ANALYSIS

nature genetics

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[3,14], Eliza Cerveira[13], Tomas W Fitzgerald[3], Erik Garrison[3], Anthony Marcketta[15], David Mittelman[16,17], Mallory Romanovitch[13], Chengsheng Zhang[13], Xiangqun Zheng-Bradley[14], Gonçalo R Abecasis[18], Steven A McCarroll[19], Paul Flicek[14], Peter A Underhill[2], Lachlan Coin[11], Daniel R Zerbino[14], Fengtang Yang[3], Charles Lee[13,20], Laura Clarke[14], Adam Auton[15], 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]

We report the sequences of 1,244 human Y chromosomes randomly ascertained from 26 worldwide populations by the 1000 Genomes Project. We discovered more than 65,000 variants, including single-nucleotide variants, multiple-nucleotide variants, insertions and deletions, short tandem repeats, and copy number variants. Of these, copy number variants contribute the greatest predicted functional impact. We constructed a calibrated phylogenetic tree on the basis of binary single-nucleotide variants and projected the more complex variants onto it, estimating the number of mutations for each class. Our phylogeny shows bursts of extreme expansion in male numbers that have occurred independently among each of the five continental superpopulations examined, at times of known migrations and technological innovations.

The Y chromosome bears a unique record of human history owing to its male-specific inheritance and the absence of crossover for most of its length, which together link it completely to male phenotype and behavior[1]. Previous studies have demonstrated the value of full sequences for characterizing and calibrating the human Y-chromosome phylogeny[2,3]. These studies have led to insights into male demography, but further work is needed to more comprehensively describe the range of Y-chromosome variation, including classes of variation more complex than single-nucleotide variants (SNVs); to investigate the mutational processes operating in the different classes; and to determine the relative roles of selection[4] and demography[5] in shaping Y-chromosome variation. The role of demography has risen to prominence with reports of male-specific bottlenecks in several geographical areas after 10 thousand years ago (kya)[5–7], at times putatively associated with the spread of farming[5] or Bronze Age culture[6]. With improved calibration of the Y-chromosome SNV mutation rate[8–10] and, consequently, more secure dating of relevant features of the Y-chromosome phylogeny, it is now possible to hone such interpretations.

We have conducted a comprehensive analysis of Y-chromosome variation using the largest extant sequence-based survey of global genetic variation—phase 3 of the 1000 Genomes Project[11]. We have documented the extent of and biological processes acting on five types of genetic variation, and we have generated new insights into the history of human males.

RESULTS

Data set

Our data set comprises 1,244 Y chromosomes sampled from 26 populations (**Supplementary Table 1**) and sequenced to a median haploid coverage of 4.3×. Reads were mapped to the GRCh37 human reference assembly used by phase 3 of the 1000 Genomes Project[11] and to the GRCh38 reference for our analysis of short tandem repeats (STRs). We used multiple haploid-tailored methods to call variants and generate call sets containing more than 65,000 variants of five types, including SNVs (**Supplementary Fig. 1** and **Supplementary Tables 2** and **3**), multiple-nucleotide variants (MNVs), short insertions and deletions (indels), copy number variants (CNVs) (**Supplementary Figs. 2–12**), and STRs (**Supplementary Tables 4–6**). We also identified karyotype variation, which included one instance of 47,XXY and several mosaics of the karyotypes 46,XY and 45,X (**Supplementary Table 7**). We applied stringent quality control to meet the Project's requirement of a false discovery rate (FDR) <5% for SNVs, indels and MNVs, and CNVs. In our validation analysis with independent data sets, the genotype concordance was greater than 99% for SNVs and was 86–97% for more complex variants (**Table 1**).

To construct a set of putative SNVs, we generated six distinct call sets, which we input to a consensus genotype caller. In an iterative

A full list of authors and affiliations appears at the end of the paper.

Received 8 November 2015; accepted 1 April 2016; published online 25 April 2016; doi:10.1038/ng.3559

Table 1 Y-chromosome variants discovered in 1,244 males

Variant type	Number	FDR (%)	Concordance (%)
SNVs	60,555	3.9	99.6
Indels and MNVs	1,427	3.6	96.4
CNVs	110	2.7	86
STRs	3,253	NA	89–97

The concordance shown is with independent genotype calls, and the CNVs considered were those computationally inferred using GenomeSTRiP. FDR, false discovery rate; NA, not available.

process, we leveraged the phylogeny to tune the final genotype calling strategy. We used similar methods for MNVs and indels, and we ran HipSTR to call STRs (**Supplementary Note**).

We discovered CNVs in the sequence data using two approaches, GenomeSTRiP[12] and CnvHitSeq[13] (**Supplementary Note**), and we validated calls using array comparative genomic hybridization (aCGH), supplemented by FISH on DNA fibers (fiber-FISH) in a few cases (**Supplementary Figs. 8** and **9**, and **Supplementary Note**). In **Figure 1**, we illustrate a representative large deletion, which we discovered in a single individual using GenomeSTRiP (**Fig. 1b**). We validated its presence by aCGH (**Fig. 1c**) and ascertained its structure with fiber-FISH (**Fig. 1d**). Notably, the event that gave rise to this variant was not a simple recombination between the segmental duplication elements it partially encompasses (**Fig. 1a**,**d**).

Phylogeny

We identified each individual's Y-chromosome haplogroup (**Supplementary Tables 8** and **9**, and **Supplementary Data**) and constructed a maximum-likelihood phylogenetic tree using 60,555 biallelic SNVs derived from 10.3 Mb of accessible DNA (**Fig. 2**, **Supplementary Figs. 13–17**, **Supplementary Note**, and **Supplementary Data**). Our tree recapitulates and refines the expected structure[2,3,5], with all but two major haplogroups from A0 through T represented. The only haplogroups absent are M and S, both subgroups of K2b1 that are largely specific to New Guinea, which was not included in the 1000 Genomes Project. Notably, the branching patterns of several lineages suggest extreme expansions ~50–55 kya and also within the last few millennia. We investigated these later expansions in some detail and describe our findings below.

