Title

Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males Versus Females

Permalink
https://escholarship.org/uc/item/1425g034

Journal
Science, 341(6145)

ISSN
0036-8075

Authors
Poznik, G David
Henn, Brenna M
Yee, Muh-Ching
et al.

Publication Date
2013-08-02

DOI
10.1126/science.1237619

Peer reviewed

Published in final edited form as:
Science. 2013 August 2; 341(6145): 562–565. doi:10.1126/science.1237619.

Sequencing Y Chromosomes Resolves Discrepancy in Time to Common Ancestor of Males versus Females

G. David Poznik[1,2], **Brenna M. Henn**[3,4], **Muh-Ching Yee**[3], **Elzbieta Sliwerska**[5], **Ghia M. Euskirchen**[3], **Alice A. Lin**[6], **Michael Snyder**[3], **Lluis Quintana-Murci**[7,8], **Jeffrey M. Kidd**[3,5], **Peter A. Underhill**[3], and **Carlos D. Bustamante**[3,*]

[1]Program in Biomedical Informatics, Stanford University School of Medicine, Stanford, CA

[2]Department of Statistics, Stanford University, Stanford, CA

[3]Department of Genetics, Stanford University School of Medicine, Stanford, CA

[4]Department of Ecology and Evolution, Stony Brook University, Stony Brook, NY

[5]Department of Human Genetics and Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[6]Department of Psychiatry, Stanford University, Stanford, CA

[7]Institut Pasteur, Unit of Human Evolutionary Genetics, 75015 Paris, France

[8]Centre National de la Recherche Scientifique, URA3012, 75015 Paris, France

Abstract

The Y chromosome and the mitochondrial genome (mtDNA) have been used to estimate when the common patrilineal and matrilineal ancestors of humans lived. We sequenced the genomes of 69 males from nine populations, including two in which we find basal branches of the Y chromosome tree. We identify ancient phylogenetic structure within African haplogroups and resolve a long-standing ambiguity deep within the tree. Applying equivalent methodologies to the Y and mtDNA, we estimate the time to the most recent common ancestor (T_{MRCA}) of the Y chromosome to be 120–156 thousand years and the mtDNA T_{MRCA} to be 99–148 ky. Our findings suggest that, contrary to prior claims, male lineages do not coalesce significantly more recently than female lineages.

The Y chromosome contains the longest stretch of non-recombining DNA in the human genome and is therefore a powerful tool with which to study human history. Estimates of the time to the most recent common ancestor (T_{MRCA}) of the Y chromosome have differed by approximately twofold from T_{MRCA} estimates for the mitochondrial genome. Y chromosome coalescence time has been estimated in the range 50–115 ky (1–3), although larger values have been reported (4, 5), whereas estimates for mitochondrial DNA (mtDNA) range from 150–240 ky (3, 6, 7). However, the quality and quantity of data available for these two uniparental loci have differed substantially. While the complete mitochondrial genome has

[*]Correspondence to: cdbustam@stanford.edu.

been resequenced thousands of times (6, 8), fully sequenced diverse Y chromosomes have only recently become available. Previous estimates of the Y chromosome T_{MRCA} relied on short resequenced segments, rapidly mutating microsatellites, or single nucleotide polymorphisms (SNPs) ascertained in a small panel of individuals and then genotyped in a global panel. These approaches likely underestimate genetic diversity and, consequently, T_{MRCA} (9).

We sequenced the complete Y chromosomes of 69 males from seven globally diverse populations of the Human Genome Diversity Panel (HGDP) and two additional African populations: San (Bushmen) from Namibia, Mbuti Pygmies from the Democratic Republic of Congo, Baka Pygmies and Nzebi from Gabon, Mozabite Berbers from Algeria, Pashtuns (Pathan) from Pakistan, Cambodians, Yakut from Siberia, and Mayans from Mexico (Fig. S1). Individuals were selected without regard to their Y chromosome haplogroups.

