Deep coalescent history of the hominin lineage

Trevor Cousins[†], Regev Schweigher[†], Richard Durbin*

Department of Genetics, University of Cambridge, Cambridge, UK

[†] These authors contributed equally to this work.

* Corresponding author. Email: rd109@cam.ac.uk

1 Abstract

Coalescent-based methods are widely used to infer population size histories, but existing analyses have limited resolution for deep time scales (> 2 million years ago). Here we extend the scope of such inference by re-analysing an ancient peak seen in human and chimpanzee effective population size around 5-7 million years ago, showing that coalescent-based inference can be extended much further into the past than previously thought. This peak is consistently observed across human and chimpanzee populations, but not in gorillas or orangutans. We show that it is unlikely to be an artefact of model violations, and discuss its potential implications for understanding hominin evolutionary history, in particular the human-chimpanzee speciation.

2 Introduction

The study of human population size over time provides a valuable lens through which to explore key questions about our evolutionary past [1]. Population size history can be inferred from present day genome sequences by examining patterns of genetic variation within and between populations. Various methods have been developed for such inference, using information such as the site frequency spectrum, linkage disequilibrium, or the genomic distribution of heterozygosity. One popular method is the pairwise sequentially Markovian coalescent (PSMC) [2], as well as its successors [3, 4, 5, 6, 7]. These have enabled various insights the evolutionary history of humans, including the magnitude of exponential growth in the last 20 thousand years (kya), the timing and severity of the out of Africa bottleneck 50-60kya, and divergence times between various subpopulations as well as between humans, Neanderthals and Denisovans [8, 9].

Methods to infer human population size history have typically focused on studying evolution more recent than 1-2 million years ago (Mya). The mean coalescence time in humans is ∼1Mya, indicating that there should be information within a present day genome sequence to probe more ancient time scales. Insights into population size history beyond this period would be particularly interesting, as it may elucidate the emergence of the genera *Homo* (∼3Mya [10, 11]) and *Australopithicus* (∼4.5Mya [12]), or even the ancestral human-chimpanzee population (∼6Mya [13]). The inference of $N(t)$ from some methods, including PSMC, do extend to this period, however it is typically not discussed in the existing literature. This may be because it has until now been unclear how robust inference is in those time frames, and in particular how badly the model is affected when its assumptions are violated.

Another set of methods uses cross-species analyses, in this case of present day sequences of humans, chimpanzees, and gorillas, to analyse the way in which these populations diverged. Methods such as CoalHMM and its successors employ coalescent hidden Markov model approaches, in which the hidden states correspond to the order of coalescent events, as well sometimes as ancestral coalescence times [14, 15, 16, 13, 17]. Typically the model parameters to be estimated are the divergence times and ancestral population sizes, which in turn are informative about the amount of incomplete lineage sorting (ILS) or gene flow. Some of these studies have suggested that the human-chimp speciation was "complex", in that after initial population divergence there was a period of gene flow before final separation [18, 19, 13, 20], though others have suggested a clean split scenario [21, 22]. Many of these analyses make coarse assumptions about population size changes in the ancestral populations, and typically assume these were panmictic. Typically the size of the human-chimpanzee ancestral population is estimated to be very large, for example this is estimated as 167,400 in [17] and 177,368 in [23], both of which are more than 8 times the long-term effective population size of humans or chimpanzee. It is plausible that the size of the human-chimpanzee ancestral population really was extremely large, however it may be that this population was structured, which leads to inflated estimates of population size when assuming panmixia [24]. In this article we examine the changes in size of the ancestral human/chimp population in higher resolution and with fewer assumptions, finding a similar magnitude to that inferred in recent time.

So far existing methods have offered limited population genetic insight into human evolution between 1-2Mya and the divergence of the ancestral human/chimpanzee population around 6Mya [14, 13, 25]. Here, we examine the potential to broaden the scope of coalescent-based population size history inference beyond two million years ago. Specifically, we discuss an ancient inferred peak in effective population size around 5-7Mya seen in PSMC analysis of human and chimpanzee genomes, but not in gorilla or orangutan. While this peak has been partially observed in several previous studies in humans and chimpanzees [2, 8, 9, 16, 26, 27], it has received little attention in the literature. We present a revised analysis of this peak using modern data and discuss its potential implications for understanding the evolutionary history of humans and other great apes. We take care to address potential model violations and other factors that may affect the reliability of these inferences. Overall, our study suggests that we may be able to significantly extend the applicability of coalescent-based inference much further into the past, generating new insights into the complex history of hominin evolution.

3 Results

3.1 An ancient peak in human and chimpanzee population size history

We separately inferred an effective population size ($N(t)$) curve over time for 26 single diploid genome sequences, each from a distinct population in the 1000 Genomes Project [28, 29, 27, 30] (1000GP) using PSMC (Figure 1a). Importantly, we extended previous similar analyses by fitting the model with fewer parameter constraints [31] and using the high coverage (30x) whole-genome sequencing data resource [30], allowing greater resolution at time scales >2Mya (see Methods). The resulting $N(t)$ estimates start between 5 kya and end around 10Mya, which is deeper in the past than in previous studies [2, 8, 9, 26]. In all human samples two peaks can be seen, one around the time of appearance of modern humans approximately 200kya and a second older one, starting around 2Mya going backwards in time, reaching a maximum around 5.5Mya with