When the tree is calibrated with a mutation rate estimate of 0.76×10^{-9} mutations per base pair per year[9], the time to the most recent common ancestor (TMRCA) of the tree is ~190,000 years, but we consider the implications of alternative mutation rate estimates below. Of the clades resulting from the four deepest branching events, all but one are exclusive to Africa, and the TMRCA of all non-African lineages (that is, the TMRCA of haplogroups DE and CF) is ~76,000 years (**Fig. 1**, **Supplementary Figs. 18** and **19**, **Supplementary Table 10**, and **Supplementary Note**). We saw a notable increase in the number of lineages outside Africa ~50–55 kya, perhaps reflecting

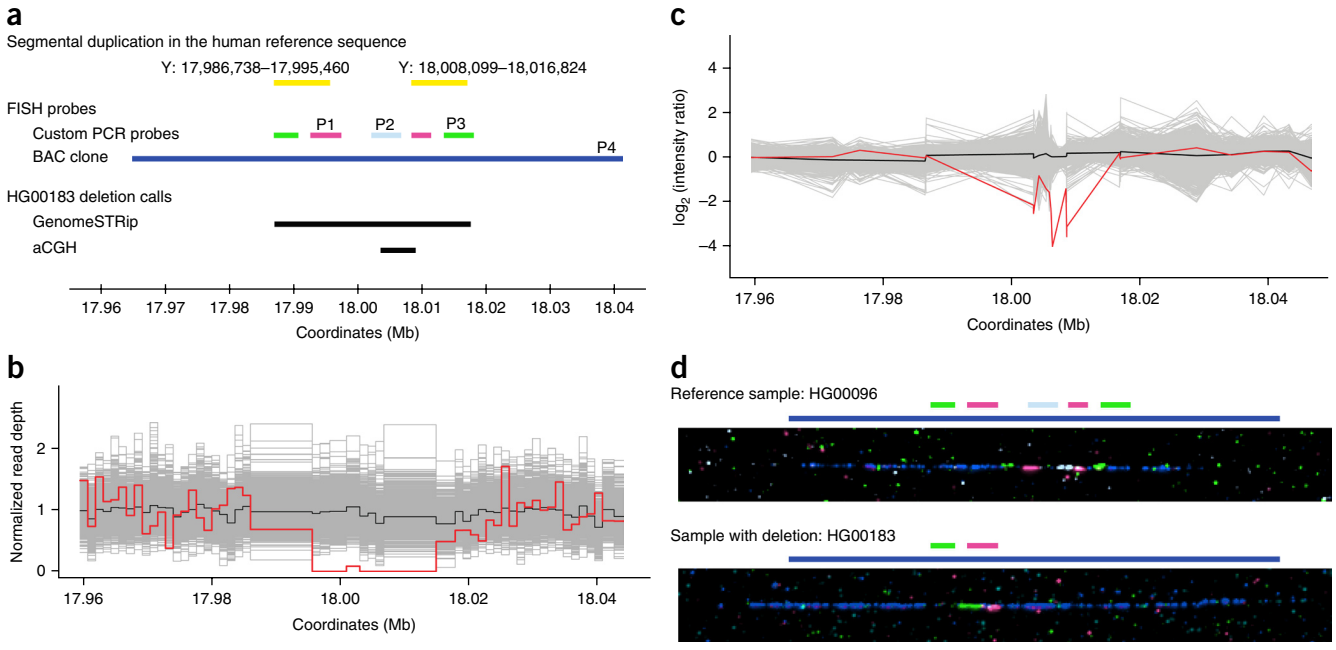

Figure 1 Discovery and validation of a representative Y-chromosome CNV. (**a**) The GRCh37 reference sequence contains an inverted segmental duplication (yellow bars) within GRCh37 Y: 17,986,738–18,016,824 bp. We designed FISH probes to target the 3′ termini of the two segments (magenta and green bars labeled P1 and P3, respectively) and the unique region between them (light-blue bar labeled P2). A fourth probe used reference sequence BAC clone RP11-12J24 (dark-blue bar labeled P4). Unlabeled green and magenta bars represent expected cross-hybridization, and black bars represent CNV events called by GenomeSTRiP and aCGH. GenomeSTRiP called a 30-kb deletion that includes the duplicated segments and the unique spacer region, whereas aCGH lacks probes in the duplicated regions. (**b**) GenomeSTRiP discovery plot. The red curve indicates the normalized read depth for sample HG00183, as compared to the read depth for 1,232 other samples (gray) and the median depth (black). (**c**) Validation by aCGH. The intensity ratio for HG00183 (red) is shown relative to that for 1,233 other samples (gray) and the median ratio (black). (**d**) Fiber-FISH validation using the probes illustrated in **a**. The reference sample, HG00096, matches the human reference sequence, with green, magenta, light-blue, magenta, and green hybridizations occurring in sequence. In contrast, we observed just one green and one magenta hybridization in HG00183, indicating deletion of one copy of the segmental duplication and the central unique region. The coordinate scale that is consistent across **a**–**c** does not apply to **d**, and, although the BAC clone hybridization (dark blue) is shorter in the sample with the deletion, it appears longer owing to the variable degree of stretching inherent to the molecular combing process.

the geographical expansion and differentiation of Eurasian populations as they settled the vast expanse of these continents. Consistent with previous proposals[14], a parsimonious interpretation of the phylogeny is that the predominant African haplogroup, haplogroup E, arose outside the continent. This model of geographical segregation within the CT clade requires just one continental haplogroup exchange (E to Africa), rather than three (D, C, and F out of Africa). Furthermore, the timing of this putative return to Africa—between the emergence of haplogroup E and its differentiation within Africa by 58 kya—is consistent with proposals, based on non–Y chromosome data, of abundant gene flow between Africa and nearby regions of Asia 50–80 kya[15].

Three new features of the phylogeny underscore the importance of South and Southeast Asia as likely locations where lineages currently distributed throughout Eurasia first diversified (**Supplementary Note**). First, we observed in a Vietnamese individual a rare F lineage that is an outgroup for the rest of the megahaplogroup (**Fig. 1** and **Supplementary Fig. 14b**). The sequence for this individual includes the derived allele for 147 SNVs shared by and specific to the 857 F chromosomes in our sample, but the lineage split off from the rest of the group ~55 kya. This finding enabled us to define a new megagroup, GHIJK-M3658, whose subclades include the vast majority of the world's non-African males[1]. Second, we identified in 12 South Asian individuals a new clade, here designated H0, that split from the rest of haplogroup H ~51 kya (**Supplementary Fig. 14b**). This new structure highlights the ancient diversity within the haplogroup and requires a more inclusive redefinition using, for example, the deeper SNV M2713, a G>A mutation at 6,855,809 bp in the GRCh37 reference. Third, a lineage carried by a South Asian Telugu individual, HG03742, enabled us to refine early differentiation within the K2a clade ~50 kya (**Fig. 1** and **Supplementary Figs. 14d** and **15**). Using the high resolving power of the SNVs in our phylogeny, we determined that this lineage split off from the branch leading to haplogroups N and O (NO) not long after the ancestors of two individuals with well-known ancient DNA (aDNA) sequences did. Ust'-Ishim[9] and Oase1 (ref. 16) lived, respectively, in western Siberia 43–47 kya and

