which includes *AMELY* (**Supplementary Note 2.1.3**).



Supplementary Figure 4

Copy-number allele calling in a repetitive region.

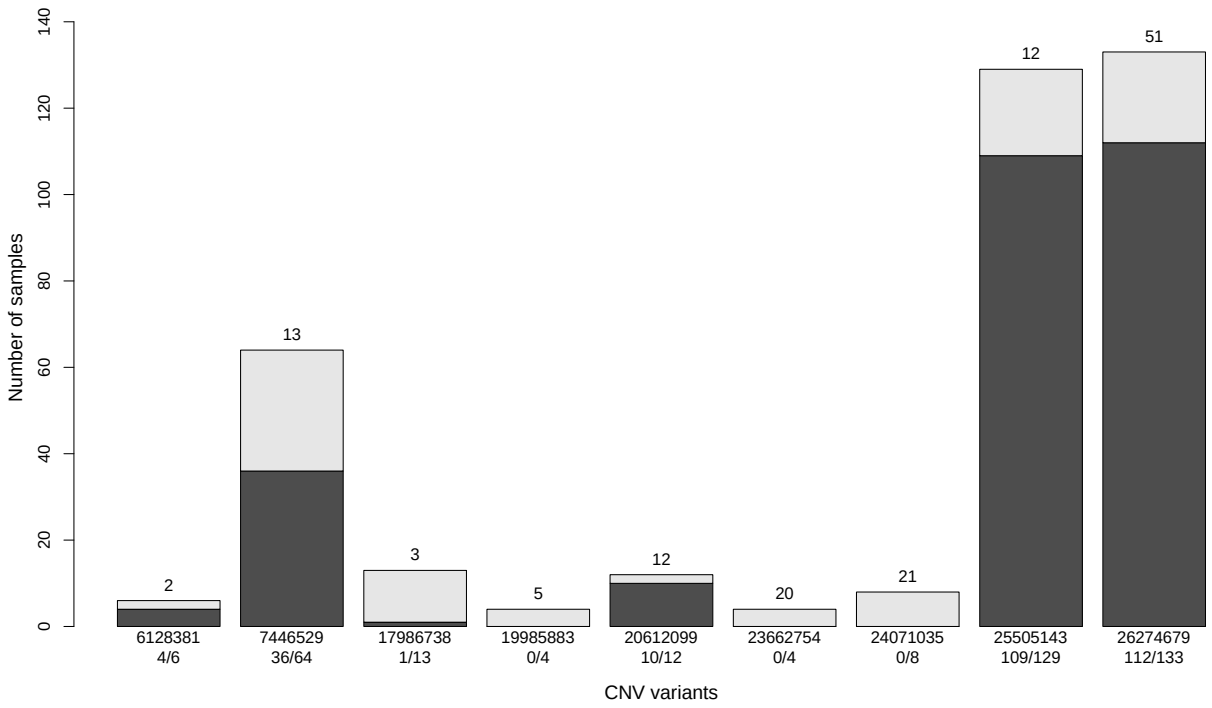
Empirical probability density of mean $\log_2$ aCGH intensity ratio for 410 non-reference calls in Y:22,218,957–22,508,011, a region for which we considered the reference allele to be two copies. Only variant samples—those for which we inferred a gain or loss of at least 0.2 copies—are shown. Peaks around –2, –0.5, and 0.4 likely correspond to 0, 1, and 3 copies, respectively (**Supplementary Note 2.1.3**).



Supplementary Figure 5

Validation of copy-number variants (CNVs).

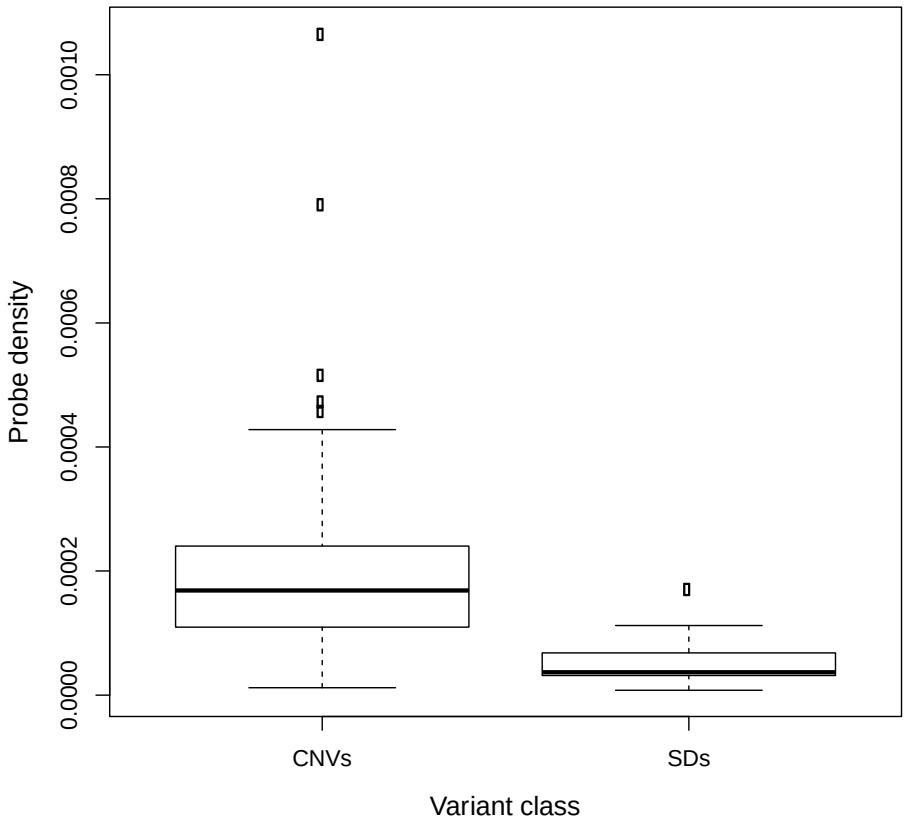
Each bar represents one variant called by Genome STRiP and covered by at least one aCGH probe. The number of probes covering each variant is reported above the bars, and the height of each bar represents the number of samples that were called as ALT by Genome STRiP and were present in the aCGH dataset. The dark gray sub-bars represent samples confirmed as ALTs, while the light gray sub-bars represent samples not confirmed as ALTs. Below each bar, the starting position of the variant and the ratio of confirmed ALTs to unconfirmed ALTs are reported, separated by a semicolon (**Supplementary Note 2.2.1**).



Supplementary Figure 6

Validation of segmental duplications (SDs).

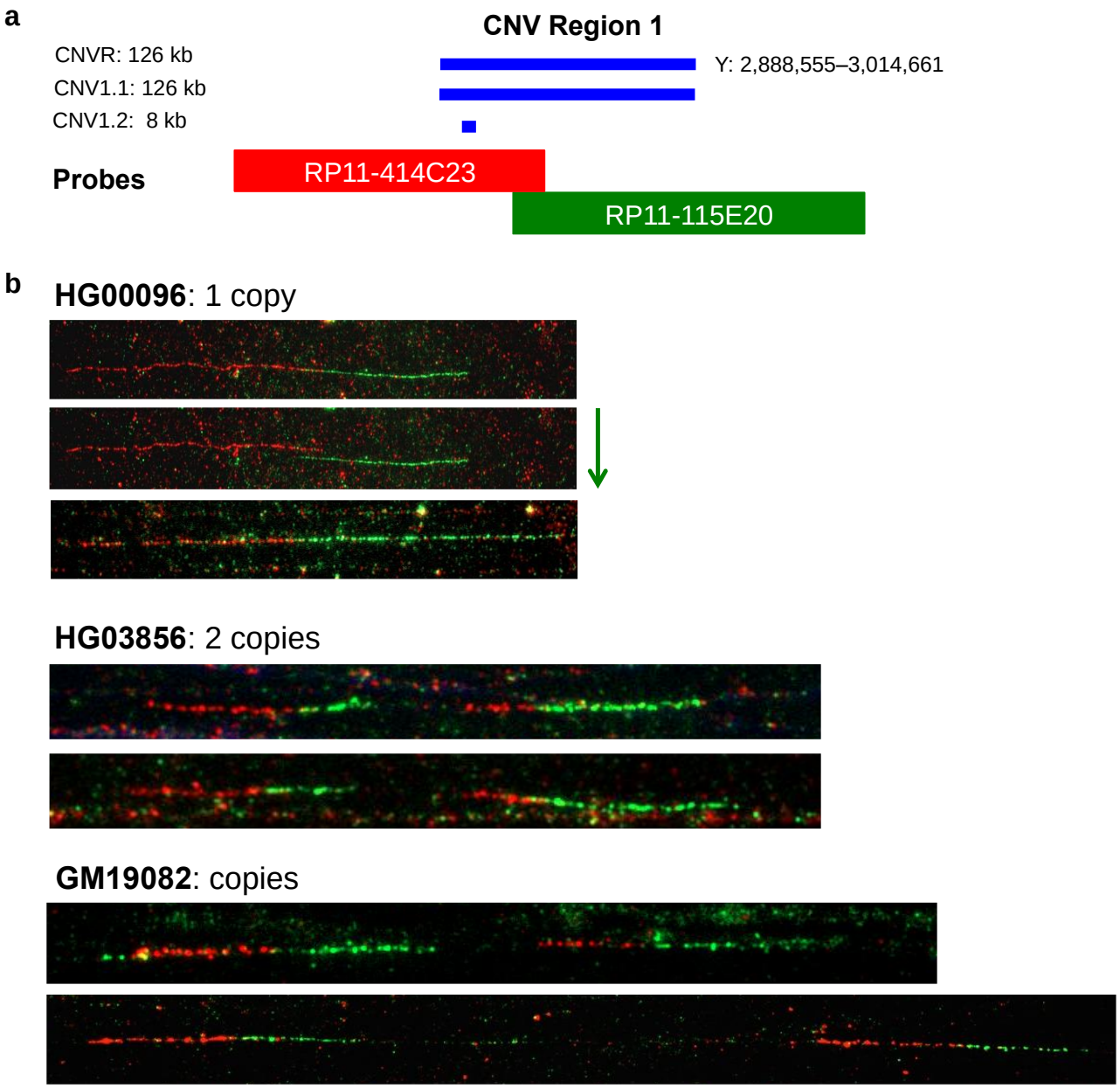
Each bar represents one variant called by Genome STRiP and covered by at least one aCGH probe. The number of probes covering each variant is reported above the bars, and the height of each bar represents the number of samples called as ALT by Genome STRiP that were present in the aCGH dataset. The dark gray sub-bars represent samples confirmed as ALTs, while the light gray sub-bars represent samples not confirmed as ALTs. Below each bar, the starting position and the ratio of confirmed to unconfirmed ALTs are reported for each variant (**Supplementary Note 2.2.1**).



Supplementary Figure 7

Array CGH probe density.

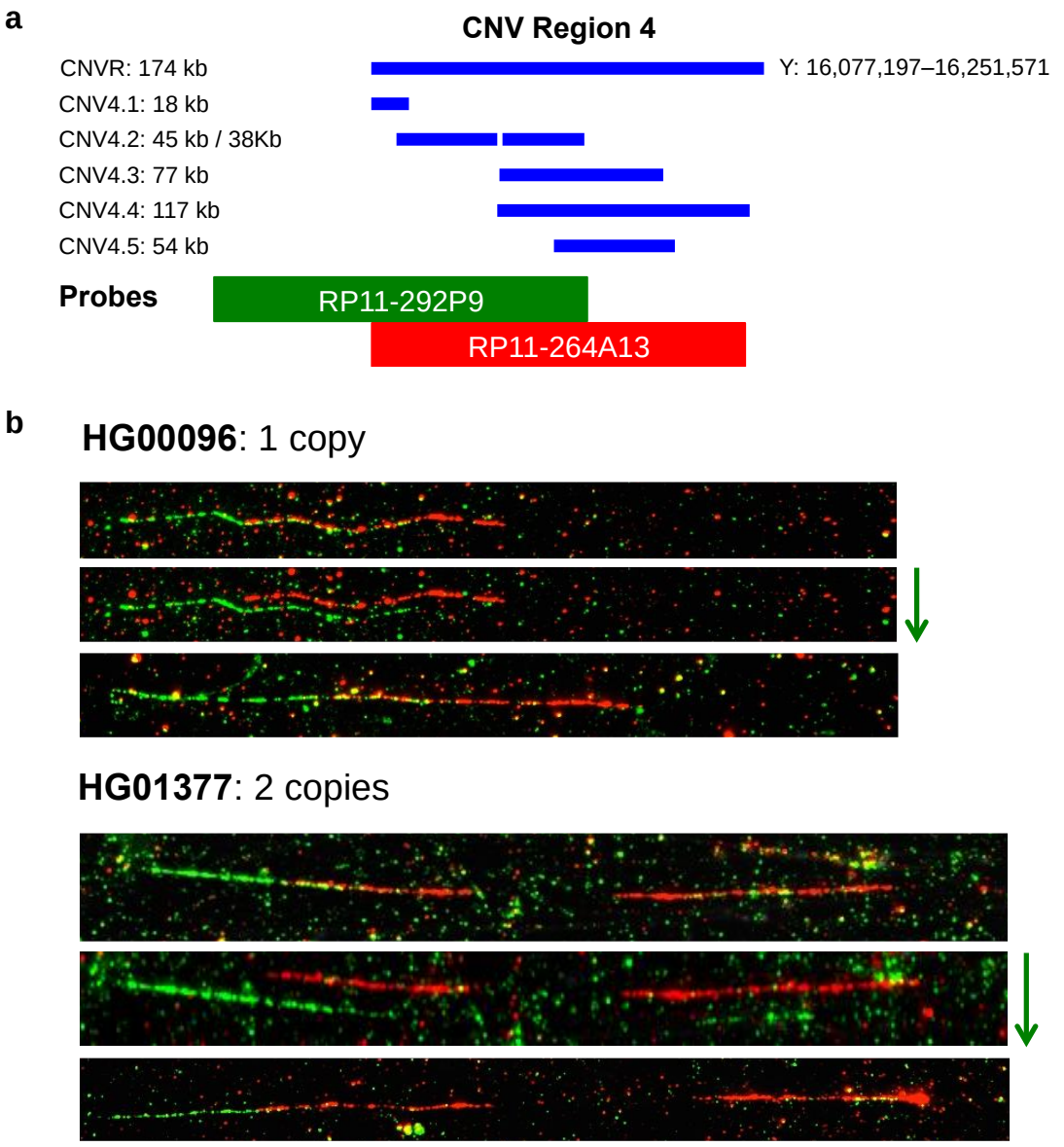
Boxplots representing the distribution of probe densities for CNVs (left) and for SDs (right) (**Supplementary Note 2.2.1**).



Supplementary Figure 8

Fibre-FISH validation of CNV region 1.

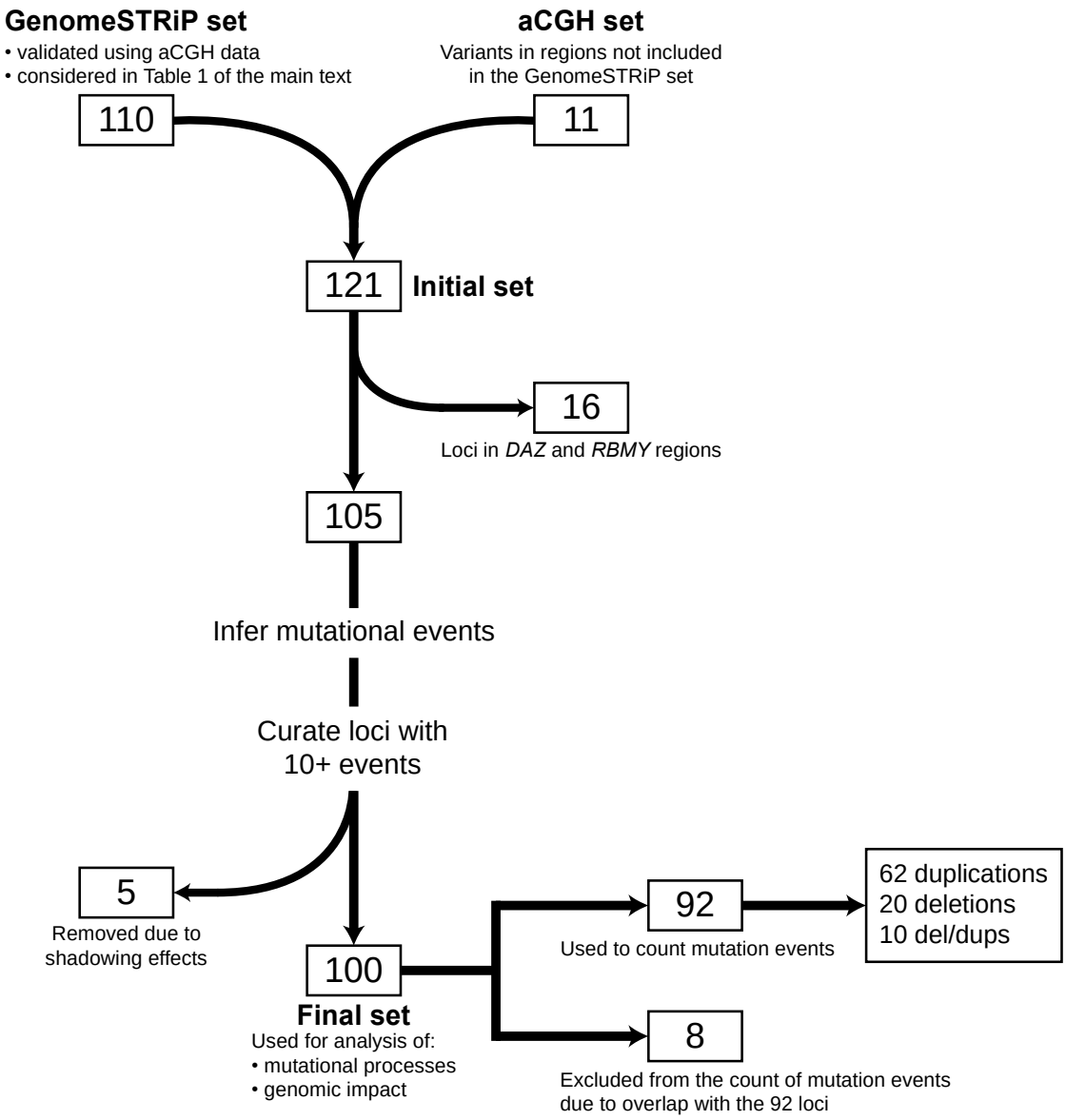
We observe one copy in HG00096 and 2 copies in HG03856 and in GM19082. (**a**) Diagrammatic representation of the CNV region and probe design, showing the relative sizes of the CNV, the BAC clones, and their overlap. (**b**) Representative fibre-FISH images, with RP-11443C23 in red and RP11-115E20 in green. The green signal has been shifted downward in the image marked with a green arrow, indicating the overlap between the two BACs (18,550 bp). (**Supplementary Note 2.2.2**)



Supplementary Figure 9

Fibre-FISH validation of CNV region 4.

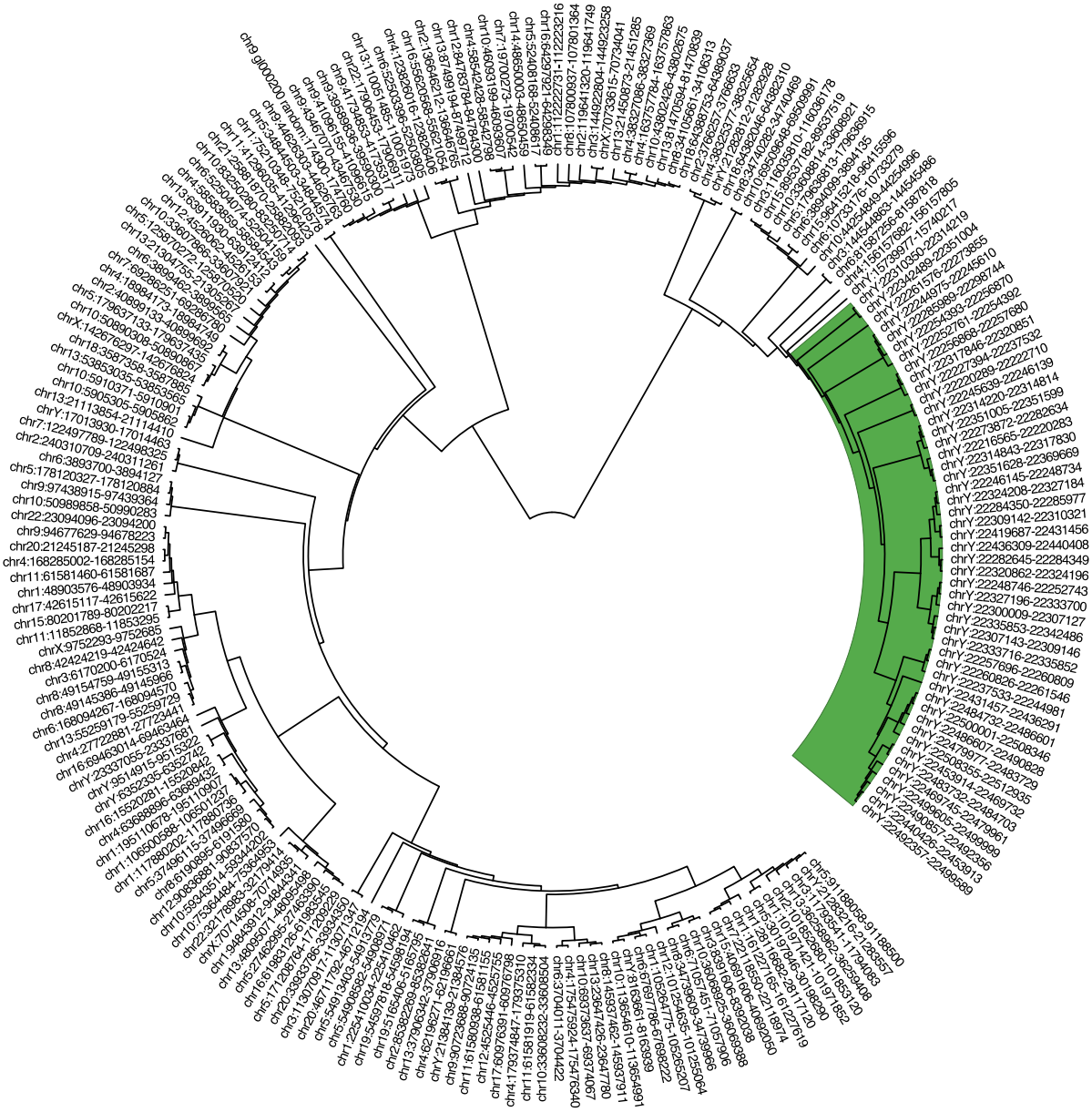
We observe one copy in HG00096 and two copies in HG01377. (**a**) Diagrammatic representation of the CNV region and probe design, showing the relative sizes of the CNV, the BAC clones, and their overlap. (**b**) Representative fibre-FISH images, with RP-292P9 in green and RP11-1264A13 in red. Green arrows mark images in which the green signal has been shifted downward. In HG00096, the signal shift shows the overlap between the two clones (~100 kb), and, in HG01377, it also demonstrates that this CNV primarily involves RP11-264A13. (**Supplementary Note 2.2.2**)



Supplementary Figure 10

Work flow for analysis of CNVs.

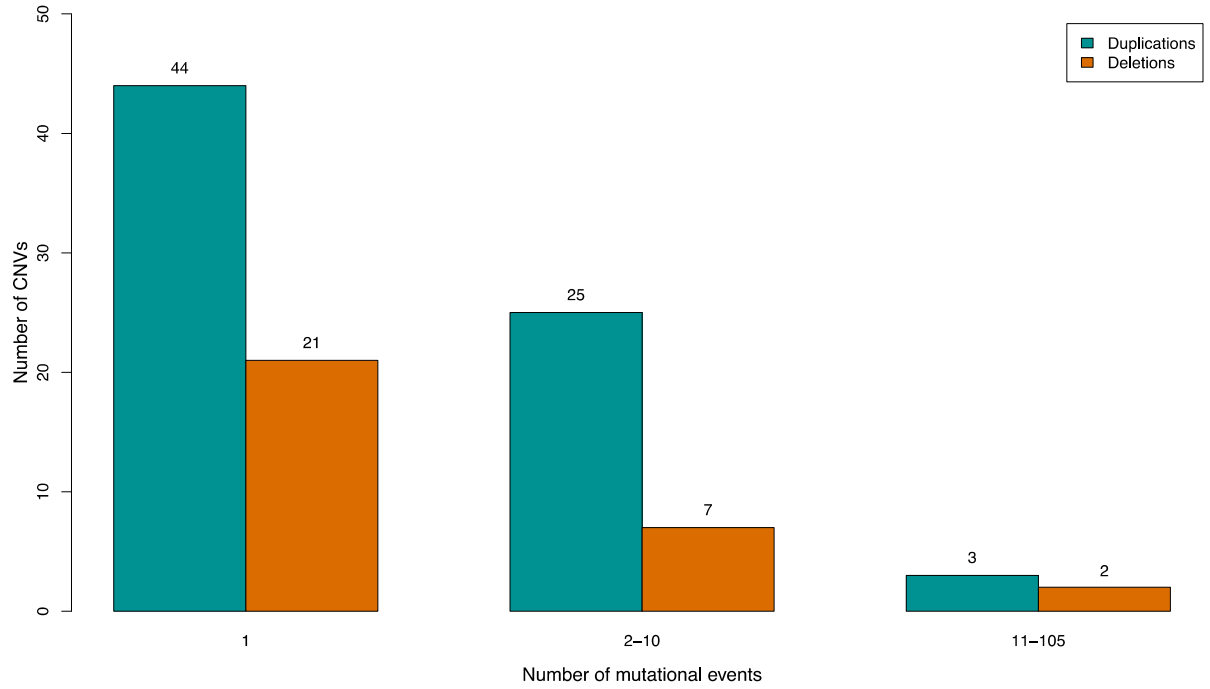
Flow chart summarizing the number of loci at each stage of the CNV analysis (**Supplementary Note 2.3.1**).



Supplementary Figure 11

Phylogeny of *LTR12B* elements in the human genome.

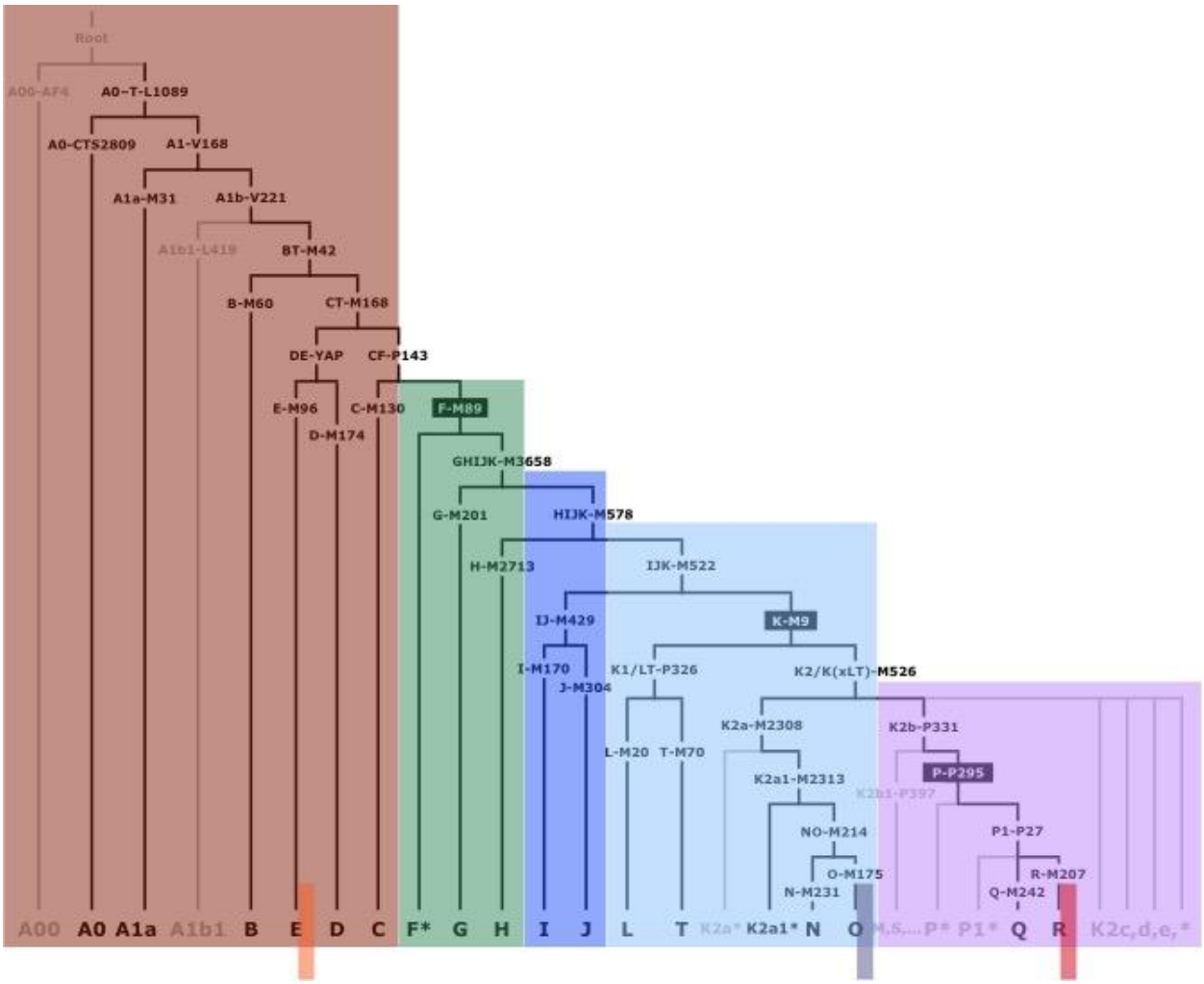
In this maximum-likelihood tree, calculated with MEGA6[1] and edited with FigTree[2], we indicate as leaf labels the coordinates (chromosome:start–end) of each of the human genome's 211 *LTR12B* elements. Branches leading to the elements in Y:22,216,565–22,369,669 and Y:22,419,687–22,512,935 are highlighted in green (**Supplementary Note 2.3.1**).



Supplementary Figure 12

CNV mutation events.

Bar plot indicating the number of CNVs associated with 1, 2–10, or 11–105 mutation events. Teal and orange bars represent duplications and deletions, respectively (**Supplementary Note 2.3.1**).



Supplementary Figure 13

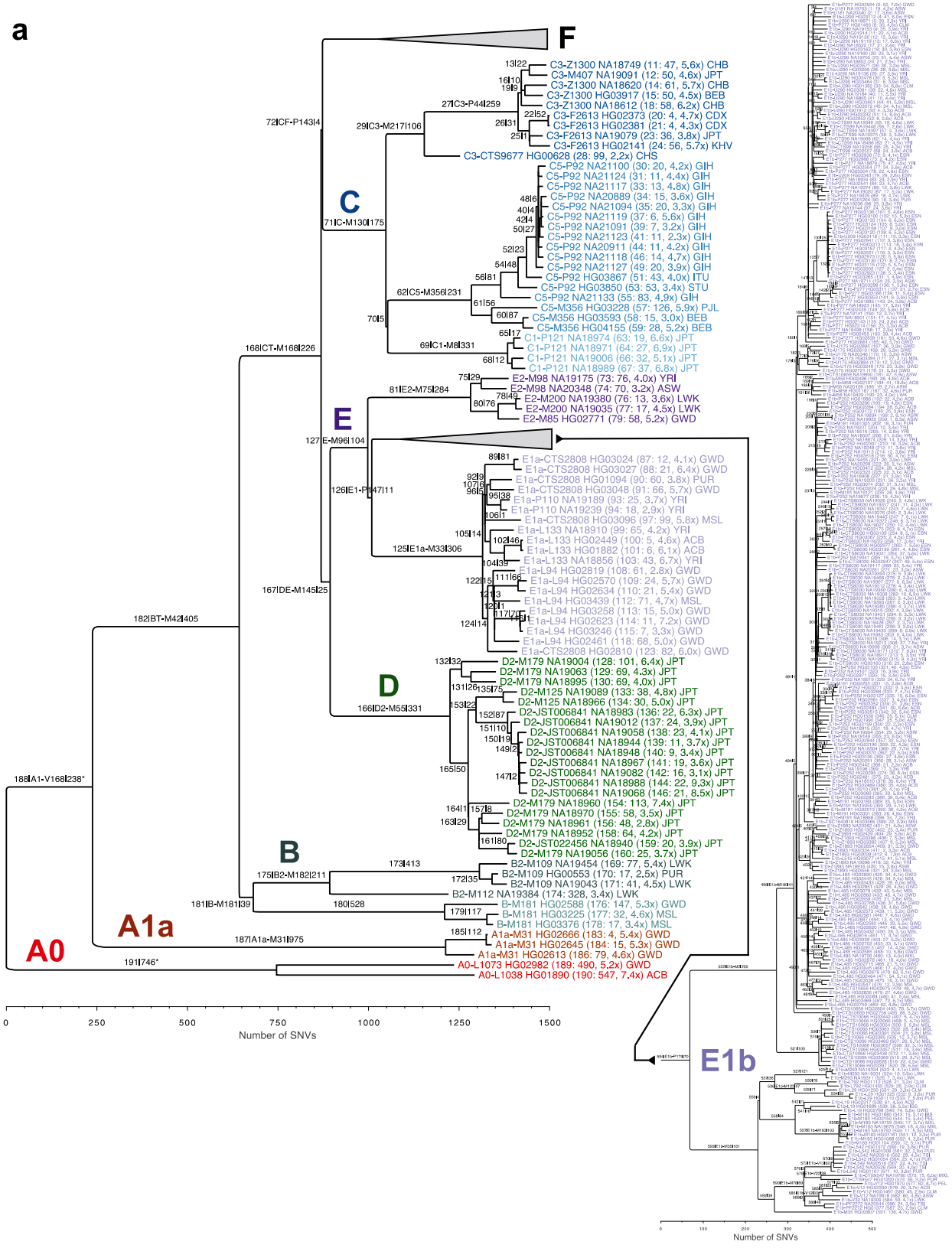
Partitioning the phylogeny.

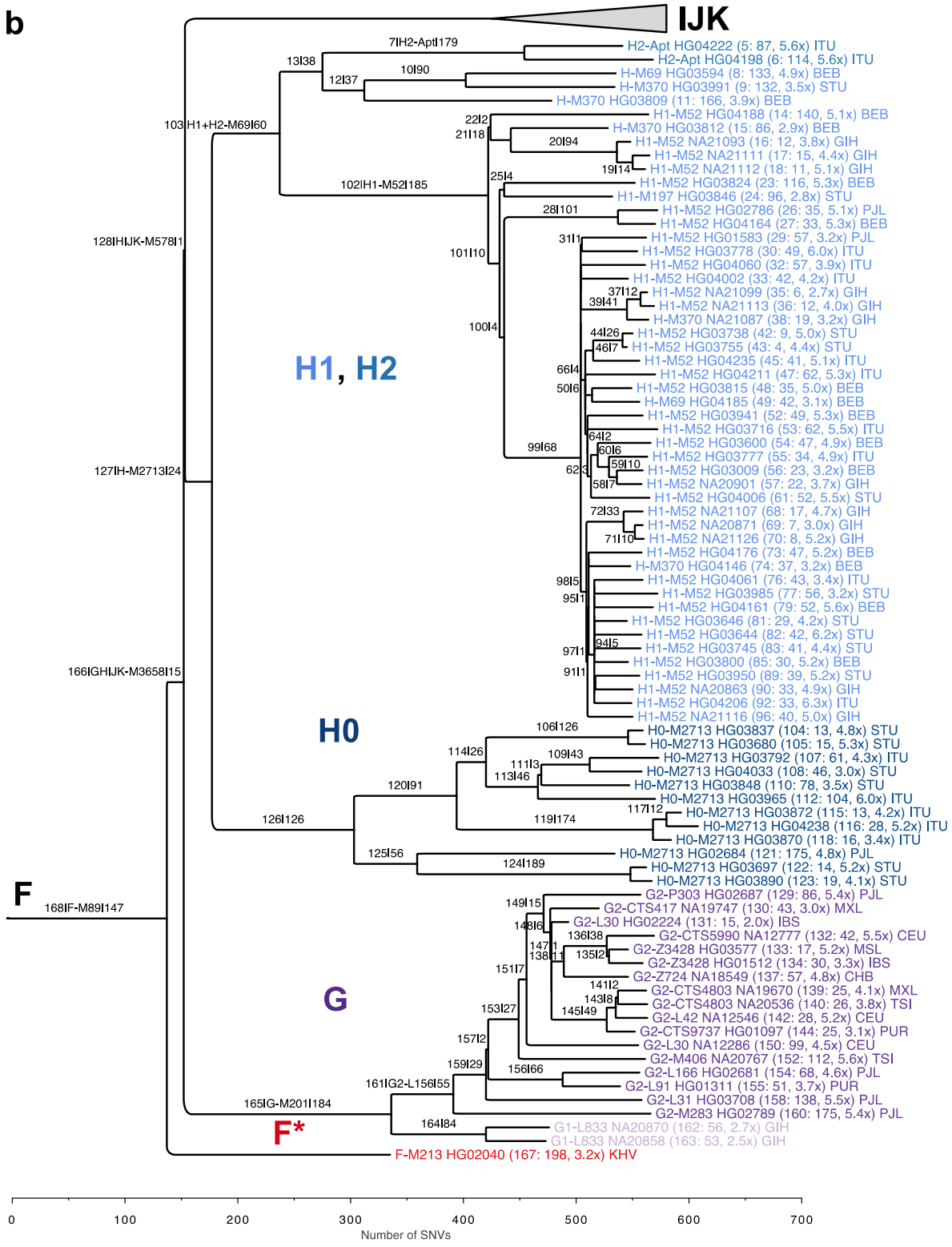
Overview of Y-chromosome tree, with major haplogroups and their defining SNPs indicated. Labels with black backgrounds mark megahaplogroups F, K, and P, and gray lineages were not sampled in this study. Colored rectangles indicate the partitioning strategy used for mapping SNVs to the tree: five main contiguous blocks and three nested components for the most frequent haplogroups: E1b, O3, and R1b (**Supplementary Note 4.3**).

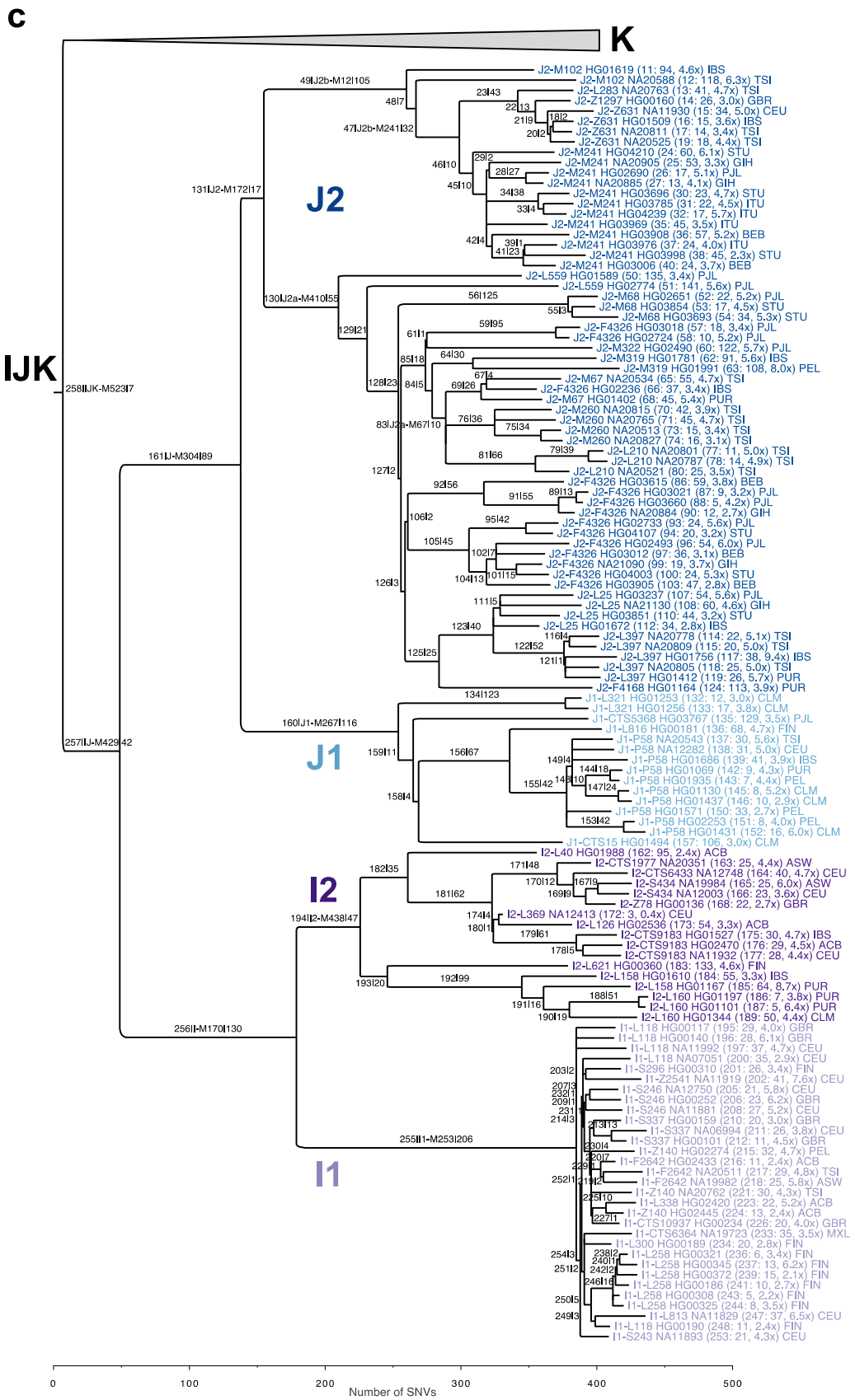
Supplementary Figure 14

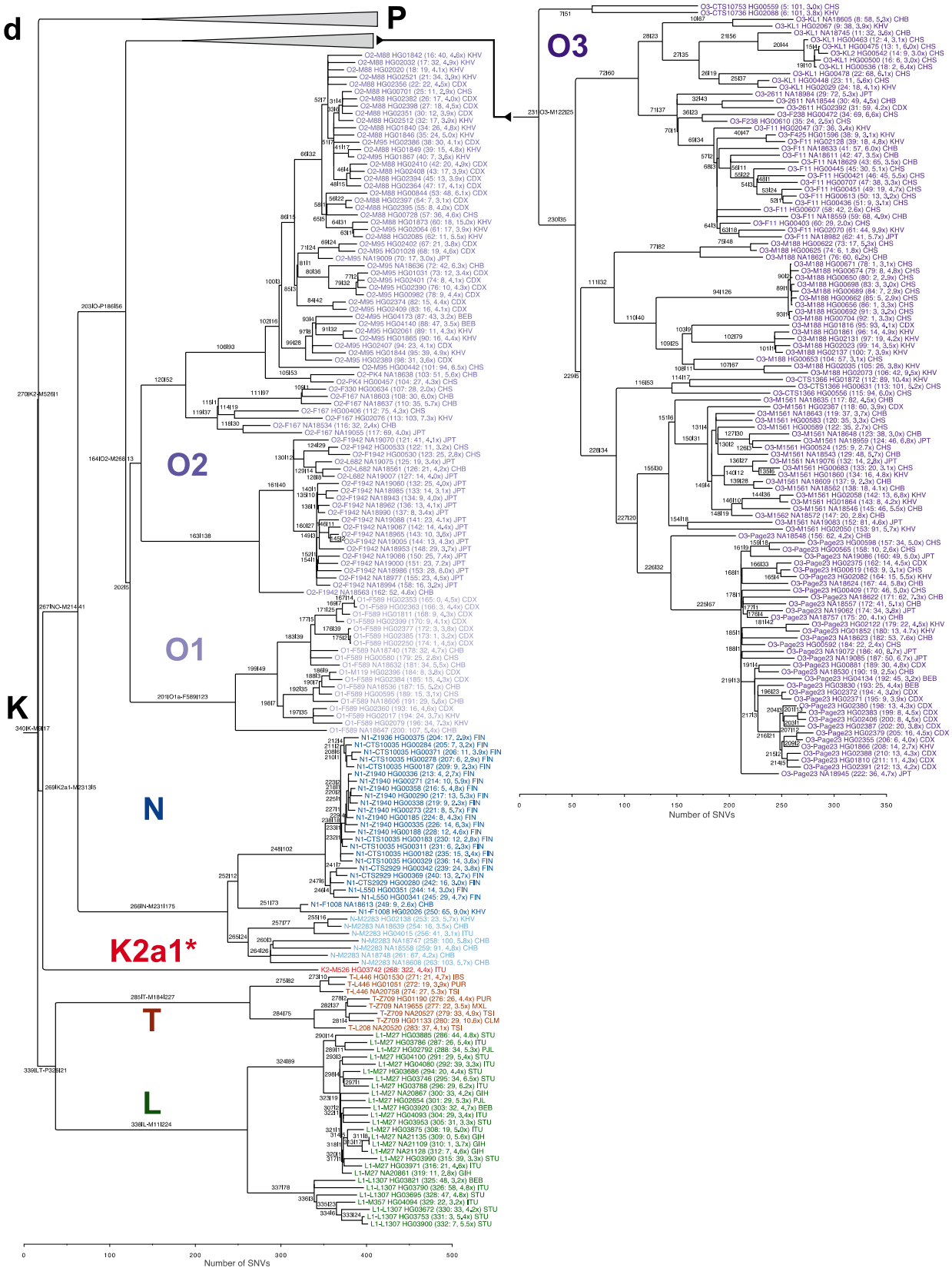
Observed phylogeny, partitioned into eight linked subtrees.