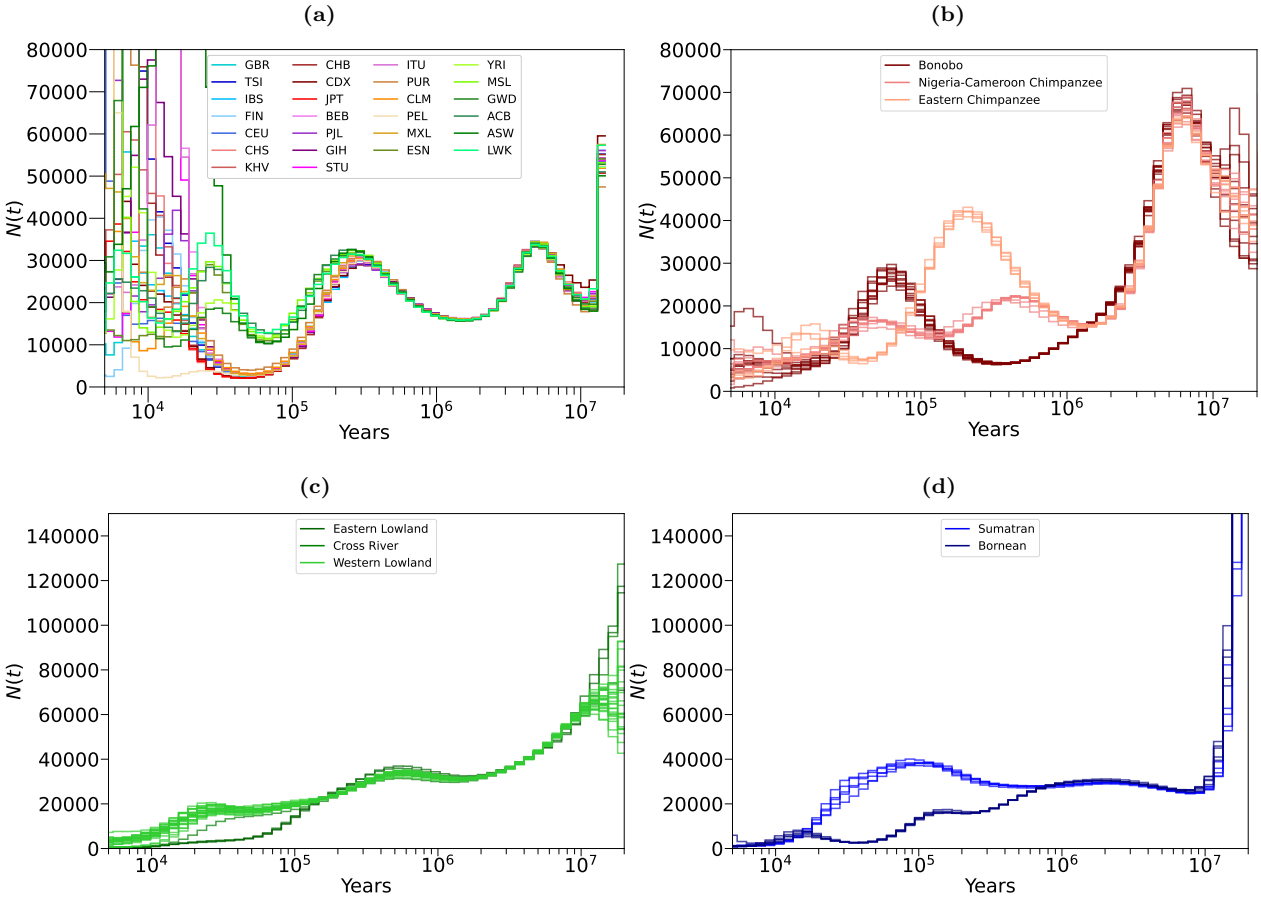

Figure 1: Inference from PSMC on humans, chimpanzees and bonobos, gorillas, and orangutans. **(a)** Estimates of population size history on humans, using 26 populations from the 1000 Genomes project, using one individual per population. **(b)** Estimates of population size history on chimpanzees and bonobos, using 8 Nigeria-Cameroon chimpanzees, 5 Eastern chimpanzees, and 11 bonobos. **(c)** Estimates of population size history on gorillas, using 3 Eastern lowland individuals, 23 Western lowland individuals, and 1 Cross River individual. **(d)** Estimates of the population size history on orangutans, using 5 Sumatran individuals and 5 Bornean individuals.

a population size of ∼37,500, then decreasing until around 10Ma. While the ancient peak is partially visible in previous whole-genome analyses [2, 9, 26], it is frequently truncated or represented with poor resolution because of the way time intervals were specified. 30 iterations of block bootstrapping reveals there is less variance in $N(t)$ in ancient time than there is recent time (Figure A1).

As the ancient peak occurs around the estimated period of human and chimpanzee species separation, it is natural to ask if a similar signal exists in chimpanzee population size history, or in other great apes in general. We therefore used PSMC to infer the population size history from 79 great ape genomes, as in [16] (see Methods). First, we recovered an ancient peak in chimpanzees and bonobos, overlapping the peak seen in the human analysis, though perhaps with its maximum slightly older (Figure 1b). This extends a similar signal further back in time, present in previous analyses but not discussed [16]. As with humans, our re-analysis produces the ancient peak more clearly because we avoid grouping old time intervals. We

note that the effective population size of the chimpanzee ancient peak is significantly higher than that of humans. The fact that the peaks observed in humans and chimpanzees align with the time frame of their estimated divergence prompts the question of whether these peaks signify a shared history between the two species. Indeed, as we discuss in the Appendix, correcting for mutation rate variations in time, as well as for variable generation times, in principle could align the two population size curves, both in their time period and their scale. However, uncertainty in these parameters, as well as in curve estimation, precludes a confident conclusion.

For gorillas (Figure 1c) and orangutans (Figure 1d), we do not detect a discernible peak in the region between 5-10Mya. This is consistent with the fact that gorillas and orangutans genetically separated from humans and chimpanzees prior to the 5-7Mya time period, and so would not be expected to share their history at that time. Moreover, it supports the claim that the ancient peak is not an artefact of the method or of the data processing, as that would result in a peak with gorillas and orangutans as well.

3.2 The fraction of uncoalesced genome as a function of time

The power of PSMC to infer $N(t)$ at 5.5Mya relies on there being sufficiently many places in the genome where the two input lineages coalesce more anciently than this time. To assess whether this is reasonable to expect, we note that under a simple panmictic model the observed range of heterozygosity in humans (e.g., $\sim$0.00069 for PEL to $\sim$0.001 for MSL in 1000GP) would correspond to a constant effective diploid population size of 13,800-20,000, using a mutation rate of 1.25e-8 per base pair per generation. Given an exponential distribution of coalescence times, and a generation time of 30 years, we expect a mean coalescence time of 828,000-1,200,000 years, and that the coalescence time will be older than 5.5Mya in 0.12%-1.01% of the genome, which would be sufficient to provide plenty of information for this period, in particular taking note that older regions of coalescence are correspondingly shorter than younger ones.

As more direct confirmation from the data, we inferred the fraction of the genome that has not yet coalesced as a function of time for the 1000GP samples, using PSMC decoding (see Methods). At least 1% of the genome is estimated not to have coalesced by 5Mya, and approximately 0.1% by 10Mya (Figure 2a). Ancient regions tend to be shared among individuals (Figure 2b); e.g. positions with an inferred posterior mean TMRCA of over 3Mya show an average correlation of r^2=0.21 across individuals (Figure A2). We note that further back in time, segments of shared ancestry become increasingly shorter because of longer exposure to recombination, and that these are scattered across the genome. For example we estimate that African individuals harbor roughly 4,000 non-contiguous segments whose coalescence time is older than 4Mya (Table 1; see Methods). We note that previous simulation studies [32, 33] and theoretical analysis [34] have shown that posterior mean estimates of true ancient coalescent times substantially older than the mean tend to be significantly downward biased, which may indicate an even older age for detected ancient regions. We repeated the analysis using an independent method, Relate [35], which does not rely on the PSMC assumptions and uses multiple sequences jointly; this also identifies sufficient coalescence material exists in this time period (see Methods and Figure A3). We further note that if we alter the time discretisation parameters in PSMC, then the inference of human $N(t)$ is almost consistent in each population until $\sim$20Mya (Figure A4), at which point it begins to fray as likely the genome is fully coalesced.

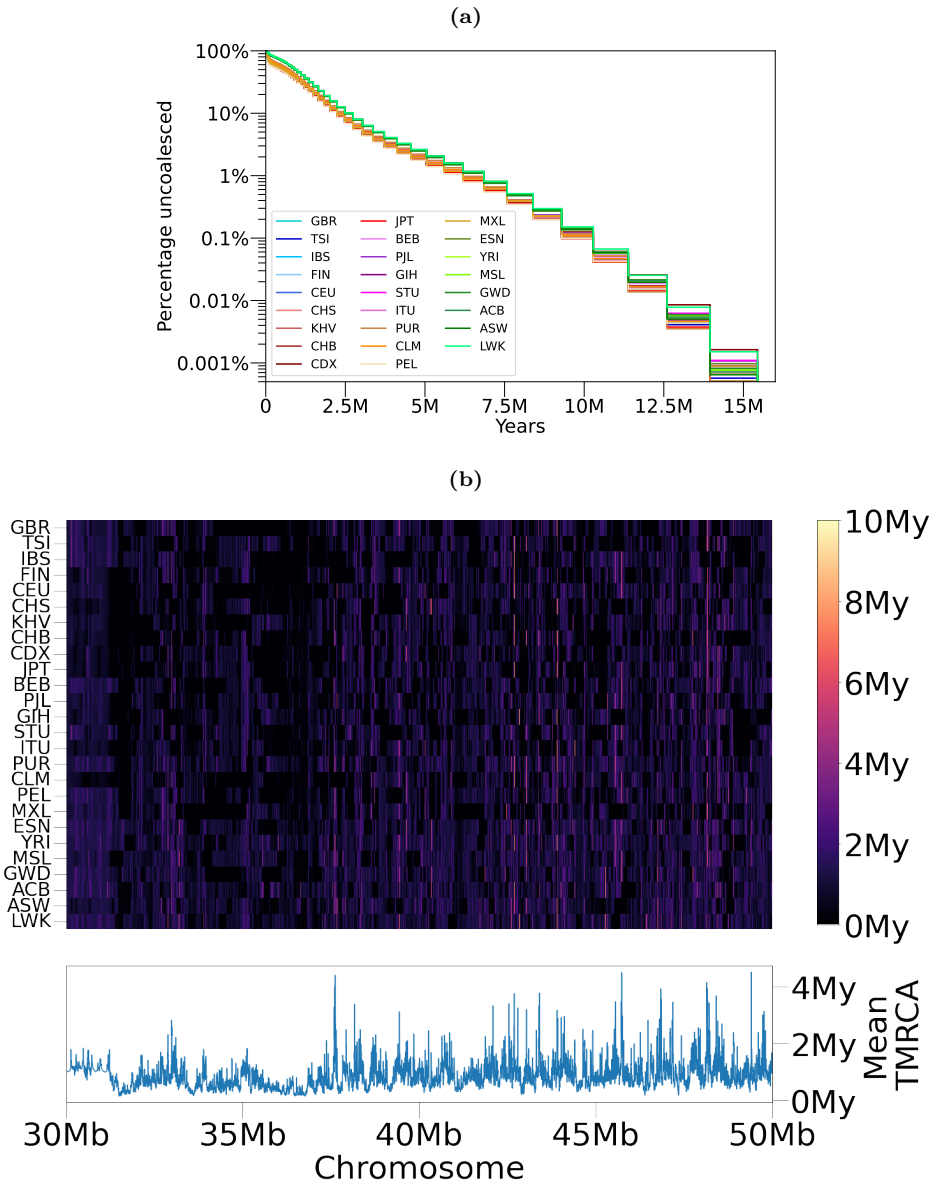

Figure 2: **(a)** Inferred fraction of uncoalesced genome as a function of time, for the 26 individuals in Figure 1a. **(b)** (Top) Posterior mean TMRCAs for each population, for 20Mb on chromosome 20. The bottom panel shows the mean of the top panel.