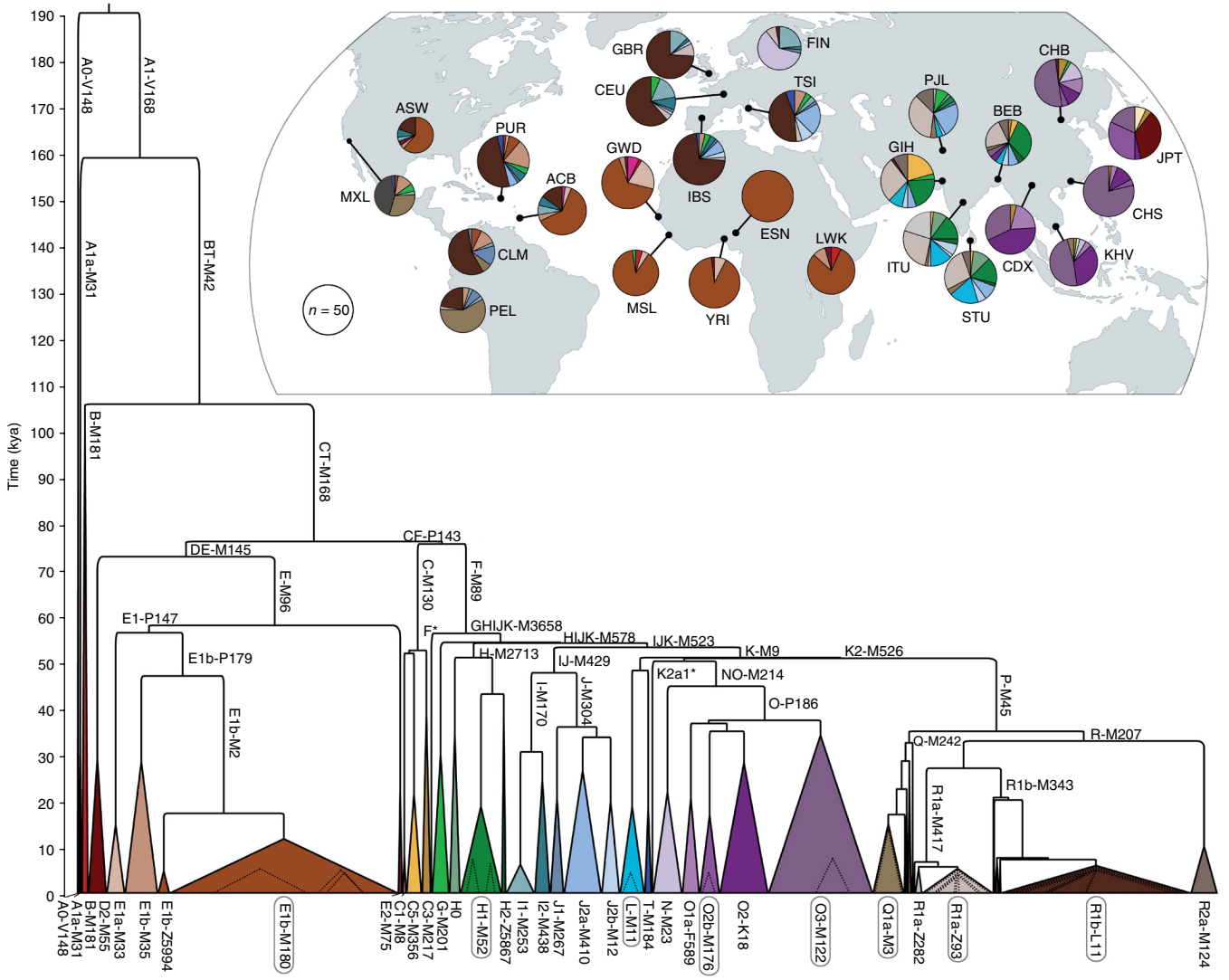

Figure 2 Y-chromosome phylogeny and haplogroup distribution. Branch lengths are drawn proportional to the estimated times between successive splits, with the most ancient division occurring ~190 kya. Colored triangles represent the major clades, and the width of each base is proportional to one less than the corresponding sample size. We modeled expansions within eight of the major haplogroups (circled) (**Fig. 4**); dotted triangles represent the ages and sample sizes of the expanding lineages. Inset, world map indicating, for each of the 26 populations, the geographic source, sample size, and haplogroup distribution.

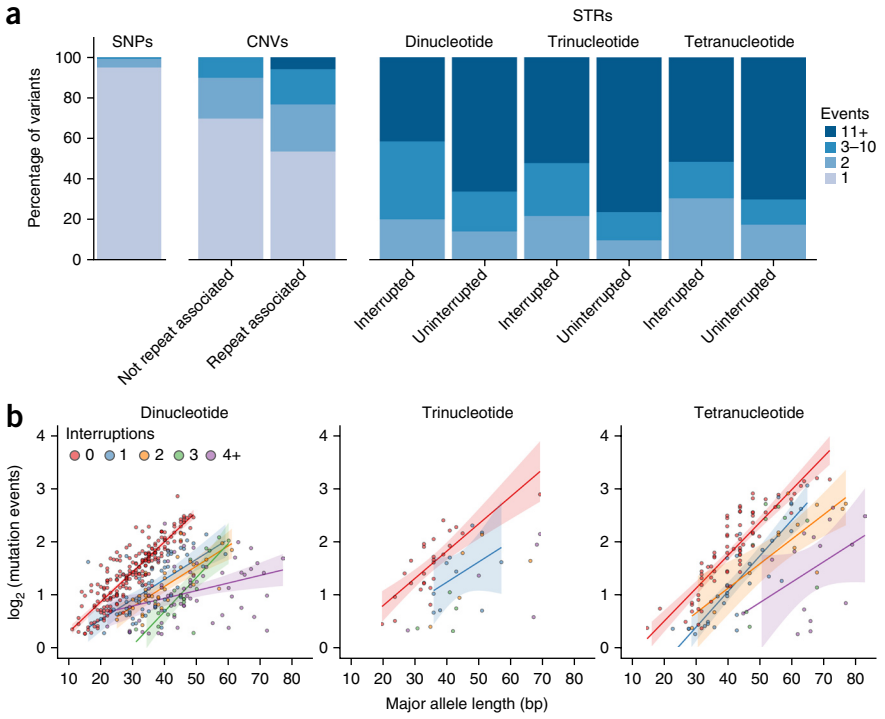

Figure 3 Mutation events. (**a**) Bar plots show the percentage of each variant type stratum associated with 1, 2, 3–10, or more mutations across the phylogeny. (**b**) For STRs, scatterplots show the logarithm of the number of mutational events versus major allele length, stratified by motif length and the number of interruptions to the repeat structure. We have plotted regression lines with shaded confidence intervals for categories with at least ten data points, and we have omitted from the plots 44 STRs with motif lengths greater than 4 bp and 91 STRs whose mutation rate estimates were equal to the minimum threshold of 1×10^{-5} mutations per generation. This figure was generated with ggplot2 (ref. 32).

in Romania 37–42 kya. The Y chromosomes of these individuals join that of HG03742 in sharing with haplogroup NO the derived T allele at M2308 (GRCh37 Y: 7,690,182 bp), and the modern sample shares just four additional mutations with the NO clade.