The Y chromosome reference sequence is 59.36 Mb, but this includes a 30 Mb stretch of constitutive heterochromatin on the q-arm, a 3 Mb centromere, 2.65 Mb and 330 kb telomeric pseudoautosomal regions (PAR) that recombine with the X chromosome, and eight smaller gaps. We mapped reads to the remaining 22.98 Mb of assembled reference sequence, which consists of three sequence classes defined by their complexity and degree of homology to the X chromosome (10): “X-degenerate,” “X-transposed,” and “ampliconic.” Both the high degree of self-identity within the ampliconic tracts and the X chromosome homology of the X-transposed region render portions of the Y chromosome ill-suited for short read sequencing. To address this, we constructed filters that reduced the data to 9.99 million sites (11) (Figs. 1, S2). We then implemented a haploid model EM algorithm to call genotypes (11).

We identified 11,640 single nucleotide variants (SNVs; Fig. S3). 2,293 (19.7%) are present in dbSNP (v135), and we assigned haplogroups on the basis of the 390 (3.4%) present in the International Society of Genetic Genealogy (ISOGG) database (12) (Fig. S4). At SNVs, median haploid coverage was 3.1× (IQR: 2.6–3.8×; Table S1, Fig. S5), and sequence validation suggests a genotype calling error rate on the order of 0.1% (11).

Because mutations accumulate over time along a single lengthy haplotype (13), the male-specific region of the Y chromosome provides power for phylogenetic inference. We constructed a maximum likelihood tree from 11,640 SNVs using the Tamura-Nei nucleotide substitution model (Fig. 2) and, in agreement with (14), observe strong bootstrap support (500 replicates) for the major haplogroup branching points. The tree both recapitulates and adds resolution to the previously inferred Y chromosome phylogeny (Fig. S6), and it characterizes branch lengths free of ascertainment bias. We identify extraordinary depth within Africa, including lineages sampled from the San hunter-gatherers that coalesce just short of the root of the entire tree. This stands in contrast to a tree from autosomal SNP genotypes (15) wherein African branches were considerably shorter than others; genotyping arrays primarily rely on SNPs ascertained in European populations and therefore undersample diversity within Africa. Two regions of reduced branch length in our tree correspond to rapid expansions: the Out of Africa event (downstream of F-M89) and the agriculture catalyzed Bantu expansions (downstream of E-M2). Among the three huntergatherer populations, we find a relatively high number of B2 lineages. Within this haplogroup, six Baka B-M192 individuals form a distinct clade that does not correspond to extant definitions (11) (Fig. S7). We estimate this previously uncharacterized structure to have arisen approximately 35 kya.

We resolve the polytomy of the Y macro-haplogroup F (16) by determining the branching order of haplogroups G, H, and IJK (Figs. 2, S6). We identified a single variant (rs73614810, a C→T transition dubbed “M578”) for which haplogroup G retains the ancestral allele, whereas its brother clades (H and IJK) share the derived allele. Genotyping M578 in a diverse panel confirmed the finding (Table S2). We thereby infer more recent common ancestry between hgH and hgIJK than between either and hgG. M578 defines an early diversification episode of the Y phylogeny in Eurasia (11).

To account for missing genotypes, we assigned each SNV to the root of the smallest subtree containing all carriers of one allele or the other and inferred that the allele specific to the subtree was derived (Fig. S8). We used the chimpanzee Y chromosome sequence to polarize 398 variants assigned to the deepest split—a task complicated by significant structural divergence (11, 17).

We estimated the coalescence time of all Y chromosomes using both a molecular clock based frequentist estimator and an empirical Bayes approach that uses a prior distribution of T_{MRCA} from coalescent theory and conducts Markov chain simulation to estimate the likelihood of parameters given a set of DNA sequences (GENETREE) (11, 18) (Table 1). To directly compare the T_{MRCA} of the Y chromosome to that of the mtDNA, we estimated their respective mutation rates by calibrating phylogeographic patterns from the initial peopling of the Americas, a recent human event with high confidence archeological dating.