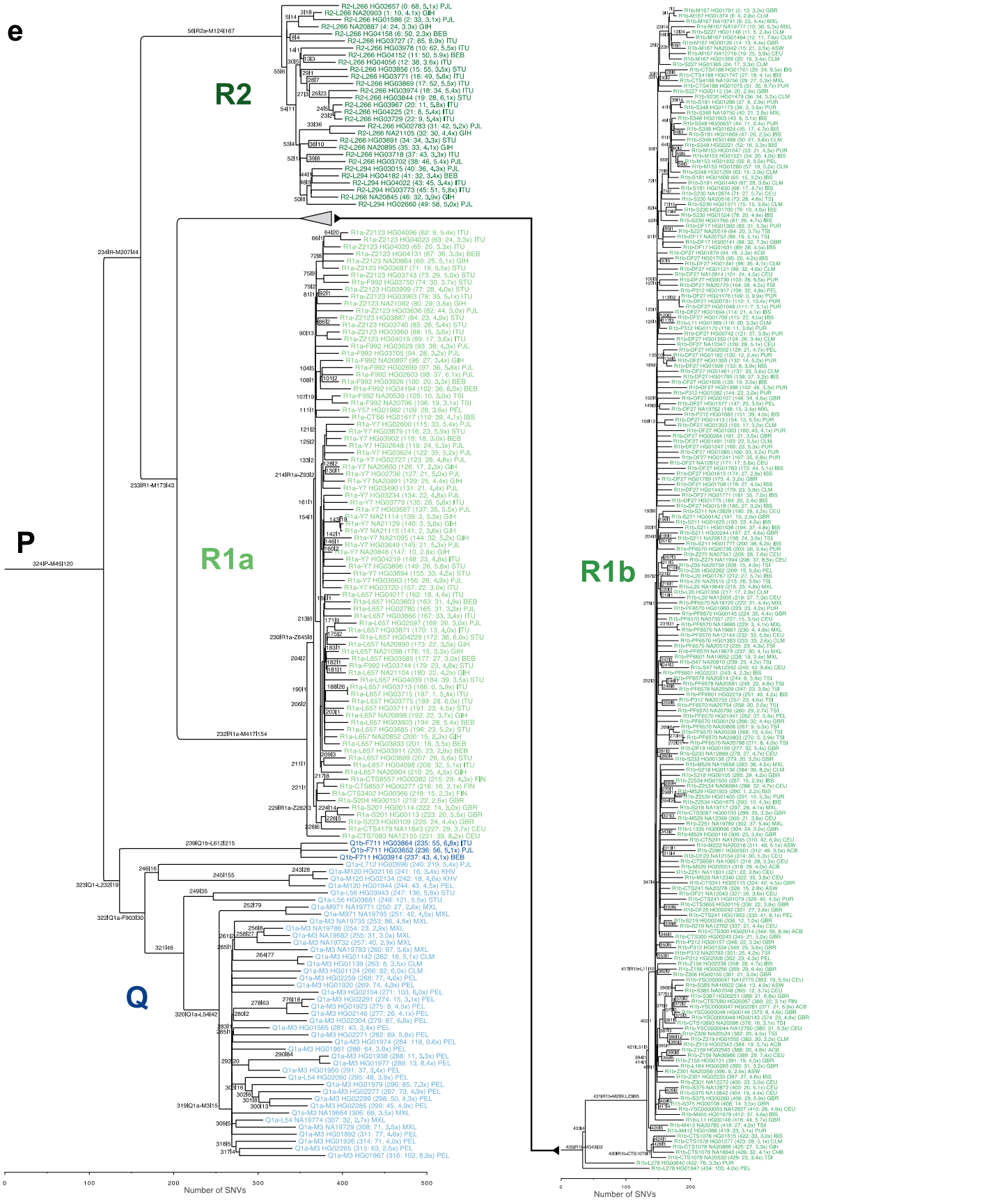
(Figure appears on the following 5 pages.) Branch lengths are drawn in proportion to the number of SNVs that map compatibly to each branch. Internal branches are labeled with an index, a canonical SNP (major branches only), and the branch length, separated by pipes (|). Terminal branches are labeled with the individual's haplogroup, most derived ISOGG SNP, and sample ID, then, in parentheses, the branch index (followed by a colon), branch length (number of singletons), and sequencing coverage. The population is indicated last. Gray triangles are place-holders for subsequent subtrees. The asterisk (*) denotes an approximate branch length due to lack of polarization at the most ancestral split, that between A0 and A1. (**a**) Haplogroups A0, A1a, B, D, E, and C (*n* = 88), with E1b inset (*n* = 298). (**b**) Haplogroups F*, G, and H (*n* = 82). (**c**) Haplogroups I and J (*n* = 124). (**d**) Haplogroups L, T, K2a1*, N, and O (*n* = 162), with O3 inset (*n* = 114). (**e**) Haplogroups Q and R (*n* = 160), with R1b inset (*n* = 216). (**Supplementary Note 4.3**)

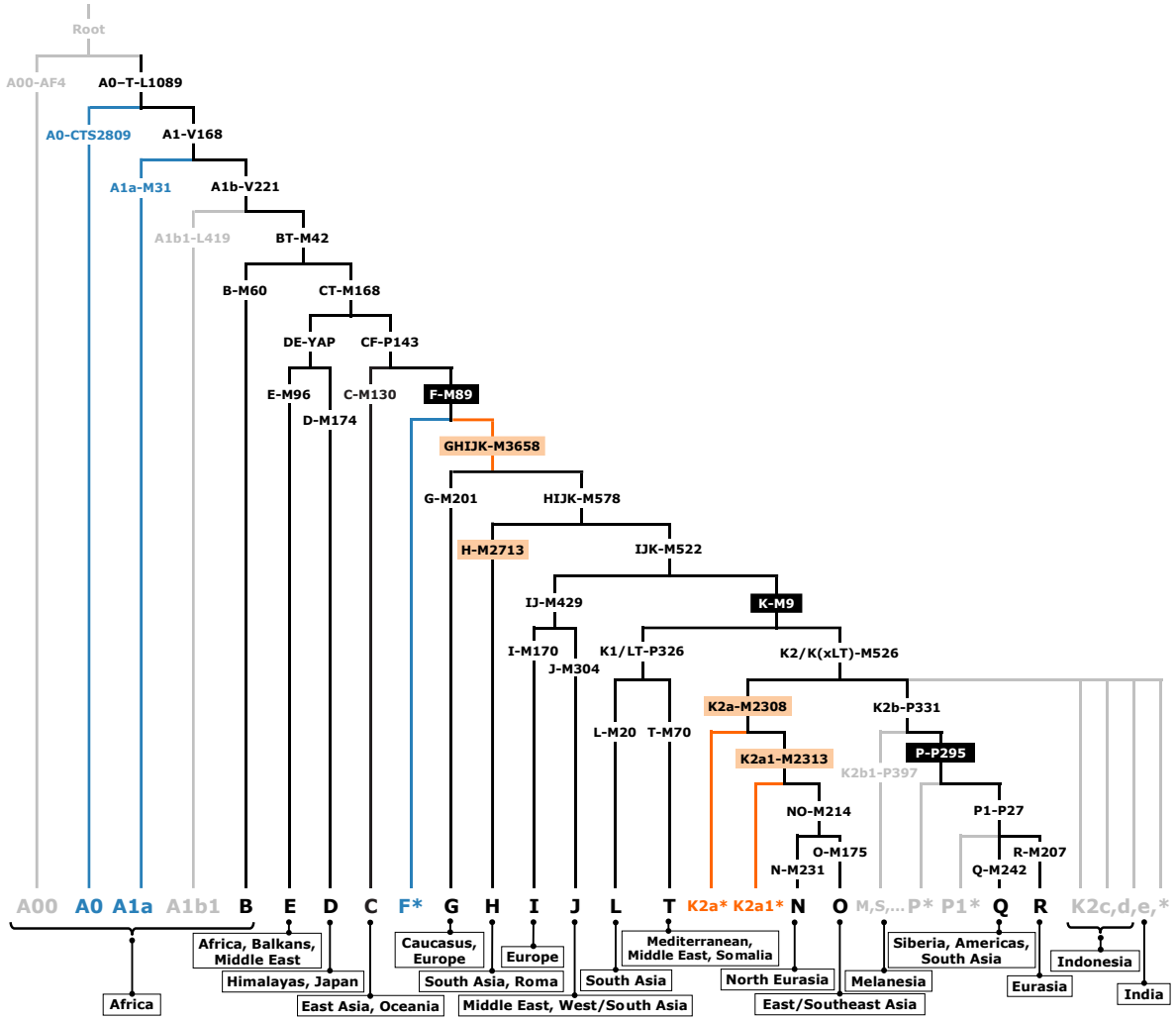

a
F
C
E
D
B
A1a
A0
E1b
C3-Z1300 NA18749 (11: 47, 5.6x) CHB
C3-M407 NA19091 (12: 50, 4.6x) JPT
C3-Z1300 NA18620 (14: 61, 5.7x) CHB
C3-Z1300 HG03917 (15: 50, 4.5x) BEB
C3-Z1300 NA18612 (18: 58, 6.2x) CHB
C3-F2613 HG02373 (20: 4, 4.7x) CDX
C3-F2613 HG02381 (21: 4, 4.3x) CDX
C3-F2613 NA19079 (23: 36, 3.8x) JPT
C3-F2613 HG02141 (24: 56, 5.7x) KHV
C3-CTS9677 HG00628 (28: 99, 2.2x) CHS
C5-P92 NA21100 (30: 20, 4.2x) GIH
C5-P92 NA21124 (31: 11, 4.4x) GIH
C5-P92 NA21117 (33: 13, 4.8x) GIH
C5-P92 NA20889 (34: 15, 3.6x) GIH
C5-P92 NA21094 (35: 20, 3.3x) GIH
C5-P92 NA21119 (37: 6, 5.6x) GIH
C5-P92 NA21091 (39: 7, 3.2x) GIH
C5-P92 NA21123 (41: 11, 2.3x) GIH
C5-P92 NA20911 (44: 11, 4.2x) GIH
C5-P92 NA21118 (46: 14, 4.7x) GIH
C5-P92 NA21127 (49: 20, 3.9x) GIH
C5-P92 HG03867 (51: 43, 4.0x) ITU
C5-P92 HG03850 (53: 53, 3.4x) STU
C5-P92 NA21133 (55: 83, 4.9x) GIH
C5-M356 HG03228 (57: 126, 5.9x) PJL
C5-M356 HG03593 (58: 15, 3.0x) BEB
C5-M356 HG04155 (59: 28, 5.2x) BEB
C1-P121 NA18974 (63: 19, 6.6x) JPT
C1-P121 NA18971 (64: 27, 6.9x) JPT
C1-P121 NA19006 (66: 32, 5.1x) JPT
C1-P121 NA18989 (67: 37, 6.8x) JPT
E2-M98 NA19175 (73: 76, 4.0x) YRI
E2-M98 NA20348 (74: 70, 3.2x) ASW
E2-M200 NA19380 (76: 13, 3.6x) LWK
E2-M200 NA19035 (77: 17, 4.5x) LWK
E2-M85 HG02771 (79: 58, 5.2x) GWD
E1a-CTS2808 HG03024 (87: 12, 4.1x) GWD
E1a-CTS2808 HG03027 (88: 21, 6.4x) GWD
E1a-CTS2808 HG01094 (90: 60, 3.8x) PUR
E1a-CTS2808 HG03048 (91: 66, 5.7x) GWD
E1a-P110 NA19189 (93: 25, 3.7x) YRI
E1a-P110 NA19239 (94: 18, 2.9x) YRI
E1a-CTS2808 HG03096 (97: 99, 5.8x) MSL
E1a-L133 NA18910 (99: 65, 4.2x) YRI
E1a-L133 HG02449 (100: 5, 4.6x) ACB
E1a-L133 HG01882 (101: 6, 6.1x) ACB
E1a-L133 NA18856 (103: 43, 6.7x) YRI
E1a-L94 HG02819 (108: 61, 2.8x) GWD
E1a-L94 HG02570 (109: 24, 5.7x) GWD
E1a-L94 HG02634 (110: 21, 5.4x) GWD
E1a-L94 HG03439 (112: 71, 4.7x) MSL
E1a-L94 HG03258 (113: 15, 5.0x) GWD
E1a-L94 HG02623 (114: 11, 7.2x) GWD
E1a-L94 HG03246 (115: 7, 3.3x) GWD
E1a-L94 HG02461 (118: 68, 5.0x) GWD
E1a-CTS2808 HG02810 (123: 82, 6.0x) GWD
D2-M179 NA19004 (128: 101, 6.4x) JPT
D2-M179 NA19063 (129: 69, 4.3x) JPT
D2-M179 NA18995 (130: 69, 4.0x) JPT
D2-M125 NA19089 (133: 38, 4.8x) JPT
D2-M125 NA18966 (134: 30, 5.0x) JPT
D2-JST006841 NA18983 (136: 22, 6.3x) JPT
D2-JST006841 NA19012 (137: 24, 3.9x) JPT
D2-JST006841 NA19058 (138: 23, 4.1x) JPT
D2-JST006841 NA18944 (139: 11, 3.7x) JPT
D2-JST006841 NA18948 (140: 9, 3.4x) JPT
D2-JST006841 NA18967 (141: 19, 3.6x) JPT
D2-JST006841 NA19082 (142: 16, 3.1x) JPT
D2-JST006841 NA18988 (144: 22, 9.3x) JPT
D2-JST006841 NA19068 (146: 21, 8.5x) JPT
D2-M179 NA18960 (154: 113, 7.4x) JPT
D2-M179 NA18970 (155: 58, 3.5x) JPT
D2-M179 NA18961 (156: 48, 2.8x) JPT
D2-M179 NA18952 (158: 64, 4.2x) JPT
D2-JST022456 NA18940 (159: 20, 3.9x) JPT
D2-M179 NA19056 (160: 25, 3.7x) JPT
B2-M109 NA19454 (169: 77, 5.4x) LWK
B2-M109 HG00553 (170: 17, 2.5x) PUR
B2-M109 NA19043 (171: 41, 4.5x) LWK
B2-M112 NA19384 (174: 328, 3.4x) LWK
B-M181 HG02588 (176: 147, 5.3x) GWD
B-M181 HG03225 (177: 32, 4.6x) MSL
B-M181 HG03376 (178: 17, 3.4x) MSL
A1a-M31 HG02666 (183: 4, 5.4x) GWD
A1a-M31 HG02645 (184: 15, 5.3x) GWD
A1a-M31 HG02613 (186: 79, 4.6x) GWD
A0-L1073 HG02982 (189: 490, 5.2x) GWD
A0-L1038 HG01890 (190: 547, 7.4x) ACB
72|CF-P143|4
29|C3-M217|106
27|C3-P44|259
71|C-M130|175
70|5
62|C5-M356|231
168|CT-M168|226
69|C1-M8|331
81|E2-M75|284
127|E-M96|104
126|E1-P147|11
125|E1a-M33|306
167|DE-M145|25
166|D2-M55|331
182|BT-M42|405
188|A1-V168|238*
175|B2-M182|211
181|B-M181|39
187|A1a-M31|975
191|746*
0
250
500
750
1000
1250
1500
Number of SNVs
100
200
300
400
500
Number of SNVs


b
IJK
7|H2-Apt|179
H2-Apt HG04222 (5: 87, 5.6x) ITU
H2-Apt HG04198 (6: 114, 5.6x) ITU
13|38
12|37
10|90
H-M69 HG03594 (8: 133, 4.9x) BEB
H-M370 HG03991 (9: 132, 3.5x) STU
H-M370 HG03809 (11: 166, 3.9x) BEB
103 H1+H2-M69|60
22|2
21|18
H1-M52 HG04188 (14: 140, 5.1x) BEB
H-M370 HG03812 (15: 86, 2.9x) BEB
20|94
H1-M52 NA21093 (16: 12, 3.8x) GIH
H1-M52 NA21111 (17: 15, 4.4x) GIH
19|14
H1-M52 NA21112 (18: 11, 5.1x) GIH
25|4
102|H1-M52|185
H1-M52 HG03824 (23: 116, 5.3x) BEB
H1-M197 HG03846 (24: 96, 2.8x) STU
28|101
H1-M52 HG02786 (26: 35, 5.1x) PJL
H1-M52 HG04164 (27: 33, 5.3x) BEB
128|HIJK-M578|1
31|1
H1-M52 HG01583 (29: 57, 3.2x) PJL
101|10
H1-M52 HG03778 (30: 49, 6.0x) ITU
H1-M52 HG04060 (32: 57, 3.9x) ITU
H1-M52 HG04002 (33: 42, 4.2x) ITU
37|12
H1-M52 NA21099 (35: 6, 2.7x) GIH
39|41
H1-M52 NA21113 (36: 12, 4.0x) GIH
H-M370 NA21087 (38: 19, 3.2x) GIH
100|4
44|26
H1-M52 HG03738 (42: 9, 5.0x) STU
46|7
H1-M52 HG03755 (43: 4, 4.4x) STU
H1, H2
H1-M52 HG04235 (45: 41, 5.1x) ITU
66|4
H1-M52 HG04211 (47: 62, 5.3x) ITU
50|6
H1-M52 HG03815 (48: 35, 5.0x) BEB
H-M69 HG04185 (49: 42, 3.1x) BEB
H1-M52 HG03941 (52: 49, 5.3x) BEB
H1-M52 HG03716 (53: 62, 5.5x) ITU
64|2
H1-M52 HG03600 (54: 47, 4.9x) BEB
99|68
60|6
H1-M52 HG03777 (55: 34, 4.9x) ITU
59|10
127|H-M2713|24
62|3
H1-M52 HG03009 (56: 23, 3.2x) BEB
58|7
H1-M52 NA20901 (57: 22, 3.7x) GIH
H1-M52 HG04006 (61: 52, 5.5x) STU
72|33
H1-M52 NA21107 (68: 17, 4.7x) GIH
H1-M52 NA20871 (69: 7, 3.0x) GIH
71|10
H1-M52 NA21126 (70: 8, 5.2x) GIH
H1-M52 HG04176 (73: 47, 5.2x) BEB
H-M370 HG04146 (74: 37, 3.2x) BEB
98|5
H1-M52 HG04061 (76: 43, 3.4x) ITU
95|1
H1-M52 HG03985 (77: 56, 3.2x) STU
H1-M52 HG04161 (79: 52, 5.6x) BEB
H1-M52 HG03646 (81: 29, 4.2x) STU
94|5
H1-M52 HG03644 (82: 42, 6.2x) STU
97|1
H1-M52 HG03745 (83: 41, 4.4x) STU
166|GHIJK-M3658|15
H1-M52 HG03800 (85: 30, 5.2x) BEB
91|1
H1-M52 HG03950 (89: 39, 5.2x) STU
H1-M52 NA20863 (90: 33, 4.9x) GIH
H1-M52 HG04206 (92: 33, 6.3x) ITU
H1-M52 NA21116 (96: 40, 5.0x) GIH
H0
106|126
H0-M2713 HG03837 (104: 13, 4.8x) STU
H0-M2713 HG03680 (105: 15, 5.3x) STU
114|26
109|43
H0-M2713 HG03792 (107: 61, 4.3x) ITU
111|3
113|46
H0-M2713 HG04033 (108: 46, 3.0x) STU
H0-M2713 HG03848 (110: 78, 3.5x) STU
120|91
H0-M2713 HG03965 (112: 104, 6.0x) ITU
117|12
126|126
119|174
H0-M2713 HG03872 (115: 13, 4.2x) ITU
H0-M2713 HG04238 (116: 28, 5.2x) ITU
H0-M2713 HG03870 (118: 16, 3.4x) ITU
125|56
H0-M2713 HG02684 (121: 175, 4.8x) PJL
124|189
H0-M2713 HG03697 (122: 14, 5.2x) STU
H0-M2713 HG03890 (123: 19, 4.1x) STU
F
168|F-M89|147
149|15
G2-P303 HG02687 (129: 86, 5.4x) PJL
G2-CTS417 NA19747 (130: 43, 3.0x) MXL
148|6
G2-L30 HG02224 (131: 15, 2.0x) IBS
136|38
G2-CTS5990 NA12777 (132: 42, 5.5x) CEU
147|1
138|11
135|2
G2-Z3428 HG03577 (133: 17, 5.2x) MSL
G2-Z3428 HG01512 (134: 30, 3.3x) IBS
151|7
G2-Z724 NA18549 (137: 57, 4.8x) CHB
G
141|2
G2-CTS4803 NA19670 (139: 25, 4.1x) MXL
143|8
G2-CTS4803 NA20536 (140: 26, 3.8x) TSI
153|27
145|49
G2-L42 NA12546 (142: 28, 5.2x) CEU
G2-CTS9737 HG01097 (144: 25, 3.1x) PUR
157|2
G2-L30 NA12286 (150: 99, 4.5x) CEU
G2-M406 NA20767 (152: 112, 5.6x) TSI
159|29
156|66
G2-L166 HG02681 (154: 68, 4.6x) PJL
161|G2-L156|55
G2-L91 HG01311 (155: 51, 3.7x) PUR
G2-L31 HG03708 (158: 138, 5.5x) PJL
165|G-M201|184
G2-M283 HG02789 (160: 175, 5.4x) PJL
F*
164|84
G1-L833 NA20870 (162: 56, 2.7x) GIH
G1-L833 NA20858 (163: 53, 2.5x) GIH
F-M213 HG02040 (167: 198, 3.2x) KHV
0
100
200
300
400
500
600
700
Number of SNVs


c
K
IJK
J2
J1
I2
I1
J2-M102 HG01619 (11: 94, 4.6x) IBS
J2-M102 NA20588 (12: 118, 6.3x) TSI
J2-L283 NA20763 (13: 41, 4.7x) TSI
J2-Z1297 HG00160 (14: 26, 3.0x) GBR
J2-Z631 NA11930 (15: 34, 5.0x) CEU
J2-Z631 HG01509 (16: 15, 3.6x) IBS
J2-Z631 NA20811 (17: 14, 3.4x) TSI
J2-Z631 NA20525 (19: 18, 4.4x) TSI
J2-M241 HG04210 (24: 60, 6.1x) STU
J2-M241 NA20905 (25: 53, 3.3x) GIH
J2-M241 HG02690 (26: 17, 5.1x) PJL
J2-M241 NA20885 (27: 13, 4.1x) GIH
J2-M241 HG03696 (30: 23, 4.7x) STU
J2-M241 HG03785 (31: 22, 4.5x) ITU
J2-M241 HG04239 (32: 17, 5.7x) ITU
J2-M241 HG03969 (35: 45, 3.5x) ITU
J2-M241 HG03908 (36: 57, 5.2x) BEB
J2-M241 HG03976 (37: 24, 4.0x) ITU
J2-M241 HG03998 (38: 45, 2.3x) STU
J2-M241 HG03006 (40: 24, 3.7x) BEB
J2-L559 HG01589 (50: 135, 3.4x) PJL
J2-L559 HG02774 (51: 141, 5.6x) PJL
J2-M68 HG02651 (52: 22, 5.2x) PJL
J2-M68 HG03854 (53: 17, 4.5x) STU
J2-M68 HG03693 (54: 34, 5.3x) STU
J2-F4326 HG03018 (57: 18, 3.4x) PJL
J2-F4326 HG02724 (58: 10, 5.2x) PJL
J2-M322 HG02490 (60: 122, 5.7x) PJL
J2-M319 HG01781 (62: 91, 5.6x) IBS
J2-M319 HG01991 (63: 108, 8.0x) PEL
J2-M67 NA20534 (65: 55, 4.7x) TSI
J2-F4326 HG02236 (66: 37, 3.4x) IBS
J2-M67 HG01402 (68: 45, 5.4x) PUR
J2-M260 NA20815 (70: 42, 3.9x) TSI
J2-M260 NA20765 (71: 45, 4.7x) TSI
J2-M260 NA20513 (73: 15, 3.4x) TSI
J2-M260 NA20827 (74: 16, 3.1x) TSI
J2-L210 NA20801 (77: 11, 5.0x) TSI
J2-L210 NA20787 (78: 14, 4.9x) TSI
J2-L210 NA20521 (80: 25, 3.5x) TSI
J2-F4326 HG03615 (86: 59, 3.8x) BEB
J2-F4326 HG03021 (87: 9, 3.2x) PJL
J2-F4326 HG03660 (88: 5, 4.2x) PJL
J2-F4326 NA20884 (90: 12, 2.7x) GIH
J2-F4326 HG02733 (93: 24, 5.6x) PJL
J2-F4326 HG04107 (94: 20, 3.2x) STU
J2-F4326 HG02493 (96: 54, 6.0x) PJL
J2-F4326 HG03012 (97: 36, 3.1x) BEB
J2-F4326 NA21090 (99: 19, 3.7x) GIH
J2-F4326 HG04003 (100: 24, 5.3x) STU
J2-F4326 HG03905 (103: 47, 2.8x) BEB
J2-L25 HG03237 (107: 54, 5.6x) PJL
J2-L25 NA21130 (108: 60, 4.6x) GIH
J2-L25 HG03851 (110: 44, 3.2x) STU
J2-L25 HG01672 (112: 34, 2.8x) IBS
J2-L397 NA20778 (114: 22, 5.1x) TSI
J2-L397 NA20809 (115: 20, 5.0x) TSI
J2-L397 HG01756 (117: 38, 9.4x) IBS
J2-L397 NA20805 (118: 25, 5.0x) TSI
J2-L397 HG01412 (119: 26, 5.7x) PUR
J2-F4168 HG01164 (124: 113, 3.9x) PUR
J1-L321 HG01253 (132: 12, 3.0x) CLM
J1-L321 HG01256 (133: 17, 3.8x) CLM
J1-CTS5368 HG03767 (135: 129, 3.5x) PJL
J1-L816 HG00181 (136: 68, 4.7x) FIN
J1-P58 NA20543 (137: 30, 5.6x) TSI
J1-P58 NA12282 (138: 31, 5.0x) CEU
J1-P58 HG01686 (139: 41, 3.9x) IBS
J1-P58 HG01069 (142: 9, 4.3x) PUR
J1-P58 HG01935 (143: 7, 4.4x) PEL
J1-P58 HG01130 (145: 8, 5.2x) CLM
J1-P58 HG01437 (146: 10, 2.9x) CLM
J1-P58 HG01571 (150: 33, 2.7x) PEL
J1-P58 HG02253 (151: 8, 4.0x) PEL
J1-P58 HG01431 (152: 16, 6.0x) CLM
J1-CTS15 HG01494 (157: 106, 3.0x) CLM
I2-L40 HG01988 (162: 95, 2.4x) ACB
I2-CTS1977 NA20351 (163: 25, 4.4x) ASW
I2-CTS8433 NA12748 (164: 40, 4.7x) CEU
I2-S434 NA19984 (165: 25, 6.0x) ASW
I2-S434 NA12003 (166: 23, 3.6x) CEU
I2-Z78 HG00136 (168: 22, 2.7x) GBR
I2-L369 NA12413 (172: 3, 0.4x) CEU
I2-L126 HG02536 (173: 54, 3.3x) ACB
I2-CTS9183 HG01527 (175: 30, 4.7x) IBS
I2-CTS9183 HG02470 (176: 29, 4.5x) ACB
I2-CTS9183 NA11932 (177: 28, 4.4x) CEU
I2-L621 HG00360 (183: 133, 4.6x) FIN
I2-L158 HG01610 (184: 55, 3.3x) IBS
I2-L158 HG01167 (185: 64, 8.7x) PUR
I2-L160 HG01197 (186: 7, 3.8x) PUR
I2-L160 HG01101 (187: 5, 6.4x) PUR
I2-L160 HG01344 (189: 50, 4.4x) CLM
I1-L118 HG00117 (195: 29, 4.0x) GBR
I1-L118 HG00140 (196: 28, 6.1x) GBR
I1-L118 NA11992 (197: 37, 4.7x) CEU
I1-L118 NA07051 (200: 35, 2.9x) CEU
I1-S296 HG00310 (201: 26, 3.4x) FIN
I1-Z2541 NA11919 (202: 41, 7.6x) CEU
I1-S246 NA12750 (205: 21, 5.8x) CEU
I1-S246 HG00252 (206: 23, 6.2x) GBR
I1-S246 NA11881 (208: 27, 5.2x) CEU
I1-S337 HG00159 (210: 20, 3.0x) GBR
I1-S337 NA06994 (211: 26, 3.8x) CEU
I1-S337 HG00101 (212: 11, 4.5x) GBR
I1-Z140 HG02274 (215: 32, 4.7x) PEL
I1-F2642 HG02433 (216: 11, 2.4x) ACB
I1-F2642 NA20511 (217: 29, 4.8x) TSI
I1-F2642 NA19982 (218: 25, 5.8x) ASW
I1-Z140 NA20762 (221: 30, 4.3x) TSI
I1-L338 HG02420 (223: 22, 5.2x) ACB
I1-Z140 HG02445 (224: 13, 2.4x) ACB
I1-CTS10937 HG00234 (226: 20, 4.0x) GBR
I1-CTS6364 NA19723 (233: 35, 3.5x) MXL
I1-L300 HG00189 (234: 20, 2.8x) FIN
I1-L258 HG00321 (236: 6, 3.4x) FIN
I1-L258 HG00345 (237: 13, 6.2x) FIN
I1-L258 HG00372 (239: 15, 2.1x) FIN
I1-L258 HG00186 (241: 10, 2.7x) FIN
I1-L258 HG00308 (243: 5, 2.2x) FIN
I1-L258 HG00325 (244: 8, 3.5x) FIN
I1-L813 NA11829 (247: 37, 6.5x) CEU
I1-L118 HG00190 (248: 11, 2.4x) FIN
I1-S243 NA11893 (253: 21, 4.3x) CEU
0
100
200
300
400
500
Number of SNVs


d
P
O3
O2
O1
N
K2a1*
T
L
K
Number of SNVs
Number of SNVs

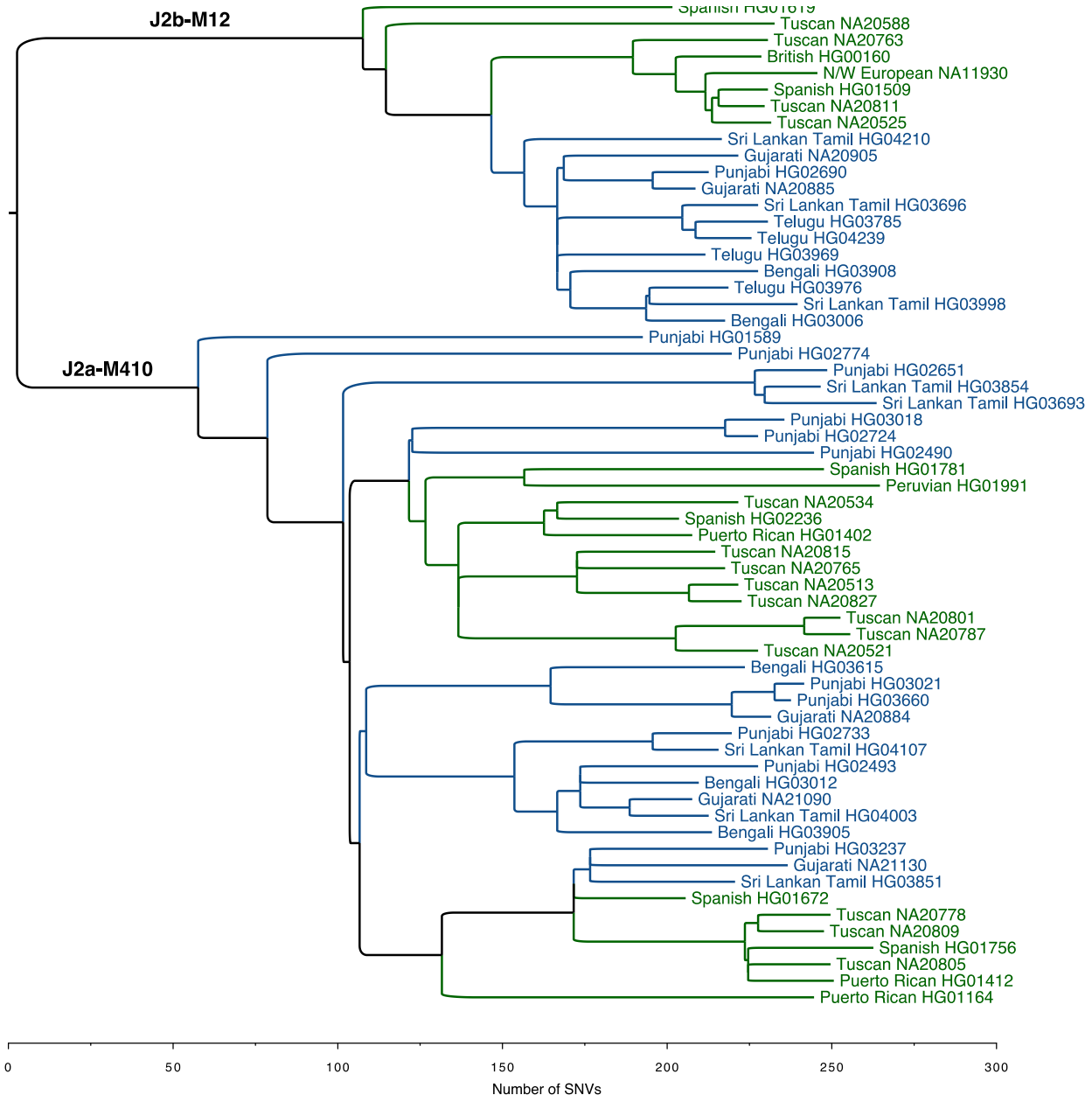

e
P
R2
R1a
R1b
Q
Number of SNVs
Number of SNVs



Supplementary Figure 15

Phylogeny updates.

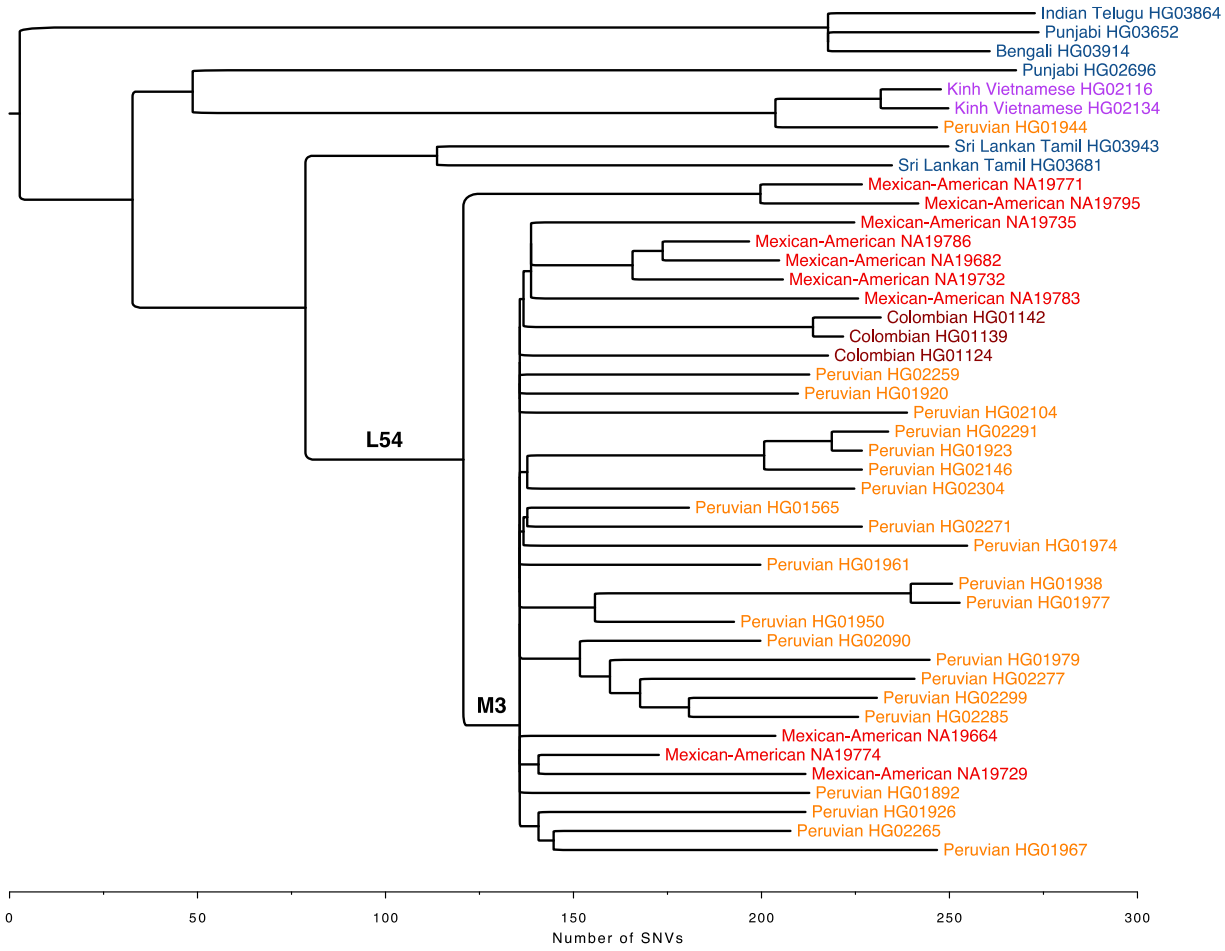
Orange branches indicate new structures identified in this study, and SNPs labeled with orange backgrounds define new lineages or redefine extant lineages. Blue branches indicate haplogroups sequenced fully for the first time, and gray lineages were not sampled. Labels with black backgrounds mark megahaplogroups F, K, and P. Boxed text indicates primary geographic distributions of the major Y-chromosome haplogroups[3–5] (**Supplementary Note 4.4**).



Supplementary Figure 16

Haplogroup J2 tree.

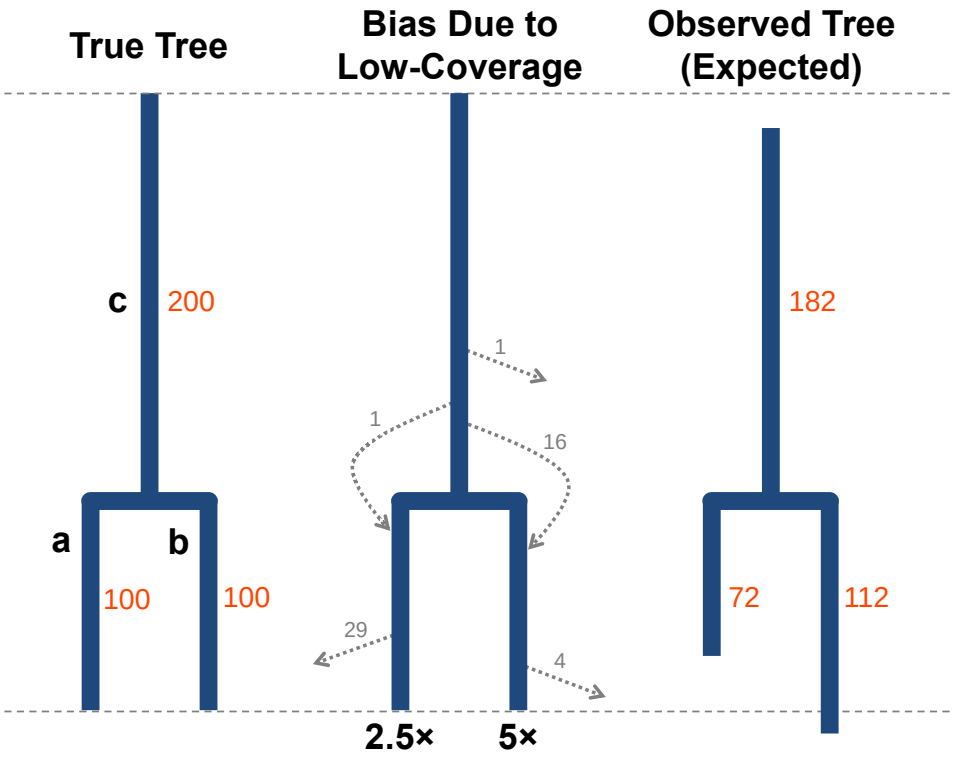
Haplogroup J2-M172 is distributed roughly evenly between South-Asian lineages (blue) and those carried by Europeans and Admixed Americans (green), but sublineages cluster by superpopulation. The distance scale is number of SNVs (**Supplementary Note 4.4.9**).



Supplementary Figure 17

Haplogroup Q tree.