Mutations

To map each SNV to a branch (or branches) of the phylogeny, we first partitioned the tree into eight overlapping subtrees (**Supplementary Fig. 13**). Within each subtree, we provisionally assigned each SNV to the internal branch constituting the minimum superset of carriers of one allele or the other, designating the derived state to the allele that was specific to this clade. When no member of the clade bore the ancestral allele, we deemed the site compatible with the subtree and assigned the SNV to the branch (**Supplementary Note** and **Supplementary Data**). Most SNVs (94%) mapped to a single branch of the phylogeny, corresponding to a single mutation event during the Y-chromosome history captured by this tree. We projected the other variants onto the tree to infer the number of mutations associated with each (**Fig. 3a**).

Our workflow to count the number of independent mutation events associated with each CNV is summarized in **Supplementary Figure 10** (**Supplementary Note**). We found that 39% of CNVs have mutated multiple times, a much higher proportion than for SNVs (**Fig. 3a** and **Supplementary Data**). CNVs can arise by several different mutational mechanisms, one of which is homologous recombination between misaligned repeated sequences. This mechanism is particularly susceptible to recurrent mutations[17], but, in comparing CNVs associated with repeated sequences to those that are not repeat associated, we did not observe a significant difference in the proportion that have mutated multiple times (Mann–Whitney two-sided test). We did, however, observe that repeat-associated CNVs tend to be longer ($P = 0.01$).

We inferred more than six independent mutation events for each of three CNVs. One CNV in particular stood out with 154 events. An apparent CNV hotspot spans a gene-free stretch of the chromosome's long arm at GRCh37 Y: 22,216,565–22,512,935 bp. The region includes two arrays of long-terminal repeat 12B (LTR12B) elements that together harbor 48 of the genome's 211 copies of this element (23%). In principle, our inference of numerous independent mutations could have been due to a 'shadowing' effect from LTR12B elements elsewhere in the genome. That is, mismapping sequencing reads and cross-hybridizing aCGH probes can lead to false inference of variation. But, in a phylogenetic analysis of all 211 LTR12B elements (**Supplementary Fig. 11**), those within the putative CNV hotspot formed a pure monophyletic clade, demonstrating that the copy number signal was genuine. The CNV has no predicted functional consequence.

STRs constituted the most mutable variant class, with a median of 16 mutations per locus and an average mutation rate of 3.9×10^{-4} mutations per generation. Assuming a generation time of 30 years, this equates to 1.3×10^{-5} mutations per year. Allele length explains more than half of the variance in the log-transformed mutation rate for uninterrupted STRs. Longer STRs mutate more rapidly, and, conditional on allele length, mutability decreases when the repeat structure is interrupted, with a general trend toward slower mutation rates for STRs with more interruptions (**Fig. 3b**). Further details are provided in our companion paper on Y-STRs[18].

Functional impact

A small proportion of SNVs have a predicted functional impact (**Supplementary Figs. 20–23**, **Supplementary Tables 11–14**, **Supplementary Note**, and **Supplementary Data**). Among 60,555 SNVs, we observed 2 singleton premature stop codons, one each in *AMELY* and *USP9Y*, and one splice-site SNV that affects all known transcripts of *TBL1Y*. Among 94 missense SNVs with SIFT[19] scores, all 30 deleterious variants were singletons or doubletons, whereas 17 of 64 tolerated variants were present at higher frequency ($P = 0.001$), underscoring the impact of purifying selection on variation in protein-coding genes. No STRs overlapped protein-coding regions, but, in contrast to the SNVs, a high proportion of CNVs have a predicted functional impact.

Twenty of 100 CNVs in our final call set overlapped 27 protein-coding genes from 17 of the 33 Y-chromosome gene families. In our analysis of 1000 Genomes Project autosomal data, we observed that the ratio of the proportion of deletions overlapping protein-coding genes to the proportion of duplications overlapping protein-coding genes was 0.84. Whereas on the autosomes deletions are less likely to overlap protein-coding genes than duplications, as others have also reported[20], we found the reverse to be true for the Y chromosome. Despite the Y chromosome's haploidy, we calculated its ratio of proportions to be 1.5, indicating a surprising increased tolerance

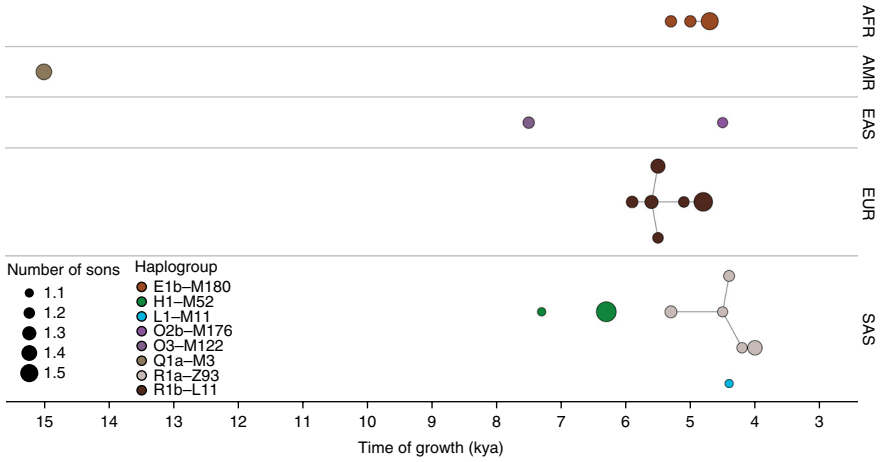

Figure 4 Explosive male-lineage expansions of the last 15,000 years. Each circle represents a phylogenetic node whose branching pattern suggests rapid expansion. The horizontal axis indicates the timings of the expansions, and circle radii reflect growth rates—the minimum number of sons per generation, as estimated by our two-phase growth model. Nodes are grouped by continental superpopulation (AFR, African; AMR, admixed American; EAS, East Asian; EUR, European; SAS, South Asian) and colored by haplogroup. Line segments connect phylogenetically nested lineages. This figure was generated with ggplot2 (ref. 32).

of gene loss as compared with the diploid genes on autosomes.