3.3 Model violations and their effects

As with all statistical frameworks, the PSMC method relies on specific modelling assumptions that may not hold true in reality. Deviations from these assumptions could have varying impacts on inference across different temporal scales, and could potentially account for some or all of the observed ancient peak. We therefore considered how departures from the modelling assumptions could plausibly affect inference of population size history in the deep past.

To this end, we considered the following possible issues. First, repeats may manifest as an artifactual heterozygous signal, leading to long segments with high heterozygosity, which then leads to an apparent

Population	Sample	1Mya	2Mya	3Mya	4Mya	5Mya	6Mya
ESN	HG03515	54,169	17,213	7,811	3,952	1,841	472
YRI	NA18488	54,486	17,229	7,668	3,841	1,752	426
MSL	HG03212	53,817	17,243	7,661	3,910	1,947	494
GWD	HG02568	54,493	16,893	7,358	3,667	1,721	397
ACB	HG01882	55,368	17,590	7,883	4,129	1,909	457
ASW	NA19625	54,974	17,395	7,798	3,984	1,852	439
LWK	NA19017	54,644	17,709	8,105	4,050	1,874	496

Table 1: The number of ancient non-contiguous segments inferred in different African populations from the 1000GP project. We count segments by looking for regions where the posterior probability of coalescence as old or older than the time given in each column is greater than 0.9.

excess of ancient coalescence events. Second, variation in mutation rates across different genomic regions could cause some regions to have a higher density of mutations leading to an interpretation as having a more ancient TMRCA, and hence systematic biases in the inference of accurate evolutionary histories. Third, the human mutation rate is estimated to have slowed down over time [36, 37, 38, 39], while in PSMC it is assumed constant. Fourth, variability in recombination rates across the genome (in particular at recombination hotspots) is also a departure from model assumptions [40, 41, 42, 43, 44]. Fifth, at longer timescales there may become a significant probability of recurrent mutations at the same site, which are not modelled in PSMC. Sixth, background selection, which has been estimated to be pervasive in humans [45, 46], has been shown to distort the inference of the coalescence rate [47, 48, 49]. Finally, balancing selection can also manifest as an excess in ancient coalescence events.

We conducted an extensive set of simulation studies to analyse how these violations may affect inference, which we describe fully in the Appendix). We concluded that it is unlikely any of these issues or violations could cause a false ancient peak of the type we see in real data. Notably, we observed that if there are large variations in the mutation rate across the genome, then false inflations of inferred $N(t)$ are observed. However, the degree of rate variation required to generate this is larger than the degree of this inferred in humans, and moreover their ancient inferred $N(t)$ has high variance, contrasting the low variance we have shown in humans in the ∼5Mya peak.

4 Discussion

We have demonstrated the applicability of coalescent-based inference to examine deep evolutionary history (five million years ago and beyond) in the hominin lineage. Specifically, we highlighted and extended the analysis of a peak in ancient population size history inferred by PSMC, which is reliably estimated from human, chimpanzee and bonobo data, though not found in gorillas or orangutans. The human, chimpanzee and bonobo peaks span approximately the same time period, suggesting that they may reflect the same event prior to, or during, human-chimpanzee speciation. We discussed several ways in which the data may violate the model underlying our analysis, and concluded that none are likely to generate a false ancient peak.

A straightforward interpretation of an ancient peak is that it reflects true changes in population size. Indeed, for humans, changes in inferred $N(t)$ since ∼80kya have been associated with the out-of-Africa event followed by more recent recent population size increases [3, 6]. The ancient peak may similarly reflect ancient

hominin population expansion and decline in the late miocene and the pliocene. To the extent the same ancient peak exists in chimpanzees and bonobos, it may reflect common environmental conditions affecting similar evolutionary processes in the different species.

However, there are other possible explanations. The separation and reconnection of local populations alters the amount of possible gene flow between them, which affects the inferred $N(t)$ [2, 50, 24, 51]. In a structured population, the coalescence rate decreases relative to a panmictic population of the same size, thus the effective population size (taken here as the inverse of the coalescence rate) increases. Indeed, in [24] the authors show that the inferred human effective population size as inferred by PSMC, including the ancient peak, can be equally well obtained by a set of populations with varying migration rates between them without any changes in size. Therefore, it is possible that the ancient peak is a signature of ancient population structure or other departures from panmixia.

One such possible structured event may be complex speciation between humans and chimpanzees, in which after an initial split the two populations rejoined or exchanged genes for a period of time. Previous publications have put forward evidence for this type of model [18, 19, 13, 20], although other analysis has favoured a model with a clean split [21, 22]. Innan and Watanabe proposed there was no strong support for gene flow after divergence, and that a clean split best fit the data [21]. This result was reinforced by Yamamichi, whose maximum likelihood model also favoured a clean split [22]. Patterson et al. proposed that gene flow did occur, because of differences in the estimated divergence time between the autosomes and X chromosome [18], although this approach has been criticised [52, 53, 22]. Using a likelihood ratio test considering divergence times across the genome, Yang rejected the null hypothesis of an absence of gene flow [19]. With an HMM that explicitly tests for immediate or prolonged separation, Mailund et al. proposed that a model with an extended period of gene flow better fits the data [13]. Galtier introduced Aphid, which distinguishes between discordant coalescent trees (ones that do not match the species trees) generated by ILS or gene flow [20]. It leverages the fact that multispecies coalescent trees affected by gene flow tend to have shorter branches, and coalescent trees affected by incomplete lineage sorting longer branches, than the average coalescent tree. This information provided strong support for a model in which there was ancient gene flow between human, chimpanzee, and even gorilla, after the initial human-chimpanzee split.

If the peak we observed in chimpanzees and bonobos is indeed reliable and aligned with the ancient peak observed in humans, this may be consistent with a period of ongoing gene flow after initial human and chimpanzee separation. Qualitatively, there are other observations which make this type of model attractive. Notably, CoalHMM [23] and TRAILS [17] assume a panmictic ancestral human-chimpanzee population and estimate its size as 177,368 and 167,400, respectively. Both of these estimates are more than 8 times the long-term effective population size of humans and chimpanzees (in both species, $\theta = 4N\mu$ is roughly 0.001, which assuming μ=1.25e-8 gives N=20,000), and are not well aligned with the PSMC inference in this period (Figure 1). A structured ancestral population could explain the large CoalHMM and TRAILS estimates, as falsely assuming panmixia can inflate estimates of population size [24]. Moreover, Aphid [20], which provides strong support for human-chimp-gorilla gene flow, estimates the size of the ancestral human-chimpanzee population to be of similar size to present day human, chimpanzee and bonobo populations. Additionally, chimpanzees and bonobos have been estimated to split cleanly around 2Mya [13], and PSMC does not infer a unique peak around this period. Finally, widespread structure in the ancestral populations of

humans, chimpanzees/bonobos, and gorillas, may plausibly explain why fossils such as *Sahelanthropus* from this period are hard to assign [54, 55, 56], and as a consequence morphological analyses can not confidently determine separation times [57].

We examined the sensitivity of PSMC to various modelling violations such as variation in mutation and recombination rates, and selection. It will be useful to incorporate these as extensions to the PSMC model, modifying the transition and emission probabilities within the underlying HMM accordingly. Indeed, some work has already been done towards this goal [58, 59], and future modification could further address slowdown or genomic variability in mutation rate, recombination hotspots, and/or selection, so as to better understand the ancient speciation process in primates [17] and indeed other species.