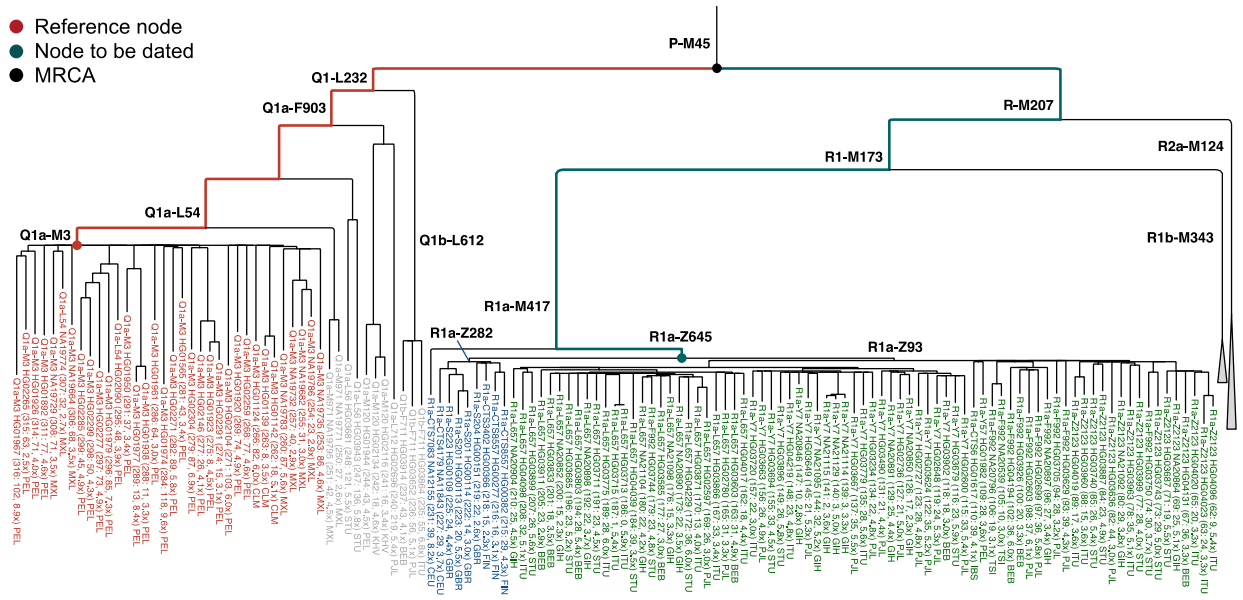
Red tones indicate Admixed-American individuals, purple indicates Vietnamese samples, and blue indicates South-Asians. Q-M3 is a star-like phylogeny. HG01944 is Peruvian, but this individual's paternal lineage is an East-Asian, rather than Native-American, branch of hgQ (**Supplementary Note 4.4.14**).



Supplementary Figure 18

Branch-length bias due to low coverage.

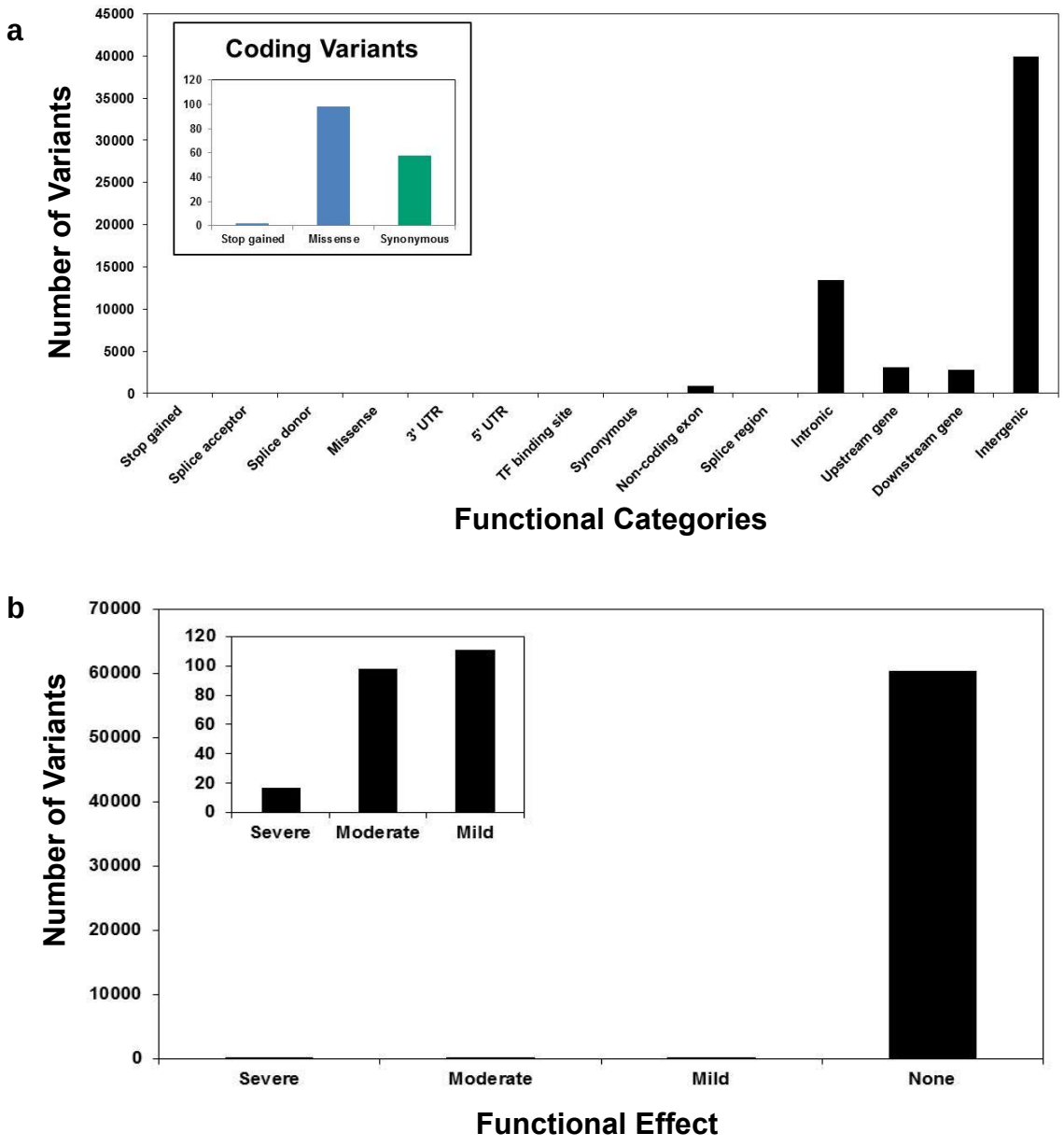
(**Left**) In this example, we assume a true tree with the lengths of each branch (number of SNVs) indicated in orange. (**Center**) With 2.5× sequencing coverage of branch *a* and 5× coverage of branch *b*, we expect 29 unobserved branch-*a* singletons and 4 unobserved branch-*b* singletons (gray arrows). In addition, we expect that 16 branch-*c* doubletons will appear to be branch-*b* singletons, whereas just 1 will appear to be a branch-*a* singleton. (**Right**) The net effect is a negative bias for the observed length of branch *a* but a positive bias for branch *b*. (**Supplementary Note 4.5.1**)



Supplementary Figure 19

Traversing high-coverage internal branches to estimate split times.

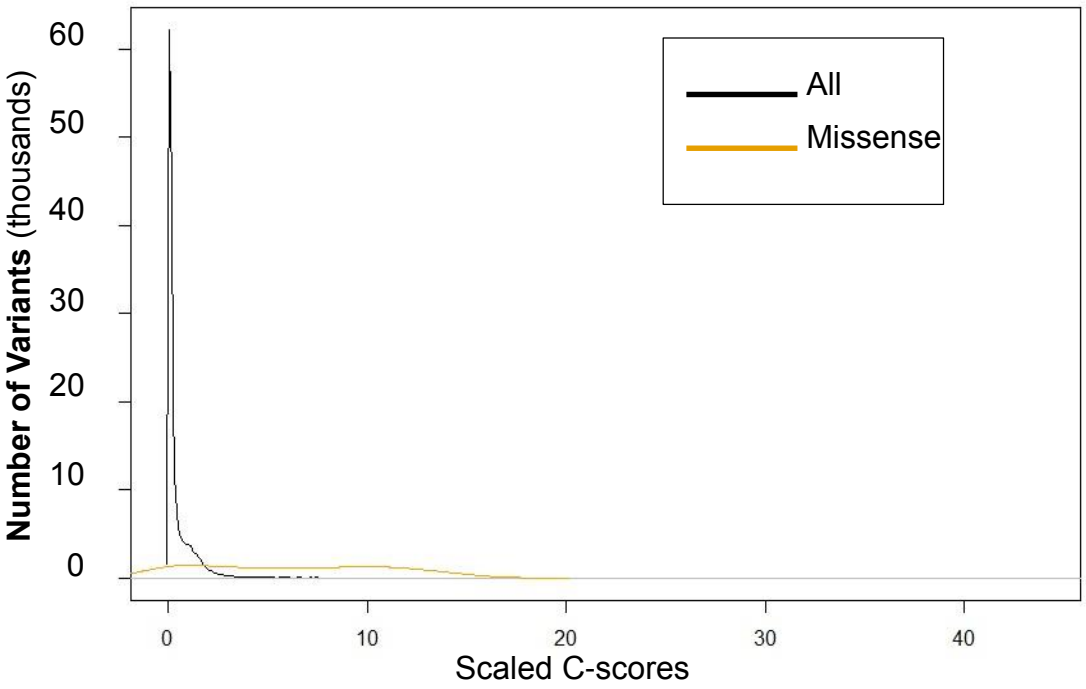
This figure illustrates our procedure to estimate the age of an internal node by measuring its height, relative to a reference node with a known age. In this example, we date the split between the European and Asian branches of R1a: R1a-Z282 (blue) and R1a-Z93 (green), respectively. The node to be dated, R1a-Z645, is indicated with a blue-green point. To estimate its age, we start from the reference node (red point; Q1a-M3) and traverse internal branches (red lines), counting the number of SNPs between the reference node and the common ancestor of the red and blue-green nodes (black point; P-M45). We then traverse the branches from the common ancestor down to the node of interest (blue-green lines; R1a-Z645) and convert the path-length difference to units of time, yielding the age difference between the reference node and the node of interest (**Supplementary Note 4.5.3**).



Supplementary Figure 20

Distributions of functional annotations.

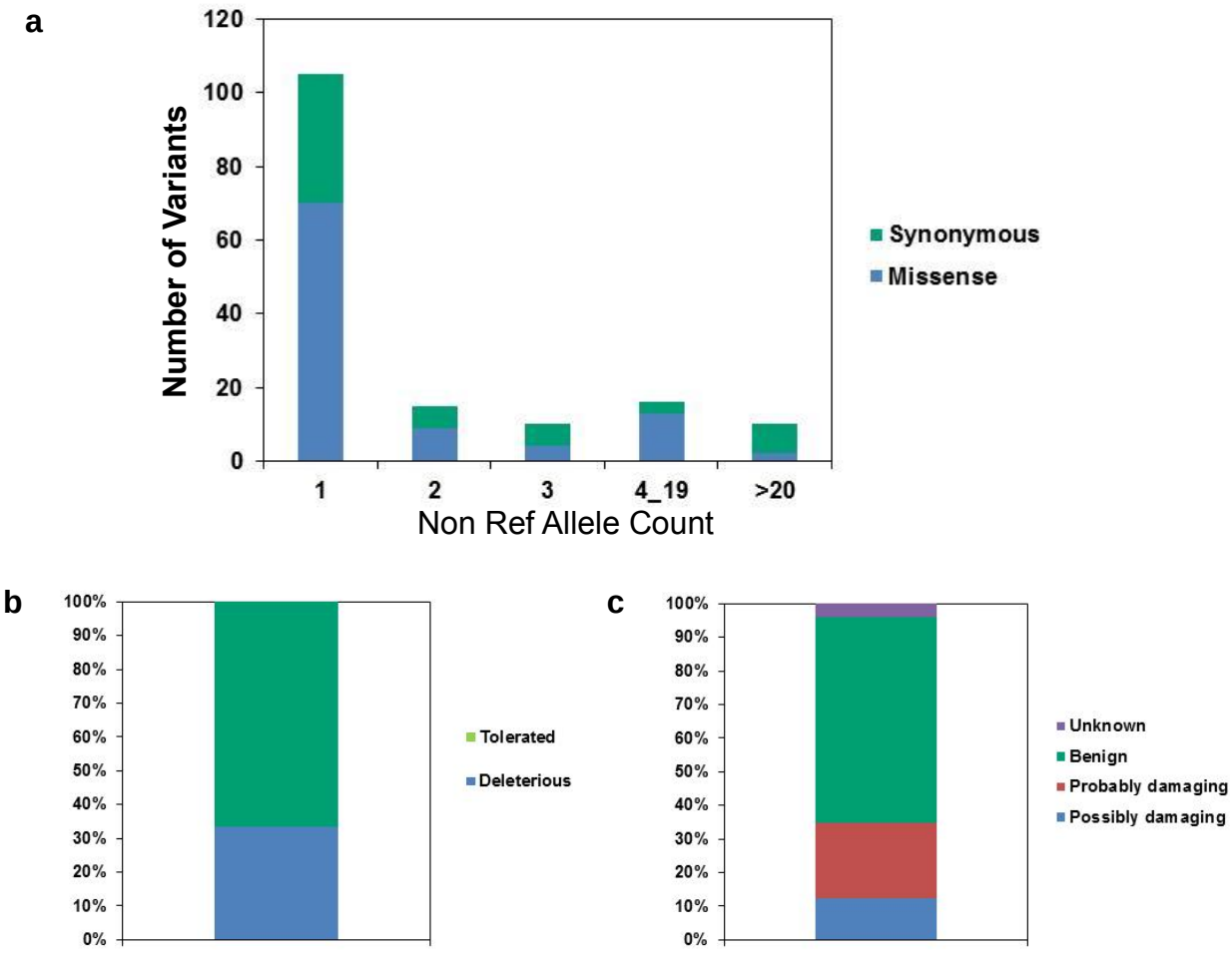
(**a**) Functional categories for all variants and for coding variant (inset). (**b**) Functional effects. (**Supplementary Note 5**)



Supplementary Figure 21

Distribution of CADD-based scores (C-scores)[6].

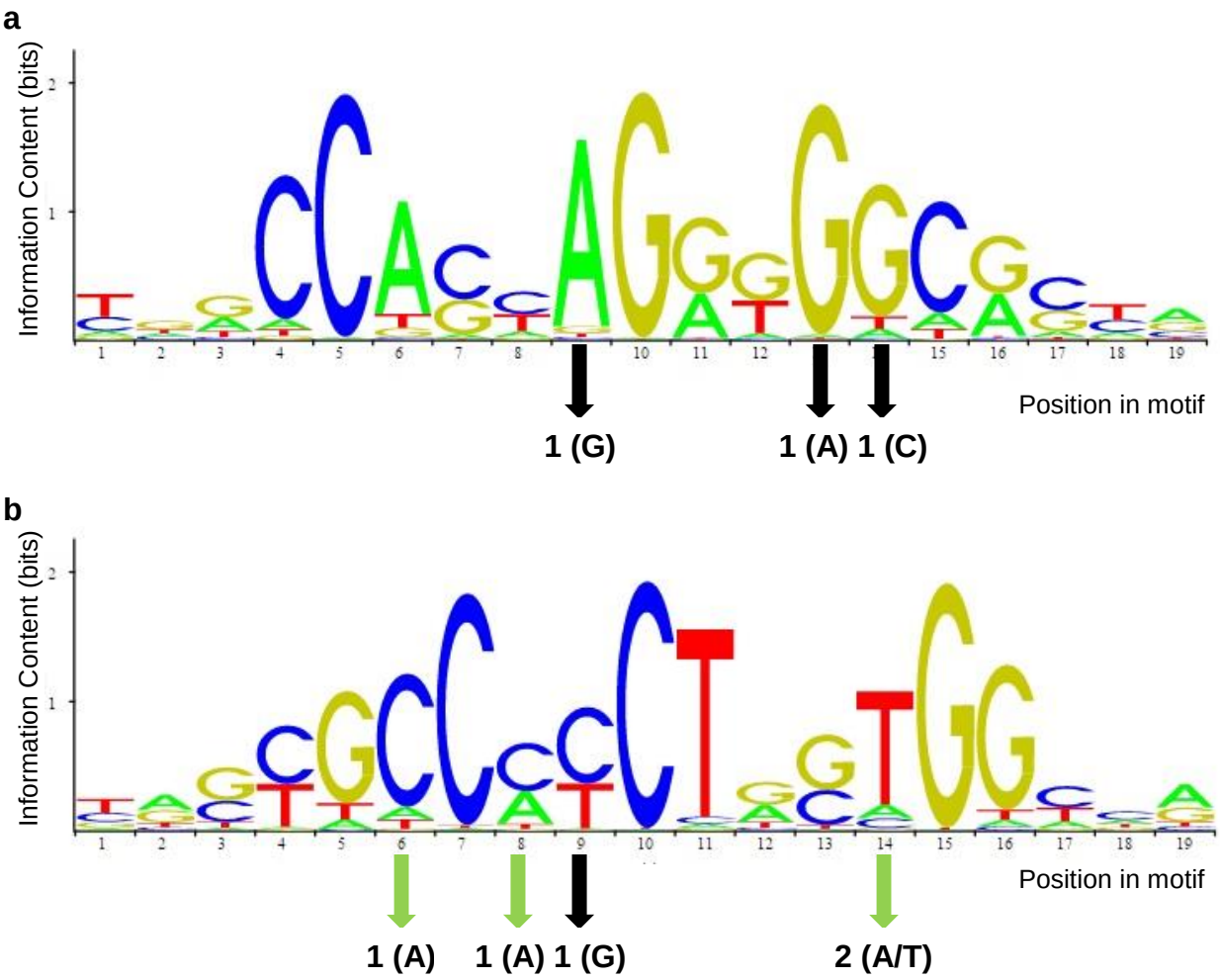
Distribution for all SNVs (black) and for missense SNVs (orange) (**Supplementary Note 5**).



Supplementary Figure 22

Coding-variant annotations.

(**a**) Distribution of synonymous and missense variants, stratified by non-reference allele count. (**b**) Percentages of missense variants predicted to be tolerated or deleterious, according to SIFT[7]. (**c**) Percentages of missense variants predicted to be benign, possibly damaging, or probably damaging, according to PolyPhen[8]. (**Supplementary Note 5**)

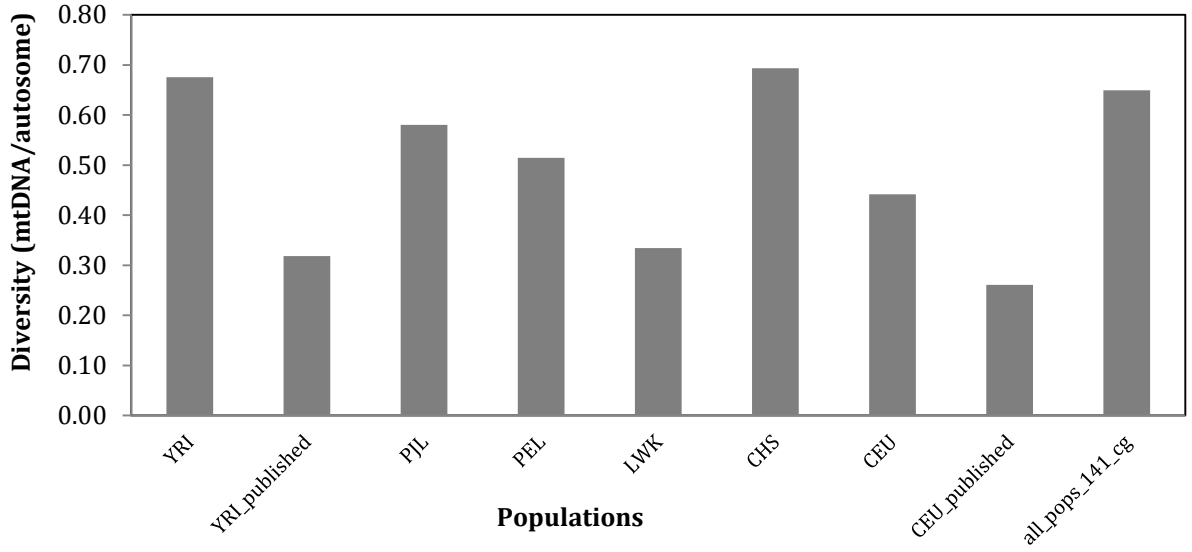


Supplementary Figure 23

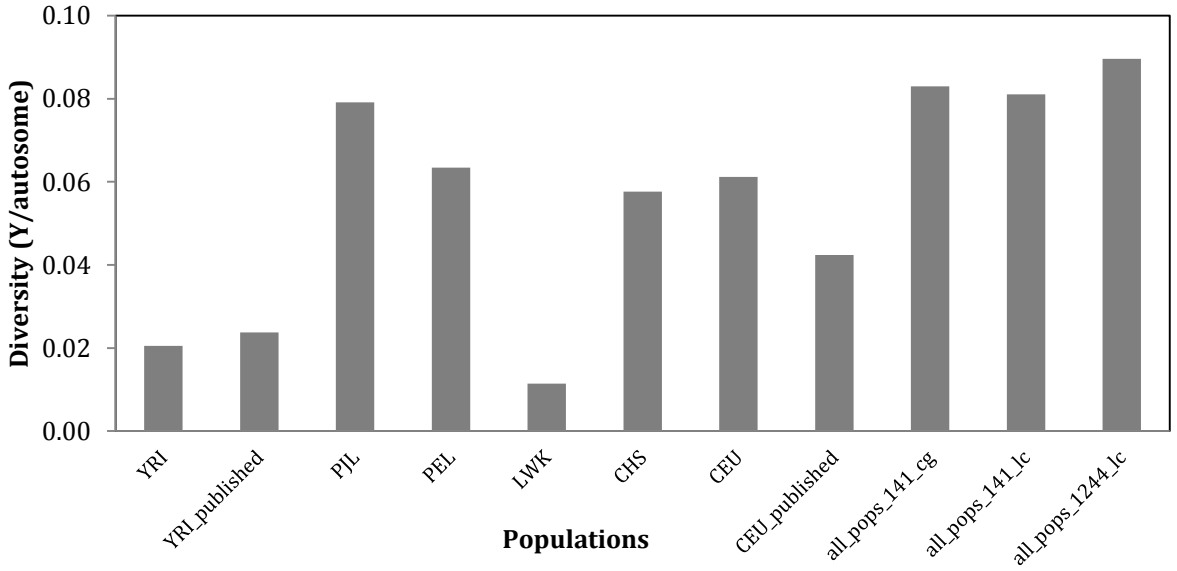
Variants affecting CTCF binding.

These sequence logos represent the CTCF binding matrix model for the (**a**) forward and (**b**) reverse strands. The *x*-axes indicate motif base positions, and the height of stacked letters indicates the total information content for a given position, with 0 corresponding to no base preference and 2 indicating a single base used. The relative sizes of the individual letters represent their relative occurrences within the motif, and black (green) arrows represent motif-destroying (motif-enhancing) variants. The number of individuals observed with the non-reference allele is indicated below each arrow. The CTCF binding motif MA0139.1 was obtained from Jaspar[9] (**Supplementary Note 5**).

a



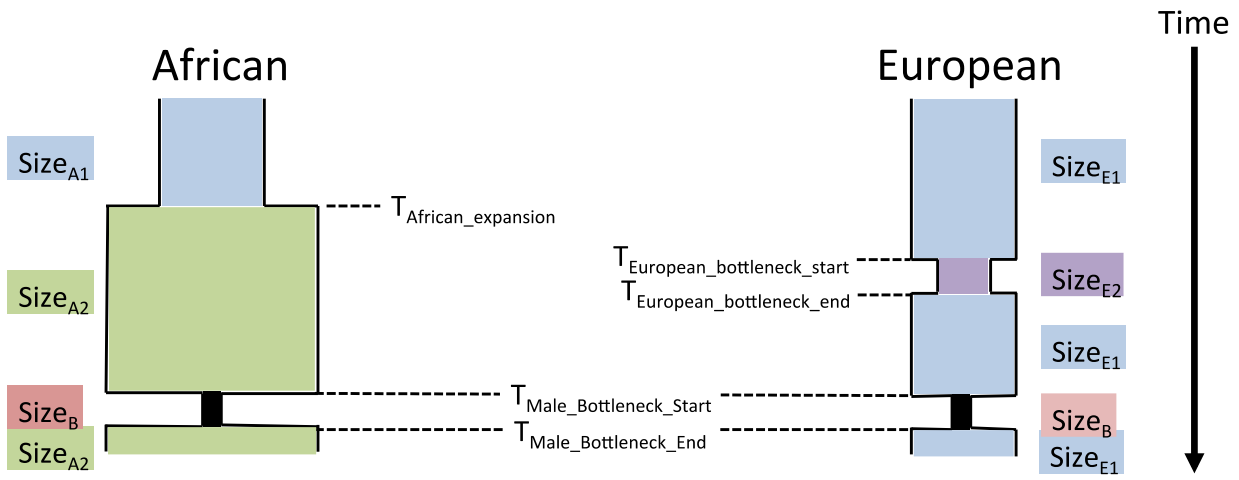
b



Supplementary Figure 24

Observed relative diversity.

Observed levels of diversity relative to those on the autosomes, stratified by population, for (**a**) mtDNA and (**b**) the Y chromosome. Values are corrected for the mutation rate specific to each genomic region. YRI_published and CEU_published, values from Sayres et al.[10]; all_pops_141_cg, values computed from the 141 male Complete Genomics sequences that overlapped with the 1000 Genomes Project phase 3 low-coverage sample; all_pops_141_lc, based on the low-coverage sequences of the same 141 individuals; all_pops_1244_lc, includes all 1,244 low-coverage sequences (**Supplementary Note 7**).

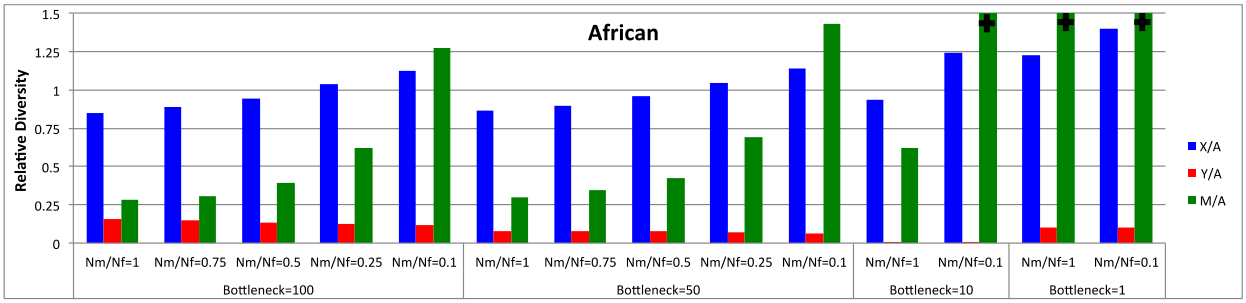


Supplementary Figure 25

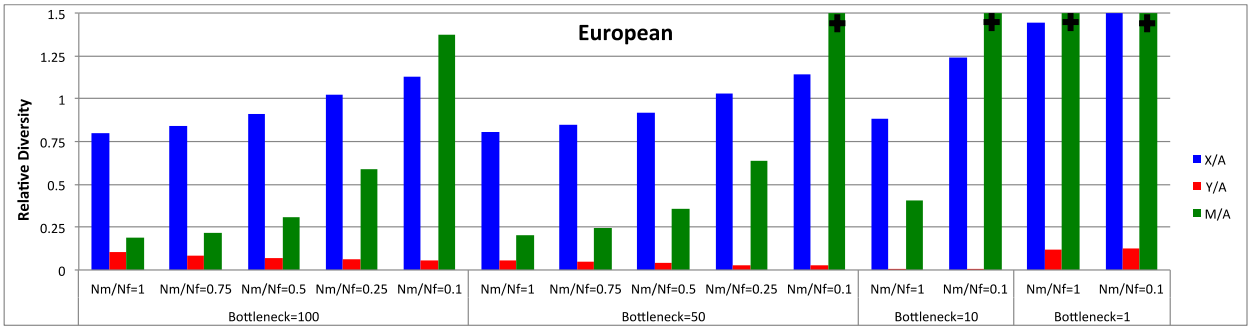
African and European demographic models.

Non-red features represent a previously developed model[11,12]. In red, $Size_B$, refers to the size of the bottleneck specific to the male lineage (**Supplementary Note 7.1**).

a



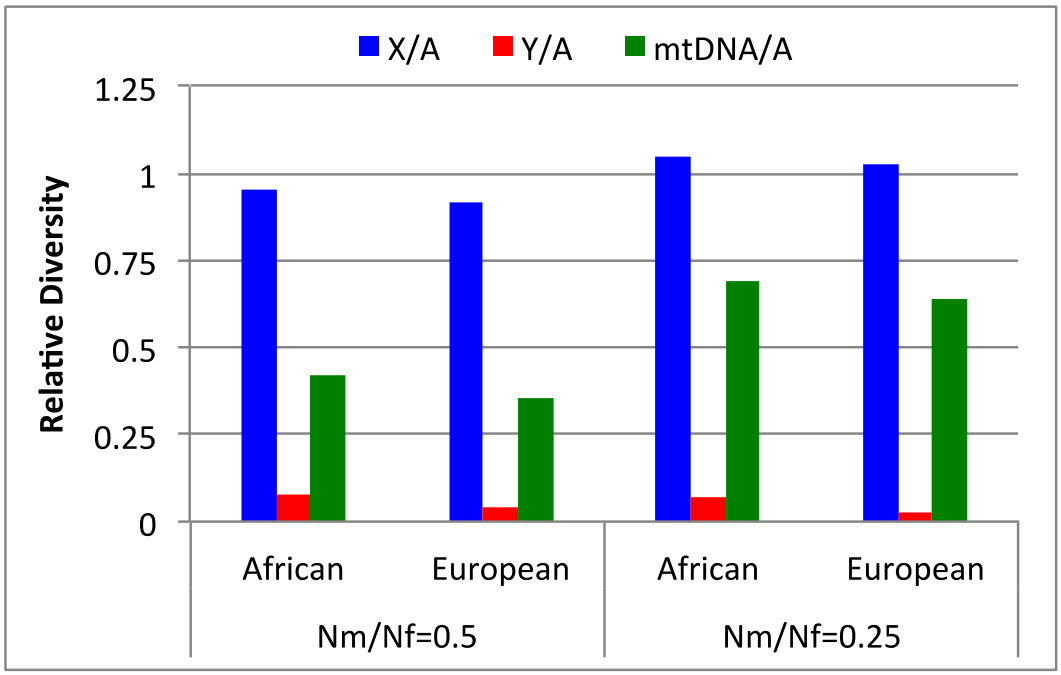
b



Supplementary Figure 26

Relative diversities for a range of demographic models.

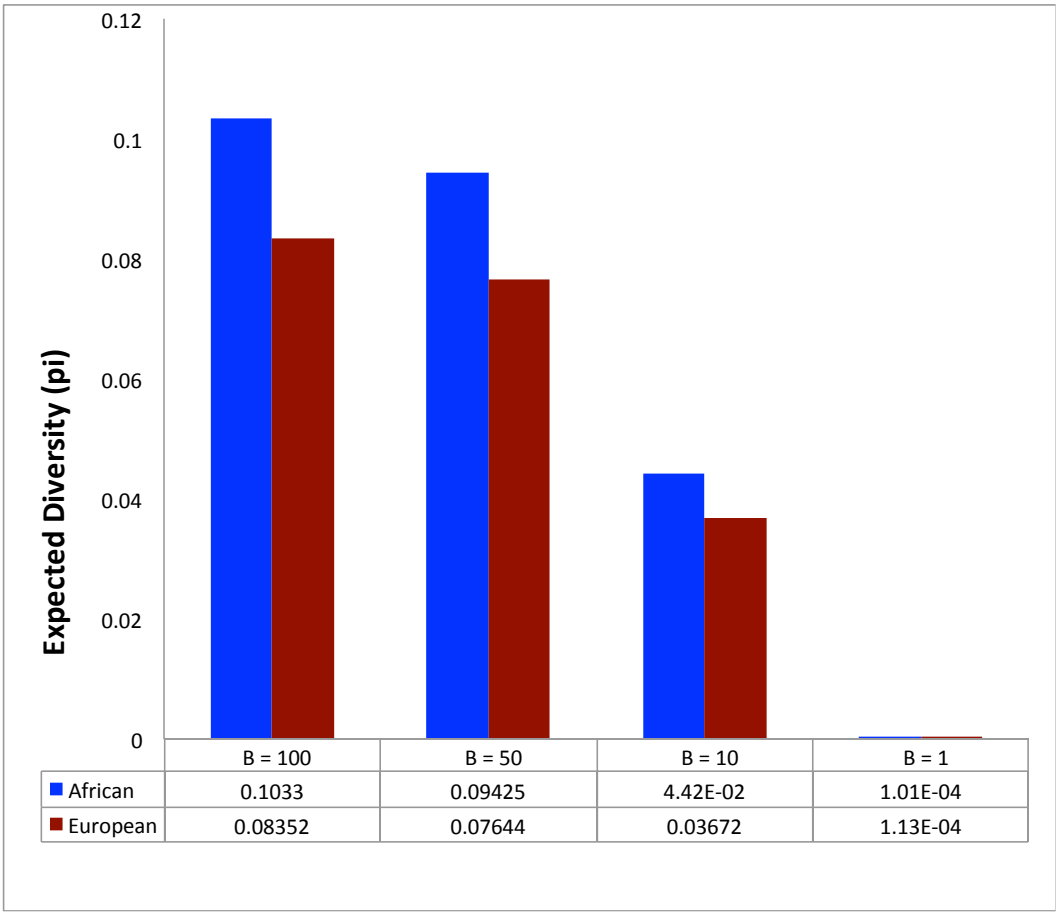
Patterns of relative diversity for X versus autosome (blue), Y versus autosome (red), and mtDNA versus autosome (green) under (**a**) African demographic history and (**b**) European demographic history. Values are corrected for differing mutation rates. Black plus signs indicate values far in excess of the plot range (**Supplementary Note 7.4**).



Supplementary Figure 27

Relative diversities for the two best-fitting demographic models.

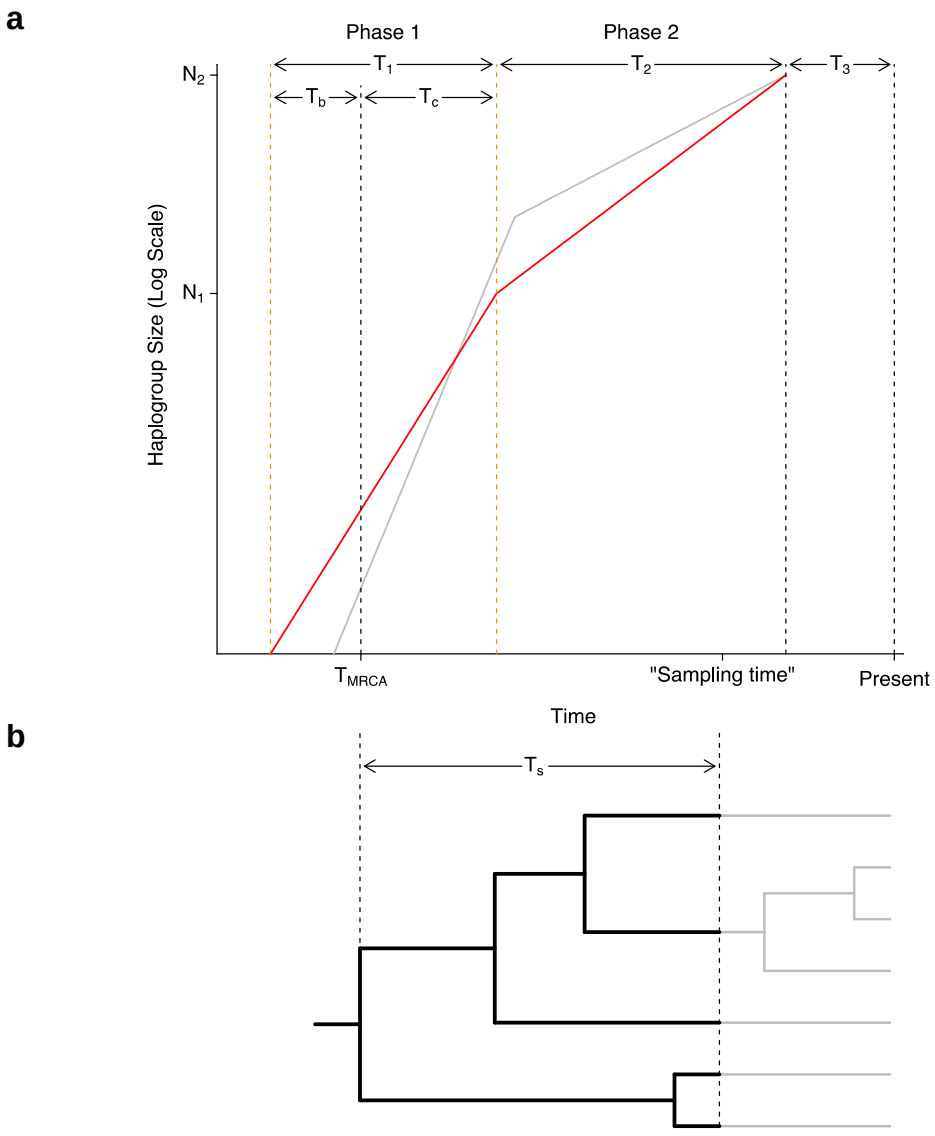
Levels of diversity, relative to the autosomes and corrected for differing mutation rates, for the X (blue), Y (red), and mtDNA (green). The models assume N_m/N_f equal to 0.5 or 0.25, a bottleneck of 50 males starting 150 generations ago and lasting for 50 generations, and 30 years per generation (**Supplementary Note 7.4**).



Supplementary Figure 28

Expected autosomal diversity versus male bottleneck size.

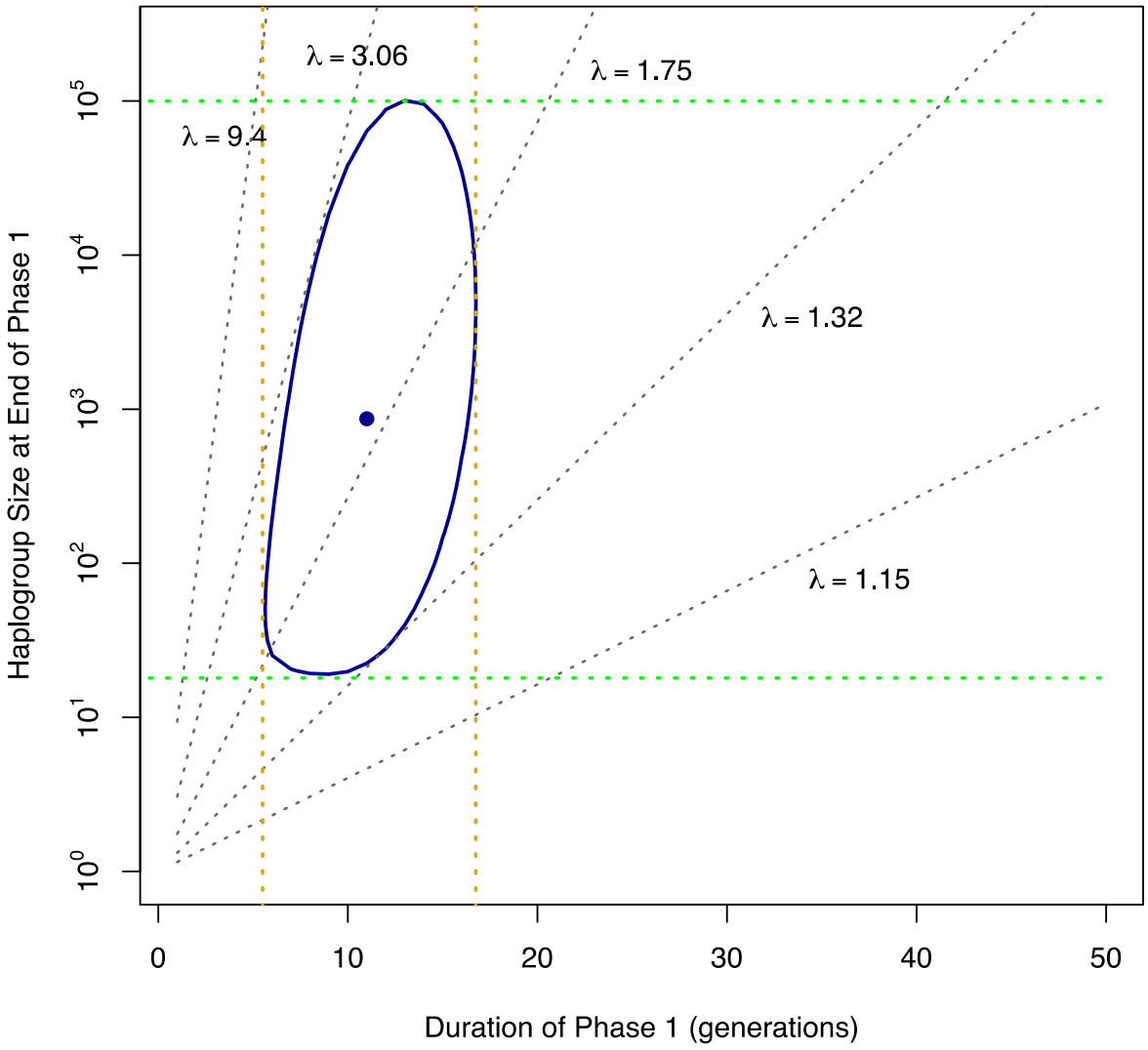
Expected autosomal diversity amongst Africans (blue) and Europeans (red) for N_m bottleneck sizes (B) of 100, 50, 10, and 1 (**Supplementary Note 7.4**).



Supplementary Figure 29

Two-phase exponential-growth model for an observed subtree with known T_{MRCA}.

(**a**) The piecewise-linear red curve represents an example growth trajectory with haplogroup population sizes N_1 and N_2 at the ends of growth phases lasting T_1 and T_2 generations, respectively. We partition T_1 into the portion before coalescence of the subtree (T_b) and the portion after (T_c), and we set the second phase to conclude T_3 generations prior to the present. (**b**) We consider up to 10 "sampling" times, each of which defines the coalescence time (T_s) of a pruned version of the observed subtree (black branches). Fixed constraints include T_3, N_2, and T_{MRCA}, the coalescence time measured from the present. For a given T_s, free parameters include T_1 and N_1, as we can estimate T_b, T_c, and T_2 from the other parameters. The solid gray curve indicates one of many other possible growth trajectories that satisfy the constraints (**Supplementary Note 8.1**).

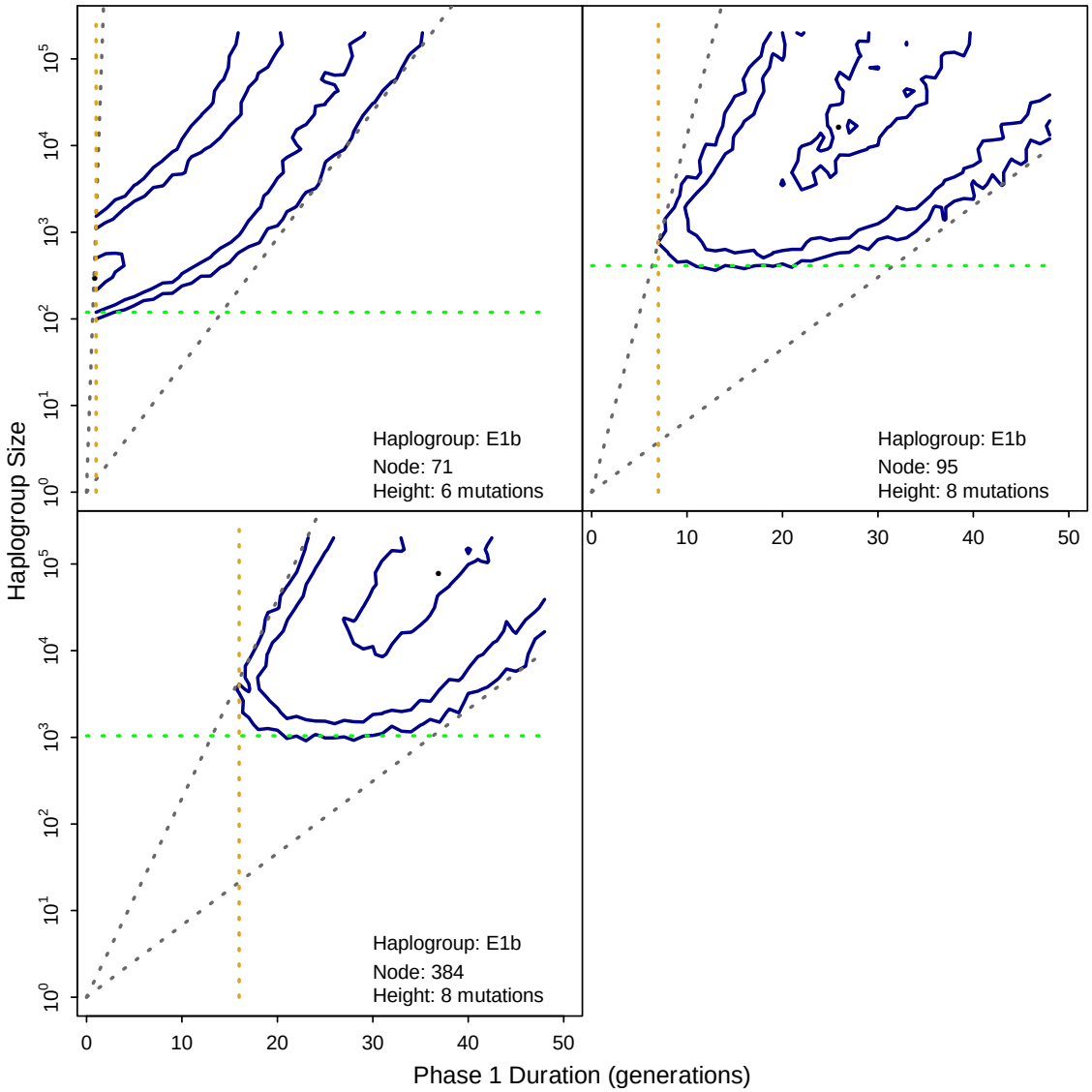


Supplementary Figure 30

Inference of phase-1 growth rate and duration.