Archeological evidence indicates that humans first colonized the Americas approximately 15 kya via a rapid coastal migration that reached Monte Verde II in southern Chile by 14.6 kya (19). The two Native American Mayans represent Y chromosome hgQ lineages, Q-M3 and Q-L54*(xM3), that likely diverged at about the same time as the initial peopling of the continents. Q is defined by the M242 mutation that arose in Asia. A descendent haplogroup, Q-L54, emerged in Siberia and is ancestral to Q-M3. Because the M3 mutation appears to be specific to the Americas (20), it likely occurred subsequent to the initial entry, and the prevalence of M3 in South America suggests that it emerged prior to the southward migratory wave. Consequently, the divergence between these two lineages provides an appropriate calibration point for the Y mutation rate. The large number of variants that have accumulated since divergence, 120 and 126, contrasts with the pedigree-based estimate of the Y chromosome mutation rate, which is based on just 4 mutations (21). Using entry to the Americas as a calibration point, we estimate a mutation rate of 0.82×10^{-9} /bp/yr (95% CI: $0.72–0.92 \times 10^{-9}$ /bp/yr; Table S3). False negatives have minimal effect on this estimate due to the low probability, at 5.7× and 8.5× coverage, of observing fewer than two reads at a site (observed proportions: 3.1%, 0.6%) and due to the fact that the number of unobserved singletons possessed by one individual is offset by a similar number of Q doubletons unobserved in the same individual and thereby misclassified as singletons possessed by the other (11) (Figs. S9, S10). This calibration approach assumes approximate coincidence between the expansion throughout the Americas and the divergence of Q-M3 and Q-L54*(xM3), but we consider deviation from this assumption and identify a strict lower bound on the point of divergence using sequences from the 1000 Genomes Project (11). As a comparison point, we consider the Out of Africa expansion of modern humans, which dates to approximately 50 kya (22) and yields a similar mutation rate of 0.79×10^{-9} /bp/yr.

We constructed an analogous pipeline for high coverage (>250×) mtDNA sequences from the 69 male samples and an additional 24 females from the seven HGDP populations (11) (Fig. S11). As in the Y chromosome analysis, we calibrated the mtDNA mutation rate using divergence within the Americas. We selected the pan-American hgA2, one of several initial founding haplogroups amongst Native Americans. The star-shaped phylogeny of hgA2 subclades suggests that its divergence was coincident with the rapid dispersal upon the initial colonization of the continents (23). Calibration on 108 previously analyzed hgA2 sequences (11) (Fig. S12) yields a point estimate equivalent to that from our seven Mayan mtDNAs, but within a narrower confidence interval. From this within-human calibration, we estimate a mutation rate of 2.3×10^{-8} /bp/yr (95% CI: $2.0–2.5 \times 10^{-8}$ /bp/yr), higher than that from human-chimpanzee divergence but similar to other estimates using within-human calibration points (24, 25).

The global T_{MRCA} estimate for any locus constitutes an upper bound for the time of human population divergence under models without gene flow. We estimate the Y chromosome T_{MRCA} to be 138 ky (120–156 ky) and the mtDNA T_{MRCA} to be 124 ky (99–148 ky; Table 1) (11). Our mtDNA estimate is more recent than many previous studies, the majority of which used mutation rates extrapolated from between-species divergence. However, mtDNA mutation rates are subject to a time-dependent decline, with pedigree-based estimates on the faster end of the spectrum and species-based estimates on the slower. Because of this time dependency and the need to calibrate the Y and mtDNA in a comparable manner, it is more appropriate here to use within-human clade estimates of the mutation rate.

Rather than assume the mutation rate to be a known constant, we explicitly account for the uncertainty in its estimation by modeling each T_{MRCA} as the ratio of two random variables. We estimate the ratio of the mtDNA T_{MRCA} to that of the Y chromosome to be 0.90 (95% CI: 0.68–1.11; Fig. S13). If, as argued above, the divergence of the Y chromosome Q lineages occurred at approximately the same time as that of the mtDNA A2 lineages, then the T_{MRCA} ratio is invariant to the specific calibration time used. Regardless, the conclusion of parity is robust to possible discrepancy between the divergence times within the Americas (11). Using comparable calibration approaches, the Y and mtDNA coalescence times are not significantly different. This conclusion would hold whether or not an alternative approach would yield more definitive T_{MRCA} estimates.

Our observation that the T_{MRCA} of the Y chromosome is similar to that of the mtDNA does not imply that the effective population sizes of males and females are similar. In fact, we observe a larger N_e in females than in males (Table 1). While, due to its larger N_e, the distribution from which the mitochondrial T_{MRCA} has been drawn is right-shifted with respect to that of the Y T_{MRCA}, the two distributions have large variances and overlap (Fig. 3).