Acknowledgements

We thank Aylwyn Scally for constructive comments. T.C. was funded by a Wellcome Postgraduate Studentship 108864/B/15/Z, and R.D. and R.S. by Wellcome Investigator Award 207492/Z/17/Z. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

References

[1] Brian Charlesworth. Effective population size and patterns of molecular evolution and variation. *Nature Reviews Genetics*, 10(3):195–205, 2009.

[2] Heng Li and Richard Durbin. Inference of human population history from individual whole-genome sequences. *Nature*, 475(7357):493–496, 2011.

[3] Stephan Schiffels and Richard Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature genetics*, 46(8):919–925, 2014.

[4] Sara Sheehan, Kelley Harris, and Yun S Song. Estimating variable effective population sizes from multiple genomes: a sequentially markov conditional sampling distribution approach. *Genetics*, 194(3):647–662, 2013.

[5] Matthias Steinrücken, Joshua S Paul, and Yun S Song. A sequentially markov conditional sampling distribution for structured populations with migration and recombination. *Theoretical population biology*, 87:51–61, 2013.

[6] Jack Kamm, Jonathan Terhorst, Richard Durbin, and Yun S Song. Efficiently inferring the demographic history of many populations with allele count data. *Journal of the American Statistical Association*, 115(531):1472–1487, 2020.

[7] Donna Henderson, Sha Zhu, Christopher B Cole, and Gerton Lunter. Demographic inference from multiple whole genomes using a particle filter for continuous markov jump processes. *Plos one*, 16(3):e0247647, 2021.

[8] Matthias Meyer, Martin Kircher, Marie-Theres Gansauge, Heng Li, Fernando Racimo, Swapan Mallick, Joshua G Schraiber, Flora Jay, Kay Prüfer, Cesare De Filippo, et al. A high-coverage genome sequence from an archaic denisovan individual. *Science*, 338(6104):222–226, 2012.

[9] Kay Prüfer, Fernando Racimo, Nick Patterson, Flora Jay, Sriram Sankararaman, Susanna Sawyer, Anja Heinze, Gabriel Renaud, Peter H Sudmant, Cesare De Filippo, et al. The complete genome sequence of a neanderthal from the altai mountains. *Nature*, 505(7481):43–49, 2014.

[10] Brian Villmoare, William H Kimbel, Chalachew Seyoum, Christopher J Campisano, Erin N DiMaggio, John Rowan, David R Braun, J Ramón Arrowsmith, and Kaye E Reed. Early homo at 2.8 ma from ledi-geraru, afar, ethiopia. *Science*, 347(6228):1352–1355, 2015.

[11] Erin N DiMaggio, Christopher J Campisano, John Rowan, Guillaume Dupont-Nivet, Alan L Deino, Faysal Bibi, Margaret E Lewis, Antoine Souron, Dominique Garello, Lars Werdelin, et al. Late pliocene fossiliferous sedimentary record and the environmental context of early homo from afar, ethiopia. *Science*, 347(6228):1355–1359, 2015.

[12] Jason E Lewis, Carol V Ward, William H Kimbel, Casey L Kidney, Frank H Brown, Rhonda L Quinn, John Rowan, Ignacio A Lazagabaster, William J Sanders, Meave G Leakey, et al. A 4.3-million-year-old australopithecus anamensis mandible from ileret, east turkana, kenya, and its paleoenvironmental context. *Journal of Human Evolution*, 194:103579, 2024.

[13] Thomas Mailund, Anders E Halager, Michael Westergaard, Julien Y Dutheil, Kasper Munch, Lars N Andersen, Gerton Lunter, Kay Prüfer, Aylwyn Scally, Asger Hobolth, et al. A new isolation with migration model along complete genomes infers very different divergence processes among closely related great ape species. *PLoS genetics*, 8(12):e1003125, 2012.

[14] Asger Hobolth, Ole F Christensen, Thomas Mailund, and Mikkel H Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS genetics*, 3(2):e7, 2007.

[15] Julien Y Dutheil, Ganesh Ganapathy, Asger Hobolth, Thomas Mailund, Marcy K Uyenoyama, and Mikkel H Schierup. Ancestral population genomics: the coalescent hidden markov model approach. *Genetics*, 183(1):259–274, 2009.

[16] Javier Prado-Martinez, Peter H Sudmant, Jeffrey M Kidd, Heng Li, Joanna L Kelley, Belen Lorente-Galdos, Krishna R Veeramah, August E Woerner, Timothy D O'connor, Gabriel Santpere, et al. Great ape genetic diversity and population history. *Nature*, 499(7459):471–475, 2013.

[17] Iker Rivas-González, Mikkel H Schierup, John Wakeley, and Asger Hobolth. Trails: Tree reconstruction of ancestry using incomplete lineage sorting. *Plos Genetics*, 20(2):e1010836, 2024.

[18] Nick Patterson, Daniel J Richter, Sante Gnerre, Eric S Lander, and David Reich. Genetic evidence for complex speciation of humans and chimpanzees. *Nature*, 441(7097):1103–1108, 2006.

[19] Ziheng Yang. A likelihood ratio test of speciation with gene flow using genomic sequence data. *Genome biology and evolution*, 2:200–211, 2010.

[20] Nicolas Galtier. An approximate likelihood method reveals ancient gene flow between human, chimpanzee and gorilla. *Peer Community Journal*, 4, 2024.

[21] Hideki Innan and Hidemi Watanabe. The effect of gene flow on the coalescent time in the human-chimpanzee ancestral population. *Molecular biology and evolution*, 23(5):1040–1047, 2006.

[22] Masato Yamamichi, Jun Gojobori, and Hideki Innan. An autosomal analysis gives no genetic evidence for complex speciation of humans and chimpanzees. *Molecular biology and evolution*, 29(1):145–156, 2012.

[23] Iker Rivas-González, Marjolaine Rousselle, Fang Li, Long Zhou, Julien Y Dutheil, Kasper Munch, Yong Shao, Dongdong Wu, Mikkel H Schierup, and Guojie Zhang. Pervasive incomplete lineage sorting illuminates speciation and selection in primates. *Science*, 380(6648):eabn4409, 2023.

[24] Olivier Mazet, Willy Rodríguez, Simona Grusea, Simon Boitard, and Lounès Chikhi. On the importance of being structured: instantaneous coalescence rates and human evolution—lessons for ancestral population size inference? *Heredity*, 116(4):362–371, 2016.

[25] Naoyuki Takahata and Yoko Satta. Evolution of the primate lineage leading to modern humans: phylogenetic and demographic inferences from dna sequences. *Proceedings of the National Academy of Sciences*, 94(9):4811–4815, 1997.

[26] Qiaomei Fu, Heng Li, Priya Moorjani, Flora Jay, Sergey M Slepchenko, Aleksei A Bondarev, Philip LF Johnson, Ayinuer Aximu-Petri, Kay Prüfer, Cesare De Filippo, et al. Genome sequence of a 45,000-year-old modern human from western siberia. *Nature*, 514(7523):445–449, 2014.

[27] 1000 Genomes Project Consortium et al. A global reference for human genetic variation. *Nature*, 526(7571):68, 2015.

[28] 1000 Genomes Project Consortium et al. A map of human genome variation from population scale sequencing. *Nature*, 467(7319):1061, 2010.

[29] 1000 Genomes Project Consortium et al. An integrated map of genetic variation from 1,092 human genomes. *Nature*, 491(7422):56, 2012.

[30] Marta Byrska-Bishop, Uday S Evani, Xuefang Zhao, Anna O Basile, Haley J Abel, Allison A Regier, André Corvelo, Wayne E Clarke, Rajeeva Musunuri, Kshithija Nagulapalli, et al. High-coverage whole-genome sequencing of the expanded 1000 genomes project cohort including 602 trios. *Cell*, 185(18):3426–3440, 2022.

[31] Leon Hilgers, Shenglin Liu, Axel Jensen, Thomas Brown, Trevor Cousins, Regev Schweiger, Katerina Guschanski, and Michael Hiller. Avoidable false psmc population size peaks occur across numerous studies. *bioRxiv*, pages 2024–06, 2024.

[32] Débora YC Brandt, Xinzhu Wei, Yun Deng, Andrew H Vaughn, and Rasmus Nielsen. Evaluation of methods for estimating coalescence times using ancestral recombination graphs. *Genetics*, 221(1):iyac044, 2022.

[33] Regev Schweiger and Richard Durbin. Ultrafast genome-wide inference of pairwise coalescence times. *Genome Research*, 33(7):1023–1031, 2023.