Diversity

Given observed diversity levels for the autosomes, X chromosome, and mitochondrial genome (mtDNA) (**Supplementary Table 15**, **Supplementary Note**, and **Supplementary Data**), Y-chromosome diversity was reported to be lower than expected from simple population genetic models that assume a Poisson-distributed number of offspring[4], and the role of selection in this disparity is debated. We confirmed that Y-chromosome diversity in our sample was low (**Supplementary Fig. 24**) and found that positing extreme male-specific bottlenecks in the last few millennia could lead to a good fit between modeled and observed relative diversity levels for the autosomes, X chromosome, Y chromosome, and mtDNA (**Supplementary Figs. 25–28**, **Supplementary Table 16**, and **Supplementary Note**). Therefore, we conclude that Y-chromosome diversity may be shaped primarily by neutral demographic processes.

Haplogroup expansions

To investigate punctuated bursts within the phylogeny and estimate growth rates, we modeled haplogroup growth as a rapid phase followed by a moderate-rate phase and applied this model to lineages showing rapid expansions (**Supplementary Figs. 29–31**, **Supplementary Tables 17–19**, **Supplementary Note**, and **Supplementary Data**), noting that such extreme expansions are seldom seen in the mtDNA phylogeny, here or in other studies[5]. We examined 20 nodes of the tree whose branching patterns were well fit by this model. These nodes were drawn from eight haplogroups and included at least one lineage from each of the five continental regions surveyed (**Fig. 4**). As the haplogroup expansions we report are among the most extreme yet observed in humans, we think it more likely than not that such events correspond to historical processes that have also left archaeological footprints. Therefore, in what follows, we propose links between genetic and historical or archaeological data. We caution that, especially in light of as yet imperfect calibration, these connections remain unproven. But they are testable, for example, using aDNA.

First, in the Americas, we observed expansion of Q1a-M3 (**Supplementary Figs. 14e** and **17**) at ~15 kya, the time of the initial colonization of the hemisphere[21]. This correspondence, based on one of the most thoroughly examined dates in human prehistory, attests to the suitability of the calibration we have chosen. Second, in sub-Saharan Africa, two independent E1b-M180 lineages expanded ~5 kya (**Supplementary Fig. 14a**), in a period before the numerical and geographical expansions of Bantu speakers, in whom E1b-M180 now predominates[22]. The presence of these lineages in non-Bantu speakers (for example, Yoruba and Esan) indicates an expansion predating the Bantu migrations, perhaps triggered by the development of ironworking[23]. Third, in Western Europe, related lineages within R1b-L11 expanded ~4.8–5.9 kya (**Supplementary Fig. 14e**), most markedly around 4.8 and 5.5 kya. The earlier of these times, 5.5 kya, is associated with the origin of the Bronze Age Yamnaya culture. The Yamnaya have been linked by aDNA evidence to a massive migration from the Eurasian Steppe, which may have replaced much of the previous European population[24,25]; however, the six Yamnaya with informative genotypes did not bear lineages descending from or ancestral to R1b-L11, so a Y-chromosome connection has not been established. The later time, 4.8 kya, coincides with the origins of the Corded Ware (Battle Axe) culture in Eastern Europe and the Bell–Beaker culture in Western Europe[26].

Potential correspondences between genetics and archeology in South and East Asia have not been investigated as extensively. In South Asia, we detected eight lineage expansions dating to ~4.0–7.3 kya and involving haplogroups H1-M52, L-M11, and R1a-Z93 (**Supplementary Fig. 14b**,**d**,**e**). The most striking were expansions within R1a-Z93, occurring ~4.0–4.5 kya. This time predates by a few centuries the collapse of the Indus Valley Civilization, associated by some with the historical migration of Indo-European speakers from the Western Steppe into the Indian subcontinent[27]. There is a notable parallel with events in Europe, and future aDNA evidence may prove to be as informative as it has been in Europe. Finally, East Asia stands out from the rest of the Old World for its paucity of sudden expansions, perhaps reflecting a larger starting population or the coexistence of multiple prehistoric cultures wherein one lineage could rarely dominate. We observed just one notable expansion within each of the O2b-M176 and O3-M122 clades (**Supplementary Fig. 14d**).

DISCUSSION

The 1000 Genomes Project data set provides a rich and unparalleled resource of Y-chromosome variation coupled with open access to DNA and cell lines that will facilitate diverse further investigations. By cataloging the phylogenetic position of ~60,000 SNVs, we have constructed a database of diagnostic variants with which one can assign Y-chromosome haplogroups to DNA samples (**Supplementary Data**). This resource is particularly valuable for SNP chip design and for aDNA studies, in which sequencing coverage is often quite low, as exemplified by our reanalysis of the Ust'-Ishim and Oase1 Y chromosomes.

The variants we report have well-calibrated FDRs. Nevertheless, because of the modest sequencing coverage, data missingness was a principal concern. Small CNVs and long STRs are largely undetected,

and low-frequency variants in general, including SNVs, are underrepresented. We therefore took great care to minimize the impact of missing variants. In particular, we designed the relevant downstream analyses to only use information from higher-frequency, shared variation, corresponding to mutations on internal branches of the tree.

Because many DNA samples were extracted from lymphoblastoid cells, another potential concern was variation that has arisen during cell culture[28]. However, such false discoveries are inherently not shared. Therefore, the precautions we took to minimize the impact of missingness also precluded *in vitro* mutations influencing our findings. We discuss additional caveats to the mapping of SNVs to branches in the **Supplementary Note**.