Schematic contour plot of the joint likelihood of T_1 and N_1, given the frequency spectrum. The blue point indicates the maximum-likelihood combination, and the blue curve indicates the acceptance region. Dotted gold vertical lines and green horizontal lines indicate the marginal confidence intervals for T_1 and N_1, respectively. Each dotted gray line originates at (0, 1) and represents the exponential growth trajectory for a specified number of sons per male per generation (λ). Trajectories that are tangent to the curve defining the acceptance region correspond to the most extreme values consistent with the data (**Supplementary Note 8.1**).

a

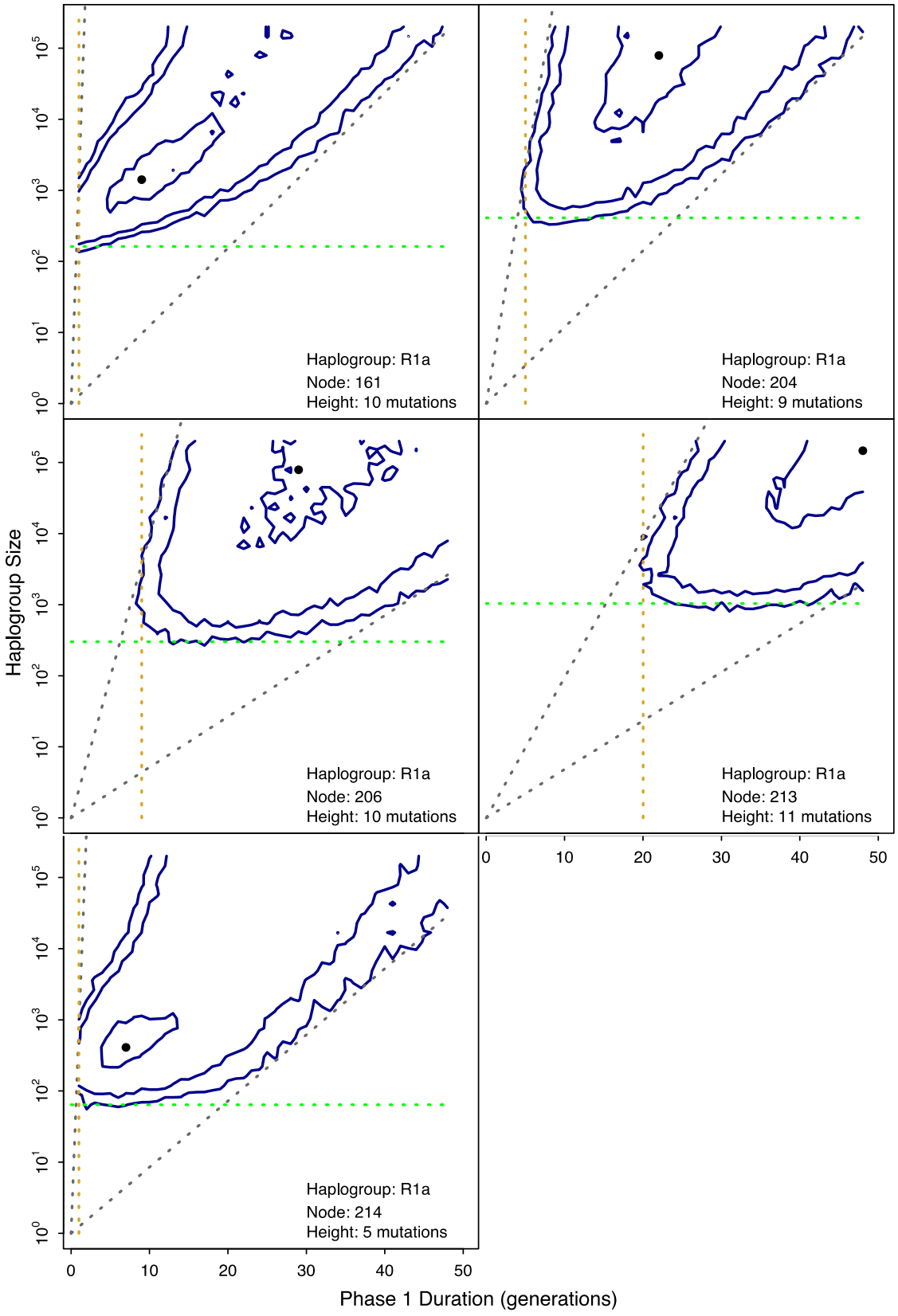


Supplementary Figure 31

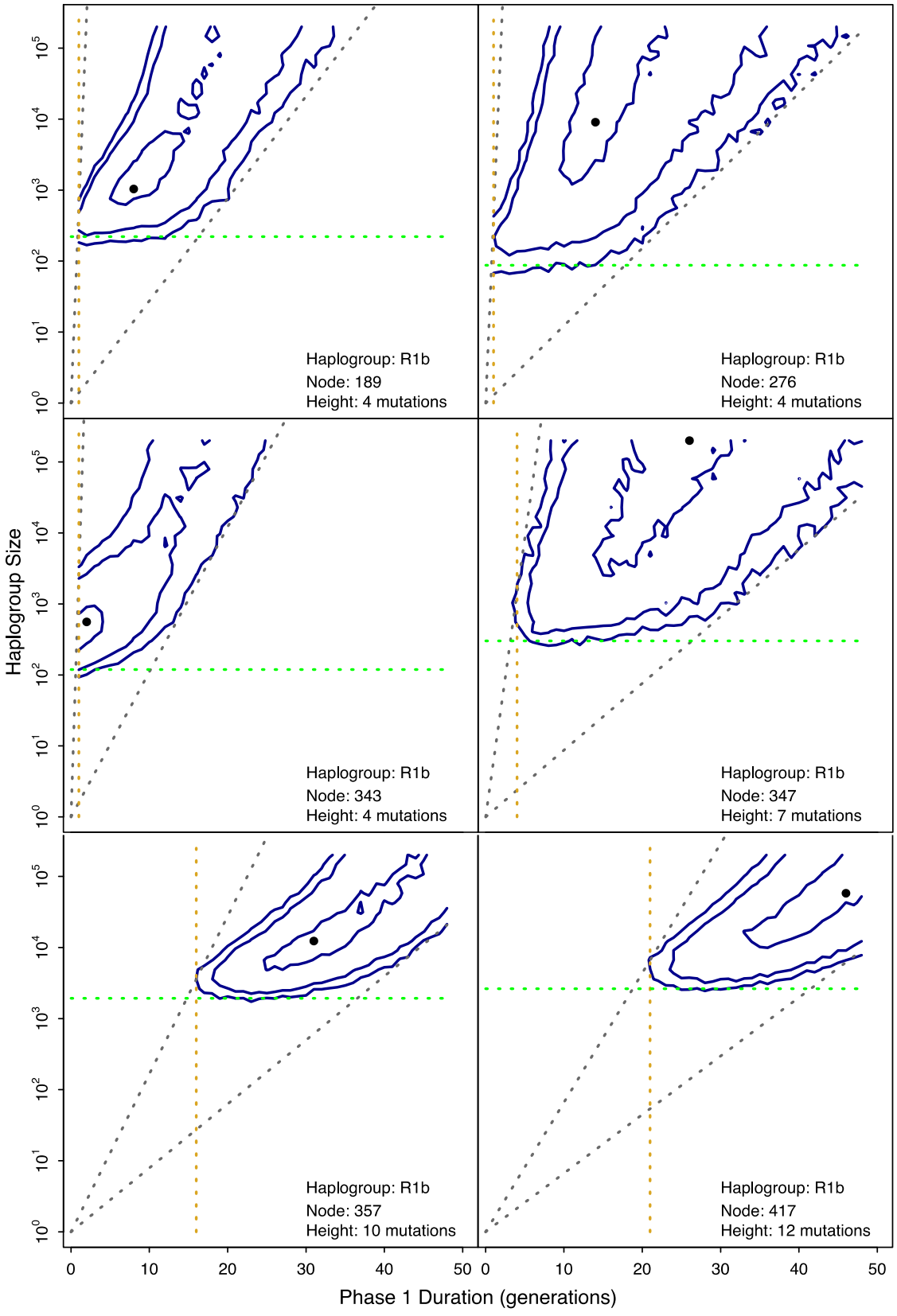
Likelihood contours for phase-1 growth parameters.

(Figure continues on the following two pages.) For each grid point, we computed the joint likelihood of the duration of the first phase of growth (T_1) and the number of individuals possessing the lineage at its conclusion (N_1), given the observed site frequency spectrum. We evaluated T_1 values ranging from 1 to 48 generations and N_1 values in a geometric progression from ~10 to 200,000 individuals. Contour lines correspond to 95%, 90%, and 50% confidence regions. Within each haplogroup, we studied a number of nodes of interest, and for each node, we include a plot for one of ten possible subtree sampling heights. See **Supplementary Figure 30** for further details and **Supplementary Data File 8b** for the full set of plots. (**a**) African haplogroup E1b. (**b**) South-Asian haplogroup R1a. (**c**) European haplogroup R1b. (**Supplementary Note 8.2**)

b


Haplogroup Size
Phase 1 Duration (generations)
Haplogroup: R1a
Node: 161
Height: 10 mutations
Haplogroup: R1a
Node: 204
Height: 9 mutations
Haplogroup: R1a
Node: 206
Height: 10 mutations
Haplogroup: R1a
Node: 213
Height: 11 mutations
Haplogroup: R1a
Node: 214
Height: 5 mutations

c


Haplogroup Size
Phase 1 Duration (generations)
Haplogroup: R1b
Node: 189
Height: 4 mutations
Haplogroup: R1b
Node: 276
Height: 4 mutations
Haplogroup: R1b
Node: 343
Height: 4 mutations
Haplogroup: R1b
Node: 347
Height: 7 mutations
Haplogroup: R1b
Node: 357
Height: 10 mutations
Haplogroup: R1b
Node: 417
Height: 12 mutations

Supplementary Tables

Supplementary Table 1

Populations of the 1000 Genomes Project. Superpopulations, codes, and descriptions of each of the 26 populations sampled (**Supplementary Note 1.1**).

Superpopulation Code	Population Code	Description
EAS	CHB	Han Chinese in Beijing, China
EAS	JPT	Japanese in Tokyo, Japan
EAS	CHS	Southern Han Chinese
EAS	CDX	Chinese Dai in Xishuangbanna, China
EAS	KHV	Kinh in Ho Chi Minh City, Vietnam
EUR	CEU	Utah Residents with Northern and Western European Ancestry
EUR	TSI	Toscani in Italy
EUR	FIN	Finnish in Finland
EUR	GBR	British in England and Scotland
EUR	IBS	Iberian in Spain
AFR	YRI	Yoruba in Ibadan, Nigeria
AFR	LWK	Luhya in Webuye, Kenya
AFR	GWD	Gambian in Western Division, The Gambia
AFR	MSL	Mende in Sierra Leone
AFR	ESN	Esan in Nigeria
AFR	ASW	Americans of African Ancestry in Southwest USA
AFR	ACB	African Caribbeans in Barbados
AMR	MXL	Mexican Ancestry from Los Angeles, USA
AMR	PUR	Puerto Ricans from Puerto Rico
AMR	CLM	Colombians from Medellin, Colombia
AMR	PEL	Peruvians from Lima, Peru
SAS	GIH	Gujarati Indian from Houston, Texas
SAS	PJL	Punjabi from Lahore, Pakistan
SAS	BEB	Bengali from Bangladesh
SAS	STU	Sri Lankan Tamil from the UK
SAS	ITU	Indian Telugu from the UK

Superpopulation Code	Description
AFR	African
AMR	Admixed American
EAS	East Asian
EUR	European
SAS	South Asian

Supplementary Table 2

SNV alternative-allele concordance. Comparison between 143 high-coverage Complete Genomics sequences and low-coverage sequences of the same individuals. Results are stratified by genotype frequency (**Supplementary Note 1.3.3**).

Alt Allele Count	Concordant	Discordant	Concordance
1	4,431	7	0.998
2	1,244	21	0.983
3	1,004	8	0.992
4	288	6	0.980
5	423	7	0.984
> 5	81,218	299	0.996
All	**88,608**	**348**	**0.996**

Supplementary Table 3

SNV derived-allele concordance. Comparison between 143 high-coverage Complete Genomics sequences and low-coverage sequences of the same individuals. Results are stratified by genotype frequency (**Supplementary Note 1.3.3**).

Derived Allele Count	Concordant	Discordant	Concordance
1	4,050	26	0.994
2	1,095	10	0.991
3	724	2	0.997
4	328	8	0.976
5	374	1	0.997
> 5	124,832	116	0.999
All	**131,403**	**163**	**0.999**

Supplementary Table 4

STR Mendelian consistency. Discrepancies between father and son STR genotypes for loci with major-allele frequencies below 95% (**Supplementary Note 3.1.2**).

Base-Pair Difference	Motif Length	Number of Errors	% of Errors
-5	5	1	1.4
-4	4	2	2.8
-3	3	1	1.4
-2	2	28	38.9
2	2	33	45.8
3	3	1	1.4
4	2	5	6.9
6	2	1	1.4

Supplementary Table 5

STR genotype concordance. Concordance between HipSTR and capillary STR genotypes for PowerPlex Y23 loci. In contrast to the SNP, MNP, indel, and CNV analyses, which used GRCh37 coordinates, we list STR start and stop positions according to GRCh38 (**Supplementary Note 3.1.2**).

Locus	Start	Stop	Correct Calls	Total Calls	% Correct
DYS481	8,558,337	8,558,402	138	158	87.3
DYS570	6,993,190	6,993,257	138	145	95.2
DYS576	7,185,318	7,185,385	175	182	96.2
DYS438	12,825,899	12,825,948	294	304	96.7
DYS392	20,471,987	20,472,025	185	191	96.9
DYS456	4,402,919	4,402,978	255	263	97.0
DYS458	7,999,839	7,999,902	197	203	97.0
DYS19	9,684,380	9,684,443	174	179	97.2
DYS549	19,358,338	19,358,389	318	326	97.5
DYS391	11,982,077	11,982,132	363	370	98.1
DYS389I	12,500,448	12,500,495	388	394	98.5
DYS439	12,403,473	12,403,564	356	361	98.6
DYS437	12,346,267	12,346,326	335	339	98.8
DYS533	16,281,349	16,281,396	357	360	99.2
DYS643	15,314,132	15,314,186	282	283	99.6
Total			**3,955**	**4,058**	**97.5**

Supplementary Table 6

STR genotype discrepancies. Base-pair differences and motif lengths for discrepancies between HipSTR and capillary electrophoresis genotypes for the PowerPlex Y23 loci (**Supplementary Note 3.1.2**).

Base-Pair Difference	Motif Length	Number of Errors	% of Errors
-24	4	1	1.0
-8	4	2	1.9
-5	5	2	1.9
-4	4	18	17.5
-3	3	3	2.9
3	3	19	18.5
4	4	40	38.8
5	5	9	8.7
6	3	3	2.9
8	4	5	4.9
9	3	1	1.0

Supplementary Table 7

Karyotyping results. X- and Y-chromosome paints on metaphase spreads were used to determine the sex chromosome copy numbers in cell lines. For each cell line, we list the observed karyotypes and the percentage of metaphase spreads in which each karyotype was observed. Each cell-line identifier with a "GM" prefix corresponds to the "NA"-prefixed sample ID of the same number (**Supplementary Note 2.2.3**).

Cell line	Population	Karyotype	Percentage
HG02372	CDX	47,XXY	100%
GM20754	TSI	46,XY	55%
		45,X	45%
GM12413	CEU	46,XY	73%
		45,X	27%
HG00246	GBR	46,XY	38%
		45,X	62%
HG02053	ACB	46,XY	50%
		45,X	50%
HG03615	BEB	46,XY	60%
		45,X	40%
HG01967	PEL	46,XY	60%
		45,X	40%
HG00251	GBR	46,XY	100%
HG01182	PUR	46,XY	100%
HG01187	PUR	46,XY	100%
HG01506	IBS	46,XY	100%
HG00634	CHS	46,XY	100%
HG00650	CHS	46,XY	100%

Supplementary Table 8

Cross-tabulation of major haplogroups by population. (**a**) Counts of individuals from each population observed to possess each haplogroup. Colors indicate superpopulations: African (AFR) (red), Admixed American (AMR) (orange), European (EUR) (green), South Asian (SAS) (blue), and East Asian (EAS) (purple). (**b**) Row percentages: distribution of haplogroups within each population. (**c**) Column percentages: distribution of populations within haplogroup clusters. (**Supplementary Note 4.1**)

a

	1244	A0 2	A1a 3	B 3	B2 4	E1a 20	E1b 298	E2 5	D2 20	C1 4	C3 10	C5 17	F* 1	G1 2	G2 17	H 7	H0 12	H1 41	H2 2	I1 31	I2 17	J1 15	J2 61	L1 27	T 8	K2* 1	N 7	N1 25	O1 19	O2 75	O3 114	Q1a 42	Q1b 3	R1a 86	R1b 216	R2 29	
GWD	55	1	3	1		11	38	1													3	1	3		2												GWD
MSL	42			2		2	37								1																						MSL
YRI	52					4	47	1																													YRI
ESN	53						53																														ESN
LWK	45				3		40	2																													LWK
ACB	47	1				2	31													3	3													7		ACB
ASW	26						17	1												1	2													5		ASW
PUR	54				1	1	14								2						3	1	3		2									27		PUR
CLM	43						8														1	6			1							3		24		CLM
MXL	33						5								2					1					1							10		14		MXL
PEL	41						2													1		3	1									24		1	9		PEL
IBS	54						2								2						2	1	6		1									1	39		IBS
TSI	53						4								2					2		1	15		3									2	24		TSI
GBR	46																			6	1		1											4	34		GBR
CEU	49														3					8	4	1	1											2	30		CEU
FIN	38																			9	1	1					23							3	1		FIN
PJL	48											1			4		1	2					12	2								1	1	17		6	PJL
GIH	58											12		2		1		11					5	5										16	1	5	GIH
BEB	42										1	2				5		10					5	2						2	2		1	9		3	BEB
STU	55											1				1	6	9					8	10								2		15		3	STU
ITU	60											1					5	9	2				4	8		1	1						1	16		12	ITU
KHV	46										1		1														1	1	2	16	22	2					KHV
CDX	50										2																		10	22	16						CDX
CHS	52										1																		2	8	41						CHS
CHB	46										3				1												5	1	5	7	23				1		CHB
JPT	56								20	4	2																			20	10						JPT
		A0	A1a	B	B2	E1a	E1b	E2	D2	C1	C3	C5	F*	G1	G2	H	H0	H1	H2	I1	I2	J1	J2	L1	T	K2*	N	N1	O1	O2	O3	Q1a	Q1b	R1a	R1b	R2	

b

	A0	A1a	B	B2	E1a	E1b	E2	D2	C1	C3	C5	F*	G1	G2	H	H0	H1	H2	I1	I2	J1	J2	L1	T	K2*	N	N1	O1	O2	O3	Q1a	Q1b	R1a	R1b	R2	
GWD	2	5	2		20	69	2																													GWD
MSL			5		5	88								2																						MSL
YRI					8	90	2																													YRI
ESN						100																														ESN
LWK				7		89	4																													LWK
ACB	2				4	66													6	6														15		ACB
ASW						65	4												4	8															19		ASW
PUR				2	2	26								4						6	2	6		4											50		PUR
CLM						19														2	14			2							7			56		CLM
MXL						15								6					3					3							30			42		MXL
PEL						5													2		7	2									59		2	22		PEL
IBS						4								4						4	2	11		2										2	72		IBS
TSI						8								4					4		2	28		6										4	45		TSI
GBR																			13	2		2												9	74		GBR
CEU														6					16	8	2	2												4	61		CEU
FIN																			24	3	3						61							8	3		FIN
PJL											2			8		2	4				2	25	4								2	2	35		12	PJL
GIH										21			3		2		19					9	9										28	2	9	GIH
BEB									2	5				12			24					12	5							5	5		2	21		7	BEB
STU										2				2	11		16					15	18								4		27		5	STU
ITU										2					8		15	3				7	13		2	2						2	27		20	ITU
KHV										2		2													2	2	2	4	35	48	4					KHV
CDX										4																		20	44	32						CDX
CHS										2																		4	15	79						CHS
CHB										7				2												11	2	11	15	50				2		CHB
JPT								36	7	4																			36	18						JPT
	A0	A1a	B	B2	E1a	E1b	E2	D2	C1	C3	C5	F*	G1	G2	H	H0	H1	H2	I1	I2	J1	J2	L1	T	K2*	N	N1	O1	O2	O3	Q1a	Q1b	R1a	R1b	R2	

c

	A0	A1a	B	B2	E1a	E1b	E2	D2	C1	C3	C5	F*	G1	G2	H	H0	H1	H2	I1	I2	J1	J2	L1	T	K2*	N	N1	O1	O2	O3	Q1a	Q1b	R1a	R1b	R2	
GWD	50	100	33		55	13	20																													GWD
MSL			67		10	12								6																						MSL
YRI					20	16	20																													YRI
ESN						18																														ESN
LWK				75		13	40																													LWK
ACB	50				10	10													10	18														3		ACB
ASW						6	20												3	12															2		ASW
PUR				25	5	5								12						18	7	5		25												12		PUR
CLM						3														6	40			12							7			11		CLM
MXL						2								12					3					12							24			6		MXL
PEL						1													3		20	2									57		1	4		PEL
IBS						1								12						12	7	10		12										1	18		IBS
TSI						1								12					6		7	25		38										2	11		TSI
GBR																			19	6		2												5	16		GBR
CEU														18					26	24	7	2												2	14		CEU
FIN																			29	6	7						92							3			FIN
PJL											6			24		8	5				7	20	7								2	33	20		21	PJL
GIH											71		100		14		27					8	19										19		17	GIH
BEB										10	12				71		24					8	7						3	2		33	10		10	BEB
STU											6				14	50	22					13	37								5		17		10	STU
ITU											6					42	22	100				7	30		100	14						33	19		41	ITU
KHV										10		100														14	4	11	21	19	5					KHV
CDX										20																		53	29	14						CDX
CHS										10																		11	11	36						CHS
CHB										30				6												71	4	26	9	20						CHB
JPT								100	100	20																			27	9						JPT
	A0	A1a	B	B2	E1a	E1b	E2	D2	C1	C3	C5	F*	G1	G2	H	H0	H1	H2	I1	I2	J1	J2	L1	T	K2*	N	N1	O1	O2	O3	Q1a	Q1b	R1a	R1b	R2	

Supplementary Table 9

Sex-biased admixture in the Americas. Among Admixed-American populations, the proportion of Native-American ancestry across the autosomes is significantly greater than that of the Y chromosome. Binomial *p*-values computed under a null hypothesis of no sex-bias (**Supplementary Note 4.4.14**).

	Percent Native American		
	Autosomal	**Y**	***P***
Peruvians	76	56	0.0040
Mexicans	46	30	0.049
Colombians	26	7	0.0016
Puerto Ricans	13	0	0.00054

Supplementary Table 10

Split-time estimates. Branch, haplogroup and canonical SNP or haplogroup and branch index in the corresponding subtree of **Supplementary Figure 14**; **Left Child** and **Right Child**, haplogroups of the two offspring branches, if a well-known label exists; ***T***, estimated split time, assuming the aDNA-based mutation rate estimate of Fu et al. ($\mu = 0.76 \times 10^{-9}$ per bp per year)[13]; ***T'***, estimated split time using the pedigree-based mutation rate estimate of Helgason et al. ($\mu = 0.888 \times 10^{-9}$ per bp per year)[14] (**Supplementary Note 4.5**).

Branch	**Left Child**	**Right Child**	***T*** (ky)	***T'*** (ky)
A0–T (Root)	A0-L991	A1-V168	190.4	163.0
A1-V168	A1a-M31	BT-M42	159.0	136.1
BT-M42	B-M181	CT-M168	105.8	90.5
B-M181	.	B2-M182	100.6	86.1
CT-M168	DE-M145	CF-P143	76.0	65.0
DE-M145	D2-M55	E-M96	72.7	62.2
E-M96	E1-P147	E2-M75	57.8	49.5
E1-P147	E1a-M33	E1b1-P179	56.3	48.2
E1a-M33	.	.	14.7	12.6
E1b1-P179	E1b1b1-M35	E1b1a1-M2	46.9	40.1
E1b1b1-M35	.	.	28.4	24.3
E1b1a1-M2	.	E1b1a1a1-M180	17.2	14.7
E1b1a1a1-M180	.	.	11.8	10.1
E1b.384	.	.	5.3	4.5
E1b.95	.	.	5.0	4.3
CF-P143	C-M130	F-M89	75.5	64.6
C-M130	C1+C5	C3-M217	52.5	44.9
C1+C5	C1-M8	C5-M356	51.8	44.3
GHIJK-M3658	G-M201	HIJK-M578	54.2	46.4
HIJK-M578	H-M2713	IJK-M523	54.0	46.2
H-M2713	H0	H1+H2-M69	50.9	43.6
H1+H2-M69	H1-M52	H2	43.0	36.8
H1.94	.	.	6.3	5.4
H1.66	.	.	7.3	6.2
IJK-M523	IJ-M429	K-M9	53.1	45.4
IJ-M429	I-M170	J-M304	47.6	40.7
I-M170	I1-M253	I2-M438	30.5	26.1
J-M304	J1-M267	J2-M172	35.9	30.7
J2-M172	J2a-M410	J2b-M12	33.7	28.8
K-M9	LT-P326	K2-M526	50.9	43.6
LT-P326	L-M11	T-M184	48.1	41.2
L1.323	.	.	4.4	3.8
K2-M526	K2a1-M2313	P-M45	50.8	43.5
NO-M214	N-M231	O-P186	44.7	38.3
N-M231	.	.	21.7	18.6
O-P186	O1+O2	O3-M122	35.0	30.0
O1+O2	O1a-F589	O2-M268	34.1	29.2
O2.160	.	.	4.5	3.9
O3.225	.	.	7.5	6.4
P-M45	Q1-L232	R-M207	35.0	30.0
Q1-L232	Q1a-F903	Q1b-L612	32.5	27.8
Q1a2a1-L54	Q1a2a1a1-M3	.	16.9	14.5
Q1a2a1a1-M3	.	.	15.0	12.8
R-M207	R1-M173	R2a-M124	32.9	28.2
R1-M173	R1a1a1-M417	R1b-M343	27.0	23.1
R1a1a1b-Z645	R1a1a1b1a-Z282	R1a1a1b2-Z93	5.6	4.8
R1a1a1b2-Z93	.	.	5.3	4.5
R1b1a2a1a-L11	.	.	5.9	5.0

Supplementary Table 11

Functional annotation of SNVs. Counts of variant effect types, ranked by severity (**Supplementary Note 5**).

Consequence	Effect	Rank	Count
Stop gain	Severe	1	2
Splice acceptor	Severe	2	7
Splice donor	Severe	3	8
Missense	Moderate	4	98
3’ UTR	Mild	5	74
5’ UTR	Mild	6	27
TF binding site	Mild	7	11
Synonymous	None	8	59
Non-coding exon	None	9	643
Splice region	None	10	32
Intronic	None	11	12,694
Upstream gene	None	12	3,502
Downstream gene	None	13	3,437
Intergenic	None	14	39,961
TOTAL			**60,555**

Supplementary Table 12

Allele rarity versus presence of a functional effect. Fisher’s exact test indicates an enrichment of rare variants among those with functional effects (**Supplementary Note 5**).

Allele Count	With Effect	No Effect	Total	*P*
1, 2	181	40,993	41,174	0.0001
>2	46	19,335	19,381	
Total	**227**	**60,328**	**60,555**	

Supplementary Table 13

Allele rarity versus CADD-based score (C-score) category[6]. Fisher's exact test indicates no enrichment of rare variants among those with elevated C-scores (**Supplementary Note 5**).

Allele Count	C-Scores		Total	*P*
	≥ 10	< 10		
1, 2	66	41,108	41,174	0.91
>2	30	19,351	19,381	
Total	**96**	**60,459**	**60,555**	

Supplementary Table 14

Missense SNVs. Allele rarity versus CADD[6], PolyPhen[8], and SIFT[7] annotations. *P*, Fisher's exact test *p*-value (**Supplementary Note 5**).

Allele Count	C-Scores		Total	*P*
	≥ 10	< 10		
1, 2	24	53	77	0.036
> 2	1	16	17	
Total	**30**	**64**	**94**	

Allele Count	PolyPhen		Total	*P*
	Damaging	Benign		
1, 2	31	46	77	0.099
> 2	3	14	17	
Total	**34**	**60**	**94**	

Allele Count	SIFT		Total	*P*
	Deleterious	Tolerated		
1, 2	30	47	77	0.001
> 2	0	17	17	
Total	**30**	**64**	**94**	

Supplementary Table 15

Mitochondrial heteroplasmy. Samples with four or more heteroplasmic mtDNA sites (**Hets**) prior to filtration (**Supplementary Note 6.2**).

ID	Hets
HG03644	34
HG02696	22
HG03478	14
HG03953	13
HG02442	10
HG03716	10
HG02134	6
HG01088	5
HG01161	5
HG01176	5
HG02250	5
HG00148	4
HG00536	4
HG01518	4
HG01974	4
HG02420	4
HG02433	4
HG02645	4
HG03786	4
NA19027	4
NA19376	4
NA19448	4
NA19466	4
NA19703	4

Supplementary Table 16

Mutation rates used to model expected diversity. We assumed a 30-year generation time to convert per-generation rate estimates from the literature to per-year rates (**Supplementary Note 7.3**).

Genomic region	Mutations per bp per generation	Citation
Autosomes	1.3×10^{-8}	Fu et al.[13]
X chromosome	0.97×10^{-8}	**Supplementary Note 7.3**
Y chromosome	2.3×10^{-8}	Fu et al.[13]
mtDNA	39×10^{-8}	Rebolledo-Jaramillo et al.[15]

Supplementary Table 17

Nodes with evidence for growth. Node, index in **Supplementary Figure 14**; **SNP**, Defining mutation; T_2, Number of generations from the node to the end of phase 2; **Rate**, Inferred average percentage growth; N_2, Population size (in thousands) at the end of phase 2 (**Supplementary Note 8**).

Haplogroup	**Node**	**SNP**	**Age** (ky)	$T_c + T_2$	**Rate** (%)	N_2 (10^3)
E1b	71	U290	4.7	90	15	236
E1b	95	.	5.0	100	13	315
E1b	384	.	5.3	109	13	637
H1	66	M2854	7.3	238	6	2,640
H1	94	Z5890	6.3	203	7	2,030
L1	323	.	4.4	134	12	3,030
O2b	160	.	4.5	123	11	286
O3	225	.	7.5	183	8	1,930
Q1a	319	M3	15.0	483	3	567
ßR1a	161	Y7	4.4	142	12	6,000
R1a	204	Y6	4.0	127	13	5,500
R1a	206	.	4.2	135	12	5,700
R1a	213	L657	4.5	144	12	12,500
R1a	214	Z93	5.3	172	10	13,000
R1b	189	DF27	5.5	115	12	638
R1b	276	U152	5.5	115	13	1,250
R1b	343	DF13	4.8	92	15	450
R1b	347	M529	5.1	102	14	476
R1b	357	P312	5.6	119	13	2,560
R1b	417	P311	5.9	127	12	3,000

Supplementary Table 18

Observed branching structure used in expansion analysis. Counts of the number of branches descending from each node with evidence for growth, at each sampling time (**Supplementary Note 8**).

	Tree Height (SNVs)									
Node	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
E1b, 71	25	27	27	27	27	27	27	28	30	31
E1b, 95	26	29	33	35	35	35	35	35	35	36
E1b, 384	40	46	50	54	58	58	58	60	60	60
H1, 66	.	.	.	8	9	9	9	9	10	10
H1, 94	10	10	10	10	10	10	10	10	10	10
L1, 323	.	12	14	14	14	14	15	15	15	15
O2b, 160	12	12	12	12	12	12	12	12	13	13
O3, 225	14	14	14	15	15	15	16	16	16	16
Q1a, 319	20	20	22	22	22	.	.	.	.	.
R1a, 161	.	18	19	22	22	22	22	22	22	22
R1a, 204	17	20	20	20	20	20	20	20	20	20
R1a, 206	.	16	18	21	21	21	21	21	21	21
R1a, 213	.	22	35	38	43	46	46	46	46	46
R1a, 214	.	16	16	20	22	25	33	46	59	62
R1b, 189	42	47	49	51	53	54	57	60	60	61
R1b, 276	25	26	29	32	34	36	37	38	38	38
R1b, 343	24	25	26	29	29	29	29	29	29	29
R1b, 347	24	25	26	27	28	31	31	31	31	31
R1b, 357	70	74	82	106	114	119	123	128	133	136
R1b, 417	62	71	83	87	98	122	130	136	140	146

Supplementary Table 19

Inferred parameters of phase-1 growth. N_1, Number of individuals at the end of phase 1; T_1, Duration of phase 1 (generations). Omitted values could not be calculated (**Supplementary Note 8.2**).

	Expected Sons Per Man			
Node	**Min**	**Max**	**Min N_1**	**Min T_1**
E1b, 71	1.46	.	93	1
E1b, 95	1.22	2.43	504	9
E1b, 384	1.22	1.66	1930	19
H1, 66	1.10	.	17	1
H1, 94	1.57	.	178	1
L1, 323	1.10	.	14	1
O2b, 160	1.17	.	41	1
O3, 225	1.22	.	83	1
Q1a, 319	1.40	.	199	1
R1a, 161	1.20	1.91	301	13
R1a, 204	1.34	4.45	317	6
R1a, 206	1.18	2.56	370	9
R1a, 213	1.17	1.57	1040	20
R1a, 214	1.24	.	64	1
R1b, 189	1.33	7.76	330	5
R1b, 276	1.19	.	42	1
R1b, 343	1.52	.	99	1
R1b, 347	1.19	7.71	199	3
R1b, 357	1.29	3.55	1466	15
R1b, 417	1.23	2.10	2253	18

Supplementary Note

Table of Contents

1 Single-Nucleotide Variants, Multiple-Nucleotide Variants, Indels

1.1 Single-Nucleotide Variants

G. David Poznik and Shane McCarthy

The 1000 Genomes Project Consortium sequenced 2,535 individuals from 26 populations representing five global super-populations (**Supplementary Table 1**)[16]. For each of the 1,244 males in the sample, we used SAMtools[17] to download binary sequence alignment/map (BAM) files containing reads mapping to the GRCh37 Y-chromosome reference sequence. Please see the main publication for details on upstream processing.

Confining our attention to the 10.3 Megabases (Mb) of the Y chromosome within which one can reliably call genotypes using short-read sequencing[4], we applied six distinct genotype calling methods to these data to identify putative single-nucleotide variants (SNVs). We ran SAMtools[17] at the Sanger Institute, FreeBayes[18] at Boston College, Platypus[19] and Cortex_var[20] at Oxford, and GATK Unified Genotyper[21,22] in haploid mode at Cornell and in diploid mode at Stanford.

To construct a preliminary consensus callset, we input the list of all putative sites to FreeBayes[18]. We then imposed six filters, restricting to (a) biallelic SNVs with (b) genotype quality (QUAL) greater than one; (c) filtered-read depth across all samples in the range 2000 to 6000 (1.6× to 4.8×), which represents a six-median-absolute-deviation interval centered at the median depth across sites; (d) no more than 10% of reads with mapping quality scores of zero; (e) no more than 400 samples (approximately one third of the total) with zero high-quality reads mapping to the site; and (f) no more than 200 samples whose maximum-likelihood genotype state was heterozygous.

Upon conducting a phylogenetic analysis and mapping SNVs to branches of the tree (**section 4.3**), we observed that a greater than expected proportion of sites were incompatible with the phylogeny. We found that incompatibilities were often traceable either to (a) reference genotype calls for which read data were contradictory and a no-call would have been most appropriate; or to (b) cases where the read data supported a non-reference genotype call but were not sufficient to surmount the strong prior induced by over 1,000 reference genotype calls in the sample. Concluding that the genotype calls in this preliminary consensus callset were marred by reference bias, we replaced the FreeBayes calls by the maximum-likelihood genotype state for each sample, subject to the condition that the likelihoods for reference and non-reference states differed by two log units. When the absolute difference in likelihoods was less than or equal to two log units, we assigned a no-call.

This approach yielded a genotype callset of 59,675 SNVs. We then identified additional biallelic SNVs by splitting complex sites into biallelic components using BCFtools[23]. We applied identical filters and added the remaining 880 sites to complete a final callset, numbering 60,555 SNVs.

We used this final callset to construct a tree (**section 4.2**). Then, as described in Poznik et al.[4], we leveraged the inferred phylogeny to impute the 6.3% of genotypes that were missing (**section 4.3**).

1.2 Multiple-Nucleotide Variants and Indels

Shane McCarthy, G. David Poznik, Yali Xue

We processed multiple-nucleotide variants (MNVs) and small insertions/deletions (indels) similarly to SNVs. First, we split sites into biallelic components and normalized (left-aligned) representations. Second, we applied the same filters and maximum-likelihood genotyping calling approach as described for the SNVs to yield an initial set of 2,706 biallelic indels and MNVs. Third, we imposed additional filters, excluding 1,279 sites that, according to the UCSC genome browser repeat-mask track[24], overlapped simple-repeat sequences (n = 290), SINEs (n = 367) or LINEs (n = 622). The false-discovery rate among the remaining 1,427 variants was 3.6%. Finally, we mapped the remaining sites to the phylogeny inferred from the SNVs and imputed genotypes accordingly.

1.3 Validation

Yali Xue, Yuan Chen, and Chris Tyler-Smith

1.3.1 False Discovery Rate

To measure the false discovery rate (FDR), we adapted the method described in the project's main paper[16], using high-coverage PCR-free genome sequences, which were available for 11 of the 1,244 males. We used these data to construct a "truth" set with which to test genotype calls based on the corresponding low-coverage sequences.

The number of non-reference calls per chromosome varies greatly as a function of haplogroup, as most of the Y-chromosome reference sequence is derived from a single R1b individual. Therefore, rather than basing our FDR estimate on alternative allele calls, we instead used derived alleles. For each derived genotype call in the 11 low-coverage Y chromosomes, we assessed whether or not it was supported in the corresponding high-coverage data. We report the FDR as the proportion of low-coverage genotypes that were not supported. For biallelic SNVs, the FDR was 3.9%, and for the combined set of indels and MNVs, the FDR was 3.6%. Both values meet the project target of less than 5%.

High-coverage Complete Genomics (CG) sequences were available for 143 males, so we also used these sequences to measure the SNV FDR. Using these data, we estimated an FDR of 1.6% (249/15,376).

1.3.2 False Negative Rate

We used the 143 high-coverage CG sequences to estimate the false negative rate, comparing the variable sites called in these sequences to those based on the corresponding low-coverage data. We observed 17,194 sites in the CG data. Of these, 13,360 were called in both datasets and 3,834 (22%) were not called in the low-coverage data. Most of these (3,343; 87%) appear to be false negative singletons (**Supplementary Figure 1**). The 85 common sites among them, those with more than ten instances of each allele ($10 < AC < 133$), largely overlapped with the set of sites that did not meet our filtration criteria.

In a second round of curation, we observed four individuals with high numbers of false negative singleton sites: NA12413, with 632 unobserved singletons, had by far the lowest coverage (0.4×); HG00628, with 209 unobserved singletons, had the tenth lowest coverage overall (2.2×) and carried an isolate lineage of haplogroup C3, so would be expected to have a large number of singletons; likewise, HG00559 carries an isolated O3 lineage and missed 130 singletons; and HG02090 in hgQ had 120 singletons unique to the CG data, but this would render the sample's branch length far longer than phylogenetically proximal samples of 8× or greater coverage, so we suspect false positives in the CG data for this sample. We deemed these four outliers, as the number of singleton false negatives among the other samples ranged from 2 to 67, with a median of 12, and with fewer than 25 false negative sites in 81.5% of samples (115/143). A total of 535 sites, mostly common false negative sites, had another SNV or indel within 30 base pairs (bp). Upon excluding from this calculation the four outlier individuals and these 535 sites, the overall rate of unobserved sites was 14.2% (2,208/15,568). We retained the four individuals in down-stream analyses.

1.3.3 Genotype Concordance

We measured alternative allele concordance (**Supplementary Table 2**) and derived allele concordance (**Supplementary Table 3**) between the phase 3 low-coverage callset and that of the high-coverage CG data, stratifying by allele counts in the 143 samples. Concordance overall was greater than 99%, but we observed a surprising pattern, where the concordance for singletons was greater than that for more frequent variants. This is most likely due to the fact that the calling algorithm we employed requires strong evidence to call singletons but is able to call genotypes with weaker evidence at sites for which the existence of the SNV is strongly supported by reads from other samples. In contrast, less strong evidence is required to call an allele for which another sample has strong evidence. Therefore, the existence of a high-quality variant called in one sample can drive miscalling of genotypes samples with one or two erroneous reads. We manually checked read depth at the 21 doubleton sites at which a discordancy was observed and found that, indeed, the average depth among the discordant individuals was 2.3, whereas the average depth among the confirmed carriers was 5.3.

By querying the associated BAM files, we measured genotype concordance with the high-coverage PCR-free genomes to be 95%. This low concordance is due to one of the samples having a 3-Mb *AMELY* deletion (into which some reads were mismapped in the high-coverage genome) and to three samples belonging to haplogroup R1b. The R1b individuals had very few alternative allele counts, in which case a single discordance would represent a large proportion. Upon excluding these four samples, alternative allele concordance was 97%.

The genotype concordance for the indels and MNVs, as compared with the 11 high-coverage PCR-free genomes, was 96.4% (319/331).