Dogma has held that the common ancestor of human patrilineal lineages, popularly referred to as the Y chromosome "Adam," lived considerably more recently than the common ancestor of female lineages, the so-called mitochondrial "Eve." However, we conclude that the mitochondrial coalescence time is not substantially greater than that of the Y chromosome. Indeed, due to our moderate-coverage sequencing and the existence of additional rare divergent haplogroups, our analysis may yet underestimate the true Y T_{MRCA}.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank O. Cornejo, S. Gravel, D. Siegmund, and E. Tsang for helpful discussions; M. Sikora and H. Costa for mapping reads from Gabonese samples; and H. Cann for assistance with HGDP samples. Supported by NLM training grant LM-07033 and NSF graduate research fellowship DGE-1147470 (GDP); NIH grant 3R01HG003229 (BMH and CDB); NIH grant DP5OD009154 (JMK and ES); Institut Pasteur, a CNRS "MIE" Grant, and a Foundation Simone et Cino del Duca Research Grant (LQM). PAU consulted on, PAU and BMH have stock in, and CDB is on the advisory board of a project at 23andMe. CDB is on the Scientific Advisory boards of Personalis, Inc.; InVitae (formerly Locus Development, Inc.); and Ancestry.com. MS is a Scientific Advisory member and founder of Personalis, a Scientific Advisory member for Genapsys Former, and a consultant for Illumina and Beckman Coulter Society for American Medical Pathology. BMH formerly had a paid consulting relationship with Ancestry.com. Genotype and variant data are provided in the supporting online material. Variants have been deposited to dbSNP (ss825679106–825690384). Individual level genetic data are available, through a data access agreement to respect the privacy of the participants for transfer of genetic data, by contacting C.D.B.

References and Notes

1. Pritchard JK, Seielstad MT, Perez-Lezaun A, Feldman MW. Population growth of human Y chromosomes: a study of Y chromosome microsatellites. Mol Biol Evol. 1999; 16:1791–8. [PubMed: 10605120]
2. Thomson R, Pritchard JK, Shen P, Oefner PJ, Feldman MW. Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc Natl Acad Sci U S A. 2000; 97:7360–5. [PubMed: 10861004]
3. Tang H, Siegmund DO, Shen P, Oefner PJ, Feldman MW. Frequentist estimation of coalescence times from nucleotide sequence data using a tree-based partition. Genetics. 2002; 161:447–59. [PubMed: 12019257]
4. Hammer MF. A recent common ancestry for human Y chromosomes. Nature. 1995; 378:376–8. [PubMed: 7477371]
5. 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. 2011; 88:814–8. [PubMed: 21601174]
6. Ingman M, Kaessmann H, Pääbo S, Gyllensten U. Mitochondrial genome variation and the origin of modern humans. Nature. 2000; 408:708–13. [PubMed: 11130070]
7. Cann RL, Stoneking M, Wilson AC. Mitochondrial DNA and human evolution. Nature. 1987; 325:31–6. [PubMed: 3025745]
8. Underhill PA, Kivisild T. Use of y chromosome and mitochondrial DNA population structure in tracing human migrations. Annu Rev Genet. 2007; 41:539–64. [PubMed: 18076332]
9. Jobling MA, Tyler-Smith C. The human Y chromosome: an evolutionary marker comes of age. Nat Rev Genet. 2003; 4:598–612. [PubMed: 12897772]
10. Skaletsky H, et al. The male-specific region of the human Y chromosome is a mosaic of discrete sequence classes. Nature. 2003; 423:825–37. [PubMed: 12815422]
11. Materials and methods are available as supplementary material on *Science* Online
12. ISOGG: International Society of Genetic Genealogy. 2013. available at http://www.isogg.org/

13. Underhill PA, et al. The phylogeography of Y chromosome binary haplotypes and the origins of modern human populations. Ann Hum Genet. 2001; 65:43–62. [PubMed: 11415522]

14. Wei W, et al. A calibrated human Y-chromosomal phylogeny based on resequencing. Genome Res. 2013; 23:388–95. [PubMed: 23038768]

15. Li JZ, et al. Worldwide human relationships inferred from genome-wide patterns of variation. Science. 2008; 319:1100–4. [PubMed: 18292342]

16. Karafet TM, et al. New binary polymorphisms reshape and increase resolution of the human Y chromosomal haplogroup tree. Genome Res. 2008; 18:830–8. [PubMed: 18385274]