Our findings illustrate unique properties of the Y chromosome. Foremost, the abundance of extreme male-lineage expansions underscores differences between male and female demographic histories. A caveat to our expansion analysis is that our inference method assumed that population structure did not affect the branching patterns immediately downstream of the particular phylogenetic node under investigation. This is reasonable because population structure is unlikely when a very rapid expansion is in progress, but, to accommodate this strong assumption, we limited all analyses to pruned internal subtrees short enough for it to hold. A second caveat relates to the choice of calibration metric, which is relevant to the links we have suggested between expansions and historical or archaeological events. Present-day geographical distributions provide strong support for the correspondences we proposed for the initial peopling of most of Eurasia by fully modern humans ~50–55 kya and for the first colonization of the Americas ~15 kya. For later male-specific expansions, we should consider the consequences of alternative mutation rate estimates, as pedigree-based methods relying on variation from the most recent several centuries[8,10,28] may be more relevant. The pedigree-based estimate from the largest set of mutations[8] would lead to a ~15% decrease in expansion times, increasing the precision of the correspondences proposed for E1b and R1a. For R1b, a 15% decrease would suggest an expansion postdating the Yamnaya migration. Using either mutation rate estimate, the lineage expansions seem to have followed innovations that may have elicited increased variance in male reproductive success[29], innovations such as metallurgy, wheeled transport, or social stratification and organized warfare. In each case, privileged male lineages could undergo preferential amplification for generations. We find that rapid expansions are not confined to extraordinary circumstances[30,31] and that the Y chromosome resulting from these rapid expansions can predominate on a continental scale and do so in some of the populations most studied by medical geneticists. Inferences incorporating demography may benefit from taking these male–female differences into account.

URLs. 1000 Genomes Project, http://www.1000genomes.org/using-1000-genomes-data; International Society of Genetic Genealogy (ISOGG), http://www.isogg.org/; FigTree, http://tree.bio.ed.ac.uk/software/figtree/; HipSTR, https://github.com/tfwillems/HipSTR.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

We thank the 1000 Genomes Project sample donors for making this work possible, all Project members for their contributions, and A. Martin for ADMIXTURE results. The tree in **Figure 2** was drawn using FigTree. G.D.P. was supported by the National Science Foundation (NSF) Graduate Research Fellowship under grant DGE-1147470 and by National Library of Medicine training grant LM-007033. Work at the Wellcome Trust Sanger Institute (Q.A., R.B., M.C., Y.C., S.L., A. Massaia, S.A. McCarthy, C.T.-S., Y.X., and F.Y.) was supported by Wellcome Trust grant 098051. F.L.M. was supported by National Institutes of Health (NIH) grant 1R01GM090087, by NSF grant DMS-1201234, and by a postdoctoral fellowship from the Stanford Center for Computational, Evolutionary and Human Genomics (CEHG). T.F.W. was supported by an AWS Education Grant, and the work of T.F.W., M.G., and Y.E. was supported in part by NIJ award 2014-DN-BX-K089. M.C. is supported by a Fundación Barrié Fellowship. H.S. and L. Coin are supported by Australian Research Council grants DP140103164 and FT110100972, respectively. M.G. was supported by a National Defense Science and Engineering Graduate Fellowship. G.R.S.R. was supported by the European Molecular Biology Laboratory and the Sanger Institute through an EBI–Sanger Postdoctoral Fellowship. X.Z.-B., P.F., D.R.Z., and L. Clarke were supported by Wellcome Trust grants 085532, 095908, and 104947 and by the European Molecular Biology Laboratory. P.A.U. was supported by SAP grant SP0#115016. C.L. was supported in part by NIH grant U41HG007497. Y.E. holds a Career Award at the Scientific Interface from the Burroughs Wellcome Fund. C.D.B. was supported by NIH grant 5R01HG003229-09.

AUTHOR CONTRIBUTIONS

G.D.P., Y.X., C.D.B., and C.T.-S. conceived and designed the project. R.B., S.L., and F.Y. generated FISH data. A. Malhotra, M.R., E.C., C.Z., and C.L. generated aCGH data. G.D.P., Y.X., F.L.M., T.F.W., A. Massaia, M.A.W.S., Q.A., S.A. McCarthy, A.N., S.K., Y.C., J.L.R.-F., M.C., H.S., M.G., R.D., G.R.S.R., T.W.F., E.G., A. Marcketta, D.M., X.Z.-B., G.R.A., S.A. McCarroll, P.F., P.A.U., L. Coin, D.R.Z., L. Clarke, A.A., Y.E., R.E.H., C.D.B., and C.T.-S. analyzed the data. G.D.P., Y.X., F.L.M., T.F.W., A. Massaia, M.A.W.S., Q.A., and C.T.-S. wrote the manuscript. All authors reviewed, revised, and provided feedback on the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare competing financial interests: details are available in the online version of the paper.

1. Jobling, M.A. & Tyler-Smith, C. The human Y chromosome: an evolutionary marker comes of age. *Nat. Rev. Genet.* **4**, 598–612 (2003).
2. Wei, W. *et al.* A calibrated human Y-chromosomal phylogeny based on resequencing. *Genome Res.* **23**, 388–395 (2013).
3. Poznik, G.D. *et al.* Sequencing Y chromosomes resolves discrepancy in time to common ancestor of males versus females. *Science* **341**, 562–565 (2013).
4. Wilson Sayres, M.A., Lohmueller, K.E. & Nielsen, R. Natural selection reduced diversity on human Y chromosomes. *PLoS Genet.* **10**, e1004064 (2014).
5. Karmin, M. *et al.* A recent bottleneck of Y chromosome diversity coincides with a global change in culture. *Genome Res.* **25**, 459–466 (2015).
6. Batini, C. *et al.* Large-scale recent expansion of European patrilineages shown by population resequencing. *Nat. Commun.* **6**, 7152 (2015).
7. Sikora, M.J., Colonna, V., Xue, Y. & Tyler-Smith, C. Modeling the contrasting Neolithic male lineage expansions in Europe and Africa. *Investig. Genet.* **4**, 25 (2013).
8. Helgason, A. *et al.* The Y-chromosome point mutation rate in humans. *Nat. Genet.* **47**, 453–457 (2015).
9. Fu, Q. *et al.* Genome sequence of a 45,000-year-old modern human from western Siberia. *Nature* **514**, 445–449 (2014).
10. Balanovsky, O. *et al.* Deep phylogenetic analysis of haplogroup G1 provides estimates of SNP and STR mutation rates on the human Y-chromosome and reveals migrations of Iranic speakers. *PLoS One* **10**, e0122968 (2015).
11. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
12. Handsaker, R.E. *et al.* Large multiallelic copy number variations in humans. *Nat. Genet.* **47**, 296–303 (2015).
13. 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).
14. Hammer, M.F. *et al.* Out of Africa and back again: nested cladistic analysis of human Y chromosome variation. *Mol. Biol. Evol.* **15**, 427–441 (1998).
15. Groucutt, H.S. *et al.* Rethinking the dispersal of *Homo sapiens* out of Africa. *Evol. Anthropol.* **24**, 149–164 (2015).
16. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
17. Zhang, F., Gu, W., Hurles, M.E. & Lupski, J.R. Copy number variation in human health, disease, and evolution. *Annu. Rev. Genomics Hum. Genet.* **10**, 451–481 (2009).