1.3.4 Transition-Transversion Ratio and the False-Positive Singleton Rate

G. David Poznik and Fernando L. Mendez

In **section 4.3**, we describe mapping the 60,555 SNVs to branches of the phylogeny that we had partitioned into eight components. Allowing each site to map to branches in multiple components, we observed 63,230 mutation events: 24,027 shared on internal branches of the tree and 39,203 singletons. Among SNVs mapping to internal branches of the tree, we observed a transition-

transversion ratio (ti/tv) of 1.73 (15,213 / 8,814), a value lower than expected for autosomes but consistent with two recent literature estimates for the portion of the Y-chromosome under consideration. Helgason et al. estimated this ratio to be 1.74 for de novo mutations[14], and Scozzari et al. estimated a ratio of 1.72[25], noting that this value is within the range of genome-wide estimates for de novo events[26–28]. Among singletons, we observed 24,358 transitions (s_i) and 14,845 transversions (s_v), yielding a lower ratio of 1.64, which suggests that singletons are enriched for false-positive sites, as expected.

We leveraged the ti/tv differential between singletons and shared SNVs to gain insight into the singleton false-positive rate. First, define s_1 and s_2 to represent the number of true- and false-positive singletons, respectively. Then, let γ_1 and γ_2 represent the ti/tv among true positives and false positives, respectively, and let $\alpha_i = \gamma_i / (1 + \gamma_i)$ and $\beta_i = 1 / (1 + \gamma_i)$ represent the transition and transversion proportions, respectively for true positives ($i = 1$) and false positives ($i = 2$). Decomposing the observed counts of each mutation type into contributions from true and false positives, we have:

$$\begin{bmatrix} s_i \\ s_v \end{bmatrix} = \begin{bmatrix} \alpha_1 & \alpha_2 \\ \beta_1 & \beta_2 \end{bmatrix} \begin{bmatrix} s_1 \\ s_2 \end{bmatrix}$$

As we expect the false-positive rate to be quite low on the internal branches, we take $\gamma_1 = 1.73$, and we take $\gamma_2 = 0.5$, assuming errors are uniformly allocated amongst the three alternative bases, one of which is a transition and two of which are transversions. Using these values, we can solve the system of equations to estimate the false-positive rate among singletons (s_1 / s_2): 3.9%.

2 Copy-Number Variants

2.1 Discovery and Genotyping

2.1.1 Genome STRiP

Robert Handsaker, Seva Kashin, and Steven McCarroll

We performed copy-number variation (CNV) discovery and genotyping using Genome STRiP[29], analyzing the Y chromosome in 1,234 male individuals with a pre-release version of the program (r1.04.1447). We excluded 10 samples for which the read depth across the chromosome (after normalization and correction for GC-bias) was either less than 0.8 times the expected coverage or greater than 1.2 times the expected coverage (based on genome-wide read depth), suggesting the potential presence of cell-line-specific clonal aneuploidy.

We ascertained CNVs by two methods. In the first method (discovery set 1, targeting uniquely alignable sequence), we ran the standard Genome STRiP CNV pipeline to find CNVs using read depth in uniquely alignable regions of the genome. We ran this CNV pipeline twice, once with an initial window size of 5 kb (overlapping windows by 2.5 kb) and once with an initial window size of 10 kb (overlapping windows by 5 kb). Other parameters were set to default values in each run. For both runs, the raw CNV calls were filtered using the following criteria:

Minimum call rate: 0.8
Minimum density of alignable positions: 0.3
Minimum cluster separation: 5.0 (standard deviations)

In addition, for the 5-kb run, sites were excluded if they were called only in samples with high numbers of variants (more than 45 variants per sample).

We estimated the false discovery rate (FDR) for these CNV calls using the intensity rank-sum (IRS) method[29] and probe intensity data from Affymetrix 6.0 SNP arrays that were run on the same individuals. For sites longer than 20 kb, the estimated FDR was zero. We included in the callset all sites longer than 20 kb and those shorter sites (under 20 kb) that contained at least one array probe and had an IRS estimated p-value < 0.01.

Calls from the 5-kb and 10-kb runs were merged and re-genotyped, and duplicate calls were removed using the standard Genome STRiP duplicate-removal filters. The sites were then manually reviewed and 27 calls were eliminated as being either (a) likely duplicate calls that were not detected by the default filters or (b) sites with weak evidence of copy-number variation.

The second method used for CNV ascertainment (discovery set 2) targeted regions of segmental duplication. In this method, segmental duplications annotated on the UCSC genome browser were prospectively genotyped for total copy number, using an expected reference copy number of two copies. The raw CNV calls were filtered using the following criteria, which were chosen based on manual review of the genotyped sites:

Minimum call rate: 0.8

Minimum density of alignable positions: 0.25
Minimum cluster separation: 5.0 (standard deviations)

In the VCF file containing the CNV calls, the CNVs detected from segmental duplication analysis (discovery set 2) have site identifiers that start with "GS_SD_M2". For the segmental duplication calls, the locations of the two segmental duplication intervals are encoded in the site ID. The POS/END attributes in the VCF file specify the leftmost of these two segmental duplication intervals.

The final CNV callset consisted of 97 sites called from the first method (discovery set 1) and 13 segmental duplication sites called from the second method (discovery set 2). Copy-number genotypes were encoded in the VCF file using the GT field, assuming that the reference allele has one copy for sites from discovery set 1 and that the reference allele has two copies for segmental duplication sites from discovery set 2.

2.1.2 CnvHitSeq

Haojing Shao and Lachlan Coin

We use a modified version of the cnvHitSeq algorithm to identify deletions and duplications on the Y chromosome. cnvHitSeq integrates read depth, paired-end insert-size aberration, and split reads via a hidden Markov Model (HMM)[30]. Preliminary analysis indicated substantial read-depth artifacts due to difficulty in mapping reads to repetitive regions of the Y chromosome. Rather than excluding these regions, we developed a population approach to model read-depth variation in a manner robust to the presence of repetitive regions.

We divided the Y chromosome into 500-bp windows and used SAMtools to calculate the total read depth, D_{ij}, for every sample j, in each window i. Rather than modeling read depth relative to the average genome-wide value for a particular sample, as in the standard cnvHitSeq model, we instead modeled the proportion of reads in window i for sample j, relative the proportion of reads in window i in the entire population. As in the usual cnvHitSeq model, we considered the copy number of sample j in window i, CN_{ij}, as the hidden state of our HMM. The emission probability of D_{ij} was calculated using a binomial distribution conditional on: the total number of reads sequenced for sample j, S_j; the total number of reads mapping to window i across all samples, $W_i = \Sigma_j D_{ij}$; the total number of reads sequenced for all samples, $T = \Sigma_j S_j$; and the copy number, CN_{ij}. That is,

$$\mathrm{P}[D_{ij} \mid CN_{ij}, W_i, S_j, T] \sim \mathrm{Binomial}(S_j, CN_{ij} \cdot W_i / T),$$

with S_j the number of trials and $(CN_{ij} \cdot W_i / T)$ the probability of success. We assumed CN_{ij} can take any value in {0, 1, 2} and used these emission probabilities in the cnvHitSeq HMM with a transition model as previously described[30]. We found the most likely path through the HMM with the Viterbi algorithm.

In order to detect mosaic copy number variation, we modified this algorithm to allow the copy number variant to affect just a fraction c of cells, resulting in a mean copy number of:

$$CN_{ij} = c \cdot CN_{ij*} + (1 - c) \cdot 1 = 1 + c \cdot (CN_{ij*} - 1),$$

where CN_{ij*} represents the number of copies of window i in the affected cells of sample j. The emission probability of D_{ij} under this generalization is therefore given by:

$$P[D_{ij} \mid CN_{ij}, W_i, S_j, T, c] \sim \text{Binomial}(S_j, (1 + c \cdot (CN_{ij*} - 1)) \cdot W_i / T).$$

We trained the parameter c within the expectation-maximization procedure used to fit the model. Using this approach, we identified four samples with mosaic deletions of the entire Y chromosome: HG02053 (c = 0.45), NA12413 (c = 0.94); NA20754 (c = 0.60), and HG00246 (c = 0.78).

2.1.3 Array Comparative Genomic Hybridization

Andrea Massaia, Ankit Malhotra, Charles Lee, Ruby Banerjee, Fengtang Yang, Qasim Ayub, Yali Xue, and Chris Tyler-Smith

To expand the set of SVs discovered with the Genome STRiP approach, we conducted structural variant (SV) discovery on array comparative genomic hybridization (aCGH) data. Log_2 intensity ratios were produced for 1,243 males in the 1000 Genomes Project phase 3 set, using sample NA10851 as reference. Each sample was analyzed on a separate array.

We excluded from the analysis 9 samples showing abnormally high or low median of log_2 intensity ratios (**Supplementary Figure 2**). The karyotype of these samples was also tested by 24-color fluorescence *in situ* hybridization (FISH), and six (HG00246, HG01967, HG02053, HG03615, NA12413, NA20754) appeared as mosaics (**Supplementary Table 7**).

Intensities for the 1,234 remaining samples were normalised by subtracting each sample's median log_2 intensity ratio across all probes from its measured value at each probe. We called CNVs from the same set of 2,714 probes that was employed for SV validation, as described in the **section 2.2**. Segmentation was performed using the algorithm GADA, as implemented for the R statistical computing environment[31], with arguments `estim.sigma2=TRUE`, `aAlpha=0.8` for the sparse Bayesian learning step, and with `T=5.5`, `MinSegLen=10` for the backward elimination procedure[32]. We used stringent calling criteria in order to minimise oversegmentation and false discovery rate.

The calling algorithm identified a total of 5,240 segments. To classify them into gains and losses, we employed an additive background model for the log_2 ratios, relating intensity to copy number in a manner similar to that of Conrad et al.[33]. According to this model, it is possible to estimate the copy number of the test and reference samples associated with the log_2 ratio of a segment and a given reference copy number, i.e.,

$$y = \log_2 \left(\frac{a + c}{b + c} \right),$$

where y is the segment's mean log_2 ratio, a is the intensity of the target, b the intensity of the reference, and c is the noise. Both the segment's mean log_2 ratio and the noise are estimated by the calling algorithm. For b, we assumed a reference intensity of 1, unless a segment overlapped for 80% of its length with a segmental duplication (SD), as annotated in the UCSC genome browser, in which case we assumed a reference intensity of 2.

We refined the initial segmentation by removing segments that appeared to be highly similar to the reference (± 0.2 copies, as estimated using the additive background model explained above) and segments that included the centromere. Moreover, as uneven distribution and low density of aCGH probes can result in false, overly long, calls, we manually checked segments longer than 1 Mb by inspecting the $\log_2$ intensity ratios plots. Of 305 such segments, we removed 24 as false positives. This resulted in a final callset of 1,892 segments identified in 857 samples.

To characterize the subset of aCGH-discovered variants that were not identified by the Genome STRiP analysis of sequencing data, we used the Bedtools intersect tool[34,35] with the `-v` option, comparing the two callsets. In the set of 1,892 aCGH calls, 613 did not overlap with any call in the Genome STRiP set. These 613 segments clustered into 15 CNV regions (CNVRs) across the male-specific Y (MSY).

We curated the calls that overlapped any of the 15 CNVRs, inspecting the $\log_2$ intensity ratio plots in order to distinguish between false positives and true variants. One of the 15 CNVRs was identified by one call in just one sample, and it seemed likely to be a false positive. A second, within the *AZFc* region, was supported by a large number of calls but was unreliable due to the low density of aCGH probes in that region.

The 13 remaining CNVRs include 11 variants, with two appearing in two CNVRs each. This set includes a ~3-Mb deletion on the short arm of the Y chromosome (Y:6,103,728–9,397,666), encompassing the Y-linked amelogenin gene (*AMELY*) (**Supplementary Figure 3**). This polymorphic deletion is well known in literature[36] and was also detected by Genome STRiP, although not as a single event.

Alleles were assigned based on the distribution of mean $\log_2$ intensity ratios, as estimated by the calling algorithm. Most of the variants in the aCGH-only callset (9 of 11) do not include intrachromosomal SDs, and the mean $\log_2$ intensity ratios of calls produced by the segmentation algorithm show a unimodal distribution, indicating a single duplication or deletion event.

Based on the genomic features of the respective regions, we assigned two variants to two CNVRs each: one to Y:6,115,346–6,124,150 and Y:9,196,977–9,384,475, and the other to Y:22,218957–22,369,669 and 22,419,003–22,508,011. The first pair represents paralogous regions. Y:9,196,977–9,384,475 spans the *TSPY* array[37] and includes several highly similar copies (~95.5% identity) of Y:6,115,346–6,124,150 interspersed with single-copy spacers. This peculiar structure accounts for the vastly different lengths of the two CNVRs (8.8 kb versus 187.5 kb). As the aCGH design does not allow us to discriminate among the different copies in the two CNVRs, we treated these two regions as a single variant, pooling the calls within them. To call alleles in these regions, we considered two copies the reference allele and assigned a score of +1 to samples called as duplicated and a score of –1 to samples called as deleted, summing scores for samples called in both regions. This model may represent an oversimplification, but aCGH is unlikely to be able to give separate signals from two highly similar regions, and the model differentiates samples called as deleted or duplicated twice across the two regions from samples with just one call.

The second pair of regions for which we considered the reference allele to be two copies was Y:22,218,957–22,369,669 and Y:22,419,687–22,508,011. Although these regions are not annotated as SDs, they nearly cover two arrays of *LTR12B* repeats, at Y:22,216,565–22,369,669 and Y:22,419,687–22,512,935, respectively. These arrays, separated by a 50-kb unique spacer, contain 33 and 15 elements, respectively, and individual elements range in length from ~300 bp to ~18 kb. These regions were included in a single CNVR, and the empirical probability density function of $\log_2$ intensity ratios of calls covering this interval shows a trimodal distribution (**Supplementary Figure 4**), with peaks around –2, –0.5, and 0.4 likely corresponding to large deletions, small deletions, and duplications, respectively. We manually checked the $\log_2$ intensity ratios for these calls to confirm that the distribution of segments' mean $\log_2$ ratios represented actual gains or drops of probe intensities. Upon confirming, we assigned values of 0, 1, and 3 copies to the samples called under the respective peaks.

2.2 Validation

2.2.1 Array Comparative Genomic Hybridization

Andrea Massaia, Ankit Malhotra, Charles Lee, Yali Xue, and Chris Tyler-Smith

Structural variants identified by Genome STRiP were validated using array comparative genomic hybridization (aCGH) intensities. The aCGH experiment was performed on the 1000 Genomes Project phase 3 samples, using sample NA10851 as reference. A total of 6,250 unique probes, ranging from 44 to 61 bp in length, were hybridized to the MSY, covering a total of 341,631 bp.

We used aCGH data from the 1,291 phase 3 females to filter probes to those yielding signals specific to the Y chromosome. Y-specific probes are expected to give low or no signal in females, that is, a $\log_2$ intensity ratio < 0. Assuming the $\log_2$ ratios of probes not specific to the Y chromosome follow a normal distribution across the 1,291 female samples, we retained those probes for which the mean $\log_2$ ratio was more than two standard deviations below 0. This left in 2,714 probes (43.4%) for SV validation.

Upon filtering to Y-specific probes, we normalized the $\log_2$ ratios by subtracting each sample's median $\log_2$ intensity ratio across all probes from its measured value at each probe. We then estimated the concordance between the Genome STRiP calls and the aCGH $\log_2$ ratios.

For each variant, we selected the samples called as REF by Genome STRiP and computed, for each one of them, the median of $\log_2$ ratios for the probes covering the variant. We assumed the medians of $\log_2$ ratios across REF samples to be normally distributed, and considered the ($\mu \pm 2\sigma$) confidence interval of this distribution. We then considered, separately for each ALT sample, the median $\log_2$ ratio for the same set of probes, confirming as ALT those samples for which this median fell outside the confidence interval. Specifically, for each sample called as duplicated, we confirmed the call if the median of $\log_2$ ratios was greater than or equal to the upper limit ($\mu + 2\sigma$) of the confidence interval and rejected it otherwise. Likewise, for each sample called as deleted, we confirmed the call if the median of $\log_2$ ratios was less than or equal to the lower limit ($\mu - 2\sigma$) and rejected the call otherwise. When validating the variants called as segmental duplications (with 2 copies being the reference allele), we pooled the probes falling in both reference copies.

Overall, 76 of the 97 copy-number variants (CNVs) and 9 of the 13 SDs called by Genome STRiP were covered by at least one aCGH probe. For each of these variants, we counted the number of ALT samples whose $\log_2$ ratios were compatible with the Genome STRiP call and those that were not (**Supplementary Figure 5**). All of the 76 CNVs covered by the aCGH probes were validated in at least one sample. Notably, most of the variants with low proportions of confirmed ALT samples were covered by a low number of probes (1 to 3). The one exception is the variant in Y:17,998,351–18,007,190, with three samples called as ALT (a deletion in HG00183 and a duplication in HG00372 in HG03976), that was covered by 7 probes.

Supplementary Figure 6 summarizes SD validation. Of 13 variants covered by aCGH probes, three were not validated in any of the aCGH samples. These three were called as ALT in 4, 4, and 8 samples. Another, called as ALT in 13 samples, was validated in just one of the aCGH samples. It is worth noting, however, that the complexity of regions defined as SDs might make it difficult to design aCGH probes specific for these regions, as suggested by our observation that CNVs and SDs differ in their densities of filtered probes, where we define probe density as the number of probes for a variant divided by its length ($p = 3.5 \times 10^{-4}$, Mann-Whitney test) (**Supplementary Figure 7**).

2.2.2 Fluorescence *In Situ* Hybridization onto DNA Fibres

Alkaline Lysis Fibre-FISH

Ruby Banerjee, Sandra Louzada, and Fengtang Yang

Methods

To further validate CNV calls, we conducted fibre-FISH experiments, following a previously described protocol[38]. Briefly, stretched DNA fibres were prepared by alkaline lysis of lymphoblastoid cells purchased from Coriell Biorepository[39]. Bacterial artificial chromosome (BAC) clones that span the CNV regions of the human Y chromosome were obtained from the clone archive resource of The Wellcome Trust Sanger Institute. DNA from each BAC clone was prepared using the Phase-Prep BAC DNA kit (Sigma-Aldrich), following the manufacturer's protocol, and was labeled with digoxigenin (DIG)-11-dUTP, biotin (BIO)-16-dUTP, or dinitrophenol (DNP)-11-dUTP (Jena Bioscience). BIO-labeled probes were detected with CF543-conjugated streptavidin (Biotium) or DyLight 488 conjugated streptavidin (Vector Labs); DIG-labeled probes were detected with monoclonal mouse anti-DIG IgG (Sigma-Aldrich) and Texas red conjugated donkey anti-mouse IgG (Molecular Probes); DNP–labeled probes were visualized using rabbit anti-DNP and Alexa 488 conjugated goat anti-rabbit IgG (Molecular Probes). After detection, slides were mounted with SlowFade Gold® mounting solution containing 4',6-diamidino-2-phenylindole (DAPI) (Molecular Probes) and kept at 4 °C.

All FISH images were captured on a Zeiss Axioplan epifluorescence microscope, equipped with narrow band-pass filter sets for DAPI, fluorescein, Spectra Gold, and Texas red fluorescence and a cooled CCD camera (Hamamatsu ORCA-ER). They were processed with the SmartCapture® software (Digital Scientific UK).

Results

Fibre-FISH was used to test two duplications. We elected to test duplications, as this type of CNV has been less thoroughly investigated and validated than deletions. Furthermore, there was a greater opportunity to generate novel insights about duplications than about deletions, as fibre-FISH provides information about the location of the duplicated copy, whereas neither sequence data nor genotype array data are able to do so.

The first duplicated region we tested was "CNV region 1" (GRCh37 Y:2,888,555–3,014,661), which includes variants CNV_Y_2888555_3014661 and CNV_Y_2892962_2900130. DNA sequence analysis indicated the presence of the reference structure in most individuals, including HG00096, which we used as a control in the experiment. Duplicated alternative alleles were inferred for samples HG03856 and GM19082, the test subjects. Two BACs, overlapping by 18.5 kb, were labeled red or green and hybridized to the control (**Supplementary Figure 8**). As expected, and consistent with a structure matching the reference sequence in this individual, we observed overlapping red and green signals, best illustrated in the downward shifted image in the middle panel for this cell line. In contrast, we observed duplication structures in both individuals classified as carrying alternative alleles by the DNA sequence analyses (**Supplementary Figure 8**). In each sample, we observed two pairs of red and green signals separated by a non-hybridizing region. Though fibre stretching is non-uniform, we observed a consistent pattern of relative lengths of the red and green signals in the two duplicated structures. In both HG03856 and GM19082, the red signal is longer than the green in the first copy and shorter than the green in the second copy, and these differences are consistent across multiple fibres. This experiment confirms the presence of a local but non-tandem duplication. To elucidate the full details of the duplicated structures will require further work.

DNA sequence analysis indicated that the second region, CNV region 4 (Y:16,077,197–16,251,571), was also present in the reference structure in most individuals, including HG00096, and that it occurred as a complex set of duplicated alternative alleles (CNV_Y_16077197_16094785, GS_SD_M2_Y_16093532_16131537_Y_16134952_16172355, CNV_Y_16134600_16211237, CNV_Y_16134975_16251571, and CNV_Y_16160861 _16214056) in a few individuals, including HG01377. Again, two BACs, in this case overlapping by ~100 kb, were labeled in red or green and hybridized to HG00096. We observed extensive overlapping signals (**Supplementary Figure 9**), consistent with a structure matching the reference sequence. In HG01377, a different structure was observed: a duplication of the red signal and a trace of the green signal, separated by a non-hybridizing gap. Thus, this experiment also confirms a local but non-tandem duplication in HG01377.

Molecular Combing Fibre-FISH

Sandra Louzada, Andrea Massaia, and Fengtang Yang

In this subsubsection, we describe the laboratory methods for the experiment to validate the deletion in HG00183. We describe the results in the main text and illustrate them in **Figure 1d**.

We purchased human lymphoblastoid B-cell lines from the Coriell Biorepository[39] and prepared single-DNA-molecule fibres using the molecular combing method described in Polley et al.[40] and following instructions from the manufacturer, Genomic Vision. Briefly, we embedded the

cells in plugs of 1.2% Lonza low melting-point agarose at a density of 1 million cells/plug. We followed this with overnight proteinase K digestion at 50°C, a wash with 1× TE buffer, and overnight digestion at 42°C with β-agarase enzyme from BioLabs. The next day, we mechanically stretched the DNA fibres onto coated coverslips (Genomic Vision) using a Molecular Combing System (Genomic Vision). We followed this with baking for 4 hours at 68°C.

We used four fibre-FISH probes, including three custom probes ("P1," "P2," "P3") from the CNV region (GRCh37 Y:17,986,693–18,017,210) and one from a human BAC clone (RP11-12J24) obtained from the clone archive resource of Wellcome Trust Sanger Institute. We produced the custom probes using long-range PCR amplification with primers designed with the web-based version of Primer3[41,42] using the following parameters:

- `PRIMER_MIN_SIZE=21`
- `PRIMER_OPT_SIZE=23`
- `PRIMER_MAX_SIZE=25`
- `PRIMER_MIN_TM=57.0`
- `PRIMER_OPT_TM=61.0`
- `PRIMER_MAX_TM=63.0`
- `PRIMER_PAIR_MAX_TM_DIFF=2.0`
- `PRIMER_PRODUCT_SIZE_RANGE=4500-5500`

We amplified using the Bioline RANGER Mix following the manufacturer's instructions for the reaction set-up, but employing the following touchdown protocol:

- 1 minute initial denaturation at 95°C
- 15 cycles of denaturation for 10 seconds at 98°C, followed by annealing/extension for 5 minutes, starting at 63°C and decreasing the temperature by 0.5°C each cycle
- 30 cycles of denaturation for 10 seconds at 98°C, followed by annealing/extension for 5 minutes at 56°C
- 10 minutes of final extension at 72°C

We purified the PCR products using a Qiagen QIAquick PCR purification kit according to the manufacturer's instructions.

We amplified the purified BAC DNA and PCR products using the GenomePlex® Whole Genome Amplification (WGA) kit from Sigma-Aldrich, following the manufacturer's protocol. We then labeled them using a modified WGA reamplification kit from Sigma-Aldrich, as described by Carpenter et al.[43]. We labeled the BAC clone with biotin-16-dUTP, P1 with digoxigenin-11-dUTP, P2 with DNP-11-dUTP, and P3 with Fluorescein-12-dUTP. The first three labels were from Jena Bioscience, and the fourth was from ThermoScientific.

We followed the fibre-FISH protocol of Carpenter et al.[43] with slight changes. After dehydrating through a 70%, 80%, and 100% ethanol series, we aged the combed coverslips in 100% ethanol for 30 seconds at 65°C and incubated them in alkaline denaturing solution (Sigma-Aldrich) for 8 minutes, followed by three washes in 1× PBS (Invitrogen) and dehydration through a 70%, 80% and 100% ethanol series. We denatured the probe mix at 65 °C for 10 minutes, added it to the coverslip, and then hybridized overnight at 37°C. Post-hybridization washes consisted of two rounds of 50% formamide/2×SSC (v/v), followed by two additional washes in 2×SSC. All

washes were at 25°C for 5 minutes with gentle agitation. We detected digoxigenin-labeled probes using a 1:100 dilution of monoclonal mouse anti-DIG antibody (Sigma-Aldrich) and a 1:100 dilution of Texas Red-X-conjugated goat anti-mouse IgG (Invitrogen). To detect DNP-labeled probes we used a 1:100 dilution of Alexa 488-conjugated rabbit anti-DNP IgG and 1:100 Alexa Fluor® 647 donkey anti-rabbit IgG (Abcam). For biotin-labeled probes, we detected with 1:100 of Cy3-streptavidin (Sigma-Aldrich) and 1:50 anti-streptavidin CF543. For this customized antibody, we labeled streptavidin (Vector Laboratories) with CF543 using the Mix-n-stain antibody labeling kit (Biotium) according to the manufacturer's instructions. Lastly, we detected fluorescein-labeled probes with 1:100 sheep anti-FITC (Southern Biotech) and donkey anti-sheep IgG Alexa Fluor® 488 (Thermo Scientific). After detection, we mounted slides with SlowFade Gold® mounting solution containing 4', 6-diamidino-2-phenylindole (Invitrogen). We viewed images on a Zeiss AxioImager D1 fluorescent microscope equipped with narrow band-pass filters for DAPI, FITC, Cy3, Cy5, and Texas Red fluorescence and on an ORCA-EA CCD camera (Hamamatsu). We used SmartCapture software (Digital Scientific, UK) to capture and process the digital images.

2.2.3 Karyotyping for Sex-Chromosome Aneuploidies

Ruby Banerjee and Fengtang Yang

Methods

Following a standard protocol[44], we prepared metaphase chromosomes from lymphoblastoid cell lines purchased from Coriell Biorepository[39]. Briefly, the cultures were first treated with 0.01 μg/mL colemid for 1 hour to arrest the dividing cells at the metaphase stage and were then treated with hypotonic solution (75 mM KCl) for 15 minutes. After two rounds of fixation and wash in methanol/acetic acid (3:1) fixative, the metaphase preparations were resuspended in fixative and stored at −20 °C until use.

Chromosome-specific paint probes for the human X and Y chromosomes were generated from 5,000 copies of flow-sorted chromosomes, provided by the Flow Cytometry Core Facility of The Wellcome Trust Sanger Institute, using the GenomePlex® Whole Genome Amplification kit (Sigma-Aldrich). X-chromosome probes were labeled with Aminoallyl-dUTP-Texas red and Y-chromosome probes were labeled with Aminoallyl-dUTP-XX-ATTO488 (Jena Bioscience). Fluorescence *in situ* hybridization (FISH) with the X and Y paint probes followed the multi-color strategy described in Gribble et al.[45].

Results

We selected 13 cell lines for karyotype analysis. One (HG02372) had been observed by Handsaker and colleagues to show female-level intensities of X-chromosome SNPs, as well as male-level intensities of Y-chromosome SNPs. The other 12 samples had lower levels of reads mapping to the Y chromosome than expected from median genome coverage or had median $\log_2$ ratios that differed from the mean value across all samples by more than 3 standard deviations (**Supplementary Figure 2**).

We observed a 47,XXY karyotype in all HG02372 cells we examined (**Supplementary Table 7**). This karyotype, known as Klinefelter Syndrome, is the most common constitutive chromosomal anomaly. It is reported to affect 1 in 500 to 1 in 1000 males and is often undiagnosed[46]. Therefore, it is not surprising to find one example in a sample of 1,244 males.

Of the 12 other cell lines, six exhibited karyotypes mosaic for the presence of 45,X at frequencies ranging from 27% to 62% (**Supplementary Table 7**). This mosaicism explains the low Y-chromosome read coverage or low aCGH signal in almost half of the samples showing unusual patterns. Our data do not indicate the timing of the Y-chromosome losses; they may have occurred after the establishment of the cell lines.

2.3 Analysis

Andrea Massaia, Yali Xue, Yuan Chen, and Chris Tyler-Smith

2.3.1 Mutation Events

To estimate the distribution of mutation events associated with Y-chromosome CNVs, we mapped the variants onto the phylogeny and counted the minimum number of mutation events in the history of each locus. We considered CNVs discovered from both sequence data and aCGH intensities (**section 2.1**) but excluded 16 variants within Y:23,632,158–27,687,750, as this region includes the *RMBY* and *DAZ* gene clusters and spans a set of well-known structural variants[47,48]. Although we have no reason to doubt the variation we observed in this region, its peculiar structure could cause oversegmentation and confound allele assignment and breakpoint definition. This left 105 loci (**Supplementary Figure 10**). Of these, 7 overlapped physically with and yielded similar genotype calls to other loci in the set. As these 7 loci may represent true gain or loss of genetic material oversegmented by the calling algorithms, we excluded them in this subsection but retained them when investigating the sequence context (**subsection 2.3.2**) and genomic impact (**subsection 2.3.3**) of the variants.

For the remaining 98 loci, we inferred the minimum number of duplication and deletion events that would lead to the observed genotypes. We inferred exclusively duplication events at 63 loci, exclusively deletion events at 22, and both (del/dups) at 13 loci. Thirty-eight of the 76 total duplications and 14 of the 35 total deletions were singletons—sites for which just one individual possessed the alternative allele, corresponding to a single event. Overall, we inferred a total of 360 duplication events, 1 to 105 per locus (mean: 4.74, median: 1), and 275 deletion events, 1 to 58 per locus (mean: 7.61, median: 1).

Nine loci required more than ten events to explain the observed genotypes. We manually checked these loci for sequence features that could explain such variability. Two of these loci, segmental duplications at (Y:6,543,373–6,559,148, Y:6,559,149–6,574,923) and (Y:17,986,738–17,995,460, Y:18,008,099–18,016,824), are within regions with high X-Y similarity and therefore may reflect X-chromosome variation randomly mapped to the Y chromosome in some samples. However, we retained the deletion call of the second locus in HG00183, as we validated it by fibre-FISH. One locus, a segmental duplication at (Y:7,446,529–7,540,962, Y:24,803,840–24,900,423), partially includes variation in the *RMBY* and *DAZ* regions[47,48] and may be confounded by the repetitive structure of these regions. Two loci, Y:13,135,703–14,045,110 and Y:28,783,131–28,814,512, lie adjacent to the centromeric heterochromatin and to the large *q*-arm heterochromatic block, respectively[49]. These loci may therefore reflect expansion or contraction of heterochromatic regions, or they may carry signal from similar heterochromatic blocks on different chromosomes. Moreover, locus Y:28,783,131–28,814,512 includes the *DYZ18* locus[49], the size of which is unknown, as it hasn't been fully sequenced. Therefore, only the absence of

the locus could be stated with accuracy. Finally, the locus Y:15,590,833–15,592,476, does not correspond to any element usually associated with high copy number variability nor with any feature suggesting a false positive. To validate the variant at this site, we manually checked read depth in the samples inferred to possess the alternative allele, as well as in 30 samples called as carrying the reference allele, including the sample used as a reference in the aCGH experiment (NA10851). As read-level data did not support the simple deletion called via aCGH and instead suggested a more complex mode of variation, we removed this locus and the events associated with it from further analyses.

We excluded five of the six loci described above, ascribing them to limitations in the calling techniques or to a "shadowing" effect, wherein reads from CNV sites elsewhere in the genome mismap to the Y chromosome and CGH probes cross-hybridize. We re-annotated the sixth as a deletion in a single sample (HG00183), but we excluded it due to overlap with another call in a larger region (Y:17,986,693–18,017,210) in the same sample. This exclusion brought to 8 the number of loci excluded due to overlap.

We retained the remaining three loci with more than ten events. Two of the three are located in the *TSPY* arrays, which are well known for being prone to frequent rearrangements[37]. The third locus entails an array of highly similar *LTR12B* elements on the long arm of the chromosome—a plausible hotspot of structural variation. This locus is the most variable in our entire dataset, with 154 inferred mutation events. It encompasses two arrays of *LTR12B* repeats at Y:22,216,565–22,369,669 and Y:22,419,687–22,512,935.

To test whether or not we could exclude a shadowing effect for the most variable CNV, we constructed a maximum-likelihood tree of all 211 *LTR12B* elements in the human genome. To do so, we downloaded each FASTA sequence from the UCSC database[24] and ran MEGA6[1] with a Jukes-Cantor model and the default options (**Supplementary Figure 11**). The *LTR12B* elements in the Y-CNV arrays form a monophyletic clade that does not include any other element, so it is unlikely that reads from elsewhere in the genome mismapped here. Furthermore, the tree indicates that the elements in these arrays are quite homogeneous and may therefore be highly prone to CNV events. Indeed, the two arrays contain 48 of the genome's 211 *LTR12B* elements, more than any autosome.

Following the manual check, 92 CNV loci remained for analysis, including 62 duplications, 20 deletions, and 10 del/dups. We inferred a low number of mutation events for most CNV loci, with just one event for 44 of 72 duplications and for 21 of 31 deletions (**Supplementary Figure 12**, **Supplementary Data File 1**).

2.3.2 Mutation Processes

To gain insight into the mutation processes underlying CNV events on the Y chromosome, we analysed the regions around the inferred breakpoints for 100 loci—the 92 above plus the 8 excluded due to overlap with other loci (5 duplications and 3 deletions). We grouped them into 48 partially overlapping regions of ~4.5 kb to ~3.7 Mb, with each region containing 1 to 12 variants. To investigate sequence self-similarity, we aligned each region against itself using Dotter[50] with the default parameters. We detected self-similarity near the inferred breakpoints in 56 of 100 cases. In each case, we manually confirmed the presence of repetitive elements, including Alu elements, LINEs, ERVs, satellite repeats, or intrachromosomal segmental

duplications. To do so, we used the UCSC genome browser[24] and BLAT[51] to align the sequences revealed by Dotter against the full human genome. We confirmed for each case that both sequences involved mapped to one of the repetitive elements listed above. Overall, we observed these elements near 10 of 23 deletions (43%), 38 of 67 duplications (57%), and 8 of 10 del/dups (80%). These observations indicate that repetitive sequences influence the majority of Y-chromosome.

As previously observed by Conrad et al.[33], we found that repeat-mediated variants are significantly longer than those that are not repeat-mediated. The mean lengths are 169,391 bp and 72,678 bp, respectively, and the median lengths are 48,689 bp and 15,138 bp, respectively (p = 0.026, Mann-Whitney two-sided test). The result holds when restricting to the pruned set of 92 variants (p = 0.013, Mann-Whitney two-sided test) and when excluding from this set the ~3 Mb repeat-mediated *AMELY* deletion (p = 0.019, Mann-Whitney two-sided test).

2.3.3 Genomic Impact

We investigated the potential impact of CNVs by considering their overlap with genes, as annotated in the Ensembl 75 database[52]. We found that 60 overlap at least one gene: 5 overlap protein-coding genes alone, 15 overlap both protein-coding genes and non-protein-coding genes (including genes annotated as miRNAs, pseudogenes, lincRNAs, antisense, miscRNAs, rRNAs, or snRNAs), and 40 overlap only non-protein-coding genes. This set of 60 CNVs includes 12 deletions, 43 duplications and 5 del/dups, among which 6 deletions, 12 duplications, and 2 del/dups overlap at least one protein-coding gene. We compared the distribution of duplications and deletions among the 119 mutation events associated with variants overlapping protein-coding genes (75 duplications versus 44 deletions) to the distribution among the 286 mutation events associated with CNVs not overlapping protein-coding genes (205 duplications versus 81 deletions), but we did not observe a significant difference ($p > 0.05$, Fisher's exact test).

For both the Y chromosome and the autosomes, we compared the proportion of deletions overlapping genes to the corresponding proportion for duplications. For the Y chromosome, we restricted the set of CNVs to include those with only duplications or with only deletions. Of the 22 deletion-only CNVs, 6 (27%) overlapped protein-coding genes, and 12 of 68 (18%) duplications did so. The ratio of these proportions is 1.5. To compare this observation to autosomal data, we extracted the number of deletions and duplications from the 1000 Genomes Project phase 3 data[16] and used Ensembl's Variant Effect Predictor (VEP)[53] to determine which variants overlapped genes. We found that of 38,258 autosomal deletions, 25,524 (66.7%) overlapped protein-coding genes, and 4,657 of 5,896 (79.0%) duplications did so. The ratio of these proportions is 0.84. Whereas, on the autosomes, deletions are less likely to overlap protein-coding genes than duplications are, we found the reverse to be true for the Y chromosome.

We considered the position of the events in the phylogeny, classifying each as terminal or internal, and found that 108 of the 119 events associated with CNVs overlapping protein-coding genes occurred on terminal branches. This proportion was similar to that among CNVs not overlapping protein-coding genes ($p > 0.05$, Fisher's exact test). In this class, 243 of 286 events were on terminal branches. Stratifying by variant subtype (deletion or duplication) yielded no additional insight.

2.3.4 Comparison to Prior Work

Finally, we compared the variants discovered in our study with the variants reported in the recent studies on Y-chromosome CNVs[54–56]. Espinosa et al.[54] studied structural variants in 70 males from the pilot phase of the 1000 Genomes Project and reported 19 variants across the MSY, 15 of which overlap with those reported here. Wei et al.[55] reported 34 raw CNV events (rawCNVEs) in a cohort of 411 healthy UK males and merged these events into 21 curated CNV events. We found that 32 of their 34 raw events overlap with our discovery set. Johansson et al.[56] reanalysed a set of 1,718 males, collected from several studies, and reported 25 CNV patterns on the MSY, but they did not report exact start and end positions. Fifty of the 121 genomic regions in our initial dataset are novel with respect to these three papers.

3 Short Tandem Repeats

Thomas Willems, Melissa Gymrek, and Yaniv Erlich

To estimate Y-STR mutation rates, we used an approach that we have fully described in a second manuscript, "Population-Scale Sequencing Data Enables Precise Estimates of Y-STR Mutation Rates"[57]. We include a brief synopsis in the following sections, but we encourage readers to consult the companion paper for the full details and for updated mutation rate estimates.

3.1 Callset

3.1.1 Generation

For our STR analyses, it was particularly important to use indel-sensitive alignments. These became available from the 1000 Genomes Project FTP site in the later stages of our work, but only relative to the GRCh38 reference. Thus, in contrast to our analyses of the other variant classes which were based on GRCh37 alignments, we downloaded BWA-MEM[58] alignments to the GRCh38 Y-chromosome reference sequence. Next, we obtained from our companion paper[57], a set of Y-STR regions to genotype. These regions were generated using both the Tandem Repeats Finder program[59] and published primers for Y-STR markers. To ensure high quality genotypes, we only selected regions that passed a series of stringent quality control filters for a higher coverage dataset. After merging the individual BAMs using SAMtools[17], we ran GitHub version g853f1a1 of HipSTR[60] using these regions and the following options: `min-reads=100`, `haploid-chrs=chrY`, `hide-allreads`. To mitigate genotyping errors, we removed all homopolymers, loci with more than 15 genotyped females, loci with DFLANKINDEL/DP > 0.075, and individual calls with DFLANKINDEL/DP > 0.1, where DP indicates the total number of reads, and DFLANKINDEL indicates the number of reads with indels in the regions flanking the STR. We further removed loci at which more than 5% of samples had out-of-frame STR calls. Lastly, to generate the final callset, we removed out-of-frame calls and calls with low posterior quality scores (Q < 0.66), and we selected only multiallelic Y-STRs with at least 100 genotyped males.

3.1.2 Quality Assessment

To assess the quality of the callset, we compared STR genotypes across 3 father-son pairs. As even the most polymorphic STRs typically mutate at rates less than 10^{-2} mutations per generation (mpg), the fraction of concordant genotypes should exceed 99% in the absence of genotyping errors. Of the nearly 1,711 pairs of father-son Y-STR calls, we observed a concordance rate of 95.8%. Restricting to STRs with a major allele frequency below 95%, the concordance rate fell to 88.9%, but the overwhelming majority of these errors were likely the result of PCR stutter—1 or 2 repeat-unit differences in loci with dinucleotide motifs (**Supplementary Table 4**).

We also compared HipSTR calls to those generated by capillary electrophoresis for 15 of the loci in the PowerPlex Y23 panel. Routinely used for forensic and paternity-related analyses, this set of loci is highly polymorphic and therefore provides a challenging validation set. Encouragingly, 97.5% of the 4,058 resulting comparisons were concordant (**Supplementary Table 5**), and the bulk of the discrepancies again involved single repeat-unit differences (**Supplementary Table 6**).

3.2 Mutation Rates

3.2.1 Y-STR Mutation Model

We assumed that STR mutations stem from a length-dependent variant of a generalized stepwise mutation model. This model is characterized by a per-generation mutation rate μ, a geometric step size distribution with parameter ρ_m, and a spring-like length constraint β that causes alleles to mutate back towards a central allele. The flexible nature of this model captures many of the salient features of microsatellite mutations. In particular, decreasing the value of the geometric step size parameter can alter the model from a single-step only model to one allowing multi-step changes. Furthermore, the length constraint controls the extent to which shorter STRs preferentially expand and longer STRs preferentially contract, a known feature of many STR mutation models. For more details on the model, please refer to the Methods section of our companion paper[57].

3.2.2 Estimating Mutation Rates

The fundamental idea behind our approach is that a Y-SNP phylogeny is sufficiently detailed and precise to estimate a Y-STR's mutational dynamics. As a result, our approach begins by building a single phylogeny relating all samples using only Y-SNP genotypes. Next, for each Y-STR, it learns an error model to account for PCR stutter artifacts and alignment errors that are problematic in the 1000 Genomes Project's low-coverage data. To model these artifacts, we assume that their sizes follow a geometric distribution with parameter ρ_s and that they increase or decrease a read's STR size with probabilities u and d, respectively. For each Y-STR, our approach learns these locus-specific models by analyzing reads across all samples and applying an expectation-maximization algorithm. It then uses a uniform prior and the learned stutter model to compute each sample's genotype posteriors, which correspond to the probabilities for the leaves of the phylogeny. Because the likelihood of a given mutation model can be efficiently evaluated using a variant of Felsenstein's tree-pruning algorithm[61], we initialize each mutation model and use numerical optimization to iteratively improve its likelihood until convergence, resulting in an estimate for the mutation rate. For full details, please refer to Figure 1 and the Methods section of the companion paper[57].

3.2.3 Mutation Rate Simulations

To validate our approach, we simulated various STR mutation models using the 1000 Genomes Project phylogeny. Each of these simulations resulted in a set of known STR genotypes, for which we then simulated reads under various stutter models. When we applied our estimation method to these simulated reads, we obtained unbiased mutation rate estimates for nearly all scenarios we considered. In contrast, estimates obtained without accounting for stutter resulted in marked upward biases. These findings are summarized in Figure 2 and Supplemental Figures 4 and 5 of our companion paper[57].

3.2.4 Results

We estimated the mutation rates of 702 Y-STRs (**Supplementary Data File 2**). To validate, we compared our estimates for 106 loci to those from a large-scale father-son study and obtained an R^2 of 0.64. We also applied our method to an orthogonal high coverage dataset and found that the resulting estimates were remarkably correlated with those generated in this study (R^2 = 0.92), lending further support to the robustness and accuracy of our method. These comparisons are outlined in extensive detail in Figure 3 of the companion paper[57].