[34] Yun S Song. Lecture notes on computational and mathematical population genetics. 2021.

[35] Leo Speidel, Marie Forest, Sinan Shi, and Simon R Myers. A method for genome-wide genealogy estimation for thousands of samples. *Nature genetics*, 51(9):1321–1329, 2019.

[36] Priya Moorjani, Ziyue Gao, and Molly Przeworski. Human germline mutation and the erratic evolutionary clock. *PLoS biology*, 14(10):e2000744, 2016.

[37] Aylwyn Scally and Richard Durbin. Revising the human mutation rate: implications for understanding human evolution. *Nature Reviews Genetics*, 13(10):745–753, 2012.

[38] V Yi Soojin. Morris goodman's hominoid rate slowdown: the importance of being neutral. *Molecular phylogenetics and evolution*, 66(2):569–574, 2013.

[39] Manjusha Chintalapati and Priya Moorjani. Evolution of the mutation rate across primates. *Current opinion in genetics & development*, 62:58–64, 2020.

[40] S Myers, C C A Spencer, A Auton, L Bottolo, C Freeman, P Donnelly, and G McVean. The distribution and causes of meiotic recombination in the human genome. *Biochem. Soc. Trans.*, 34(Pt 4):526–530, 2006.

[41] International HapMap Consortium. The International HapMap Project. *Nature*, 426(6968):789–796, 2003.

[42] Thomas D Petes. Meiotic recombination hot spots and cold spots. *Nature Reviews Genetics*, 2(5):360–369, 2001.

[43] KT Nishant, Chetan Kumar, and MRS Rao. Humhot: a database of human meiotic recombination hot spots. *Nucleic acids research*, 34(suppl_1):D25–D28, 2006.

[44] Laure Ségurel, Ellen Miranda Leffler, and Molly Przeworski. The case of the fickle fingers: how the prdm9 zinc finger protein specifies meiotic recombination hotspots in humans. *PLoS biology*, 9(12):e1001211, 2011.

[45] Graham McVicker, David Gordon, Colleen Davis, and Phil Green. Widespread genomic signatures of natural selection in hominid evolution. *PLoS genetics*, 5(5):e1000471, 2009.

[46] David A Murphy, Eyal Elyashiv, Guy Amster, and Guy Sella. Broad-scale variation in human genetic diversity levels is predicted by purifying selection on coding and non-coding elements. *Elife*, 12:e76065, 2022.

[47] Daniel R Schrider, Alexander G Shanku, and Andrew D Kern. Effects of linked selective sweeps on demographic inference and model selection. *Genetics*, 204(3):1207–1223, 2016.

[48] Parul Johri, Kellen Riall, Hannes Becher, Laurent Excoffier, Brian Charlesworth, and Jeffrey D Jensen. The impact of purifying and background selection on the inference of population history: problems and prospects. *Molecular biology and evolution*, 38(7):2986–3003, 2021.

[49] Simon Boitard, Armando Arredondo, Lounès Chikhi, and Olivier Mazet. Heterogeneity in effective size across the genome: effects on the inverse instantaneous coalescence rate (iicr) and implications for demographic inference under linked selection. *Genetics*, 220(3):iyac008, 2022.

[50] Olivier Mazet, Willy Rodríguez, and Lounès Chikhi. Demographic inference using genetic data from a single individual: Separating population size variation from population structure. *Theoretical Population Biology*, 104:46–58, 2015.

[51] Trevor Cousins, Aylwyn Scally, and Richard Durbin. A structured coalescent model reveals deep ancestral structure shared by all modern humans. *bioRxiv*, pages 2024–03, 2024.

[52] John Wakeley. Complex speciation of humans and chimpanzees. *Nature*, 452(7184):E3–E4, 2008.

[53] Daven C Presgraves and V Yi Soojin. Doubts about complex speciation between humans and chimpanzees. *Trends in ecology & evolution*, 24(10):533–540, 2009.

[54] Michel Brunet, Franck Guy, David Pilbeam, Hassane Taisso Mackaye, Andossa Likius, Djimdoumalbaye Ahounta, Alain Beauvilain, Cécile Blondel, Hervé Bocherens, Jean-Renaud Boisserie, et al. A new hominid from the upper miocene of chad, central africa. *Nature*, 418(6894):145–151, 2002.

[55] Milford H Wolpoff, Brigitte Senut, Martin Pickford, and John Hawks. Sahelanthropus or'sahelpithecus'? *Nature*, 419(6907):581–582, 2002.

[56] Michel Brunet. Sahelanthropus or'sahelpithecus'? *Nature*, 419(6907):582–582, 2002.

[57] Bernard Wood and Terry Harrison. The evolutionary context of the first hominins. *Nature*, 470(7334):347–352, 2011.

[58] Trevor Cousins, Daniel Tabin, Nick Patterson, David Reich, and Arun Durvasula. Accurate inference of population history in the presence of background selection. *bioRxiv*, 2024.

[59] Gustavo V. Barroso, Nataša Puzović, and Julien Y Dutheil. Inference of recombination maps from a single pair of genomes and its application to ancient samples. *PLoS genetics*, 15(11):e1008449, 2019.

[60] Petr Danecek, James K Bonfield, Jennifer Liddle, John Marshall, Valeriu Ohan, Martin O Pollard, Andrew Whitwham, Thomas Keane, Shane A McCarthy, Robert M Davies, et al. Twelve years of samtools and bcftools. *Gigascience*, 10(2):giab008, 2021.

[61] Heng Li, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. The sequence alignment/map format and samtools. *Bioinformatics*, 25(16):2078–2079, 2009.

[62] Heng Li. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, 27(21):2987–2993, 2011.

[63] Moisès Coll Macià, Laurits Skov, Benjamin Marco Peter, and Mikkel Heide Schierup. Different historical generation intervals in human populations inferred from neanderthal fragment lengths and mutation signatures. *Nature Communications*, 12(1):5317, 2021.

[64] Jared C Roach, Gustavo Glusman, Arian FA Smit, Chad D Huff, Robert Hubley, Paul T Shannon, Lee Rowen, Krishna P Pant, Nathan Goodman, Michael Bamshad, et al. Analysis of genetic inheritance in a family quartet by whole-genome sequencing. *Science*, 328(5978):636–639, 2010.

[65] Philip Awadalla, Julie Gauthier, Rachel A Myers, Ferran Casals, Fadi F Hamdan, Alexander R Griffing, Mélanie Côté, Edouard Henrion, Dan Spiegelman, Julien Tarabeux, et al. Direct measure of the de novo mutation rate in autism and schizophrenia cohorts. *The American Journal of Human Genetics*, 87(3):316–324, 2010.

[66] Raheleh Rahbari, Arthur Wuster, Sarah J Lindsay, Robert J Hardwick, Ludmil B Alexandrov, Saeed Al Turki, Anna Dominiczak, Andrew Morris, David Porteous, Blair Smith, et al. Timing, rates and spectra of human germline mutation. *Nature genetics*, 48(2):126–133, 2016.

[67] Augustine Kong, Michael L Frigge, Gisli Masson, Soren Besenbacher, Patrick Sulem, Gisli Magnusson, Sigurjon A Gudjonsson, Asgeir Sigurdsson, Aslaug Jonasdottir, Adalbjorg Jonasdottir, et al. Rate of de novo mutations and the importance of father's age to disease risk. *Nature*, 488(7412):471–475, 2012.

[68] Jacob J Michaelson, Yujian Shi, Madhusudan Gujral, Hancheng Zheng, Dheeraj Malhotra, Xin Jin, Minghan Jian, Guangming Liu, Douglas Greer, Abhishek Bhandari, et al. Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, 151(7):1431–1442, 2012.

[69] 1000 Genomes Project Consortium et al. Variation in genome-wide mutation rates within and between human families. *Nature genetics*, 43(7):712–714, 2011.

[70] Aylwyn Scally. The mutation rate in human evolution and demographic inference. *Current opinion in genetics & development*, 41:36–43, 2016.

[71] Hans Ellegren, Nick GC Smith, and Matthew T Webster. Mutation rate variation in the mammalian genome. *Current opinion in genetics & development*, 13(6):562–568, 2003.

[72] Alan Hodgkinson and Adam Eyre-Walker. Variation in the mutation rate across mammalian genomes. *Nature reviews genetics*, 12(11):756–766, 2011.