18. Willems, T. *et al.* Population-scale sequencing data enables precise estimates of Y-STR mutation rates. *Am. J. Hum. Genet.* http://dx.doi.org/10.1016/j.ajhg.2016.04.001 (2016).
19. 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).
20. Sudmant, P.H. *et al.* Global diversity, population stratification, and selection of human copy-number variation. *Science* **349**, aab3761 (2015).
21. Raghavan, M. *et al.* Genomic evidence for the Pleistocene and recent population history of Native Americans. *Science* **349**, aab3884 (2015).
22. de Filippo, C., Bostoen, K., Stoneking, M. & Pakendorf, B. Bringing together linguistic and genetic evidence to test the Bantu expansion. *Proc. R. Soc. B Biol. Sci.* **279**, 3256–3263 (2012).
23. Jobling, M.A., Hollox, E., Hurles, M., Kivisild, T. & Tyler-Smith, C. *Human Evolutionary Genetics* 2nd edn (Garland Science, 2014).
24. Allentoft, M.E. *et al.* Population genomics of Bronze Age Eurasia. *Nature* **522**, 167–172 (2015).
25. Haak, W. *et al.* Massive migration from the steppe was a source for Indo-European languages in Europe. *Nature* **522**, 207–211 (2015).
26. Harding, A.F. *European Societies in the Bronze Age* (Cambridge University Press, 2000).
27. Bryant, E.F. & Patton, L.L. *The Indo-Aryan Controversy: Evidence and Inference in Indian History* (Routledge, 2005).
28. Xue, Y. *et al.* Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. *Curr. Biol.* **19**, 1453–1457 (2009).
29. Betzig, L. Means, variances, and ranges in reproductive success: comparative evidence. *Evol. Hum. Behav.* **33**, 309–317 (2012).
30. Zerjal, T. *et al.* The genetic legacy of the Mongols. *Am. J. Hum. Genet.* **72**, 717–721 (2003).
31. Balaresque, P. *et al.* Y-chromosome descent clusters and male differential reproductive success: young lineage expansions dominate Asian pastoral nomadic populations. *Eur. J. Hum. Genet.* **23**, 1413–1422 (2015).
32. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis* (Springer, 2009).

[1]Program in Biomedical Informatics, Stanford University, Stanford, California, USA. [2]Department of Genetics, Stanford University, Stanford, California, USA. [3]Wellcome Trust Sanger Institute, Wellcome Genome Campus, Hinxton, 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, 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, Queensland, Australia. [12]Harvard–MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA. [13]Jackson Laboratory for Genomic Medicine, Farmington, Connecticut, USA. [14]European Molecular Biology Laboratory, European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, UK. [15]Department of Genetics, Albert Einstein College of Medicine, Bronx, New York, USA. [16]Virginia Bioinformatics Institute, Virginia Tech, Blacksburg, Virginia, USA. [17]Department of Biological Sciences, Virginia Tech, Blacksburg, Virginia, USA. [18]Department of Biostatistics, School of Public Health, University of Michigan, Ann Arbor, Michigan, USA. [19]Department of Genetics, Harvard Medical School, Boston, Massachusetts, USA. [20]Department of Life Sciences, Ewha Womans University, Seoul, Republic of 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).

ONLINE METHODS

Study samples. The 1000 Genomes Project Consortium sequenced the genomes of 2,535 individuals from 26 populations representing five global superpopulations (**Supplementary Table 1**). The Project's phase 3 analysis included 2,504 of these[11], and we used the Y-chromosome reads from the 1,244 males for this study.

SNVs, MNVs, and indels. To identify putative SNVs within the 10.3 Mb of the Y chromosome that is amenable to short-read sequencing[3], we generated six call sets using SAMtools[33], FreeBayes[34], Platypus[35], Cortex_var[36], and GATK UnifiedGenotyper[37,38] in both haploid and diploid modes. We used FreeBayes to construct a preliminary consensus call set and imposed filters for the number of alleles, genotype quality, read depth, mapping quality, missingness, and called heterozygosity. Finally, we called each genotype as the maximum-likelihood allele whenever a two-log-unit difference in likelihoods existed between the two possible states. For MNVs and indels, we imposed additional filters to exclude repetitive regions of the genome.

We used 11 high-coverage PCR-free genome sequences to estimate the FDR and 143 high-coverage Complete Genomics sequences to estimate the false negative rate and genotype concordance. We also estimated the singleton false positive rate by comparing the transition–transversion ratio among singletons to the corresponding ratio among shared SNVs.

CNVs. We discovered and genotyped CNVs using aCGH and two computational methods, Genome STRiP[12] and CnvHitSeq[13], across the entire euchromatic region. We ran GenomeSTRiP separately for uniquely alignable sequences and segmental duplications, using 5-kb and 10-kb windows and filtering calls on the basis of call rate, density of alignable positions, cluster separation, and manual review to assess duplication of findings and strength of evidence. We excluded ten samples with evidence of cell-line-specific clonal aneuploidy. To estimate FDR, we used the intensity rank-sum method[12] and probe intensity data from Affymetrix 6.0 SNP arrays.

We generated a second call set using the CnvHitSeq algorithm, which we modified to model read depth variation in a manner robust to the presence of repetitive regions and to estimate mosaicism. For the third call set, we used intensity ratios from 2,714 aCGH probes, with sample NA10851 as the reference. We segmented with the GADA algorithm[39,40], called genotypes on the basis of the distribution of mean $\log_2$-transformed intensity ratios using the additive background model of Conrad *et al.*[41], and imposed stringent criteria to minimize the FDR.

To validate the computational call sets, we used aCGH; alkaline-lysis fiber-FISH, following the protocol of Perry *et al.*[42]; and molecular combing fiber-FISH, following Polley *et al.*[43], Carpenter *et al.*[44], and instructions from the manufacturer, Genomic Vision.

Karyotyping for sex-chromosome aneuploidies. Metaphase chromosome spreads were prepared from lymphoblastoid cell lines (Coriell Biorepository) according to a standard protocol[45]. Chromosome-specific paint probes for the human X and Y chromosomes were generated from 5,000 copies of flow-sorted chromosomes, using the GenomePlex Whole-Genome Amplification kit (Sigma-Aldrich). Probes were labeled and FISH was performed following the strategy described in Gribble *et al.*[46].