4 Phylogeny

4.1 Haplogroup Classification and Distribution

G. David Poznik

We assigned a haplogroup affiliation to each individual (**Supplementary Data File 3** and **Supplementary Table 8**) using the definitions within the January 18, 2014 version of the SNP Compendium maintained by the International Society of Genetic Genealogy (ISOGG)[62]. To do so, we probed each sample for derived alleles at any site and then removed all inconsistencies due to homoplasy, genotype error, or database misspecification. Left with a consistent path through the phylogenetic decision tree for each individual, we called the haplogroup based on the most derived SNP remaining.

Initially, based on the ISOGG Compendium alone, we could not classify 13 individuals more precisely than haplogroup F (hgF), the megahaplogroup that includes most non-African lineages. Therefore, to supplement the ISOGG resource, we constructed a list of 20 SNPs we had found to be present in the derived state in both a hgH individual[4] and a hgF3 lineage[63]. We used this shared branch to define a new subgroup of hgH, which we provisionally dub "H0." We found that 12 of the 13 putatively F* individuals possessed the derived allele for these sites, so we classified them as belonging to hgH0.

4.2 Tree Inference

4.2.1 Total-Evidence Maximum-Likelihood Tree

Apurva Narechania, G. David Poznik, Juan Rodriguez-Flores, Rob Desalle

To construct a total-evidence maximum-likelihood (ML) tree, we converted genotype calls for the 60,555 biallelic SNVs to nexus format and ran RAxML8[64] using the ASC_GTRGAMMA model. We then conducted 100 ML bootstraps and mapped these to the total-evidence tree.

Using prior knowledge[65], we rooted the tree to the midpoint of the split between A0 and A1 and then used MEGA5[66] to manually rotate internal nodes to conform to the canonical representation. We used FigTree[2] to plot (**Supplementary Data File 4a**).

4.2.2 Rooted Tree

Yuan Chen, G. David Poznik, and Yali Xue

Because the "chimpanzee and human Y chromosomes are remarkably divergent"[67], we were not able to use a non-human outgroup for the full 10.3 Mb under analysis. However, using Enredo Pecan/Ortheus (EPO) alignment[68] within the Ensembl genome browser[69], we identified regions where the human and chimpanzee Y chromosomes align one-to-one. Upon restricting to these 9.58 Mb, we again used RAxML8[64] to construct a second, rooted, version of the phylogeny, which indicates the relative lengths of the roots of A0 and A1 (**Supplementary Data File 4b**).

4.3 Mapping SNVs to the Tree

G. David Poznik

We mapped each SNV to one or more branches of the inferred topology. Doing so has four benefits. First, the number of SNVs mapping to a given branch is an interpretable distance measure that we can use to estimate split times (**section 4.5**). Second, once mapped, each SNV becomes a diagnostic marker with which one can classify future samples. Such sequencing-based rosters of phylogenetically placed SNVs are particularly valuable for ancient DNA (aDNA) studies, in which sequencing coverage can be quite low, such as Schroeder et al.[70]. Third, we can identify the ancestral state for each SNV, and fourth, we can impute missing genotypes.

4.3.1 Results

Prior to mapping SNVs to the phylogeny, we partitioned the ML tree into eight overlapping subtrees (**Supplementary Figure 13**). For each subtree, we defined a set of SNVs that were variable within it and assigned each site to the internal branch constituting the minimum superset of carriers of one allele or the other (**Supplementary Data File 5**). Let *M* represent this minimum superset. We designated the derived state to the allele that was observed only within *M* and the ancestral state to the other allele. When the dichotomy was clean (i.e., no ancestral alleles were observed within *M*), we deemed the site compatible with the subtree and imputed missing genotypes accordingly. Otherwise, for sites incompatible with the subtree, we did not impute missing genotypes.

Of the 60,555 sites, 56,714 (93.7%) mapped compatibly to exactly one of the eight subtrees, 2,518 (4.2%) mapped to two, and 443 (0.7%) mapped to three or more, whereas 880 (1.5%) did not map compatibly to any. This set of 880 sites differs substantially from the set of 880 sites referred to in **section 1.1**; the equivalent cardinalities of these two sets is merely coincidental. In total, we observed 63,230 mutation events. One thousand fifty-two sites (1.7%) mapped incompatibly with one subtree, and 649 (1.1%) mapped incompatibly with two or more subtrees. These calculations exclude branches duplicated between subtrees and count each of the 2,487 ($2n - 1$) branches of the full phylogeny exactly once.

Supplementary Figure 14 shows the eight subtrees with each branch length drawn proportional to the number of SNVs mapping compatibly. We comment on noteworthy features of this observed phylogeny in **section 4.4**.

4.3.2 Discussion

Partitioning the Sample Prior to Tree Construction

We extracted the subtrees from the total-evidence tree for consistency, but another approach would be to leverage preexisting information about the phylogeny and partition the sample based on observed haplogroups prior to tree construction. To do so would lead to no information loss and would be more computationally efficient. Initially, this was the primary motivation for the subtree-based analysis. However, another benefit of mapping SNVs to branches on a subtree-by-subtree basis is that doing so reduces the probability that any given SNV will be rendered incompatible due either to recurrent mutations (i.e., homoplasy) or to genotype error at a disparate location in the tree.

Inferring Ancestral States

By including in each subtree a small set of samples that overlapped with neighboring subtrees, we polarized ancestral and derived states for all branches but the most basal of the global phylogeny; we could not polarize SNVs mapping to the two branches separating hgA0 from the rest of the tree, as no outgroup was available for this most ancient split. Due to reversion mutations, alleles that are ancestral in one subtree may be derived in another, so we determined the globally ancestral allele based on the outermost subtree in which we observed a SNV (**Supplementary Data File 5**).

Isolate Lineages

A caveat to our SNV-to-branch mapping procedure is that though it works well for well-balanced regions of the tree, where the superposition of lineages elicits high effective coverage on the internal branches of the tree, it breaks down in instances where the outgroup of a clade is represented by just one or two low-coverage samples. When an outgroup lacks data for a given site, the site cannot be assigned to the branch immediately upstream of the outgroup. Instead, these sites will be misassigned to the root of the sister clade. Therefore, the lengths of branches adjacent to isolate lineages (such as the hgF* individual in **Supplementary Figure 14b** or the hgK2a1* individual in **Supplementary Figure 14d**) must be interpreted with caution, as must ancestral allele imputations in samples from isolated regions of the tree.

Topology Refinements

In rapidly diversifying regions of the tree, with few SNPs to support branching events, tree-inference errors occur. In order to identify and resolve such errors, we manually curated those regions of the tree analyzed for signals of growth in chapter **8** and made minor rearrangements to local topologies when doing so increased the number of SNVs compatible with the tree. We deemed a putative rearrangement permissible if and only if the set of compatible SNPs after the arrangement was a proper superset of the set preceding it. Through this effort, we were able to place 98 initially incompatible SNPs onto the tree.

Missingness and In Vitro Mutations

Due to modest sequencing coverage, data missingness was a principal concern, but the impact on our downstream analysis was minimal. Type 2 errors primarily affect low frequency variants, at which missing genotypes lead to unobserved singletons and, to a lesser degree, doubletons misclassified as singletons and missing doubletons. In contrast, we can accurately infer the phylogenetic placement of higher-frequency variants using only those samples with data for any given site. We therefore eschewed the use of information from low-frequency variants—those corresponding to the tips of the tree—in all our downstream analyses and instead leveraged information from internal branches of the tree.

In vitro mutations likewise had little impact on our downstream analyses. These de novo mutations are generally unique to a given sample and therefore present as singletons—at the tips of the phylogenetic tree. The fact that we observed clusters of nearly identical lineages within, for example, haplogroup E1b (**Supplementary Figure 14a**) indicates that the impact is minimal. But because we did not use singleton branch lengths in our downstream analyses due to the greater bias from missing data, in vitro mutations did not affect our results.

4.4 Features of the Tree

G. David Poznik, Fernando L. Mendez, and Peter A. Underhill

In this section, we describe the haplogroup distribution of our sample and indicate noteworthy features of the phylogenetic tree revealed by the sequencing and analysis undertaken in this work. In particular, we highlight novel structures (**Supplementary Figure 15**) and describe instances of short internal branching, a phenomenon whose extreme is a "star-like" phylogeny.

We comment on five star-like phylogenies within: E1b in Africa, R1b and I1 in Europe, R1a in South Asia, and Q1a in the Americas. Under a Wright-Fisher model, one expects longer branches toward the root of the tree, where there are fewer lineages and, consequently, longer coalescence waiting times. Short internal branches and, in particular, star-like phylogenies may reflect the breaking of one or both of the Wright-Fisher assumptions of: (a) constant population size, which is violated by population growth; or (b) exchangeability, which is violated when populations cease exchanging genes (e.g., due to migration). We model haplogroup expansions in chapter **8**.

We have arranged **subsections 4.4.1–4.4.15** in phylogenetic, rather than alphabetic, order (**Supplementary Figure 13**). Of the 20 haplogroups lettered A though T, 16 are terminal monophyletic clades. The letters F, K, and P refer to a nested set of megahaplogroups: F is the ancestor to haplogroups G, H, I, J, and K; K includes L, T, N, O, M, S, and P; and P is the parent of Q and R. Finally, A is paraphyletic, encompassing four distinct clades. At its highest level, the known Y-chromosome phylogeny is (A00, (A0, (A1a, (A1b1, BT)))), where hgBT is the ancestor to all haplogroups lettered B through T. The overwhelming majority of living men carry lineages that descend from this clade.

When appropriate, we compare our tree to those of the seven studies that have previously evaluated at least 1 Mb of Y-chromosome sequence in at least 30 individuals (*n*): Wei et al. 2013 (8.97 Mb, $n = 36$)[71], Poznik et al. 2013 (9.99 Mb, $n = 69$)[4], Francalacci et al. 2013 (8.97 Mb, $n = 1208$)[63], Scozzari et al. 2014 (1.5 Mb, $n = 68$)[25], Yan et al. 2014 (3.9 Mb, $n = 78$)[72], Hallast et al. 2015 (3.7 Mb, $n = 448$)[73], and Karmin et al. 2015 (8.8 Mb, $n = 456$)[74].

4.4.1 Haplogroups A0 and A1a: Rare African Lineages

Of the four previous studies to evaluate 8 Mb or more of Y-chromosome sequence, each included at least one A1b1 lineage, and one (Karmin et al.[74]) sequenced two A00 individuals, but none had sequenced representatives of A0 or A1a. Though our sample did not include the rare A00 haplogroup, a product of the most ancient known split that was first reported in 2013[75], it did include the first two full sequences of hgA0 (**Supplementary Figure 14a**), which itself arose from the second-most ancient known split[65]. We observed both A0 lineages within individuals of West-African ancestry: one Gambian and one African Caribbean from Barbados, and we note a deep split between the two. These A0 sequences enabled us to infer ancestral and derived states for SNVs that occur within the rest of the tree. The sample also includes the first three full A1a sequences, each of which we observed in a Gambian individual.

4.4.2 Haplogroup B: A Novel Subgroup, B3

In the extant phylogeny, African haplogroup B bifurcates into the rare B1-M236 and the more common B2-M182. Four of the seven Y-chromosome sequencing studies included lineages

descending from B2-M182[4,25,73,74], and one included B1[25]. We observed seven B lineages: four B2 and three (two Mende and one Gambian) that form a distinct clade (**Supplementary Figure 14a**). Based on high-coverage capture-sequencing of ~1.5 Mb, Scozzari et al.[25] report 23 SNPs on the branch leading to B1. Because each of the three individuals in the 1000 Genomes Project sample carries exclusively ancestral alleles at each of the 23 sites, we conclude that they are not B1. Rather, these lineages constitute a novel subclade, which we provisionally name "B3."

To define the new topology of haplogroup B, we again use SNPs from[25]. Of the 33 SNPs they report on the branch leading to B2, we observe two in the derived state in each of the three B3 individuals: M8711 (8,139,185 A→G) and M8719 (8,481,949 C→G). This sharing indicates that B3 split with B2 relatively quickly after their parent lineage split with B1, thus yielding the following new topology: (B1, (B2, B3)).

4.4.3 Haplogroup D: Specific to Japanese Samples

Haplogroup D was the only major clade with perfect population specificity within our sample. We observed each of the 20 D lineages among Japanese individuals (**Supplementary Figure 14a**), though the haplogroup is also known to occur in Tibet and Southeast Asia[3].

4.4.4 Haplogroup E: The Predominant Haplogroup of Africa

Haplogroup E includes three major branches: E1a, E1b, and E2. We observe 20 E1a, exclusively among individuals with West-African ancestry, and our sample includes the first five full sequences of E2 (**Supplementary Figure 14a**).

Half of all individuals sequenced belonged to E1b, R1b, or O3, with E1b the single most common group, accounting for 24% (n = 298) of the sample. At least 65% of each African population were E1b, including 100% (53/53) of the Esan. Within E1b, a deep split separates E1b1a-M2 from E1b1b-M35 (**Supplementary Figure 14a**). All six European E1b were E1b1b-M35, and, but for two Gambians and four Luhya, all African E1b were E1b1a-M2. In this African E1b1a-M2 branch, we observe a large star-like phylogeny, as seen in Poznik et al.[4]. This structure likely reflects rapid growth associated with the Bantu expansion.

4.4.5 Haplogroup C: Asian Haplogroup with Recently Resolved Internal Structure

Our sample includes 31 haplogroup-C sequences (**Supplementary Figure 14a**): four Japanese C1-M8; ten C3-M217, of which nine occurred in East-Asian populations; and 17 C5-M356 that were exclusive to South Asia, with at least one representative in each of the five populations. The structure at the base of haplogroup C was unresolved until recently, with these three major subgroups forming a trichotomy. We identified five SNPs for which haplogroups C1-M8 and C5-M356 share the derived alleles and for which C3-M217 retains the ancestral allele, confirming the topology worked out in Karmin et al.[74]: ((C1-M8, C5-M356), C3-M217).

The parent of hgC and megahaplogroup F, CF-P143, is another short internal branch. These clades share just four SNPs, including the previously known P143.

4.4.6 Paragroup F*: An Isolate Lineage Reveals New Internal Structure, GHIJK

We observed one lineage, carried by the Vietnamese sample HG02040, that is derived for F-M89 and most other SNPs shared by the rest of megahaplogroup F. However, the individual carried

the ancestral allele for four SNPs present in the derived state in all other hgF representatives. These SNPs include M3658 (85,89,031, C→T), M3680 (14,237,670, C→T), M3684 (14,367,181 G→A), and 14,565,310 C→A. The lineage therefore constitutes an outgroup to the rest of the clade and is the only one that we classify as belonging to paragroup F*. As haplogroup G was the most basal subgroup of F in any of the seven aforementioned studies, this lineage is novel.

This first full sequence of an F* lineage enabled us to define new internal structure of the phylogeny. The 4 SNPs for which the lineage retains the ancestral allele define a new internal branch of the tree, GHIJK, that is immediately upstream of the one-SNP HIJK branch identified by Poznik et al.[4] (**Supplementary Figure 14b** and **Supplementary Figure 15**). No HG02040 reads were available for an additional 11 hgF SNPs. Due to low coverage and the isolated position of this lineage (please see the caveat in **subsection 4.3.2**), the greedy algorithm assigned these 11 SNPs to the GHIJK branch as well (for a total of 15), but it is unlikely that the F* lineage actually retains the ancestral allele at more than one of these additional sites (p = 0.03, binomial test), given that it carried the derived allele at 147/151 observed hgF SNPs.

Karmin et al.[74] recently noted in the high-coverage sequences of Wong et al.[76] the presence of an F2 lineage, carried by the Malay sample SSM072, that also lies outside the main GHIJK subgroup of F.

4.4.7 Haplogroup G

Haplogroup G is relatively rare in the sample, with just 19 lineages, primarily from European and northern South-Asian populations, plus one Mende individual clustering most closely with an Iberian lineage (**Supplementary Figure 14b**), which likely resulted from recent gene flow.

4.4.8 Haplogroup H: Twelve Individuals Define a Novel Subgroup, H0

Twelve samples, all of South-Asian ancestry, initially appeared to be F*. They possessed derived alleles for M89-equivalent SNPs and no downstream ISOGG markers. However, we identified 24 SNPs shared between these lineages and haplogroup H (**Supplementary Figure 14b**). Therefore, we propose to redefine hgH by one of these 24 SNPs (e.g., M2713, a G→A mutation at coordinate 6,855,809) (**Supplementary Figure 15**) rather than M69, which occurs downstream. As the group represented by these twelve lineages should be considered a proper subgroup of H, we have provisionally labeled this new clade "H0". Comparing this new clade to the two most recently published Y-chromosome sequencing studies, we found that each included one representative. Hallast et al. 2015[73] included an isolate Nepali lineage "nep-0186" and Karmin et al. 2015[74] included an isolate Malayali lineage from Southern India, "16806." Both of these lineages are most closely related to that of our H0 Punjabi sample HG02684.

We observe considerable structure within the poorly characterized haplogroup H1-M52. Just two H2-Apt lineages occur within the sample (both Telugu), but a sister clade of three lineages shares 38 SNPs (**Supplementary Figure 14b**).

4.4.9 Haplogroups I and J: Star-Like I1 and European/South-Asian Structure in J2

We observe haplogroups I1, I2, and J1 across European populations and populations with known European admixture, as well as in a single non-European, a Punjabi carrier of J1. Haplogroup J2, on the other hand, is evenly split between Europe (27/61) and South Asia (34/61). Lineages cluster by superpopulation, but not as two distinct clades. Rather, there are several

superpopulation-specific clusters (**Supplementary Figure 14c** and **Supplementary Figure 16**). In I1, we see a star-like phylogeny that is mirrored in R1b (**Supplementary Figure 14e**).

4.4.10 Haplogroups L and T (K1)

Within megahaplogroup K, the first bifurcation is between K1, also known as "LT", and K2, also known as "K(xLT)." K1, in turn, splits into haplogroups L and T (**Supplementary Figure 14d**). We observed 27 L lineages, with at least two in each South-Asian population and zero elsewhere. The eight hgT individuals in our sample occur in European and Admixed-American populations.

4.4.11 Paragroup K2a1*: An Isolate Lineage Reveals Novel Substructure That Informs Reanalysis of Ust'-Ishim and Oase1

Within haplogroup K2, we observed one Y chromosome that did not fit into the known phylogeny. The lineage, carried by Telugu sample HG03742, is derived for the SNPs that define megahaplogroup K and for M526, which defines K(xLT), recently renamed K2[5]. Though the lineage was ancestral for M214, which defines hgK2a/NO, it shared five derived alleles with the NO branch: M2308 (7,690,182 A→T), M2313 (8,674,808 C→T), M2335 (19,513,070 C→T), M2339 (21,797,754 T→C), and M2346 (23,617,006 G→A). Thus, we initially classified it as a novel NO* lineage requiring redefinition of hgNO (**Supplementary Figure 14d**). No such lineage occurs in any of the seven prior Y-chromosome sequencing studies, however Karmin et al.[74] identified the presence of such a lineage, carried by the Malay sample SSM016, within the sequences of Wong et al.[76].

Ust'-Ishim, "a 45,000-year-old modern human from western Siberia"

The Telugu lineage became particularly informative with the publication of the "Genome sequence of a 45,000-year-old modern human from western Siberia"[13]. We checked the "Ust'-Ishim" K2 lineage reported therein at each of the five sites listed above. The ancient human carried the derived T, with 22 reads of support, at M2308 and the ancestral allele at the other four sites. Thus, upon redefining K2a with M2308, we classify Ust'-Ishim as K2a*. We then define the parent branch of NO to be K2a1-M2313 and classify the isolate Telugu lineage, HG03742, as K2a1* (**Supplementary Figure 15**).

We checked the allele status in both HG03742 and Ust'-Ishim for the five known sites on lineages K2c-P261, K2c-P263, K2d-P402, and K2d-P403, and we checked the single-base insertion K2e-M147 that has been observed exclusively in two Indian samples[77]. We observed the reference (ancestral) allele at all sites in both individuals.

Oase1, "an early modern human from Romania with a recent Neanderthal ancestor"

The new branches that we defined above as ancestral to NO—K2a-M2308 and K2a1-M2313, which are united as branch #269 in **Supplementary Figure 14d—**inform our understanding of the Y-chromosome lineage of Oase1, "An early modern human from Romania with a recent Neanderthal ancestor"[78]. Based on ISOGG SNPs alone, the authors could not ascribe the lineage to any specific haplogroup of megagroup F, but we have increased granularity by reanalyzing these data.

Three Oase1 genotypes overlap with the IJK branch, and all were derived: 7,702,973 (T→A), 7,792,789 (G→A), and 21,571,895 (G→A). Further, one genotype overlaps with the K branch, and Oase1 also carried the derived allele for this SNP: 15,842,844 (G→A). Next, Oase1 carried

the derived T at the M2308 transversion shared by Ust'-Ishim. Finally, data were available for just one of the four K2a1 SNPs, M2346, and Oase1 possessed an ancestral G. Therefore, the Oase1 lineage branches from our phylogeny before the Telugu lineage split with NO, but no earlier than the emergence of the Ust'-Ishim lineage.

4.4.12 Haplogroup N: A North-Eurasia Connection

We observe two principle clades within haplogroup N (**Supplementary Figure 14d**). Five of seven instances of the smaller branch occur in Han Chinese in Beijing. The larger branch, N1, itself divides into two distinct clades, one with two members in Asia and the other with 23 Finnish individuals. Rootsi et al.[79] have ascribed this pattern to a hypothesized late Pleistocene–Holocene migration toward Northwestern Europe from an ancestral East-Asian source.

4.4.13 Haplogroup O: The Predominant Haplogroup of East Asia

Haplogroup O is the most common in Asia, and we observed it almost exclusively within East-Asian populations, with four (of 208) exceptions occurring in Bengali individuals. O2 and O3 were especially common, with 75 and 114 representatives, respectively. There were at least seven O2 and at least ten O3 in each East-Asian population. In contrast to R1a and R1b, there is an abundance of relatively deep structure within O2 and O3 (**Supplementary Figure 14d**).

4.4.14 Haplogroup Q: The Predominant Haplogroup of the Americas

Within haplogroup Q, the first bifurcation we observe is between three South-Asian Q1b-L275 and 42 carriers of Q1a-L472 (**Supplementary Figure 14e** and **Supplementary Figure 17**). Q1a is an ancestor of Q-L54, the most common indigenous American haplogroup. Each of the 36 instances of Q-L54 occurs within the Americas. Most also carry the M3 SNP, which is known to be present in Siberia and predominant among Native-American paternal lineages[80,81]. The Q-M3 subtree exhibits a star-like pattern, which may coincide with the initial colonization of the Americas. Subclusters of Q-M3 are population-specific (**Supplementary Figure 17**), a fact that may reflect the halting of gene flow subsequent to the founding of the groups from which these modern populations descend.

Interestingly, one Peruvian Q lineage, that of HG01944, does not belong to the Q-L54 subgroup. Rather, the lineage clusters most closely with two Vietnamese samples (**Supplementary Figure 17**). We hypothesized that this individual carries hgQ not due to Native-American paternal inheritance, but rather, due to post-Columbian Asian admixture. Peru has in fact been home to Asian immigrant communities since the early 17th century[82].

Admixture analysis confirmed this hypothesis. Upon merging 1000 Genomes Affymetrix 6.0 array genotypes with data from indigenous Americans[83] and running the program ADMIXTURE[84] with $K = 4$ clusters, we observe the following ancestry proportions across the autosomes of HG01944: 51% Native American, 39% East Asian, 9% European, and 1% African.

The distribution of Y chromosomes among Admixed-American populations reflects a significantly gender-biased admixture process, wherein Native-American Y chromosomes are underrepresented with respect to genome-wide ancestry proportions (**Supplementary Table 9**). In particular, 0 of 54 Puerto Rican men carry an indigenous Y chromosome, despite 13% Native-American ancestry (binomial $p = 0.0005$).

4.4.15 Haplogroup R: The Predominant Haplogroup of Europe

We observed 216 instances of hgR1b, the most common haplogroup in Europe. Men from each of the European populations, and from each of the Admixed-American and Caribbean populations, carried R1b. Within the clade, we observe a massive star-like phylogeny (**Supplementary Figure 14e**), which likely reflects recent rapid growth within Europe.

Haplogroup R1a is also present, though at far lower frequency, in all sampled European populations, but we observe R1a primarily among South Asians, with at least ten instances in each of the five South-Asian populations. R1a is also star-like (**Supplementary Figure 14e**). R2 (n = 29) was specific to, and represented across, South Asia.

4.5 Split Times

G. David Poznik

4.5.1 Unsuitability of Terminal Branches

Low sequencing coverage leads to missing genotypes and to undetected SNVs, which, in turn, lead to unreliable branch-length measurements toward the tips of the tree. But the bias for any particular branch is not a simple function of the true branch length and its sequencing coverage. Rather, it is a complex interaction of these factors along with the true lengths of upstream branches and the sequencing coverages of lineages descending from them.

To illustrate, consider the simplest subtree: two lineages, a and b, with one shared branch, c, and let the true lengths of the three branches be d_a, d_b, and d_c, respectively (**Supplementary Figure 18**). For the purposes of this example, we make two assumptions:

1. We detect a SNV if and only if we observe at least two (high-quality) sequencing reads between the two samples.
2. The number of reads observed in an individual at an arbitrary site is Poisson distributed, with mean equal to the average sequencing coverage of the individual across all sites, λ_i.

Finally, let i_0, i_1, and i_{2+} be the probabilities of observing zero, one, and two or more reads in lineage i, where $i \in \{a, b\}$.

Under the first assumption, we will fail to detect a singleton whenever we observe fewer than two sequencing reads from the one sample possessing it. This leads to an expected shortening of observed singleton branch i by an amount $d_i(i_0 + i_1)$. Furthermore, we expect to misclassify branch-c doubletons as singletons when we observe zero reads in one of the two individuals and at least two in the other. This leads to an expected lengthening of branch a by an amount $d_c b_0 a_{2+}$ and of branch b by $d_c a_0 b_{2+}$. Finally, we will miss a doubleton entirely when we observe fewer than two reads between the two samples, and we will correctly call doubletons if and only if one or more reads are observed in each sample, an event with probability $(1 - a_0)(1 - b_0)$. Therefore, our observations of the lengths of each branch will be biased by an amount δ_i, with:

$$\delta_a = d_c b_0 a_{2+} - d_a(a_0 + a_1),$$
$$\delta_b = d_c a_0 b_{2+} - d_b(b_0 + b_1),$$
$$\delta_c = -d_c(a_0 + b_0 - a_0 b_0).$$

Under the second assumption, with $\lambda_a = 2.5$ and $\lambda_b = 5$, we have $a_0 b_{2+} \approx 0.08$ and $b_0 + b_1 \approx 0.04$. Therefore, if $d_c > d_b/2$, then $\delta_b > 0$. That is, we expect to infer more SNVs on branch b than its true length, and the number we expect to infer grows linearly with the true length of c. In contrast, because b_0 is small (~0.007), $\delta_a \approx -d_a(a_0 + a_1)$. That is, the number of SNVs we expect to infer on branch a is approximately independent of the true lengths of the other branches.

This example illustrates the fact one cannot simply "correct for missing singletons." Furthermore, as the subtree size and complexity grows, it becomes intractable to model the ensemble of interactions. Consequently, we cannot use the terminal branch lengths of the full tree to estimate the ages of its internal nodes. Instead, we used two alternative approaches to estimate split times, and **Supplementary Table 10** lists point estimates for the major nodes of the tree.

4.5.2 Approach 1: Pruning Sample to Higher-Coverage Sequences

We re-ran all phylogenetic analysis for the three most represented haplogroups: E (n = 323), O (n = 208), and R (n = 331). This time, we restricted to sequences with 5× or greater coverage (**Supplementary Data File 5**), a level that reduced singleton missingness rates but left a sufficient number of descendants to yield reliable estimates for the important subclades. According to the assumptions outlined in **subsection 4.5.1**, singleton missingness at 5× should be under 5%. For each node of interest, we calculated the mean tip-to-root height of the subclade defined by that node and scaled by the mutation period (**subsection 4.5.4**) to estimate the split time.

4.5.3 Approach 2: Traversing Internal Branches

For the less represented haplogroups, we instead used exclusively internal branches. Branches in the interior of the tree have high sequencing coverage, as we effectively sequence an internal branch each time we sequence an individual who descends from it[4]. Consequently, we expect little bias, and we can avoid the complications of low-coverage sequencing by traversing internal branches to estimate split times.

Supplementary Figure 19 outlines a procedure to estimate a split time by measuring the height of a node to be dated, relative to a reference node with a known age. First, we measure the SNV-count distance, d_{ra}, from a reference node, r, to the most recent common ancestor, a, of r and the node to be dated, n. Then, we subtract from d_{ra} the distance between a and n, d_{an}, and convert the resulting height difference, d_{rn}, to an age difference by multiplying by an estimate of the mutation period, μ^{-1}. With t_r representing the known age of the reference node, we estimate the age of n, T_n, as:

$$T_n = d_{rn}\mu^{-1} + t_r.$$

4.5.4 Mutation Rate

To calibrate, we used the Y-chromosome mutation rate estimate of Fu et al. Based on their sequence analysis of the 45,000-year-old Ust'-Ishim sample, they estimated 0.76×10^{-9} SNV mutations per bp per year[13]. Given that there are approximately 10 million callable positions within the 10.3-Mb region we analyzed[4], this estimate corresponds to a mutation period of ~131.6 years per SNV $(0.76 \times 10^{-9}$ SNVs per bp per year $\times 10^7$ positions$)^{-1}$.

The Fu et al. estimate is a bit lower than that of Helgason et al.[14]. Using 274 Icelandic patrilines, Helgason et al. estimated 0.888×10^{-9} per bp per year for X-degenerate sequence. It is important to note that, due to uncertainties in the age of the Ust'-Ishim fossil and other factors, the pedigree-based estimate is more precise. However, it may be less applicable to estimating deep split times, as it averages over a far more recent time-scale, and it is unclear how close the rate of spontaneous mutations is to the rate of accumulation of mutations over evolutionary time periods. In light of proposals that the mutation rate per year has changed over the course of Hominidae evolution[85,86], we chose to use an estimate that incorporates information from deeper history, with the goal of minimizing bias, albeit with greater estimation variance. For comparison, we include in **Supplementary Table 10** split-time point estimates implied by both mutation rate estimates.

We chose the Q-M3 clade (node 319 of **Supplementary Figure 14e**) as the reference point for the traversal-based split-time estimation outlined in **subsection 4.5.3**. Therefore, to set t_r, we had to estimate the T_{MRCA} of Q-M3. To do so, we restricted to the four sequences with at least 7× coverage, yielding predicted singleton missingness below 1%. These lineages (HG01974, HG01977, HG01979, and HG01967) descend independently from Q-M3 and have accumulated 118, 117, 109, and 111 SNVs, respectively, with a sample mean of 113.75. Scaling by the mutation period estimate yields an estimated T_{MRCA} equal to 15.0 ky (113.75 SNVs × 0.1316 ky per SNV).

This estimate, 15.0 ky for the T_{MRCA} of Q-M3, provides a good sanity check for the mutation rate estimate we have chosen to use. Our sample includes 34 Native Americans within this clade. Together, they form a star-like phylogeny (**Supplementary Figure 14e** and **Supplementary Figure 17**) that strongly suggests coincidence with the time of initial human expansion into the Americas, a time that several well-dated archaeological sites[87–89] indicate most likely occurred ~15 kya[87].

Fifteen SNPs separate Q-M3 from its sister clade that also descends from Q-L54. Setting $d_{rn} = 15$ SNPs, $\mu^{-1} = 0.1316$ ky per SNP, and $t_r = 15$ ky yields an estimated Q-L54 split time of 16.9 ky. This value is identical to the point estimate based on a transversions-only analysis of the genome of a Late Pleistocene human from a Clovis burial site[90], providing another sanity check. In prior work[4], we argued that the Q-L54 split was roughly coincident with the peopling of the Americas, but the larger sample herein has enabled us to improve upon this approximation.

5 Functional Annotation

Qasim Ayub, Yuan Chen, Graham Ritchie, Yali Xue, and Chris Tyler-Smith

We used Ensembl's Variant Effect Predictor (VEP)[53] to functionally annotate 60,555 single nucleotide variants (SNVs). Single annotations were obtained for 49,311 variants, whereas 11,244 variants had two or more associated functional annotations. The 14 annotation consequences observed in this dataset were ranked into three categories on the basis of the severity of their expected effect. For each variant with multiple annotations, only the most severe effect was considered (**Supplementary Table 11**).

As expected, the vast majority of the variants were either intergenic or intronic, with no functional effect, and only 159 were coding, two of which cause a severe loss of function (**Supplementary Figure 20**). Rare variants, described here as singletons or doubletons, were significantly enriched in functional annotation categories with severe, moderate or mild effects (**Supplementary Table 12**) ($p = 0.0001$, Fisher's exact test).

We also examined Combined Annotation-Dependent Depletion (CADD) scores (C-scores), which indicate the deleteriousness of SNVs[6] (**Supplementary Figure 21**). We downloaded C-scores[91] and interrogated the variants using custom scripts. Unexpectedly, when using a scaled C-score cut-off of 10 to include SNVs with the highest 10% of C-scores genome-wide, we do not observe an overall enrichment for rare deleterious variants ($p = 0.91$, Fisher's exact test) (**Supplementary Table 13**). This can be attributed to difficulties in Y assembly and alignment and a lack of power for conservation scores on the Y chromosome, as this chromosome has not been sequenced in many species. In addition, there is sparse ENCODE regulatory data for the Y chromosome.

The stop-gain variants were present as singletons in two males, one affecting *AMELY* in haplogroup R2 and the other *USP9Y* in an N1 individual. Approximately one third of the 98 missense variants were identified as "deleterious" by SIFT[7] or "possibly/probably damaging" by PolyPhen[8] (**Supplementary Figure 22**, **Supplementary Data File 6**). Rare deleterious missense variants (singleton or doubleton non-ref allele counts) were significantly enriched on the Y chromosome. Comparison of 94/98 missense variants annotated by all three methods (CADD, PolyPhen, and SIFT) shows a significant enrichment for rare variants designated as "deleterious" by SIFT ($p = 0.001$, Fisher's exact test) or with scaled C-scores greater than or equal to 10 ($p = 0.036$, Fisher's exact test), but not for rare variants annotated as "probably/possibly damaging" by PolyPhen ($p = 0.099$, Fisher's exact test) (**Supplementary Table 14**).

Eight of the 11 variants that affect transcription factor binding are predicted to disrupt the motif for the transcriptional repressor CCCTC-binding factor, CTCF, a zinc finger protein. There is an equal proportion of variants that enhance and destroy CTCF motifs (**Supplementary Figure 23**). The remainder change motifs for *HNF4A* in 5 males and change motifs for *REST* and *GABPA* in two separate individuals, but these are not predicted to disrupt binding.

6 Mitochondrial DNA

Maria Cerezo, G. David Poznik, Apurva Narechania, Shane A. McCarthy, Yali Xue, and Chris Tyler-Smith

6.1 Phylogenetic Analysis

To analyze the mitochondrial genomes (mtDNA) of the 1,244 males, we used the 1000 Genomes Project phase 3 SAMtools[17] callset (**subsection 9.2.3**). Coverage was high, ranging from 42× to 11,834×, with a median of 2,115× and a mean of 2,135×.

We excluded deletions, generated a FASTA file using VCFtools[23], and aligned the sequences to the revised Cambridge Reference Sequence (rCRS) using MEGA6[1]. As recommended by PhyloTree v.16[92], we did not use the following variants for phylogenetic reconstruction due of their rapid mutation rates: 309.1C(C) (an insertion of one or two cytosines after coordinate 309), 315.1C, AC indels at 515–522, 16182C, 16183C, 16193.1C(C), and 16519. We re-assigned heterozygous genotype calls to the more likely of the two nucleotides according to the PhyloTree phylogeny. We then manually added deletions identified in the Boston College callset (**subsection 9.2.3**), as they were more consistent with known phylogenetic placements than those of the SAMtools callset.

We assigned haplogroups to each sample with HaploGrep[93], inferred the mtDNA phylogeny using RAxML[64], and plotted the tree using FigTree[2], manually rotating internal nodes to conform to the canonical representation (**Supplementary Data File 7**).

6.2 Heteroplasmy

As the mean autosomal coverage was ~7×, mismapping nuclear mitochondrial DNAs (NUMTs), if present, could have contributed at most a very low proportion of reads. To yield a conservative but reliable set of heteroplasmy calls, we thresholded the proportion of reads supporting a heteroplasmy at 10%. Especially in light of the mapping quality threshold of `-C50`, this 10% cutoff should be sufficient to exclude possible NUMT contamination for all samples, with the possible exception of the one with the lowest mtDNA coverage (42×), HG03478. We checked this sample and did not observe any overrepresentation.

Using the 10% threshold, we observed 0 to 34 heteroplasmic sites per sample. Some may represent genuine heteroplasmy, and others may indicate the presence of contamination at levels not detected by the Project's standard QC. More than half of the samples (n = 758, 61%) had no heteroplasmy; 305, 108, and 49 individuals had 1, 2, or 3 heteroplasmic sites, respectively. We identified 24 samples with at least four heteroplasmic sites (**Supplementary Table 15**), including two with unusually high counts of 22 and 34. These outlying values could suggest sample contamination, but we did not find evidence for contamination in the corresponding autosomal data. Another possible explanation is cell culture mutation. Since others have extensively studied mtDNA heteroplasmy in different tissues (e.g. Hughes et al.[94]), and most of our samples are from cell lines, we deemed the biology of these heteroplasmies of limited novelty and interest, so we did not investigate further.

7 Diversity

Melissa A. Wilson Sayres, Yuan Chen, and Yali Xue

To compare diversity of the mitochondrial genome (mtDNA) to that of the Y chromosome, we used 141 high-coverage sequences generated with Complete Genomics technology. We observed fairly high ratios of mtDNA diversity to autosomal diversity, with values ranging from ~0.26 to 0.69 (**Supplementary Figure 24a**). In most populations, the observed ratio was much higher than 0.25, the expectation under a neutral model with equal variance in male and female reproductive success. In contrast, on the Y chromosome, we observed ratios that were much lower than the expected 0.25. Similar to previous estimates[10], values ranged from 0.011 to 0.083 (**Supplementary Figure 24b**).

We tested a series of models of a recent male bottleneck that may be able to explain low Y diversity relative to autosomes, high mtDNA diversity relative to autosomes, and only a slight increase in X-chromosome versus autosomal diversity[95]. In these models, we also considered their effect on absolute diversity across the autosomes, for which previous demographic models have been built[11,12]. In brief, we explore whether there is a scenario of recent and severe male bottlenecks that does not also dramatically reduce diversity on the autosomes, and so will be consistent with patterns of diversity across all genomic regions.

7.1 Demographic Model

For African and European demographic histories, we assume population-specific models that have been described in detail elsewhere[11,12]. We added to these models a male-specific bottleneck occurring 4,500 years ago (**Supplementary Figure 25**). The effect of a bottleneck on genetic diversity depends on the ratio of the bottleneck's duration (length) to the effective number of individuals within it (strength), so it is difficult to disentangle the exact values of the length and strength of a bottleneck[11]. In this set of simulations, we set the bottleneck to last for a number of generations equal to the number of males in the bottleneck. This fixed ratio is expected to reduce Y chromosome diversity to a level that is independent of the specific value of the length and strength. However, because the Y chromosome does not evolve independently of the autosomes and X chromosome, we can vary the absolute strength (and length) to attempt to explain the reduction in Y-chromosome diversity in the context of observed diversity across the rest of the genome. In each demographic model, we reduced the effective number of males, from the modern estimate to a small number and then returned it to the modern estimate. We simulated a bottleneck of size 100 Y chromosomes for 100 generations, and repeated for 50, 10, and an extreme value of 1 Y chromosome for 1 generation.

7.2 Effective Population Size

We assessed these different bottlenecks under various assumptions about the long-term effective numbers of males, N_m, and females, N_f. We computed each quantity assuming the effective population size of the autosomes, N_A, remains constant for a given set of population-specific parameters, regardless of the skew in N_m and N_f. Imposing this constraint preserves the fit of the

population-specific demographic models that were derived from autosomal data. We do not apply this assumption at the male bottleneck, for which we allow a reduction in N_m and observe that reduction's effect on the diversity of the autosomes, X, Y, and mtDNA.

We fix N_A because reducing N_m without increasing N_f leads to a severe reduction in N_A that is inconsistent with empirical observations. In previous analyses, the effects on X-, Y-, and mtDNA diversity were qualitatively the same when reducing N_m and keeping N_f fixed, allowing N_A to bottom out[10]. Our primary interests here are the ratios of N_A to the effective population sizes of the other chromosome types: N_{chrX}, N_{chrY}, and N_{mtDNA}. These ratios change equally with variations in N_m/N_f, whether or not N_A is fixed, but fixing N_A allows us to investigate the variations in N_{chrX}, N_{chrY}, and N_{mtDNA} while maintaining a reasonable approximation of N_A.

For given male and female effective population sizes, the effective population sizes for each chromosome type are[96]:

$$\begin{aligned} N_A &= 4N_mN_f / (N_m + N_f) \\ N_{\text{chrX}} &= 9N_mN_f / (4N_m + 2N_f) \\ N_{\text{chrY}} &= N_m/2 \\ N_{\text{mtDNA}} &= N_f/2. \end{aligned}$$

For a fixed ratio of males to females ($R = N_m/N_f$) and a fixed total effective population size, $N_A = 4N_mN_f / (N_m + N_f)$, we can write the male and female effective population sizes as:

$$\begin{aligned} N_f &= N_A (1 + R^{-1}) / 4 \\ N_m &= N_f \cdot R. \end{aligned}$$

Using these equations, we used standard neutral coalescent simulations implemented in the program ms[97] to simulate data for the four chromosome types while varying R, but keeping N_A constant. We keep N_A constant to mimic the real data, as the demographic parameters were originally estimated from autosomal markers.

7.3 Mutation Rates

In **Supplementary Table 16**, we list mutation rates per bp per generation for the autosomes, Y chromosome, and mtDNA. Assuming a generation time of 30 years, the Y-chromosome mutation rate of 0.76×10^{-9} mutations per bp per year equates to 2.3×10^{-8} mutations per bp per year. There is not yet a whole-genome pedigree estimate of the X-chromosome mutation rate, but if mutations are primarily due to errors occurring during replication, then mutation rates on each of the chromosomes are expected to vary with respect to the time spent in the male and female germlines:

$$\begin{aligned} \mu_{\text{A}} &= \tfrac{1}{2}\,\mu_{\text{m}} + \tfrac{1}{2}\,\mu_{\text{f}} \\ \mu_{\text{X}} &= \tfrac{1}{3}\,\mu_{\text{m}} + \tfrac{2}{3}\,\mu_{\text{f}} \\ \mu_{\text{Y}} &= \mu_{\text{m}} \end{aligned}$$

Thus, we use estimates of the autosomal and Y-chromosome mutation rates and assumptions about time spent in the male and female germlines to estimate the mutation rate on the X chromosome. To estimate μ_f, we rearrange the first equation above and substitute values for μ_A and for $\mu_Y = \mu_m$:

$$\begin{aligned}\mu_f &= 2\,\mu_A - \mu_m = 2\,\mu_A - \mu_Y \\ &= 2 \cdot 1.3 \times 10^{-8} - 2.3 \times 10^{-8} \\ &= 0.3 \times 10^{-8} \text{ mutations per bp per generation.}\end{aligned}$$

We then use this value to estimate the mutation rate on the X chromosome:

$$\begin{aligned}\mu_X &= \tfrac{1}{3}\,\mu_m + \tfrac{2}{3}\,\mu_f \\ &= \tfrac{1}{3}\,(2.3 \times 10^{-8}) + \tfrac{2}{3}\,(0.3 \times 10^{-8}) \\ &= 0.97 \times 10^{-8} \text{ mutations per bp per generation.}\end{aligned}$$

7.4 Results and Discussion

Modeling variance in male reproductive success, with an assumption of growth in African populations and a bottleneck in European populations, shows that an extreme variance in male reproductive success can lead to high diversity on the mtDNA relative to the autosomes, increased diversity on the X chromosome, and reduced diversity on the Y[10]. But this long-term reduction in N_m relative to N_f is not sufficient to explain observed levels. However, a recent and extreme bottleneck in the male lineage can further reduce diversity on the Y chromosome relative to the autosomes under all models of variance in male reproductive success (**Supplementary Figure 26**). Some of the models that combine high variance in male reproductive success and a recent extreme male bottleneck are consistent with observed levels of diversity, relative to the autosome, on the X, Y, and mtDNA. Specifically, models assuming N_m/N_f equal to 0.5 or 0.25, for a bottleneck of 50 males, starting 150 generations ago and lasting for 50 generations yielded reasonable values (**Supplementary Figure 27**).