[73] Thomas CA Smith, Peter F Arndt, and Adam Eyre-Walker. Large scale variation in the rate of germ-line de novo mutation, base composition, divergence and diversity in humans. *PLoS genetics*, 14(3):e1007254, 2018.

[74] Jerome Kelleher, Alison M Etheridge, and Gilean McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS computational biology*, 12(5):e1004842, 2016.

[75] Laurent C Francioli, Paz P Polak, Amnon Koren, Androniki Menelaou, Sung Chun, Ivo Renkens, Genome of the Netherlands Consortium, Cornelia M van Duijn, Morris Swertz, Cisca Wijmenga, et al. Genome-wide patterns and properties of de novo mutations in humans. *Nature genetics*, 47(7):822–826, 2015.

[76] Hákon Jónsson, Patrick Sulem, Birte Kehr, Snaedis Kristmundsdottir, Florian Zink, Eirikur Hjartarson, Marteinn T Hardarson, Kristjan E Hjorleifsson, Hannes P Eggertsson, Sigurjon Axel Gudjonsson, et al. Parental influence on human germline de novo mutations in 1,548 trios from iceland. *Nature*, 549(7673):519–522, 2017.

[77] Wendy SW Wong, Benjamin D Solomon, Dale L Bodian, Prachi Kothiyal, Greg Eley, Kathi C Huddleston, Robin Baker, Dzung C Thach, Ramaswamy K Iyer, Joseph G Vockley, et al. New observations on maternal age effect on germline de novo mutations. *Nature communications*, 7(1):10486, 2016.

[78] Miguel Rodriguez-Galindo, Sònia Casillas, Donate Weghorn, and Antonio Barbadilla. Germline de novo mutation rates on exons versus introns in humans. *Nature communications*, 11(1):3304, 2020.

[79] Richard R Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical population biology*, 23(2):183–201, 1983.

[80] S Myers, CCA Spencer, A Auton, L Bottolo, C Freeman, P Donnelly, and et G McVean. The distribution and causes of meiotic recombination in the human genome, 2006.

[81] Richard A Gibbs, John W Belmont, Paul Hardenbol, Thomas D Willis, Fuli L Yu, HM Yang, Lan-Yang Ch'ang, Wei Huang, Bin Liu, Yan Shen, et al. The international hapmap project. 2003.

[82] Simon Myers, Rory Bowden, Afidalina Tumian, Ronald E Bontrop, Colin Freeman, Tammie S MacFie, Gil McVean, and Peter Donnelly. Drive against hotspot motifs in primates implicates the prdm9 gene in meiotic recombination. *Science*, 327(5967):876–879, 2010.

[83] Emil D Parvanov, Petko M Petkov, and Kenneth Paigen. Prdm9 controls activation of mammalian recombination hotspots. *Science*, 327(5967):835–835, 2010.

[84] Kenneth Paigen and Petko M Petkov. Prdm9 and its role in genetic recombination. *Trends in Genetics*, 34(4):291–300, 2018.

[85] Corinne Grey, Frédéric Baudat, and Bernard de Massy. Prdm9, a driver of the genetic map. *PLoS genetics*, 14(8):e1007479, 2018.

[86] International HapMap Consortium Altshuler David altshuler@ molbio. mgh. harvard. edu Donnelly Peter donnelly@ stats. ox. ac. uk. A haplotype map of the human genome. *Nature*, 437(7063):1299–1320, 2005.

[87] International HapMap Consortium et al. A second generation human haplotype map of over 3.1 million snps. *Nature*, 449(7164):851, 2007.

[88] TH Jukes. Evolution of protein molecules. *Mammalian Protein Metabolism*, 3, 1969.

[89] Brian Charlesworth, MT Morgan, and Deborah Charlesworth. The effect of deleterious mutations on neutral molecular variation. *Genetics*, 134(4):1289–1303, 1993.

[90] Magnus Nordborg, Brian Charlesworth, and Deborah Charlesworth. The effect of recombination on background selection. *Genetics Research*, 67(2):159–174, 1996.

[91] Parul Johri, Susanne P Pfeifer, and Jeffrey D Jensen. Developing an evolutionary baseline model for humans: jointly inferring purifying selection with population history. *Molecular biology and evolution*, 40(5):msad100, 2023.

[92] Jacob I Marsh and Parul Johri. Biases in arg-based inference of historical population size in populations experiencing selection. *Molecular Biology and Evolution*, page msae118, 2024.

[93] Deborah Charlesworth. Balancing selection and its effects on sequences in nearby genome regions. *PLoS genetics*, 2(4):e64, 2006.

[94] KL Bubb, D Bovee, D Buckley, E Haugen, M Kibukawa, M Paddock, A Palmieri, S Subramanian, Y Zhou, R Kaul, et al. Scan of human genome reveals no new loci under ancient balancing selection. *Genetics*, 173(4):2165–2177, 2006.

[95] Saurabh Asthana, Steffen Schmidt, and Shamil Sunyaev. A limited role for balancing selection. *Trends in Genetics*, 21(1):30–32, 2005.

[96] Felix M Key, João C Teixeira, Cesare de Filippo, and Aida M Andrés. Advantageous diversity maintained by balancing selection in humans. *Current opinion in genetics & development*, 29:45–51, 2014.

[97] Bora E Baysal, Elizabeth C Lawrence, and Robert E Ferrell. Sequence variation in human succinate dehydrogenase genes: evidence for long-term balancing selection on sdha. *BMC biology*, 5:1–14, 2007.

[98] Aida M Andrés, Melissa J Hubisz, Amit Indap, Dara G Torgerson, Jeremiah D Degenhardt, Adam R Boyko, Ryan N Gutenkunst, Thomas J White, Eric D Green, Carlos D Bustamante, et al. Targets of balancing selection in the human genome. *Molecular biology and evolution*, 26(12):2755–2764, 2009.

[99] Ellen M Leffler, Ziyue Gao, Susanne Pfeifer, Laure Ségurel, Adam Auton, Oliver Venn, Rory Bowden, Ronald Bontrop, Jeffrey D Wall, Guy Sella, et al. Multiple instances of ancient balancing selection shared between humans and chimpanzees. *Science*, 339(6127):1578–1582, 2013.

[100] Radoje Drmanac, Andrew B Sparks, Matthew J Callow, Aaron L Halpern, Norman L Burns, Bahram G Kermani, Paolo Carnevali, Igor Nazarenko, Geoffrey B Nilsen, George Yeung, et al. Human genome sequencing using unchained base reads on self-assembling dna nanoarrays. *Science*, 327(5961):78–81, 2010.

[101] Matthew D Rasmussen, Melissa J Hubisz, Ilan Gronau, and Adam Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10(5):e1004342, 2014.

[102] W James Kent, Charles W Sugnet, Terrence S Furey, Krishna M Roskin, Tom H Pringle, Alan M Zahler, and David Haussler. The human genome browser at ucsc. *Genome research*, 12(6):996–1006, 2002.

5 Methods

5.1 Processing data

We took high-coverage whole-genome-sequence cram files for one individual in each of the 26 populations from the 1000 Genomes project. These are aligned to GRCh38. The cram files were converted to bam and indexed with samtools [60, 61]. The genotype likelihoods were calculated with bcftools mpileup [62] by skipping alignments with mapping quality less than 20, skipping bases with base alignment quality less than 20, and setting the coefficient for downgrading mapping quality to 50. SNPs were called using bcftools and all indels were excluded. Variants were then designated as uncallable if the minimum mapping quality was less than 20, the minimum consensus quality was less than 20, or the coverage was less than half or more than double the mean coverage. Finally, we designated all regions in the strict mappability mask for GRCh38 (`ftp.1000genomes.ebi.ac.uk/vol1/ftp/data_collections/1000_genomes_project/working/20160622_genome_mask_GRCh38`) as uncallable. Uncallable positions

are labelled as missing data in the HMM.

We downloaded processed primate data from `eichlerlab.gs.washington.edu/greatape/data/` [16]. We took the VCF files and masked positions according to the given bed files, which described sites where coverage was less than 5 and regions that did not pass the quality filters as discussed in their paper. After filtering, only 49% of the genome was designated as callable. For the eastern lowland gorilla, we arbitrarily chose the individual labelled "Mkubwa", and for the Nigeria-Cameroon chimpanzee we chose the individual labelled "Akwaya Jean". We note that analysis on other individuals looked similar. For the gorilla we used a mutation rate per base pair per generation of 1.43e-08, and a generation time of 19 years; for the chimpanzee, a mutation rate per base pair per generation of 1.78e-08 and a generation time of 24 years were used [36].