17. Hughes JF, et al. Chimpanzee and human Y chromosomes are remarkably divergent in structure and gene content. Nature. 2010; 463:536–9. [PubMed: 20072128]

18. Griffiths RC, Tavaré S. Sampling theory for neutral alleles in a varying environment. Philos Trans R Soc Lond B Biol Sci. 1994; 344:403–10. [PubMed: 7800710]

19. Goebel T, Waters MR, O'Rourke DH. The late Pleistocene dispersal of modern humans in the Americas. Science. 2008; 319:1497–502. [PubMed: 18339930]

20. Dulik MC, 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. 2012; 90:229–46. [PubMed: 22281367]

21. Xue Y, et al. Human Y chromosome base-substitution mutation rate measured by direct sequencing in a deep-rooting pedigree. Curr Biol. 2009; 19:1453–7. [PubMed: 19716302]

22. Klein RG. Out of Africa and the evolution of human behavior. Evol Anthropol. 2008; 17:267–281.

23. Kumar S, et al. Large scale mitochondrial sequencing in Mexican Americans suggests a reappraisal of Native American origins. BMC Evol Biol. 2011; 11:293. [PubMed: 21978175]

24. Ho SYW, Phillips MJ, Cooper A, Drummond AJ. Time dependency of molecular rate estimates and systematic overestimation of recent divergence times. Mol Biol Evol. 2005; 22:1561–8. [PubMed: 15814826]

25. Henn BM, Gignoux CR, Feldman MW, Mountain JL. Characterizing the time dependency of human mitochondrial DNA mutation rate estimates. Mol Biol Evol. 2009; 26:217–30. [PubMed: 18984905]

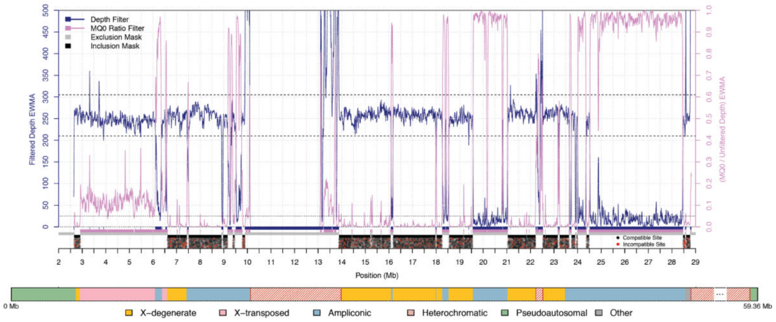

Fig. 1. Callability mask for the Y chromosome

Exponentially-weighted moving averages of read depth (blue line) and the proportion of reads mapping ambiguously (MQ0 ratio; violet line) versus physical position. Regions with values outside the envelopes defined by the dashed lines (depth) or dotted lines (MQ0) were flagged (blue and violet boxes) and merged for exclusion (gray boxes). The complement (black boxes) defines the regions within which reliable genotype calls can be made. Below, a scatter plot indicates the positions of all observed SNVs. Those incompatible with the inferred phylogenetic tree (red) are uniformly distributed. The X-degenerate regions yield quality sequence data, ampliconic sequences tend to fail both filters, and mapping quality is poor in the X-transposed region.