STRs. We called genotypes using HipSTR and assessed call quality by comparing genotypes across three father–son pairs and by measuring concordance with capillary electrophoresis for 15 loci in the PowerPlex Y23 panel. To estimate Y-STR mutation rates, we used an approach we have fully described in a companion manuscript[18]. We modeled mutations with a geometric step size distribution and a spring-like length constraint, and, to account for PCR stutter artifacts and alignment errors, we learned an error model for each locus. We then leveraged the Y-chromosome SNV phylogeny to compute each sample's genotype posteriors, used a variant of Felsenstein's tree pruning algorithm[47] to evaluate the likelihood of a given mutation model, and optimized the model until convergence. We validated our estimates with simulations and compared them to published estimates when available.

Phylogeny. We assigned haplogroups using the 18 January 2014 version of the SNP Compendium maintained by the International Society of Genetic Genealogy (ISOGG). To construct a total-evidence maximum-likelihood tree, we converted genotype calls for the 60,555 biallelic SNVs to nexus format and ran RAxML8 (ref. 48) using the ASC_GTRGAMMA model. We then conducted 100 maximum-likelihood bootstraps and mapped these to the total-evidence tree. We partitioned the maximum-likelihood tree into eight overlapping subtrees, and 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. To estimate split times, we used two approaches to account for the modest coverage of our sequences. In the first approach, we pruned the sample to sequences with 5× or greater coverage, and in the second approach we traversed exclusively internal branches of the tree, as internal branches have high effective sequencing coverage due to the superposition of descending lineages. We calibrated using two mutation rate estimates from the literature[8,9].

Functional annotation. We used Ensembl's Variant Effect Predictor[49] to functionally annotate SNVs. To evaluate deleteriousness, we used Combined Annotation-Dependent Depletion (CADD)[50], SIFT[19], and PolyPhen[51].

Mitochondrial DNA. We excluded deletions and mutations proscribed by PhyloTree v.16 (ref. 52), generated a FASTA file using VCFtools[53], and aligned mtDNA sequences to the revised Cambridge Reference Sequence (rCRS) using MEGA6 (ref. 54). We assigned haplogroups to each sample using HaploGrep[55], manually checked all variant calls, inferred the mtDNA phylogeny using RAxML[48], and plotted the tree using FigTree.

Diversity. We used 141 high-coverage Complete Genomics sequences to compare mtDNA diversity to that of the Y chromosome. Seeking to recapitulate this observed relative diversity, as well as the observed diversity of the X chromosome and the autosomes, we used standard neutral coalescent simulations implemented in the program ms[56] to simulate data for the four chromosome types under a series of demographic models. In all models, we held the autosomal effective population size fixed to values previously described for African and European demographic histories[57,58], but we varied the ratio of male/female effective population sizes.

Haplogroup expansions. To estimate male-lineage growth rates, we developed a two-phase exponential growth model wherein the first phase coincides with an apparent rapid haplogroup expansion and the second phase links the first phase to the earliest time for which reasonable estimates exist of the size of the relevant population. Our primary objective was to estimate the duration of the first phase, T_1, and the effective number of carriers of a haplogroup at its conclusion, N_1, to estimate the growth rate during this period—the mean number of sons per man per generation. To do so, we conducted maximum-likelihood inference over a grid of (T_1, N_1) points for each of a sequence of 'sampling' times, T_s, defined by pruning the subtree of a phylogenetic node of interest to a fixed root-to-tip height (number of SNPs) (**Supplementary Fig. 29**).

With N_2 fixed, we needed one additional parameter, T_2, to specify the full demographic model corresponding to each (T_1, N_1) point to simulate two-phase growth. We estimated T_2 using 10,000 ms coalescent simulations[56] constrained by the TMRCA of the node of interest. With T_2 and N_2 in hand, we simulated two-phase growth to assemble a reference distribution of site frequency spectra (SFS) against which to compare the observed data. We did so 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 up to ten possible T_s values, the lattice contained up to 15,360 points, and for each we conducted 16,384 ms simulations of two-phase growth, fixing the number of lineages equal to that of the pruned observed tree. For each T_s, we approximated the likelihood of a particular (T_1, N_1) point by comparing the SFS values of the observed tree to those of the corresponding reference distribution, using an SFS distance measure we defined. Finally, we used the resulting likelihood contours to infer the magnitude of growth in the first phase.

doi:10.1038/ng.3559

33. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
34. Garrison, E. & Marth, G. Haplotype-based variant detection from short-read sequencing. *arXiv* http://arxiv.org/abs/1207.3907 (2012).
35. Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* **46**, 912–918 (2014).
36. 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).
37. McKenna, A. *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
38. DePristo, M.A. *et al.* A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.* **43**, 491–498 (2011).
39. Pique-Regi, R. *et al.* Sparse representation and Bayesian detection of genome copy number alterations from microarray data. *Bioinformatics* **24**, 309–318 (2008).
40. 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).
41. Conrad, D.F. *et al.* Origins and functional impact of copy number variation in the human genome. *Nature* **464**, 704–712 (2010).
42. Perry, G.H. *et al.* Copy number variation and evolution in humans and chimpanzees. *Genome Res.* **18**, 1698–1710 (2008).
43. Polley, S. *et al.* Evolution of the rapidly mutating human salivary agglutinin gene (*DMBT1*) and population subsistence strategy. *Proc. Natl. Acad. Sci. USA* **112**, 5105–5110 (2015).
44. 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).
45. Verma, R.S. & Babu, A. *Human Chromosomes: Principles & Techniques* 2nd edn. (McGraw-Hill, 1995).
46. 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).
47. Felsenstein, J. Evolutionary trees from DNA sequences: a maximum likelihood approach. *J. Mol. Evol.* **17**, 368–376 (1981).
48. Stamatakis, A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**, 1312–1313 (2014).
49. McLaren, W. *et al.* Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**, 2069–2070 (2010).
50. Kircher, M. *et al.* A general framework for estimating the relative pathogenicity of human genetic variants. *Nat. Genet.* **46**, 310–315 (2014).
51. Adzhubei, I.A. *et al.* A method and server for predicting damaging missense mutations. *Nat. Methods* **7**, 248–249 (2010).
52. van Oven, M. & Kayser, M. Updated comprehensive phylogenetic tree of global human mitochondrial DNA variation. *Hum. Mutat.* **30**, E386–E394 (2009).
53. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).
54. 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).
55. Kloss-Brandstätter, A. *et al.* HaploGrep: a fast and reliable algorithm for automatic classification of mitochondrial DNA haplogroups. *Hum. Mutat.* **32**, 25–32 (2011).
56. Hudson, R.R. Generating samples under a Wright–Fisher neutral model of genetic variation. *Bioinformatics* **18**, 337–338 (2002).
57. 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).
58. 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).