One concern is that reducing the effective number of males will not only affect relative levels of diversity on the X, Y, and mtDNA versus the autosomes but will also reduce the absolute level of diversity on the autosomes. We show that bottlenecks of 1 or 10 males reduce absolute levels of diversity on the autosomes to those much lower than observed, but bottlenecks in the male lineage of 50 males for 50 generations or 100 males for 100 generations are not expected to severely reduce autosomal diversity (**Supplementary Figure 28**).

8 Haplogroup Expansion

Fernando L. Mendez and G. David Poznik

In **section 4.4**, we noted several haplogroups that appear to have diversified rapidly at some point in the past (**Supplementary Figure 14**). To gain insight into these events, we developed a model with two phases of exponential growth, the first of which corresponds with the rapid diversification. This phase could represent either a period in which the haplogroup became common within a population or a period of rapid population growth, which may have been driven by a technological or cultural innovation, social selection, or by the opportunity to expand into a previously uninhabited region. These driving factors may or may not be shared by other lineages. Because such phenomena and their abilities to drive rapid growth are necessarily transitory, we model a second phase of more moderate population growth between the period of rapid growth and a time for which reasonable estimates of historical population sizes exist.

We investigated growth within 10 haplogroups, representing each of the five superpopulations. These included: E1b in Africa; I1 and R1b in Europe; H1, L1, R1a, and R2 in South Asia; O2b and O3 in East Asia; and Q1a in the Americas. **Supplementary Table 17** describes properties of the nodes in which we observed growth. We did not observe clear signals of growth in our analysis of I1 or R2; we had greatest power in haplogroups E1b, R1a, and R1b, as the nodes suggesting rapid growth in these groups left the greatest number of descendants in our sample.

8.1 Inference Framework

8.1.1 Two-Phase Growth Model

We partition the time since the onset of growth into three intervals (**Supplementary Figure 29**): two phases that we model explicitly, followed by the most recent period leading to the present, which we do not model, primarily because very recent growth rates are known to be distinctly greater than those of the past.

Let T_j and N_j be the duration of phase j and the effective number of carriers of the haplogroup at its conclusion. We define the first phase to coincide with an apparent rapid haplogroup expansion, and our primary objective is to infer maximum likelihood values of T_1 and N_1, from which we can compute the growth rate, λ_1, the mean number of sons per man per generation. The role of the second phase is to link the period of rapid expansion to the earliest time for which reasonable estimates exist for the size of the relevant population; the quantities N_2 and T_3 are fixed constraints in our model.

We conduct maximum-likelihood inference over a grid of (T_1, N_1) points. Since N_2 is fixed, for each (T_1, N_1), we need one additional parameter, T_2, in order to specify the full demographic model for simulations of two-phase growth. We can estimate T_2 using the T_{MRCA} of the node of interest, a third fixed constraint. Because it is generally the case that the MRCA of a modern sample would have had closely related contemporaries who experienced the same growth context but who have no living male-line descendants, the first growth phase typically begins prior to the T_{MRCA} of the subtree under investigation. It is therefore convenient to partition the first phase:

$$T_1 = T_b + T_c,$$

with T_c equal to the mean coalescence time of lineages sampled at the end of phase 1 and T_b representing the mean time during which growth occurred prior to the MRCA. We then have:

$$T_{\mathrm{MRCA}} = T_c + T_2 + T_3.$$

With T_3 and T_{MRCA} fixed constraints, an estimate of T_c leads directly to an estimate of T_2.

Each (T_1, N_1) point corresponds to a single λ_1:

$$N_j = N_{j-1} \cdot \lambda_j^{T_j} \Rightarrow \lambda_1 = \sqrt[T_1]{N_j / N_{j-1}},$$

with N_0 = 1, the founder. To estimate T_c and, thereby, T_2 for a given (T_1, N_1), we ran 10,000 ms coalescent simulations[97] with a growth rate of λ_1 and a sample size of 20 chromosomes; the sample size has little influence on the mean coalescence time, except when growth rates are very low. With T_2 and N_2 in hand, we have λ_2 and could therefore simulate two-phase growth to construct a reference distribution of site frequency spectra (SFS) against which to compare the observed data.

8.1.2 Reference Distribution of Site Frequency Spectra

For each phylogenetic node of interest, we analyzed a sequence of pruned subtrees, defining each by a fixed root-to-tip height (number of SNPs) and pruning away all branches whose origins are greater than this number of SNPs downstream of the subtree root (**Supplementary Figure 29**). We consider heights (h) ranging from as few as 3 to as many 12 SNPs. Each height corresponds to a "sampling" time, with the age of the subtree at the time of sampling given by:

$$T_s = h\mu^{-1},$$

where μ^{-1}, the inverse of the mutation rate, is the mutation period—the number of generations represented by each SNP.

There are two important advantages to confining our attention to these internal regions of the tree. First, doing so reduces the impact of missing data, especially unobserved singletons, as internal braches have high coverage. Second, it reduces the effect on the genealogy of recent population structure and regional expansions, as the pruned subtrees are largely agnostic to recent phenomena.

We assembled a reference distribution of site frequency spectra for each point of a three-dimensional lattice of (T_1, N_1, T_s) values, allowing T_1 to range from 1 to 48 generations and distributing 32 N_1 values in a geometric progression between 13.6 and 200,000 individuals, with each value approximately 36% greater than the previous. With up to ten possible T_s values, the lattice contained up to 15,360 (48·32·10) points, and for each, we conducted 16,384 (2^{14}) ms simulations of two-phase growth, fixing the number of lineages equal to that of the pruned observed tree (**Supplementary Table 18**).

For a point on the lattice, T_s may or may not exceed the corresponding T_c. When it does, we simulate phase-1 growth, with rate λ_1, to last T_1 generations and phase-2 growth to endure from end of phase 1 to the time of sampling: (T_s – T_c) generations. When T_s < T_c, sampling occurs

prior to the end of phase 1. In this case, we simulate growth at rate λ_1 lasting for (T_b + T_s) generations and do not simulate the second phase of growth.

When the coalescence time of the simulated tree differed from T_s by more than one generation, we rejected the simulation and repeated. Doing so ensures that the reference distribution of SFS is consistent with the model specification. The rejection rate was low because the coalescence time of an exponentially growing population has a small variance.

We computed the SFS for each simulation, adjusting the number of singletons to achieve uniform root-to-tip height among lineages, in accord with the pruned subtrees. We use this summary statistic for likelihood-based inference. For each level of the lattice (i.e., for each T_s value), we infer joint confidence intervals for T_1 and N_1 by comparing the frequency spectrum of the observed pruned subtree to the reference distribution of spectra obtained from the simulated genealogies.

8.1.3 Distance Measure for Site Frequency Spectra

We defined a distance measure to compare the frequency spectrum of a pruned observed tree to those of the genealogies simulated with the same number of lineages and number of SNPs per lineage. For spectra f and g, each of length (n – 1), we compute a vector of differences,

$$v_{f,g} = f - g,$$

and a vector of reverse-cumulative differences, $\boldsymbol{w_{f,g}}$, with:

$$\begin{aligned}
(w_{f,g})_{n-1} &= (v_{f,g})_{n-1} \\
(w_{f,g})_{n-2} &= (w_{f,g})_{n-1} + (v_{f,g})_{n-2} \\
&\cdots \\
(w_{f,g})_i &= \sum_{j=i}^{n-1} (v_{f,g})_j
\end{aligned}$$

We then define the distance between f and g as the L^1 norm of the truncated vector $(w_{f,g})_{2\ldots(n-1)}$:

$$d(f,g) = \sum_{i=2}^{n-1} \left|(w_{f,g})_i\right| .$$

This function has the desirable property that it is small for pairs of spectra with similar frequencies. As a consequence, it is robust to minor differences in genealogies that may arise from the random sampling of lineages.

The distance function does not consider singletons because the singleton counts are constrained both in the observed subtree and in the simulations; we pruned the observed subtrees such that the number of SNPs is the same for all lineages, and we adjusted the number of singletons in the simulations to conform to this constraint.

It is clear that d is non-negative and symmetric and that $d(f, g) = 0$ if and only if $f = g$. To demonstrate that d obeys the triangle inequality and is therefore a proper distance measure, we note that:

$$v_{f,h} = v_{f,g} + v_{g,h}$$

and that:

$$\begin{aligned}(w_{f,h})_i &= \sum_{j=1}^{n-1}(v_{f,h})_j = \sum_{j=1}^{n-1}(v_{f,g} + v_{g,h})_j = \sum_{j=1}^{n-1}(v_{f,g})_j + \sum_{j=1}^{n-1}(v_{g,h})_j \\ &= (w_{f,g})_i + (w_{g,h})_i.\end{aligned}$$

Therefore,

$$\begin{aligned}d(f,h) = \sum_{i=2}^{n-1}|(w_{f,h})_i| &= \sum_{i=2}^{n-1}|(w_{f,g})_i + (w_{g,h})_i| \\ &\leq \sum_{i=2}^{n-1}|(w_{f,g})_i| + \sum_{i=2}^{n-1}|(w_{g,h})_i| = d(f,g) + d(g,h).\end{aligned}$$

8.1.4 Inference

Consider a single sampling height corresponding to one level of the three-dimensional parameter lattice. Using the SFS as a summary statistic, we could approximate the likelihood of a particular (T_1, N_1) point of the grid by calculating the proportion of simulations that yield spectra identical to that of the observed tree (i.e., d(simulated, observed) = 0). For robustness to noise, and for computational tractability, we instead deem the SFS of a simulated genealogy to "match" that of the observed genealogy if the distance between them was within the lowermost 0.5% tail of the ~25 million distances computed over the grid—16,384 simulations at each of 1,536 (48·32) grid points. We estimate the likelihood of a (T_1, N_1) pair as the fraction of simulations whose SFS "matched" that of the observed tree according to this definition. We then plot joint confidence bounds and marginalize, using the likelihood ratio criterion with one degree of freedom to estimate confidence bounds independently for T_1, N_1, and λ_1. When our estimate of λ_1 is significantly greater than the haplogroup's average growth rate over the first two phases, we reject the null hypothesis of single-phase growth, $\lambda_1 = \lambda_2$.

We defined 95% confidence intervals based on the asymptotic likelihood ratio criterion (**Supplementary Figure 30**), but a number of assumptions and approximations could potentially impact our inference. These include constraining the root-to-tip heights of trees and conditioning simulations on T_{MRCA}. To test whether our inference procedure yields appropriate coverage probabilities, we simulated 900 trees for each of 12 sets of T_1, N_1, and n values corresponding to inferred expansions. We then filtered the simulations to those for which the number of branches at a representative sampling time equaled the number in the corresponding real data. We then conducted inference as described above. Among the 101 simulations that remained, the inferred 95% confidence interval contained the true growth rate 93 times (**Supplementary Data File 8a**),

giving an estimated coverage probability of 92% (95% CI: 85–97%). In each of the 8 instances in which the true growth rate was falsely rejected, it lay slightly below the CI lower bound.

8.1.5 Considerations

In this subsection, we discuss five factors that influence analysis: the height of a pruned subtree, the assumption that growth predates the MRCA, sensitivity to SNPs incompatible with the tree, sensitivity to the mutation rate parameter, and estimation of and sensitivity to N_2.

First, inference for a given node may vary with the height of the pruned subtree. Although taller subtrees may contain more branches and therefore have greater power, they are also more likely to bear the influence of population structure. For instance, within node 71 of E1b, subtrees of heights greater than 10 contain branches specific to each of three populations: YRI, ACB, and LWK. Consequently, we determined the tree heights appropriate for analysis on a case-by-case basis.

A second consideration is that our model assumes that the onset of growth predated the MRCA. However, under this assumption, we may falsely detect a signal of growth in a node that predates the onset of expansion, but which has descendants that experienced growth, as similar patterns of genetic diversity could emerge from the two scenarios. R1a node 206 serves as an illustrative example. Of the two branches descending from this node, one leads to a subtree, rooted at node 204, with 22 representatives, and the other is a single lineage carried by HG03911 alone (branch 205). Since branch 204 is short, with just 2 SNPs, the subtrees rooted at 204 and 206 yield highly similar growth inferences. However, if the lineage defined by node 206 had been growing exponentially, it is unlikely that we would have observed such an extreme asymmetry in its descendants. Rather, it is more likely that growth commenced after the time of node 206 but within the subsequent interval during which our inference method is sensitive to growth.

Third, inference may be sensitive to the presence of common SNPs not correctly assigned to the appropriate internal branches of the phylogeny. Misassignment can occur due to recurrent or reversion mutations, genotyping errors, or incorrect reconstruction of the tree topology. We minimized tree reconstruction error through manual curation of the phylogeny (**section 4.3**).

Fourth, growth-rate estimates depend on the mutation-rate parameter, which influences both the coalescent simulations and the estimated T_{MRCA}. An under- or over-estimate of μ would lead to a corresponding under- or over-estimate of the growth rate.

Finally, our model requires estimates of the number of carriers of each haplogroup at some point prior to the onset of the extraordinary recent growth experienced across the world. To estimate N_2 for a given haplogroup, we used rough population-size estimates from the literature and scaled to account for the frequency of the haplogroup within the population. We also scaled by a factor of one-fifth to account for the facts that males generally represent approximately half of the population and that most males do not contribute to the Y-chromosome pool and effective population size.

N_2 estimates affect inference of λ_2. However, the degree of influence is mitigated by the fact that one or both of N_1 and T_2 are generally large. When N_1 is large, phase 2 is marked by a low

coalescence rate, leading to trees that are largely independent of N_2, and when T_2 is large, large errors in N_2 correspond to small errors of λ_2.

For subgroups of E1b1a, we used an estimated sub-Saharan African population of ~11 million by 1 A.D.[98] (cited in Durand et al.[99]). Of these ~11 million individuals, we estimate that ~30% lived in the area containing modern-day Nigeria and Sierra Leone, the most likely origin of the branches analyzed. For haplogroups R1b and I1, we use Beloch's estimate of 23 million individuals living in the European portion of the Roman Empire by 1 A.D.[100] (cited in Durand et al.[99]), and for haplogroups R1a, H1, L1 and R2 in South Asia, we use the 1880 census size of ~255 million[101]. For haplogroup O3 in East Asia, we used a figure of ~60 million individuals, based on a Chinese census of 2 A.D.[102]. For haplogroup O2b in Japan, we used an estimate of ~5 million individuals in the year 800 A.D.[103], and for haplogroup Q1a, we used a figure of 6 million individuals, the geometric mean of the upper and lower estimates cited in Snow et al.[104].

8.2 Results

For each node, we analyzed up to ten sampling heights, and we summarize results by combining confidence intervals across these analyses (**Supplementary Table 19**). We have plotted likelihood contours for a subset of E1b, R1b, and R1a nodes in **Supplementary Figure 31** and for all nodes in **Supplementary Data File 8b**.

8.2.1 Africa

Haplogroup E1b

Within African haplogroup E1b (**Supplementary Figure 14a**), we observed signals of expansion in nodes 71, 95 and 384 (**Supplementary Figure 31a**). Node 95 exhibits levels of growth of between 22% and 143% per generation for subtrees of height 8 or greater. An important component of this signal is due to very rapid growth in a descendant, node 71, whose growth exceeds 40% per generation. Subtrees rooted at node 384 and of height 8 or greater also exhibit a signal of growth, with rates ranging from 20% to 70% per generation. We inferred similar growth parameters for nodes 95 and 384, and the two have very similar estimated ages of about 5,000 years, so both may reflect the same event (**Supplementary Table 19**). There is no clear signal of growth in the much older node 388, for which the model may be a poor fit.

8.2.2 Europe

Haplogroup R1b

We observed evidence of growth in each of the six European R1b nodes (**Supplementary Figure 14e**) that we analyzed: 189, 276, 343, 347, 357, and 417 (**Supplementary Figure 31b**). We inferred the growth rate of the subtree rooted at 347 to be 19% to 670% per generation, but because this node differs from node 343 only by its inclusion of two additional samples (HG02014 and HG00243), the growth we observed in 347 may be entirely due to that of 343. As was the case for node 71 of haplogroup E1b, the duration of phase 1 was small, and the growth rate was large.

Haplogroup I1

In haplogroup I1 (**Supplementary Figure 14c**), we observed frequency spectra consistent with a single phase of growth. This may be due to the fact that our approach has reduced power to detect low growth rates, especially with small samples. In addition, our inference for this haplogroup may be more sensitive to assumed values of N_2 and the T_{MRCA}, given its recent age.

8.2.3 South Asia

Haplogroup R1a

In the South Asian portion of haplogroup R1a (**Supplementary Figure 14e**), nodes 161 and 204 both exhibit signals of growth, with rates ranging from 20% to 90% and 34% to 345% per generation, respectively (**Supplementary Figure 31c** and **Supplementary Table 19**). Upon combining information across sampling heights, we can reject values of T_1 below 10 generations for each node, and in both cases, the upper bounds of the T_1 confidence intervals were outside the explored range. In both subtrees, we infer that N_1 exceeded 300 and may have attained a value on the order of 10^5.

The R1a subtree rooted at node 206 includes the subtree rooted at node 204 plus one additional lineage. Since branch 204 is short, with just 2 SNPs, the two subtrees yield highly similar growth inferences. Similarly, with large sampling heights, we recover a signal of growth for the subtree rooted at node 213, which includes both subtrees 161 and 206, each of which have short roots. The tree rooted at node 214 is not exclusive to South Asia.

Haplogroup H1

We observed evidence of growth in the subtrees rooted at nodes 66 and 94 of haplogroup H1 (**Supplementary Figure 14b**). The observed signal was relatively weak in the node-66 subtree; at ~10%, it was barely enough to reject a single phase. However, the tree associated with node 94 exhibits a much stronger signal of at least 55%, with an N_1 of at least 220. We did not observe a signal of growth in the subtrees rooted at nodes 95, 97, 98 or 99.

Haplogroup L1

As for haplogroup I1, we observed a relatively weak signal of growth in L1 (**Supplementary Figure 14d**). We did not reject single-phase growth, however we had reduced power due to small sample size.

Haplogroup R2

Again, we cannot reject a single phase of growth for haplogroup R2 (**Supplementary Figure 14e**), which has an older T_{MRCA} and a lower estimated average growth rate than L1.

8.2.4 East Asia

Haplogroup O2b

We observe a phase-1 growth rate of at least 17% within the branch-160 subtree of haplogroup O2b (**Supplementary Figure 14d**), a lineage restricted to Japan. This rate is sufficiently great to reject single-phase growth.

Haplogroups O3

The subhaplogroup of O3 defined by node 225 (**Supplementary Figure 14d**) exhibits evidence of growth in East Asia, with a rate of at least 22% per generation.

8.2.5 The Americas

Haplogroup Q1a

Haplogroup Q1a is associated with Native-American populations (**section 4.1**). In particular, branch 319 (**Supplementary Figure 14e**), marked by the M3 mutation, occurs exclusively in the Americas and Siberia[105]. We observed rapid growth of at least 40% per generation in this subtree, and we can definitively reject a single phase of growth. The period of rapid growth appears to have been relatively brief; trees with more recent "sampling" times point to a T_1 of fewer than 20 generations. However, population clustering within our sample suggests that the assumption of no population structure may not be valid for the full duration of the first phase. Unfortunately, trees based on more ancient sampling times have reduced power to detect the transition between the two phases of growth.

8.3 Conclusions

Our haplogroup expansion analyses led to three key observations. First, we note that several haplogroups experienced growth consistent with two exponential phases. This finding lends insight into the causes of these expansions, as it implies that the driving processes changed over time. Haplogroup-specific expansion could in principle be driven by natural selection, social selection, or differential growth rates among subpopulations. However, our findings indicate that it is unlikely natural selection played a key role in the expansions or in the concomitant drop in genetic diversity within continental populations. Were natural selection a principal driver, the growth rate differences between selected and non-selected haplogroups would probably have persisted for more than a few tens of generations. But, in contrast, we observed that the first phase of growth was generally brief and marked by a far greater growth rate than the second phase. Furthermore, if large intra-population differences in haplogroup growth rates were primarily due to selection, and some portion of the differential selection persisted through time, we would expect the non-selected haplogroups to have been crowded out and exist at low frequencies, if at all. Instead, we observed that slowly-growing haplogroups with ancient diversification coexist alongside recently expanded ones.

Second, we inferred the presence of several explosive expansions, including those in the subtrees rooted at branch 71 of haplogroup E1b (U290) and branches 189 (DF27), 276 (U152), and 343 (DF13) of R1b. Though it is possible that strong social selection may yield such an effect, as has been proposed[106], the growth signals we observed in most cases were best explained by strong and sustained expansions rather than by explosive growth lasting very few generations.

Third, we observed multiple subtrees with similar inferred ages and growth rates within each of E1b (nodes 95 and 384), R1a (161 and 204), and R1b (189 and 276). These similarities may reflect shared demographic events.

In addition to the considerations cited in **section 8.1**, five assumptions and approximations may have influenced our inference. First, we assumed a two-phase growth model in which all lineages descending from a given node expand at the same rate. Though this model enabled us to hone in on signals of rapid growth, it may oversimplify historical demographic changes. Second, we assumed that population structure did not affect the branching patterns we observed immediately downstream of the nodes of interest. Third, to control for variances in coalescence times we conditioned on the inferred T_{MRCA} of each node. This assumption was likely conservative in that it led to overestimated likelihoods of very small growth rates. Fourth, we assumed that the number of mutations accumulated in a branch is a good proxy for its length measured in generations. This assumption may be problematic for small sampling times with small root-to-tip heights, but, on the other hand, using larger sampling times would also be problematic, as the assumed lack of population structure would be less likely to hold. Finally, though our estimation makes use of the likelihood ratio criterion, for computational efficiency we used an approximate likelihood estimated through coalescent simulations and a distance defined between frequency spectra. Though we did not thoroughly explore the general validity of this approach, it worked well with simulated data.

9 Data Availability

G. David Poznik

9.1 Supplementary Data File

A zipped archive of supporting data is available on the journal's website, as well as on the 1000 Genomes Project FTP site (**section 9.2**). The archive includes the following files and subdirectories:

`0.README.txt`
Details the contents and format of each file

`1.CNV.summary.txt`
Summary of inferred CNV mutation events

`2.STR.summary.txt`
Point estimates and confidence intervals for Y-STR mutation rates

`3.haplogroups.txt`
Short-form and long-form haplogroup calls for each individual

`4a.ML.tree/`
Total-evidence maximum-likelihood tree

`4b.rooted.tree/`
Tree based on restricted regions, with chimpanzee outgroup

`5.snp.to.branch.mappings/`

`branches/`
Files mapping each branch to a set of descendants

`snps/`
Files indicating for each branch, which SNPs map to it

`subset.with.5x+.coverage/`
As above, but with the sample pruned to sequences with 5× or greater coverage

`6.functional.analyses.xlsx`
Summary of functional analysis

`7.mtDNA.tree/`
Total evidence maximum-likelihood tree of male mtDNAs

`8.expansions/`
Likelihood contours for simulations and analysis

9.2 FTP Site

9.2.1 Information

A full description of data management and community access can be found in Clarke et al.[107].

1000 Genomes Project FTP sites

Europe: ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/
USA: ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/

Tutorials for access and use

http://www.1000genomes.org/using-1000-genomes-data

Email

Support for using the 1000 Genomes Project data can be obtained via email:

info@1000genomes.org

9.2.2 Sequence Read Alignments (BAM Files)

Alignments to the GRCh37 Reference Sequence

Index File

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/
20130502.phase3.low_coverage.alignment.index

Main Directory

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/

Example Full Path

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00160/alignment/
HG00160.mapped.ILLUMINA.bwa.GBR.low_coverage.20120522.bam

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3/data/HG00160/alignment/
HG00160.mapped.ILLUMINA.bwa.GBR.low_coverage.20120522.bam.bai

Alignments to the GRCh38 Reference Sequence (Used in STR Analysis)

Main Directory

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/
1000_genomes_project/data

Example Full Path

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/
1000_genomes_project/data/ACB/HG01890/alignment/
HG01890.alt_bwamem_GRCh38DH.20150718.ACB.low_coverage.cram

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/
1000_genomes_project/data/ACB/HG01890/alignment/
HG01890.alt_bwamem_GRCh38DH.20150718.ACB.low_coverage.cram.crai

9.2.3 Genotype Calls (VCF Files)

SNVs

Unimputed

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC.20130502.60555_biallelic_snps.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC.20130502.60555_biallelic_snps.vcf.gz.tbi

Phylogenetically Imputed

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC_phyloImputedV5.20130502.60555_biallelic_snps.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC_phyloImputedV5.20130502.60555_biallelic_snps.vcf.gz.tbi

Indels and MNVs

Unimputed

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC.20130502.biallelic_indelsAndMNPs.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC.20130502.biallelic_indelsAndMNPs.vcf.gz.tbi

Phylogenetically Imputed

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC_phyloImputedV5.20130502.biallelic_indelsAndMNPs.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY_10Mbp_mask.glia_freebayes_maxLikGT_siteQC_phyloImputedV5.20130502.biallelic_indelsAndMNPs.vcf.gz.tbi

CNVs

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/README_ALL.chrY.phase3_cnv_broad_genome_strip.20130502.cnv.low_coverage.genotypes.copy.txt

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY.phase3_cnv_broad_genome_strip.20130502.cnv.low_coverage.genotypes.copy.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY.phase3_cnv_broad_genome_strip.20130502.cnv.low_coverage.genotypes.copy.vcf.gz.tbi

STRs

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY.HipSTR.20130502.STRs.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/chrY/ALL.chrY.HipSTR.20130502.STRs.vcf.gz.tbi

MtDNA

SAMtools callset

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130723_phase3_wg/si/ALL.chromMT.samtools.20130502.snps_indels.low_coverage.genotypes.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130723_phase3_wg/si/ALL.chromMT.samtools.20130502.snps_indels.low_coverage.genotypes.vcf.gz.tbi

Boston College callset

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130723_phase3_wg/bc/ALL.chrMT.bc_haplotypes_3bp_1pct.20130502.low_coverage.genotypes.vcf.gz

ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/technical/working/20130723_phase3_wg/bc/ALL.chrMT.bc_haplotypes_3bp_1pct.20130502.low_coverage.genotypes.vcf.gz.tbi

10 The 1000 Genomes Project Consortium

Participants are arranged by project role, then by institution alphabetically, and finally alphabetically within institutions except for Principal Investigators and Project Leaders, as indicated.

Corresponding Authors: Adam Auton[1], Gonçalo R. Abecasis[2]

Steering Committee: David M. Altshuler[3] (Co-Chair), Richard M. Durbin[4] (Co-Chair), Gonçalo R. Abecasis[2], David R. Bentley[5], Aravinda Chakravarti[6], Andrew G. Clark[7], Peter Donnelly[8,9], Evan E. Eichler[10,11], Paul Flicek[12], Stacey B. Gabriel[13], Richard A. Gibbs[14], Eric D. Green[15], Matthew E. Hurles[4], Bartha M. Knoppers[16], Jan O. Korbel[12,17], Eric S. Lander[13], Charles Lee[18,19], Hans Lehrach[20,21], Elaine R. Mardis[22], Gabor T. Marth[23], Gil A. McVean[8,9], Deborah A. Nickerson[10], Jeanette P. Schmidt[24], Stephen T. Sherry[25], Jun Wang[26-30], Richard K. Wilson[22]

Production Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator)[14], Eric Boerwinkle[14], Harsha Doddapaneni[14], Yi Han[14], Viktoriya Korchina[14], Christie Kovar[14], Sandra Lee[14], Donna Muzny[14], Jeffrey G. Reid[14], Yiming Zhu[14], **BGI-Shenzhen** Jun Wang (Principal Investigator)[26-30], Yuqi Chang[26], Qiang Feng[26,27], Xiaodong Fang[26,27], Xiaosen Guo[26,27], Min Jian[26,27], Hui Jiang[26,27], Xin Jin[26], Tianming Lan[26], Guoqing Li[26], Jingxiang Li[26], Yingrui Li[26], Shengmao Liu[26], Xiao Liu[26,27], Yao Lu[26], Xuedi Ma[26], Meifang Tang[26], Bo Wang[26], Guangbiao Wang[26], Honglong Wu[26], Renhua Wu[26], Xun Xu[26], Ye Yin[26], Dandan Zhang[26], Wenwei Zhang[26], Jiao Zhao[26], Meiru Zhao[26], Xiaole Zheng[26], **Broad Institute of MIT and Harvard** Eric S. Lander (Principal Investigator)[13], David M. Altshuler[3], Stacey B. Gabriel (Co-Chair)[13], Namrata Gupta[13], **Coriell Institute for Medical Research** Neda Gharani[31], Lorraine H. Toji[31], Norman P. Gerry[31], Alissa M. Resch[31], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Jonathan Barker[12], Laura Clarke[12], Laurent Gil[12], Sarah E. Hunt[12], Gavin Kelman[12], Eugene Kulesha[12], Rasko Leinonen[12], William M. McLaren[12], Rajesh Radhakrishnan[12], Asier Roa[12], Dmitriy Smirnov[12], Richard E. Smith[12], Ian Streeter[12], Anja Thormann[12], Iliana Toneva[12], Brendan Vaughan[12], Xiangqun Zheng-Bradley[12], **Illumina** David R. Bentley (Principal Investigator)[5], Russell Grocock[5], Sean Humphray[5], Terena James[5], Zoya Kingsbury[5], **Max Planck Institute for Molecular Genetics** Hans Lehrach (Principal Investigator)[20,21], Ralf Sudbrak (Project Leader)[32], Marcus W. Albrecht[33], Vyacheslav S. Amstislavskiy[20], Tatiana A. Borodina[33], Matthias Lienhard[20], Florian Mertes[20], Marc Sultan[20], Bernd Timmermann[20], Marie-Laure Yaspo[20], **McDonnell Genome Institute at Washington University** Elaine R. Mardis (Co-Principal Investigator) (Co-Chair)[22], Richard K. Wilson (Co-Principal Investigator)[22], Lucinda Fulton[22], Robert Fulton[22], **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[25], Victor Ananiev[25], Zinaida Belaia[25], Dimitriy Beloslyudtsev[25], Nathan Bouk[25], Chao Chen[25], Deanna Church[34], Robert Cohen[25], Charles Cook[25], John Garner[25], Timothy Hefferon[25], Mikhail Kimelman[25], Chunlei Liu[25], John Lopez[25], Peter Meric[25], Chris O'Sullivan[35], Yuri Ostapchuk[25], Lon Phan[25], Sergiy Ponomarov[25], Valerie Schneider[25], Eugene Shekhtman[25], Karl Sirotkin[25], Douglas Slotta[25], Hua Zhang[25], **University of Oxford** Gil A. McVean (Principal Investigator)[8,9], **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[4], Senduran Balasubramaniam[4], John Burton[4], Petr Danecek[4], Thomas M. Keane[4], Anja Kolb-Kokocinski[4], Shane McCarthy[4], James Stalker[4], Michael Quail[4]

Analysis Group: Affymetrix Jeanette P. Schmidt (Principal Investigator)[24], Christopher J. Davies[24], Jeremy Gollub[24], Teresa Webster[24], Brant Wong[24], Yiping Zhan[24], **Albert Einstein College of Medicine** Adam Auton (Principal Investigator)[1], Christopher L. Campbell[1], Yu Kong[1], Anthony Marcketta[1] **Baylor College of Medicine** Richard A. Gibbs (Principal Investigator)[14], Fuli Yu (Project Leader)[14], Lilian Antunes[14], Matthew Bainbridge[14], Donna Muzny[14], Aniko Sabo[14], Zhuoyi Huang[14] **BGI-Shenzhen** Jun Wang (Principal Investigator)[26-30], Lachlan J.M. Coin[26], Lin Fang[26,27], Xiaosen Guo[26], Xin Jin[26], Guoqing Li[26], Qibin Li[26], Yingrui Li[26], Zhenyu Li[26], Haoxiang Lin[26], Binghang Liu[26], Ruibang Luo[26], Haojing Shao[26], Yinlong Xie[26], Chen Ye[26], Chang Yu[26], Fan Zhang[26], Hancheng Zheng[26], Hongmei Zhu[26], **Bilkent University** Can Alkan[36], Elif Dal[36], Fatma Kahveci[36], **Boston College** Gabor T. Marth (Principal Investigator)[23], Erik P. Garrison (Project Lead)[4], Deniz Kural[37], Wan-Ping Lee[37], Wen Fung Leong[38], Michael Stromberg[39], Alistair N. Ward[23], Jiantao Wu[39], Mengyao Zhang[40], **Broad Institute of MIT and Harvard** Mark J. Daly (Principal Investigator)[13], Mark A. DePristo (Project Leader)[41], Robert E. Handsaker (Project Leader)[13,40], David M. Altshuler[3], Eric Banks[13], Gaurav Bhatia[13], Guillermo del Angel[13], Stacey B. Gabriel[13], Giulio Genovese[13], Namrata Gupta[13], Heng Li[13], Seva Kashin[13,40], Eric S. Lander[13], Steven A. McCarroll[13,40], James C.

Nemesh[13], Ryan E. Poplin[13], **Cold Spring Harbor Laboratory** Seungtai C. Yoon (Principal Investigator)[42], Jayon Lihm[42], Vladimir Makarov[43], **Cornell University** Andrew G. Clark (Principal Investigator)[7], Srikanth Gottipati[44], Alon Keinan[7], Juan L. Rodriguez-Flores[45], **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator)[12,17], Tobias Rausch(Project Leader)[17,46], Markus H. Fritz[46], Adrian M. Stütz[17], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Kathryn Beal[12], Laura Clarke[12], Avik Datta[12], Javier Herrero[47], William M. McLaren[12], Graham R.S. Ritchie[12], Richard E. Smith[12], Daniel Zerbino[12], Xiangqun Zheng-Bradley[12], **Harvard University** Pardis C. Sabeti (Principal Investigator)[13,48], Ilya Shlyakhter[13,48], Stephen F. Schaffner[13,48], Joseph Vitti[13,49], **Human Gene Mutation Database** David N. Cooper (Principal Investigator)[50], Edward V. Ball[50], Peter D. Stenson[50], **Illumina** David R. Bentley (Principal Investigator)[5], Bret Barnes[39], Markus Bauer[5], R. Keira Cheetham[5], Anthony Cox[5], Michael Eberle[5], Sean Humphray[5], Scott Kahn[39], Lisa Murray[5], John Peden[5], Richard Shaw[5], **Icahn School of Medicine at Mount Sinai** Eimear E. Kenny (Principal Investigator)[51]**, Louisiana State University** Mark A. Batzer (Principal Investigator)[52], Miriam K. Konkel[52], Jerilyn A. Walker[52], **Massachusetts General Hospital** Daniel G. MacArthur (Principal Investigator)[53], Monkol Lek[53], **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Leader)[32], Vyacheslav S. Amstislavskiy[20], Ralf Herwig[20], **McDonnell Genome Institute at Washington University** Elaine R. Mardis (Co-Principal Investigator)[22], Li Ding[22], Daniel C. Koboldt[22], David Larson[22], Kai Ye[22], **McGill University** Simon Gravel[54], **National Eye Institute, NIH** Anand Swaroop[55], Emily Chew[55], **New York Genome Center** Tuuli Lappalainen (Principal Investigator)[56,57], Yaniv Erlich (Principal Investigator)[56,58], Melissa Gymrek[13,56,59,60], Thomas Frederick Willems[61], **Ontario Institute for Cancer Research** Jared T. Simpson[62], **Pennsylvania State University** Mark D. Shriver (Principal Investigator)[63], **Rutgers Cancer Institute of New Jersey** Jeffrey A. Rosenfeld (Principal Investigator)[64], **Stanford University** Carlos D. Bustamante (Principal Investigator)[65], Stephen B. Montgomery (Principal Investigator)[66], Francisco M. De La Vega (Principal Investigator)[65], Jake K. Byrnes[67], Andrew W. Carroll[68], Marianne K. DeGorter[66], Phil Lacroute[65], Brian K. Maples[65], Alicia R. Martin[65], Andres Moreno-Estrada[65,69], Suyash S. Shringarpure[65], Fouad Zakharia[65], **Tel-Aviv University** Eran Halperin (Principal Investigator)[70-72], Yael Baran[70], **The Jackson Laboratory for Genomic Medicine** Charles Lee (Principal Investigator)[18,19], Eliza Cerveira[18], Jaeho Hwang[18], Ankit Malhotra (Co-Project Lead)[18], Dariusz Plewczynski[18], Kamen Radew[18], Mallory Romanovitch[18], Chengsheng Zhang (Co-Project Lead)[18], **Thermo Fisher Scientific** Fiona C.L. Hyland[73], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Alexis Christoforides[74], Nils Homer[75], Tyler Izatt[74], Ahmet A. Kurdoglu[74], Shripad A. Sinari[74], Kevin Squire[76], **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[25], Chunlin Xiao[25], **University of California, San Diego** Jonathan Sebat (Principal Investigator)[77,78], Danny Antaki[77], Madhusudan Gujral[77], Amina Noor[77], Kenny Ye[79], **University of California, San Francisco** Esteban G. Burchard (Principal Investigator)[80], Ryan D. Hernandez (Principal Investigator)[80-82], Christopher R. Gignoux[80], **University of California, Santa Cruz** David Haussler (Principal Investigator)[83,84], Sol J. Katzman[83], W. James Kent[83], **University of Chicago** Bryan Howie[85], **University College London** Andres Ruiz-Linares (Principal Investigator)[86], **University of Geneva** Emmanouil T. Dermitzakis (Principal Investigator)[87-89], **University of Maryland School of Medicine** Scott E. Devine (Principal Investigator)[90], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator) (Co-Chair)[2], Hyun Min Kang (Project Leader)[2], Jeffrey M. Kidd (Principal Investigator)[91,92], Tom Blackwell[2], Sean Caron[2], Wei Chen[93], Sarah Emery[92], Lars Fritsche[2], Christian Fuchsberger[2], Goo Jun[2,94], Bingshan Li[95], Robert Lyons[96], Chris Scheller[2], Carlo Sidore[2,97,98], Shiya Song[91], Elzbieta Sliwerska[92], Daniel Taliun[2], Adrian Tan[2], Ryan Welch[2], Mary Kate Wing[2], Xiaowei Zhan[99] **University of Montréal** Philip Awadalla (Principal Investigator)[62,100], Alan Hodgkinson[100], **University of North Carolina at Chapel Hill** Yun Li[101], **University of North Carolina at Charlotte** Xinghua Shi (Principal Investigator)[102], Andrew Quitadamo[102], **University of Oxford** Gerton Lunter (Principal Investigator)[8], Gil A. McVean (Principal Investigator) (Co-Chair)[8,9], Jonathan L. Marchini (Principal Investigator)[8,9], Simon Myers (Principal Investigator)[8,9], Claire Churchhouse[9], Olivier Delaneau[9,87], Anjali Gupta-Hinch[8], Warren Kretzschmar[8], Zamin Iqbal[8], Iain Mathieson[8], Androniki Menelaou[9,103], Andy Rimmer[87], Dionysia K. Xifara[8,9], **University of Puerto Rico** Taras K. Oleksyk (Principal Investigator)[104], **University of Texas Health Sciences Center at Houston** Yunxin Fu (Principal Investigator)[94], Xiaoming Liu[94], Momiao Xiong[94], **University of Utah** Lynn Jorde (Principal Investigator)[105], David Witherspoon[105], Jinchuan Xing[106], **University of Washington** Evan E. Eichler (Principal Investigator)[10,11], Brian L. Browning (Principal Investigator)[107], Sharon R. Browning (Principal Investigator)[108], Fereydoun Hormozdiari[10], Peter H. Sudmant[10], **Weill Cornell Medical College**, Ekta Khurana (Principal Investigator)[109], **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[4], Matthew E. Hurles (Principal Investigator)[4], Chris Tyler-Smith (Principal Investigator)[4], Cornelis A. Albers[110,111], Qasim Ayub[4], Senduran Balasubramaniam[4], Yuan Chen[4], Vincenza Colonna[4,112], Petr Danecek[4], Luke Jostins[8], Thomas M. Keane[4], Shane McCarthy[4], Klaudia Walter[4], Yali Xue[4], **Yale University** Mark B. Gerstein (Principal Investigator)[113-115],

Alexej Abyzov[116], Suganthi Balasubramanian[115], Jieming Chen[113], Declan Clarke[117], Yao Fu[113], Arif O. Harmanci[113], Mike Jin[115], Donghoon Lee[113], Jeremy Liu[115], Xinmeng Jasmine Mu[13,113], Jing Zhang[113,115], Yan Zhang[113,115]