5.2 Mutation rates and generation times for humans and primates

In humans we set the generation time as 29 [63] and the mutation rate per generation per base pair as μ=1.25e-08 [64, 65, 66, 67, 68, 69, 70]. The rates we set for the other great apes are taken from [39]. In chimpanzees and bonobos we set μ=1.78e-08 and generation time equal to 24 years. For Gorillas we set μ=1.42e-08 and generation time equal to 19 years. For Orangutans we set μ=2.03e-08 and generation time equal to 27.

5.3 PSMC analysis

We ran PSMC as embedded in *cobraa*, `www.github.com/trevorcousins/cobraa`. We used 64 time intervals and enabled the parameters for the effective population size in each to be inferred freely. This contrasts the default settings of previous implementations of the algorithm that force adjacent intervals to be the same, which has been shown to lead to fitting problems [31]. In the 1000GP data, we fixed the scaled mutation rate θ as 0.0008 which is roughly the mean across populations. For the great apes, we fixed the θ for gorilas as 0.0014, chimpanzees and bonobos as 0.001, and orangutans as 0.0014. We found that fixing this parameter reduces noise in the estimation of ancient $N(t)$, and also forces time interval boundaries across individuals to align. The initial value for the ratio of mutation rate to recombination rate was set as 1.5, and the recombination rate was set to be inferred as part of the EM algorithm, which was iterated for 30 iterations. The time interval boundaries are controlled by equation (1) where D is the number of discrete time boundaries (we used 64 throughout), and ω and T_{max} control the spread of time interval boundaries. In Figure 1 we used ω=0.01 and T_{max}=50; in Figure A4 we used ω=0.01 and T_{max}=150.

5.4 PSMC decoding

Using the inferred effective population size parameters, scaled recombination rate, and scaled mutation rate (θ=0.0008), we used the PSMC decoding to get a posterior probability of coalescence at every position. To calculate the amount of uncoalesced genome over time, we integrated over the full posterior probability (Figure 2). To scale from time in coalescent units (equation (1)) to time in years, we used N=16,000 ($\theta = 4N\mu$= 0.0008 with μ=1.25e-08) then multiplied coalescent time by $2N\mu g$, where g is the number of years per generations. To get a point estimate from PSMC (Figure 2b), we take the posterior mean at each position using the midpoint of the time interval boundaries.

To calculate the number of ancient non-contiguous segments inferred in different African populations (Table 1), we looked for regions where the posterior probability of coalescence as old or older than 2Mya, 3Mya, or 4Mya, respectively, is greater than 0.9. We then only counted non-contiguous segments, which are all segments that are not adjacent.

5.5 Nonparametric bootstrapping

For a given individual, for each chromosome, we break the sequence up into segments of 5Mb, then reconstruct a new sequence by sampling with replacement from all 5Mb segments. These are then concatenated together to form a contiguous sequence, on which PSMC inference is performed using 30 iterations. This procedure is repeated 30 times for humans and 30 times for chimps.

5.6 Relate

We ran Relate [35] on all samples on the 1000GP for GBP, CHB, and YRI. We randomly selected 10 diploid sequences from each population. To calculate the fraction of uncoalesced genome we discretised time into intervals of 500 generations and computed a histogram based on Relate's ARG. The inferred fraction of uncoalesced genome at 5Mya was roughly 0.3%, for each sample in each population.

6 Appendix

6.1 Figures

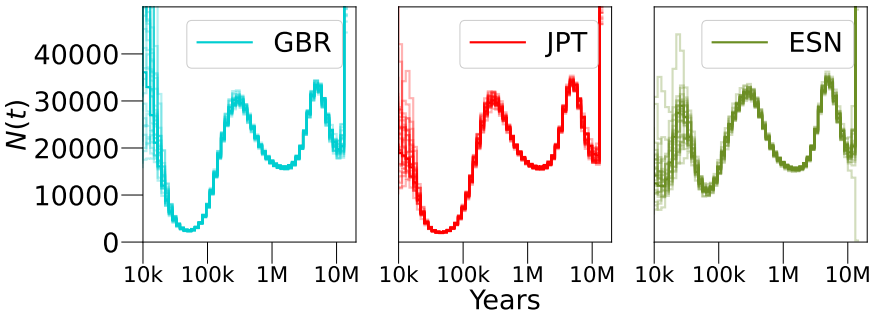

Figure A1: 30 iterations of block bootstraps on three different humans from the 1000GP project. We chopped the genome into 5Mb windows, then for each bootstrap we randomly resampled these with replacement.

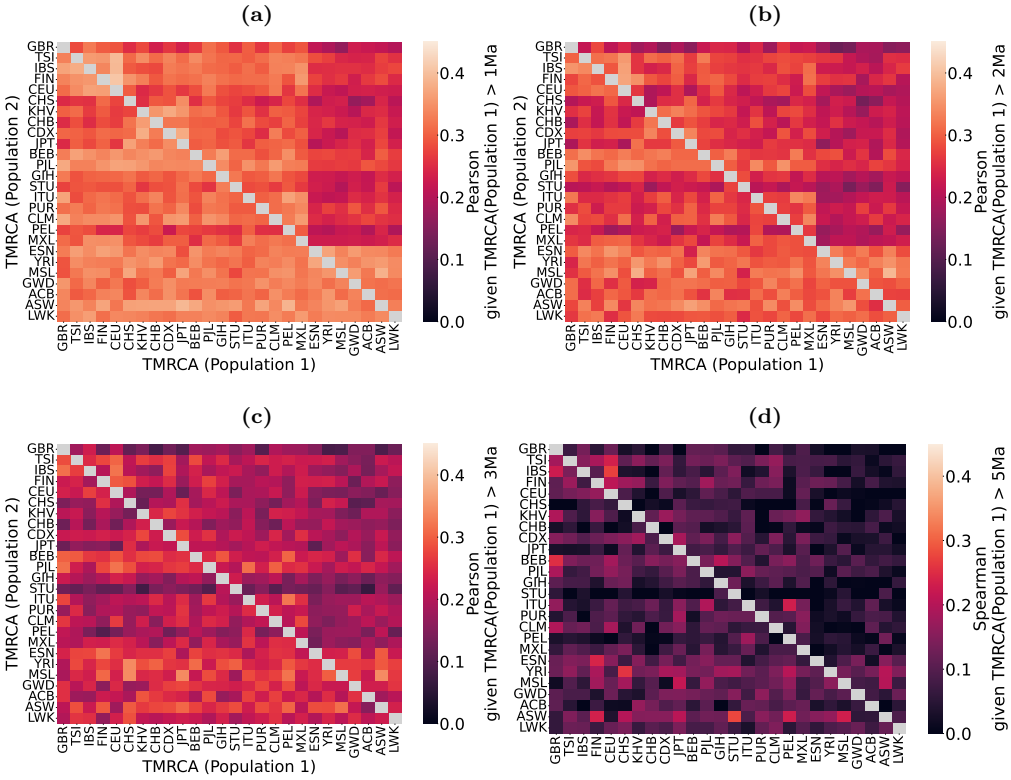

Figure A2: The Pearson correlation between inferred TMRCAs, for all pairwise combinations of two populations from the 1000GP. We calculated these for chromosome 1. To obtain point estimates at each position, we used the posterior mean. For each possible population pair, we look at positions where population 1 (x-axis) is larger than t and correlate these with the TMRCAs from population 2 (y-axis; note these are not necessarily larger than t, so the matrices are not symmetric). The value of t for **(a)**, **(b)**, **(c)**, and **(d)** is 1Mya, 2Mya, 3Mya, and 5Mya, respectively. The average Pearson r^2 for (a), (b), (c), and (d) is 0.3, 0.26, 0.21, and 0.11, respectively, and all correlations have p-value<0.05 except for 72 combinations from (d).