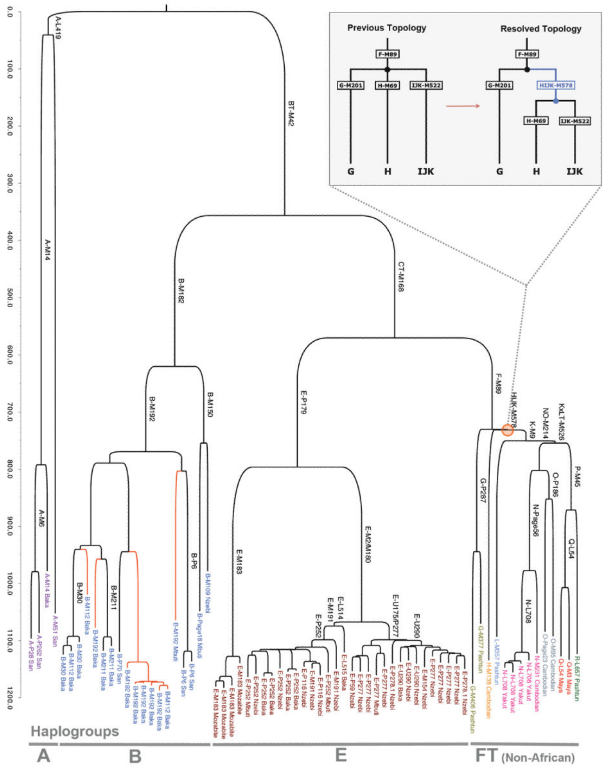

Fig. 2. Y chromosome phylogeny inferred from genomic sequencing

This tree recapitulates the previously known topology of the Y chromosome phylogeny; however, branch lengths are now free of ascertainment bias. Branches are drawn proportional to the number of derived SNVs. Internal branches are labeled with defining ISOGG variants inferred to have arisen on the branch. Leaves are colored by major haplogroup cluster and labeled with the most derived mutation observed and the population from which the individual was drawn. Previously uncharacterized structure within African hgB2 is indicated in orange. (**Inset**) Resolution of a polytomy was possible through the identification of a variant for which hgG retains the ancestral allele, whereas hgH and hgIJK share the derived allele.

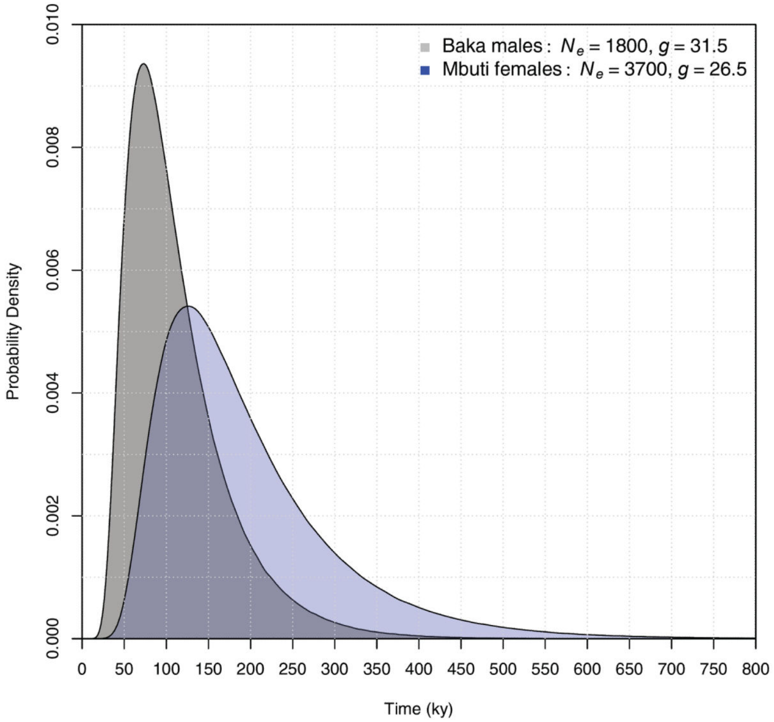

Fig. 3. Similarity of T_{MRCA} does not imply equivalent N_e of males and females

The T_{MRCA} for a given locus is drawn from a predata (i.e., prior) distribution that is a function of N_e, generation time, sample size, and demographic history. Consider the distribution of possible T_{MRCA}'s for a set of 100 uniparental chromosomes. Although the Mbuti mtDNA N_e is twice as large as that of the Baka Y chromosome, the corresponding predata T_{MRCA} distributions overlap considerably.

Table 1

T_{MRCA} and N_e estimates for the Y chromosome and mtDNA.

	Y Chromosome				mtDNA			
Method	**Pop**	***n***	T_{MRCA} [a]	N_e	**Pop**	***n***	T_{MRCA} [a]	N_e
Molecular Clock	All	69	139 (120–156)	4500[b]	All	93	124 (99–148)	9500[b]
GENETREE [3]	San	6	128 (112–146)	3800	Nzebi	18	105 (91–119)	11,500
	Baka	11	122 (106–137)	1800	Mbuti	6	121 (100–143)	3700

[a] Employs mutation rate estimated from within-human calibration point. Times measured in ky.

[b] Uses Watterson's estimator, $\hat{\theta}_w$.

[3] Each coalescent analysis restricted to a single population spanning the ancestral root (11).