Structural Variation Group: BGI-Shenzhen Yingrui Li[26], Ruibang Luo[26], Hongmei Zhu[26], **Bilkent University** Can Alkan[36], Elif Dal[36], Fatma Kahveci[36], **Boston College** Gabor T. Marth (Principal Investigator)[23], Erik P. Garrison[4], Deniz Kural[37], Wan-Ping Lee[37], Alistair N. Ward[23], Jiantao Wu[23], Mengyao Zhang[23], **Broad Institute of MIT and Harvard** Steven A. McCarroll (Principal Investigator) [13,40], Robert E. Handsaker (Project Leader)[13,40], David M. Altshuler[3], Eric Banks[13], Guillermo del Angel[13], Giulio Genovese[13], Chris Hartl[13], Heng Li[13], Seva Kashin[13,40], James C. Nemesh[13], Khalid Shakir[13], **Cold Spring Harbor Laboratory** Seungtai C. Yoon (Principal Investigator)[42], Jayon Lihm[42], Vladimir Makarov[43], **Cornell University** Jeremiah Degenhardt[7], **European Molecular Biology Laboratory** Jan O. Korbel (Principal Investigator) (Co-Chair)[12,17], Markus H. Fritz[46], Sascha Meiers[17], Benjamin Raeder[17], Tobias Rausch[17,46], Adrian M. Stütz[17], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Francesco Paolo Casale[12], Laura Clarke[12], Richard E. Smith[12], Oliver Stegle[12], Xiangqun Zheng-Bradley[12], **Illumina** David R. Bentley (Principal Investigator)[5], Bret Barnes[39], R. Keira Cheetham[5], Michael Eberle[5], Sean Humphray[5], Scott Kahn[39], Lisa Murray[5], Richard Shaw[5], **Leiden University Medical Center,** Eric-Wubbo Lameijer[118], **Louisiana State University** Mark A. Batzer (Principal Investigator)[52], Miriam K. Konkel[52], Jerilyn A. Walker[52], **McDonnell Genome Institute at Washington University** Li Ding (Principal Investigator)[22], Ira Hall[22], Kai Ye[22], **Stanford University** Phil Lacroute[65], **The Jackson Laboratory for Genomic Medicine** Charles Lee (Principal Investigator) (Co-Chair)[18,19], Eliza Cerveira[18], Ankit Malhotra[18], Jaeho Hwang[18], Dariusz Plewczynski[18], Kamen Radew[18], Mallory Romanovitch[18], Chengsheng Zhang[18], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Nils Homer[75], **US National Institutes of Health** Deanna Church[34], Chunlin Xiao[25], **University of California, San Diego** Jonathan Sebat (Principal Investigator)[77], Danny Antaki[77], Vineet Bafna[119], Jacob Michaelson[120], Kenny Ye[79], **University of Maryland School of Medicine** Scott E. Devine (Principal Investigator)[90], Eugene J. Gardner (Project Leader)[90], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[2], Jeffrey M. Kidd (Principal Investigator)[91,92], Ryan E. Mills (Principal Investigator)[91,92], Gargi Dayama[91,92], Sarah Emery[92], Goo Jun[2,94], **University of North Carolina at Charlotte** Xinghua Shi (Principal Investigator)[102], Andrew Quitadamo[102], **University of Oxford** Gerton Lunter (Principal Investigator)[8], Gil A. McVean (Principal Investigator)[8,9], **University of Texas MD Anderson Cancer Center** Ken Chen (Principle Investigator)[121], Xian Fan[121], Zechen Chong[121], Tenghui Chen[121], **University of Utah** David Witherspoon[105], Jinchuan Xing[106], **University of Washington** Evan E. Eichler (Principal Investigator) (Co-Chair)[10,11], Mark J. Chaisson[10], Fereydoun Hormozdiari[10], John Huddleston[10,11], Maika Malig[10], Bradley J. Nelson[10], Peter H. Sudmant[10], **Vanderbilt University School of Medicine** Nicholas F. Parrish[95], **Weill Cornell Medical College**, Ekta Khurana (Principal Investigator)[109], **Wellcome Trust Sanger Institute** Matthew E. Hurles (Principal Investigator)[4], Ben Blackburne[4], Sarah J. Lindsay[4], Zemin Ning[4], Klaudia Walter[4], Yujun Zhang[4], **Yale University** Mark B. Gerstein (Principal Investigator)[113-115], Alexej Abyzov[116], Jieming Chen[113], Declan Clarke[117], Hugo Lam[122], Xinmeng Jasmine Mu[13,113], Cristina Sisu[113], Jing Zhang[113,115], Yan Zhang[113,115]

Exome Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator) (Co-Chair)[14], Fuli Yu (Project Leader)[14], Matthew Bainbridge[14], Danny Challis[14], Uday S. Evani[14], Christie Kovar[14], James Lu[14], Donna Muzny[14], Uma Nagaswamy[14], Jeffrey G. Reid[14], Aniko Sabo[14], Jin Yu[14], **BGI-Shenzhen** Xiaosen Guo[26,27], Wangshen Li[26], Yingrui Li[26], Renhua Wu[26], **Boston College** Gabor T. Marth (Principal Investigator) (Co-Chair)[23], Erik P. Garrison[4], Wen Fung Leong[23], Alistair N. Ward[23], **Broad Institute of MIT and Harvard** Guillermo del Angel[13], Mark A. DePristo[41], Stacey B. Gabriel[13], Namrata Gupta[13], Chris Hartl[13], Ryan E. Poplin[13], **Cornell University** Andrew G. Clark (Principal Investigator)[7], Juan L. Rodriguez-Flores[45], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Laura Clarke[12], Richard E. Smith[12], Xiangqun Zheng-Bradley[12], **Massachusetts General Hospital** Daniel G. MacArthur (Principal Investigator)[53], **McDonnell Genome Institute at Washington University** Elaine R. Mardis (Principal Investigator)[22], Robert Fulton[22], Daniel C. Koboldt[22], **McGill University** Simon Gravel[54], **Stanford University** Carlos D. Bustamante (Principal Investigator)[65], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Alexis Christoforides[74], Nils Homer[75], Tyler Izatt[74], **US National Institutes of Health** Stephen T. Sherry (Principal Investigator)[25], Chunlin Xiao[25], **University of Geneva** Emmanouil T. Dermitzakis (Principal Investigator)[87-89], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[2], Hyun Min Kang[2], **University of Oxford** Gil A. McVean (Principal Investigator)[8,9], **Yale University** Mark B. Gerstein (Principal Investigator)[113-115], Suganthi Balasubramanian[115], Lukas Habegger[113]

Functional Interpretation Group: Cornell University Haiyuan Yu (Principal Investigator)[44], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Laura Clarke[12], Fiona Cunningham[12], Ian Dunham[12], Daniel Zerbino[12], Xiangqun Zheng-Bradley[12], **Harvard University** Kasper Lage (Principal Investigator)[13,123], Jakob Berg Jespersen[13,123,124], Heiko Horn[13,123], **Stanford University** Stephen B. Montgomery (Principal Investigator)[66], Marianne K. DeGorter[66], **Weill Cornell Medical College**, Ekta Khurana (Principal Investigator)[109], **Wellcome Trust Sanger Institute** Chris Tyler-Smith (Principal Investigator) (Co-Chair)[4], Yuan Chen[4], Vincenza Colonna[4,112], Yali Xue[4], **Yale University** Mark B. Gerstein (Principal Investigator) (Co-Chair)[113-115], Suganthi Balasubramanian[115], Yao Fu[113], Donghoon Kim[115]

Chromosome Y Group: Albert Einstein College of Medicine Adam Auton (Principal Investigator)[1], Anthony Marcketta[1], **American Museum of Natural History** Rob Desalle[125], Apurva Narechania[126], **Arizona State University** Melissa A. Wilson Sayres[127], **Boston College** Erik P. Garrison[4], **Broad Institute of MIT and Harvard** Robert E. Handsaker[13,40], Seva Kashin[13,40], Steven A. McCarroll[13,40], **Cornell University:** Juan L. Rodriguez-Flores[45], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator)[12], Laura Clarke[12], Xiangqun Zheng-Bradley[12], **New York Genome Center** Yaniv Erlich[56,58], Melissa Gymrek[13,56,59,60], Thomas Frederick Willems[61], **Stanford University** Carlos D. Bustamante (Principal Investigator)(Co-Chair)[65], Fernando L. Mendez[65], G. David Poznik[128], Peter A. Underhill[65], **The Jackson Laboratory for Genomic Medicine** Charles Lee[18,19], Eliza Cerveira[18], Ankit Malhotra[18], Mallory Romanovitch[18], Chengsheng Zhang[18], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[2], **University of Queensland** Lachlan Coin (Principal Investigator)[129], Haojing Shao[129], **Virginia Bioinformatics Institute** David Mittelman[130], **Wellcome Trust Sanger Institute** Chris Tyler-Smith (Principal Investigator)(Co-Chair)[4], Qasim Ayub[4], Ruby Banerjee[4], Maria Cerezo[4], Yuan Chen[4], Thomas W. Fitzgerald[4], Sandra Louzada[4], Andrea Massaia[4], Shane McCarthy[4], Graham R. Ritchie[4], Yali Xue[4], Fengtang Yang[4]

Data Coordination Center Group: Baylor College of Medicine Richard A. Gibbs (Principal Investigator) [14], Christie Kovar[14], Divya Kalra[14], Walker Hale[14], Donna Muzny[14], Jeffrey G. Reid[14], **BGI-Shenzhen** Jun Wang (Principal Investigator)[26-30], Xu Dan[26], Xiaosen Guo[26,27], Guoqing Li[26], Yingrui Li[26], Chen Ye[26], Xiaole Zheng[26], **Broad Institute of MIT and Harvard** David M. Altshuler[3], **European Molecular Biology Laboratory, European Bioinformatics Institute** Paul Flicek (Principal Investigator) (Co-Chair)[12], Laura Clarke (Project Lead)[12], Xiangqun Zheng-Bradley[12], **Illumina** David R. Bentley (Principal Investigator)[5], Anthony Cox[5], Sean Humphray[5], Scott Kahn[39], **Max Planck Institute for Molecular Genetics** Ralf Sudbrak (Project Lead)[32], Marcus W. Albrecht[33], Matthias Lienhard[20], **McDonnell Genome Institute at Washington University** David Larson[22], **Translational Genomics Research Institute** David W. Craig (Principal Investigator)[74], Tyler Izatt[74], Ahmet A. Kurdoglu[74], **US National Institutes of Health** Stephen T. Sherry (Principal Investigator) (Co-Chair)[25], Chunlin Xiao[25], **University of California, Santa Cruz** David Haussler (Principal Investigator)[83,84], **University of Michigan** Gonçalo R. Abecasis (Principal Investigator)[2], **University of Oxford** Gil A. McVean (Principal Investigator)[8,9], **Wellcome Trust Sanger Institute** Richard M. Durbin (Principal Investigator)[4], Senduran Balasubramaniam[4], Thomas M. Keane[4], Shane McCarthy[4], James Stalker[4]

Samples and ELSI Group: Aravinda Chakravarti (Co-Chair)[6], Bartha M. Knoppers (Co-Chair)[16], Gonçalo R. Abecasis[2], Kathleen C. Barnes[131], Christine Beiswanger[31], Esteban G. Burchard[80], Carlos D. Bustamante[65], Hongyu Cai[26], Hongzhi Cao[26,27], Richard M. Durbin[4], Norman P. Gerry[31], Neda Gharani[31], Richard A. Gibbs[14], Christopher R. Gignoux[80], Simon Gravel[54], Brenna Henn[132], Danielle Jones[44], Lynn Jorde[105], Jane S. Kaye[133], Alon Keinan[7], Alastair Kent[134], Angeliki Kerasidou[135], Yingrui Li[26], Rasika Mathias[136], Gil A. McVean[8,9], Andres Moreno-Estrada[65,69], Pilar N. Ossorio[137,138], Michael Parker[135], Alissa M. Resch[31], Charles N. Rotimi[139], Charmaine D. Royal[140], Karla Sandoval[65], Yeyang Su[26], Ralf Sudbrak[32], Zhongming Tian[26], Sarah Tishkoff[141], Lorraine H. Toji[31], Chris Tyler-Smith[4], Marc Via[142], Yuhong Wang[26], Huanming Yang[26], Ling Yang[26], Jiayong Zhu[26]

Sample Collection: British from England and Scotland (GBR) Walter Bodmer[143], **Colombians in Medellín, Colombia (CLM)** Gabriel Bedoya[144], Andres Ruiz-Linares[86], **Han Chinese South (CHS)** Zhiming Cai[26], Yang Gao[145], Jiayou Chu[146], **Finnish in Finland (FIN)** Leena Peltonen‡, **Iberian Populations in Spain (IBS)** Andres Garcia-Montero[147], Alberto Orfao[147], **Puerto Ricans in Puerto Rico (PUR)** Julie Dutil[148], Juan C. Martinez-Cruzado[104], Taras K. Oleksyk[104], **African Caribbean in Barbados (ACB)** Kathleen C. Barnes[131], Rasika A. Mathias[136], Anselm Hennis[149,150], Harold Watson[150], Colin McKenzie[151], **Bengali in Bangladesh (BEB)** Firdausi Qadri[152], Regina LaRocque[152], Pardis C. Sabeti[13,48], **Chinese Dai in Xishuangbanna, China (CDX)** Jiayong Zhu[26], Xiaoyan Deng[153], **Esan in Nigeria (ESN)** Pardis C. Sabeti[13,48], Danny Asogun[154], Onikepe Folarin[155], Christian

Happi[155,156], Omonwunmi Omoniwa[155,156], Matt Stremlau[13,48], Ridhi Tariyal[13,48], **Gambian in Western Division – Mandinka (GWD)** Muminatou Jallow[8,157], Fatoumatta Sisay Joof[8,157], Tumani Corrah[8,157], Kirk Rockett[8,157], Dominic Kwiatkowski[8,157], **Indian Telugu in the U.K. (ITU)** and **Sri Lankan Tamil in the UK (STU)** Jaspal Kooner[158], **Kinh in Ho Chi Minh City, Vietnam (KHV)** Trần Tịnh Hiền[159], Sarah J. Dunstan[159,160], Nguyen Thuy Hang[159], **Mende in Sierra Leone (MSL)** Richard Fonnie[161], Robert Garry[162], Lansana Kanneh[161], Lina Moses[162], Pardis C. Sabeti[13,48], John Schieffelin[162], Donald S. Grant[161,162], **Peruvian in Lima, Peru (PEL)** Carla Gallo[163], Giovanni Poletti[163], **Punjabi in Lahore, Pakistan (PJL)** Danish Saleheen[164,165], Asif Rasheed[164]

Scientific Management: Lisa D. Brooks[166], Adam L. Felsenfeld[166], Jean E. McEwen[166], Yekaterina Vaydylevich[166], Eric D. Green[15], Audrey Duncanson[167], Michael Dunn[167], Jeffery A. Schloss[166], Jun Wang[26-30], Huanming Yang[26,168]

Writing Group: Adam Auton[1], Lisa D. Brooks[166], Richard M. Durbin[4], Erik P. Garrison[4], Hyun Min Kang[2], Jan O. Korbel[12,17], Jonathan L. Marchini[8,9], Shane McCarthy[4], Gil A. McVean[8,9], Goncalo R. Abecasis[2]

1 Department of Genetics, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
2 Center for Statistical Genetics, Biostatistics, University of Michigan, Ann Arbor, Michigan 48109, USA.
3 Vertex Pharmaceuticals, Boston, MA 02210, USA.
4 Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Cambridge, CB10 1SA, UK.
5 Illumina United Kingdom, Chesterford Research Park, Little Chesterford, Nr Saffron Walden, Essex CB10 1XL, UK.
6 McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
7 Center for Comparative and Population Genomics, Cornell University, Ithaca, New York 14850, USA.
8 Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford OX3 7BN, UK.
9 Department of Statistics, University of Oxford, Oxford OX1 3TG, UK.
10 Dept of Genome Sciences, University of Washington School of Medicine, Seattle, Washington 98195, USA.
11 Howard Hughes Medical Institute, University of Washington, Seattle, WA 98195, USA.
12 European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD, UK.
13 The Broad Institute of MIT and Harvard, 7 Cambridge Center, Cambridge, Massachusetts 02142, USA.
14 Baylor College of Medicine, Human Genome Sequencing Center, Houston, Texas 77030, USA.
15 US National Institutes of Health, National Human Genome Research Institute, 31 Center Drive, Bethesda, Maryland 20892, USA.
16 Centre of Genomics and Policy, McGill University, Montreal, Quebec H3A 1A4, Canada.
17 European Molecular Biology Laboratory, Genome Biology Research Unit, Meyerhofstr. 1, Heidelberg, Germany.
18 The Jackson Laboratory for Genomic Medicine, 10 Discovery Drive, Farmington, Connecticut 06032, USA.
19 Department of Life Sciences, Ewha Womans University, Ewhayeodae-gil, Seodaemun-gu, Seoul, South Korea 120-750.
20 Max Planck Institute for Molecular Genetics, D-14195 Berlin-Dahlem, Germany.
21 Dahlem Centre for Genome Research and Medical Systems Biology, D-14195 Berlin-Dahlem, Germany.
22 McDonnell Genome Institute at Washington University, Washington University School of Medicine, St Louis, Missouri 63108, USA.
23 USTAR Center for Genetic Discovery & Department of Human Genetics, University of Utah School of Medicine.
24 Affymetrix, Inc., Santa Clara, California 95051, USA.
25 US National Institutes of Health, National Center for Biotechnology Information, 45 Center Drive, Bethesda, Maryland 20892, USA.
26 BGI-Shenzhen, Shenzhen 518083, China.
27 Department of Biology, University of Copenhagen, Ole Maaløes Vej 5, 2200 Copenhagen, Denmark.
28 Princess Al Jawhara Albrahim Center of Excellence in the Research of Hereditary Disorders, King Abdulaziz University, Jeddah, Saudi Arabia.
29 Macau University of Science and Technology, Avenida Wai long, Taipa, Macau 999078, China.
30 Department of Medicine and State Key Laboratory of Pharmaceutical Biotechnology, University of Hong Kong, 21 Sassoon Road, Hong Kong.
31 Coriell Institute for Medical Research, Camden, New Jersey 08103, USA.
32 European Centre for Public Heath Genomics, UNU-MERIT, Unsiversity Maastricht,PO Box 616, 6200 MD Maastricht, The Netherlands.
33 Alacris Theranostics GmbH, D-14195 Berlin-Dahlem, Germany.
34 Personalis, Inc., Menlo Park, California 94025, USA.
35 US National Institutes of Health, National Human Genome Research Institute, 50 South Drive, Bethesda, Maryland 20892, USA.
36 Dept of Computer Engineering, Bilkent University, TR-06800 Bilkent, Ankara, Turkey.
37 Seven Bridges Genomics, Inc., 1 Broadway, 14th Floor, Cambridge, MA 02142, USA.
38 University of Oklahoma.
39 Illumina, Inc., San Diego, California 92122, USA.
40 Dept of Genetics, Harvard Medical School, Cambridge, Massachusetts 02142, USA.
41 SynapDx, Four Hartwell Place, Lexington, MA 02421, USA.
42 Cold Spring Harbor Laboratory, Cold Spring Harbor, New York 11724, USA.
43 Seaver Autism Center and Dept of Psychiatry, Mount Sinai School of Medicine, New York, New York 10029, USA.
44 Dept of Biological Statistics and Computational Biology, Cornell University, Ithaca, New York 14853, USA.
45 Department of Genetic Medicine, Weill Cornell Medical College, New York, NY, 10044, USA.
46 European Molecular Biology Laboratory, Genomics Core Facility, Meyerhofstrasse 1, 69117 Heidelberg, Germany.
47 Bill Lyons Informatics Centre, UCL Cancer Institute, University College London, London WC1E 6DD, UK.
48 Center for Systems Biology and Dept Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.
49 Department of Organismic and Evolutionary Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

50 Institute of Medical Genetics, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK.
51 Icahn School of Medicine at Mount Sinai, One Gustave L. Levy Place, Box 1498, New York, NY 10029-6574, USA.
52 Dept of Biological Sciences, Louisiana State University, Baton Rouge, Louisiana 70803, USA.
53 Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
54 McGill University and Genome Quebec Innovation Centre, 740, Avenue du Dr. Penfield, Montreal, Qc, Canada.
55 National Eye Institute, National Institutes of Health, Bethesda, Maryland, 20892.
56 New York Genome Center, 101 Avenue of the Americas, 7th floor, New York, NY 10013, USA.
57 Department of Systems Biology, Columbia University, New York, NY 10032, USA.
58 Department of Computer Science, Fu Foundation School of Engineering, Columbia University, New York, NY, USA.
59 Harvard-MIT Division of Health Sciences and Technology, Cambridge, MA USA.
60 General Hospital and Harvard Medical School, Boston, MA USA.
61 Whitehead Institute for Biomedical Research, Nine Cambridge Center, Cambridge, MA 02142, USA.
62 Ontario Institute for Cancer Research, MaRS Centre, 661 University Avenue, Suite 510, Toronto, Ontario, M5G 0A3, Canada.
63 Dept of Anthropology, Penn State University, University Park, Pennsylvania 16802, USA.
64 Rutgers Cancer Institute of New Jersey, New Brunswick, NJ 08903, USA.
65 Dept of Genetics, Stanford University, Stanford, California 94305, USA.
66 Departments of Genetics and Pathology, Stanford University, Stanford, California 94305-5324, USA.
67 Ancestry.com, San Francisco, California 94107, USA.
68 DNAnexus, 1975 W El Camino Real STE 101, Mountain View CA 94040, USA.
69 Laboratorio Nacional de Genómica para la Biodiversidad (LANGEBIO), CINVESTAV, Irapuato, Guanajuato 36821, Mexico.
70 Blavatnik School of Computer Science, Tel-Aviv University, Israel, 69978.
71 Dept of Microbiology, Tel-Aviv University, Israel, 69978.
72 International Computer Science Institute, Berkeley, California 94704, USA.
73 Thermo Fisher Scientific, 200 Oyster Point Boulevard, South San Francisco, CA 94080, USA.
74 The Translational Genomics Research Institute, Phoenix, Arizona 85004, USA.
75 Life Technologies, Beverly, Massachusetts 01915, USA.
76 Dept of Human Genetics, David Geffen School of Medicine at UCLA, Los Angeles, California 90024, USA.
77 Dept of Psychiatry, University of California, San Diego, La Jolla, California 92093, USA.
78 Dept of Cellular and Molecular Medicine, University of California, San Diego, La Jolla, California 92093, USA.
79 Dept of Epidemiology and Population Health, Albert Einstein College of Medicine, Bronx, New York 10461, USA.
80 Depts of Bioengineering & Therapeutic Sciences, University of California, San Francisco, San Francisco, California 94158, USA.
81 Institute for Quantitative Biosciences (QB3), University of California, San Francisco, 1700 4th Street San Francisco, California 94158.
82 Institute for Human Genetics, University of California, San Francisco, 1700 4th Street San Francisco, California 94158.
83 Center for Biomolecular Science and Engineering, University of California-Santa Cruz, Santa Cruz, California 95064, USA.
84 Howard Hughes Medical Institute, Santa Cruz, California 95064, USA.
85 Dept of Human Genetics, University of Chicago, Chicago, Illinois 60637, USA.
86 Dept of Genetics, Evolution and Environment, University College London, London WC1E 6BT, UK.
87 Dept of Genetic Medicine and Development, University of Geneva Medical School, Geneva, 1211 Switzerland.
88 Institute for Genetics and Genomics in Geneva, University of Geneva, 1211 Geneva, Switzerland.
89 Swiss Institute of Bioinformatics, 1211 Geneva, Switzerland.
90 Institute for Genome Sciences, University of Maryland School of Medicine, Baltimore, Maryland 21201, USA.
91 Dept of Computational Medicine and Bioinfomatics, University of Michigan, Ann Arbor, Michigan 48109, USA.
92 Department of Human Genetics, University of Michigan Medical School, Ann Arbor, Michigan 48109, USA.
93 Department of Pediatrics, University of Pittsburgh, Pittsburgh, PA 15224.
94 The University of Texas Health Science Center at Houston, Houston, Texas 77030, USA.
95 Vanderbilt University School of Medicine, Nashville, Tennessee 37232, USA.
96 University of Michigan Sequencing Core, University of Michigan, Ann Arbor, Michigan 48109, USA.
97 Istituto di Ricerca Genetica e Biomedica, CNR, Monserrato, 09042 Cagliari, Italy.
98 Dipartimento di Scienze Biomediche, Università delgi Studi di Sassari, 07100 Sassari, Italy.
99 UT Southwestern Medical Center, 5323 Harry Hines Blvd, Dallas, TX 75390.
100 Dept of Pediatrics, University of Montreal, Ste. Justine Hospital Research Centre, Montreal, Quebec H3T 1C5, Canada.
101 Department of Genetics, Department of Biostatistics, Department of Computer Science, University of North Carolina, Chapel Hill 27599.
102 Department of Bioinformatics and Genomics, College of Computing and Informatics, University of North Carolina at Charlotte, 9201 University City Blvd., Charlotte, NC 28223, USA.
103 Department of Medical Genetics, Center for Molecular Medicine, University Medical Center Utrecht, Utrecht, The Netherlands.
104 Dept of Biology, University of Puerto Rico at Mayagüez, Mayagüez, Puerto Rico 00680, USA.
105 Eccles Institute of Human Genetics, University of Utah School of Medicine, Salt Lake City, Utah 84112, USA.
106 Dept of Genetics, Rutgers University, Piscataway, New Jersey 08854, USA.
107 Dept of Medicine, Division of Medical Genetics, University of Washington, Seattle, Washington 98195, USA.
108 Department of Biostatistics, University of Washington, Seattle, Washington 98195, USA.
109 Department of Physiology and Biophysics, Weill Cornell Medical College, New York, NY, 10065.
110 Department of Human Genetics, Radboud Institute for Molecular Life Sciences and Donders Centre for Neuroscience, Radboud University Medical Center, Geert Grooteplein 10, 6525 GA Nijmegen.
111 Department of Molecular Developmental Biology, Faculty of Science, Radboud Institute for Molecular Life Sciences (RIMLS), Radboud University, 6500 HB Nijmegen, The Netherlands.
112 Institute of Genetics and Biophysics, National Research Council (CNR), 80125 Naples, Italy.
113 Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.
114 Dept of Computer Science, Yale University, New Haven, Connecticut 06520, USA.
115 Dept of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
116 Department of Health Sciences Research, Mayo Clinic, Rochester, MN 55905.

117 Dept of Chemistry, Yale University, New Haven, Connecticut 06520, USA.
118 Molecular Epidemiology Section, Dept of Medical Statistics and Bioinformatics, Leiden University Medical Center 2333 ZA, The Netherlands.
119 Dept of Computer Science, University of California, San Diego, La Jolla, California 92093, USA.
120 Beyster Center for Genomics of Psychiatric Diseases, University of California-San Diego, La Jolla, California 92093, USA.
121 Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77230, USA.
122 Bina Technologies, Roche Sequencing, Redwood City, CA, 94065, USA.
123 Department of Surgery, Massachusetts General Hospital, Boston, Massachusetts 02114, USA.
124 Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, Kemitorvet Building 208, 2800 Lyngby, Denmark.
125 Sackler Institute for Comparative Genomics, American Museum of Natural History, New York, NY.
126 Department of Invertebrate Zoology, American Museum of Natural History, New York, New York 10024, USA.
127 School of Life Sciences, Arizona State University, Tempe, AZ 85287-4701, USA.
128 Program in Biomedical Informatics, Stanford University, Stanford, CA 94305, USA.
129 Institute for Molecular Bioscience, University of Queensland, St Lucia, QLD 4072, Australia.
130 Virginia Bioinformatics Institute, 1015 Life Sciences Drive, Blacksburg, VA 24061, USA.
131 Division of Allergy & Clinical Immunology, School of Medicine, Johns Hopkins University, Baltimore, Maryland 21205, USA.
132 Dept of Ecology and Evolution, Stony Brook University, Stony Brook NY 11794.
133 Centre for Health, Law and Emerging Technologies, University of Oxford, Oxford OX3 7LF, UK.
134 Genetic Alliance, London, N1 3QP, UK.
135 The Ethox Center, Nuffield Department of Population Health, University of Oxford, Old Road Campus, OX3 7LF.
136 Johns Hopkins University School of Medicine, Baltimore, Maryland 21205, USA.
137 Dept of Medical History and Bioethics, Morgridge Institute for Research, University of Wisconsin-Madison, Madison, Wisconsin 53706, USA.
138 University of Wisconsin Law School, Madison, Wisconsin 53706, USA.
139 US National Institutes of Health, Center for Research on Genomics and Global Health, National Human Genome Research Institute, 12 South Drive, Bethesda, Maryland 20892, USA.
140 Department of African & African American Studies, Duke University, Durham, North Carolina 27708, USA.
141 Dept of Genetics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, USA.
142 Department of Psychiatry and Clinical Psychobiology & Institute for Brain, Cognition and Behavior (IR3C), University of Barcelona, 08035 Barcelona, Spain.
143 Cancer and Immunogenetics Laboratory, University of Oxford, John Radcliffe Hospital, Oxford OX3 9DS, UK.
144 Laboratory of Molecular Genetics, Institute of Biology, University of Antioquia, Medellín, Colombia.
145 Peking University Shenzhen Hospital, Shenzhen, 518036, China.
146 Institute of Medical Biology, Chinese Academy of Medical Sciences & Peking Union Medical College, Kunming 650118, China.
147 Instituto de Biologia Molecular y Celular del Cancer, Centro de Investigacion del Cancer/IBMCC (CSIC-USAL), Institute of Biomedical Research of Salamanca (IBSAL) & National DNA Bank Carlos III, University of Salamanca, Salamanca, Spain.
148 Ponce Research Institute, Ponce Health Sciences University, Ponce, Puerto Rico, 00716.
149 Chronic Disease Research Centre, Tropical Medicine Research Institute, Cave Hill Campus, The University of the West Indies.
150 Faculty of Medical Sciences, Cave Hill Campus, The University of the West Indies.
151 Tropical Metabolism Research Unit, Tropical Medicine Research Institute, Mona Campus, The University of the West Indies.
152 International Centre for Diarrhoeal Disease Research, Dhaka, Bangladesh.
153 Xishuangbanna Health School, Xishuangbanna 666100, China.
154 Irrua Specialist Teaching Hospital, Edo State, Nigeria.
155 Redeemers University, Ogun State, Nigeria.
156 Harvard T.H. Chan School of Public Health, Boston, Massachusetts 02115, USA.
157 Medical Research Council Unit, The Gambia.
158 NHLI, Imperial College London, Hammersmith Hospital, London, United Kingdom.
159 Centre for Tropical Medicine, Oxford University Clinical Research Unit, Ho Chi Minh City, Viet Nam.
160 Peter Doherty Institute of Infection and Immunity, The University of Melbourne, Australia.
161 Kenema Government Hospital, Ministry of Health and Sanitation, Kenema, Sierra Leone.
162 Tulane University Health Sciences Center, New Orleans, USA.
163 Laboratorios de Investigación y Desarrollo, Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Peru.
164 Center for Non-Communicable Diseases, Karachi, Pakistan.
165 Dept of Epidemiology and Biostatistics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, Pennsylvania 19104.
166 US National Institutes of Health, National Human Genome Research Institute, 5635 Fishers Lane, Bethesda, Maryland 20892, USA.
167 Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK.
168 James D. Watson Institute of Genome Sciences, Hangzhou 310008, China.

11 References

1. Tamura, K., Stecher, G., Peterson, D., Filipski, A. & Kumar, S. MEGA6: Molecular evolutionary genetics analysis version 6.0. *Mol. Biol. Evol.* **30,** 2725–2729 (2013).
2. Rambaut, A. Figtree. (2006). Available at: http://tree.bio.ed.ac.uk/software/figtree/.
3. Chiaroni, J., Underhill, P.A. & Cavalli-Sforza, L.L. Y chromosome diversity, human expansion, drift, and cultural evolution. *Proc. Natl. Acad. Sci.* **106,** 20174–20179 (2009).
4. Poznik, G.D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341,** 562–565 (2013).
5. Karafet, T.M., Mendez, F.L., Sudoyo, H., Lansing, J.S. & Hammer, M.F. Improved phylogenetic resolution and rapid diversification of Y-chromosome haplogroup K-M526 in Southeast Asia. *Eur. J. Hum. Genet.* **23,** 369–73 (2015).
6. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46,** 310–315 (2014).
7. Kumar, P., Henikoff, S. & Ng, P.C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.* **4,** 1073–1081 (2009).
8. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7,** 248–249 (2010).
9. Mathelier, A. *et al.* JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **42,** D142–7 (2014).
10. Wilson Sayres, M.A., Lohmueller, K.E. & Nielsen, R. Natural selection reduced diversity on human Y chromosomes. *PLoS Genet.* **10,** e1004064 (2014).
11. Lohmueller, K.E., Bustamante, C.D. & Clark, A.G. Methods for human demographic inference using haplotype patterns from genomewide single-nucleotide polymorphism data. *Genetics* **182,** 217–231 (2009).
12. Lohmueller, K.E., Bustamante, C.D. & Clark, A.G. The effect of recent admixture on inference of ancient human population history. *Genetics* **185,** 611–622 (2010).
13. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514,** 445–449 (2014).
14. Helgason, A. *et al.* The Y-chromosome point mutation rate in humans. *Nat. Genet.* **47,** 453–457 (2015).
15. Rebolledo-Jaramillo, B. *et al.* Maternal age effect and severe germ-line bottleneck in the inheritance of human mitochondrial DNA. *Proc. Natl. Acad. Sci.* **111,** 15474–15479 (2014).
16. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526,** 68–74 (2015).
17. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–2079 (2009).

18. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv Prepr.* 1–9 (2012).

19. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46,** 912–918 (2014).

20. Iqbal, Z., Caccamo, M., Turner, I., Flicek, P. & McVean, G. De novo assembly and genotyping of variants using colored de Bruijn graphs. *Nat. Genet.* **44,** 226–232 (2012).

21. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20,** 1297–1303 (2010).

22. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43,** 491–498 (2011).

23. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27,** 2156–2158 (2011).

24. Kuhn, R.M., Haussler, D. & James Kent, W. The UCSC genome browser and associated tools. *Brief. Bioinform.* **14,** 144–161 (2013).

25. Scozzari, R. *et al.* An unbiased resource of novel SNP markers provides a new chronology for the human Y chromosome and reveals a deep phylogenetic structure in Africa. *Genome Res.* **24,** 535–44 (2014).

26. Campbell, C.D. *et al.* Estimating the human mutation rate using autozygosity in a founder population. *Nat. Genet.* **44,** 1277–1281 (2012).

27. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488,** 471–5 (2012).

28. Michaelson, J.J. *et al.* Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell* **151,** 1431–1442 (2012).

29. Handsaker, R.E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47,** 296–303 (2015).

30. Bellos, E., Johnson, M.R. & Coin, L.J.M. cnvHiTSeq: integrative models for high-resolution copy number variation detection and genotyping using population sequencing data. *Genome Biol.* **13,** R120 (2012).

31. Pique-Regi, R., Cáceres, A. & González, J.R. R-Gada: a fast and flexible pipeline for copy number analysis in association studies. *BMC Bioinformatics* **11,** 380 (2010).

32. Pique-Regi, R. *et al.* Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24,** 309–318 (2008).

33. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464,** 704–712 (2010).

34. Quinlan, A.R. & Hall, I.M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26,** 841–2 (2010).

35. Quinlan, A.R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Curr. Protoc. Bioinforma.* **47,** 11.12.1–11.12.34 (2014).

36. Lattanzi, W. *et al.* A large interstitial deletion encompassing the amelogenin gene on the short arm of the Y chromosome. *Hum. Genet.* **116,** 395–401 (2005).

37. Tyler-Smith, C., Taylor, L. & Müller, 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).

38. Perry, G.H. *et al.* Copy number variation and evolution in humans and chimpanzees. *Genome Res.* **18,** 1698–1710 (2008).

39. Coriell. Coriell Biorepository. (2014). Available at: https://catalog.coriell.org.

40. Polley, S. *et al.* Evolution of the rapidly mutating human salivary agglutinin gene (DMBT1) and population subsistence strategy. *Proc. Natl. Acad. Sci.* **112,** 5105–5110 (2015).

41. Untergasser, A. *et al.* Primer3--new capabilities and interfaces. *Nucleic Acids Res.* **40,** e115 (2012).

42. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23,** 1289–91 (2007).

43. Carpenter, D. *et al.* Obesity, starch digestion and amylase: association between copy number variants at human salivary (AMY1) and pancreatic (AMY2) amylase genes. *Hum. Mol. Genet.* **24,** 3472–3480 (2015).

44. Verma, R.S. & Babu, A. *Human Chromosomes: Principles & Techniques, 2nd edition*. (McGraw-Hill, Inc., 1995).

45. Gribble, S.M. *et al.* Massively Parallel Sequencing Reveals the Complex Structure of an Irradiated Human Chromosome on a Mouse Background in the Tc1 Model of Down Syndrome. *PLoS One* **8,** e60482 (2013).

46. Groth, K.A., Skakkebæk, A., Høst, C., Gravholt, C.H. & Bojesen, A. Clinical review: Klinefelter syndrome--a clinical update. *J. Clin. Endocrinol. Metab.* **98,** 20–30 (2013).

47. Jobling, M.A. *et al.* Recurrent duplication and deletion polymorphisms on the long arm of the Y chromosome in normal males. *Hum. Mol. Genet.* **5,** 1767–75 (1996).

48. 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. *Nat. Genet.* **35,** 247–51 (2003).

49. Skaletsky, H. *et al.* The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. *Nature* **423,** 825–37 (2003).

50. Sonnhammer, E.L. & Durbin, R. A dot-matrix program with dynamic threshold control suited for genomic DNA and protein sequence analysis. *Gene* **167,** GC1–10 (1995).

51. Kent, W.J. BLAT--the BLAST-like alignment tool. *Genome Res.* **12,** 656–64 (2002).

52. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–9 (2015).

53. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26,** 2069–2070 (2010).

54. Espinosa, J.R.F., Ayub, Q., Chen, Y., Xue, Y. & Tyler-Smith, C. Structural variation on the human Y chromosome from population-scale resequencing. *Croat. Med. J.* **56,** 194–207 (2015).

55. Wei, W. *et al.* Copy number variation in the human Y chromosome in the UK population. *Hum. Genet.* **134,** 789–800 (2015).

56. Johansson, M.M. *et al.* Microarray Analysis of Copy Number Variants on the Human Y Chromosome Reveals Novel and Frequent Duplications Overrepresented in Specific Haplogroups. *PLoS One* **10,** e0137223 (2015).

57. Willems, T. *et al.* Population-Scale Sequencing Data Enables Precise Estimates of Y-STR Mutation Rates. *Am. J. Hum. Genet.* **in press,** (2016).

58. Li, H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. *arXiv Prepr.* **1303.3997,** (2013).

59. Benson, G. Tandem repeats finder: A program to analyze DNA sequences. *Nucleic Acids Res.* **27,** 573–580 (1999).

60. Willems, T. HipSTR. (2015). Available at: https://github.com/tfwillems/HipSTR.

61. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17,** 368–376 (1981).

62. ISOGG. International Society of Genetic Genealogy. (2013). Available at: http://www.isogg.org/.

63. Francalacci, P. *et al.* Low-pass DNA sequencing of 1200 Sardinians reconstructs European Y-chromosome phylogeny. *Science* **341,** 565–9 (2013).

64. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 1312–1313 (2014).

65. Cruciani, F. *et al.* A revised root for the human Y chromosomal phylogenetic tree: the origin of patrilineal diversity in Africa. *Am. J. Hum. Genet.* **88,** 814–8 (2011).

66. Tamura, K. *et al.* MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol. Biol. Evol.* **28,** 2731–9 (2011).

67. Hughes, J.F. *et al.* Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. *Nature* **463,** 536–9 (2010).

68. Paten, B., Herrero, J., Beal, K., Fitzgerald, S. & Birney, E. Enredo and Pecan: Genome-wide mammalian consistency-based multiple alignment with paralogs. *Genome Res.* **18,** 1814–1828 (2008).

69. Cunningham, F. *et al.* Ensembl 2015. *Nucleic Acids Res.* **43,** D662–D669 (2014).

70. Schroeder, H. *et al.* Genome-wide ancestry of 17th-century enslaved Africans from the Caribbean. *Proc. Natl. Acad. Sci.* **112,** 3669–73 (2015).

71. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23,** 388–395 (2013).

72. Yan, S. *et al.* Y Chromosomes of 40% Chinese Descend from Three Neolithic Super-Grandfathers. *PLoS One* **9,** e105691 (2014).

73. Hallast, P. *et al.* The Y-Chromosome Tree Bursts into Leaf: 13,000 High-Confidence SNPs Covering the Majority of Known Clades. *Mol. Biol. Evol.* **32,** 661–73 (2015).

74. Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* **25,** 459–466 (2015).

75. Mendez, F.L. *et al.* An African American paternal lineage adds an extremely ancient root to the human Y chromosome phylogenetic tree. *Am. J. Hum. Genet.* **92,** 454–9 (2013).

76. Wong, L.-P. *et al.* Deep Whole-Genome Sequencing of 100 Southeast Asian Malays. *Am. J. Hum. Genet.* **92,** 52–66 (2013).

77. Underhill, P.A. *et al.* Y chromosome sequence variation and the history of human populations. *Nat. Genet.* **26,** 358–61 (2000).

78. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524,** 216–219 (2015).

79. Rootsi, S. *et al.* A counter-clockwise northern route of the Y-chromosome haplogroup N from Southeast Asia towards Europe. *Eur. J. Hum. Genet.* **15,** 204–11 (2007).

80. Dulik, M.C. *et al.* Y-chromosome analysis reveals genetic divergence and new founding native lineages in Athapaskan- and Eskimoan-speaking populations. *Proc. Natl. Acad. Sci.* **109,** 8471–6 (2012).

81. Regueiro, M., Alvarez, J., Rowold, D. & Herrera, R.J. On the origins, rapid expansion and genetic diversity of Native Americans from hunting-gatherers to agriculturalists. *Am. J. Phys. Anthropol.* **150,** 333–48 (2013).

82. Bethel, L. *The Cambridge History of Latin America*. (Cambridge University Press, 1984). doi:10.1017/CHOL9780521245166

83. Mao, X. *et al.* A genomewide admixture mapping panel for Hispanic/Latino populations. *Am. J. Hum. Genet.* **80,** 1171–8 (2007).

84. Alexander, D.H., Novembre, J. & Lange, K. Fast model-based estimation of ancestry in unrelated individuals. *Genome Res.* **19,** 1655–64 (2009).

85. Scally, A. *et al.* Insights into hominid evolution from the gorilla genome sequence. *Nature* **483,** 169–175 (2012).

86. Scally, A. & Durbin, R. Revising the human mutation rate: implications for understanding human evolution. *Nat. Rev. Genet.* **13,** 745–53 (2012).

87. Goebel, T., Waters, M.R. & O'Rourke, D.H. The late Pleistocene dispersal of modern humans in the Americas. *Science* **319,** 1497–502 (2008).

88. Dillehay, T.D. *et al.* Monte Verde: seaweed, food, medicine, and the peopling of South America. *Science* **320,** 784–6 (2008).

89. Waters, M.R. *et al.* The Buttermilk Creek complex and the origins of Clovis at the Debra L. Friedkin site, Texas. *Science* **331,** 1599–603 (2011).

90. Rasmussen, M. *et al.* The genome of a Late Pleistocene human from a Clovis burial site in western Montana. *Nature* **506,** 225–9 (2014).

91. Kircher, M. *et al.* Combined Annotation Dependent Depletion (CADD). (2013). Available at: http://cadd.gs.washington.edu/.

92. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30,** E386–E394 (2009).

93. Kloss-Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32,** 25–32 (2011).

94. Hughes, D.A. *et al.* Evaluating intra- and inter-individual variation in the human placental transcriptome. *Genome Biol.* **16,** 54 (2015).

95. Arbiza, L., Gottipati, S., Siepel, A. & Keinan, A. Contrasting X-linked and autosomal diversity across 14 human populations. *Am. J. Hum. Genet.* **94,** 827–844 (2014).

96. Hartl, D.L. & Clark, A.G. *Principles of Population Genetics, Fourth Edition*. (Sinauer Associates, Inc., 2007).

97. Hudson, R.R. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* **18,** 337–338 (2002).

98. Clark, C. *Population Growth and Land Use*. (St. Martin's Press, 1968).

99. Durand, J.D. Historical Estimates of World Population: An Evaluation. *Popul. Dev. Rev.* **3,** 253–296 (1977).

100. Beloch, J. *Die bevölkerung der griechisch-römischen Welt*. (Duncker & Humblot, 1886).

101. Plowden, W.C. *Report on the Census of British India taken on the 17th February 1881*. (Eyre and Spottiswoode, 1883).

102. Dillon, M. *China: A Cultural and Historical Dictionary*. (Routledge, 1998).

103. Davis, D.R. & Weinstein, D.E. Bones, bombs, and break points: The geography of economic activity. *Am. Econ. Rev.* **92,** 1269–1289 (2002).

104. Snow, D.R. Microchronology and demographic evidence relating to the size of pre-columbian north american Indian populations. *Science* **268,** 1601–1604 (1995).

105. Dulik, M.C. *et al.* Mitochondrial DNA and Y chromosome variation provides evidence for a recent common ancestry between Native Americans and Indigenous Altaians. *Am. J. Hum. Genet.* **90,** 229–46 (2012).

106. Zerjal, T. *et al.* The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72,** 717–721 (2003).

107. Clarke, L. *et al.* The 1000 Genomes Project: data management and community access. *Nat. Methods* **9,** 459–62 (2012).