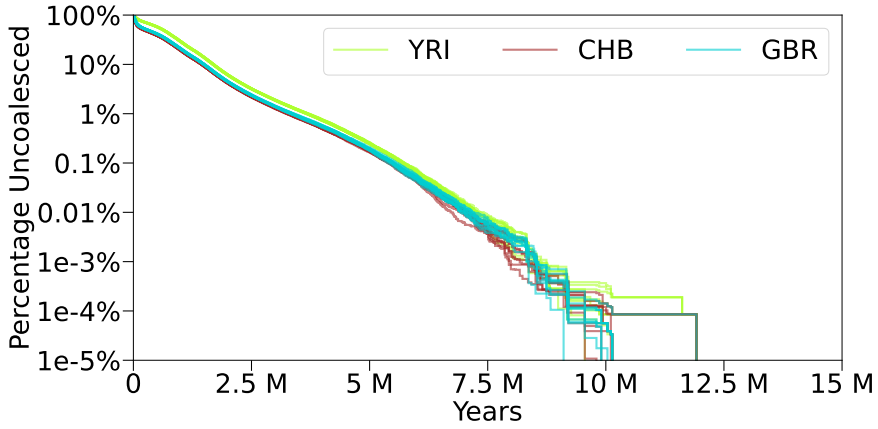

Figure A3: The Relate inferred fraction of uncoalesced genome as a function of time, for 10 diploid samples in the YRI, GBR and CHB populations from 1000GP.

(a)

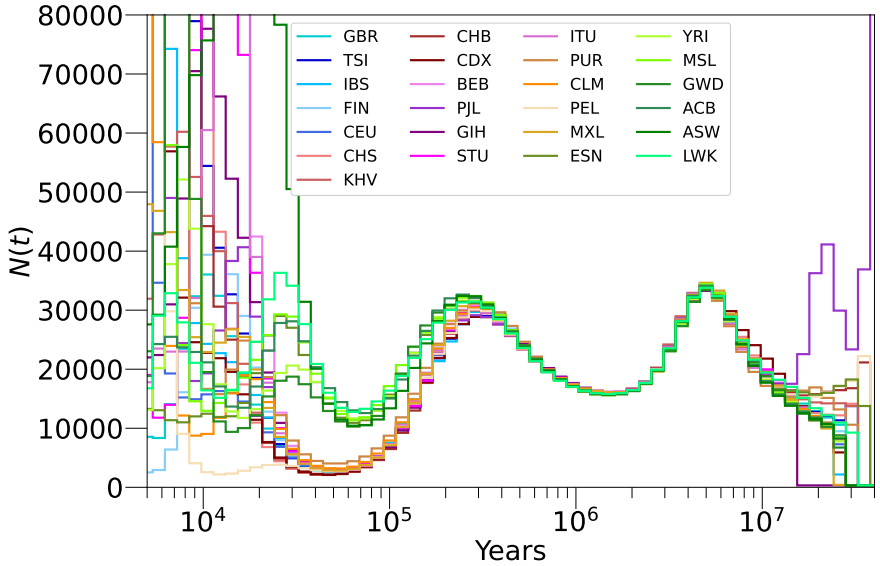

Figure A4: Using PSMC on humans with altered time interval boundaries (ω=0.01 and T_{max}=150) shows relatively consistent inference beyond 10Ma. Across African samples, the estimated coalescence rate trajectory up until ∼20Mya has relatively low variance.

6.2 Analysis of PSMC model violations

In this section, we analyse the effect of various model mispecifications on the inference from PSMC. We also analyse how the inferred coalescence times across the genome associate with various annotations, including for example repeat content, distance to coding sequence, strength of background selection, and recombination rate.

6.2.1 Repeats as a cause of artifactual heterozygous signal

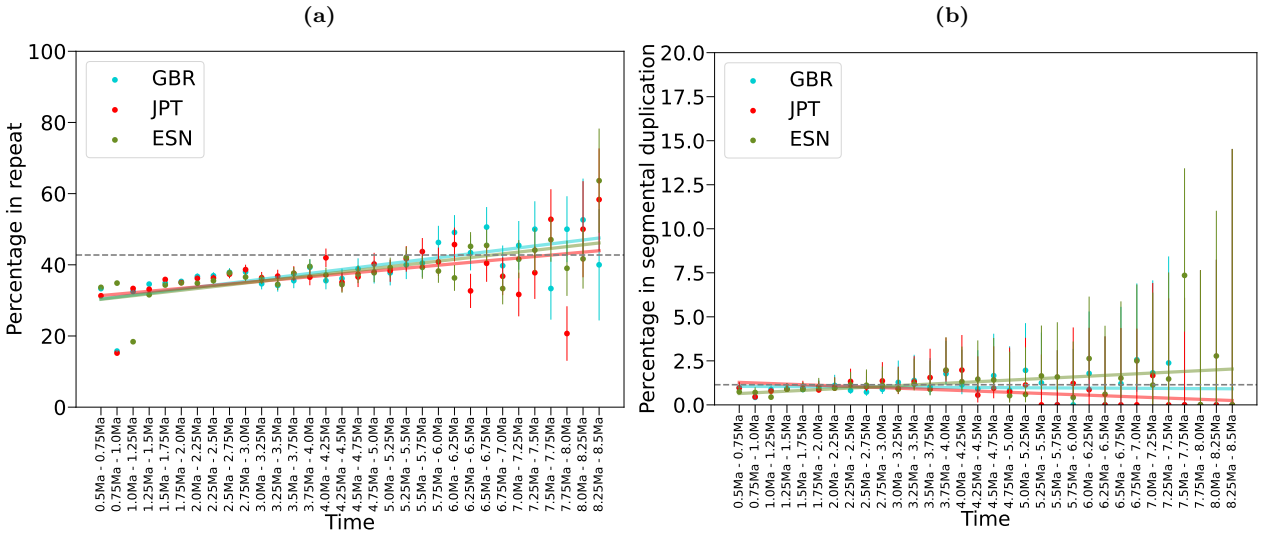

Figure A5: The relationship between the posterior mean TMRCA and the repeat content ((**a**)) or segmental duplications ((**b**)), on chromosome 1 for one individual from GBR, JPT, and ESN in the 1000GP. (a) The reported Pearson r^2 for each population is 0.73, 0.48, 0.69, with p-value<0.005 for each. The grey, dashed line indicates the chromosome wide average repeat content. The reported Pearson r^2 for GBR and ESN is not significant, though for JPT it is -0.5 with p-value 0.003.

As noted in previous work [2], false heterozygotes caused by repeated regions or segmental duplications may lead to excessively long segments with high heterozygosity, which may lead to an excess of ancient coalescent events. To examine whether this might artefactually be creating the observed signal, we analysed the fraction of the genome in repeated regions and stratified this by inferred TMRCA. We observed that the repeat content increases as a function of TMRCA (regression slope $\sim 2\%$), and that the effect is significant (Pearson's $r^2 \approx 0.63$, p-value<0.005; see Methods and Figure A5a). A similar analysis with segmental duplications revealed no significant correlation (Figure A5b). We re-ran our analysis with repeated regions and segmental duplications masked out, and observed that the ancient peak is still present, although when repeats are removed it exhibits less prominence (Figure A6a and Figure A6b, respectively). Thus we conclude that repeated regions or segmental duplications are unlikely to be the cause of the ancient peak.

6.2.2 Variable mutation rate across the genome

The PSMC model assumes a constant mutation rate throughout the genome, despite ample evidence that the mutation rate varies significantly by genomic location [71, 72]. Unfortunately, variation in mutation rates across different genomic regions can lead to systematic biases in the estimation of TMRCAs. This is because

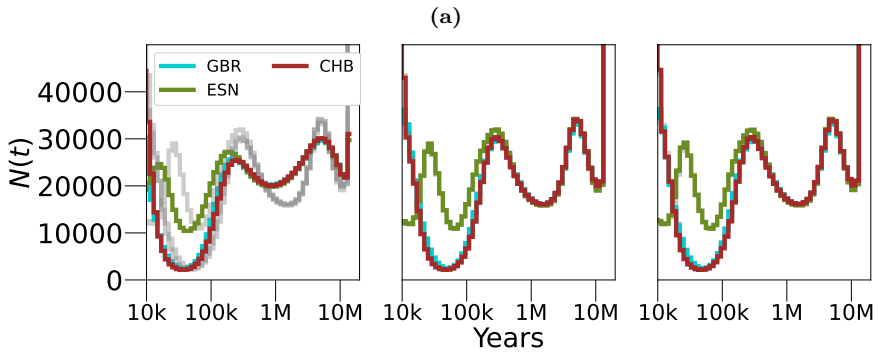

Figure A6: Running PSMC inference on human samples after masking out repeated regions (a), segmental duplications (b), and CpG islands (c).

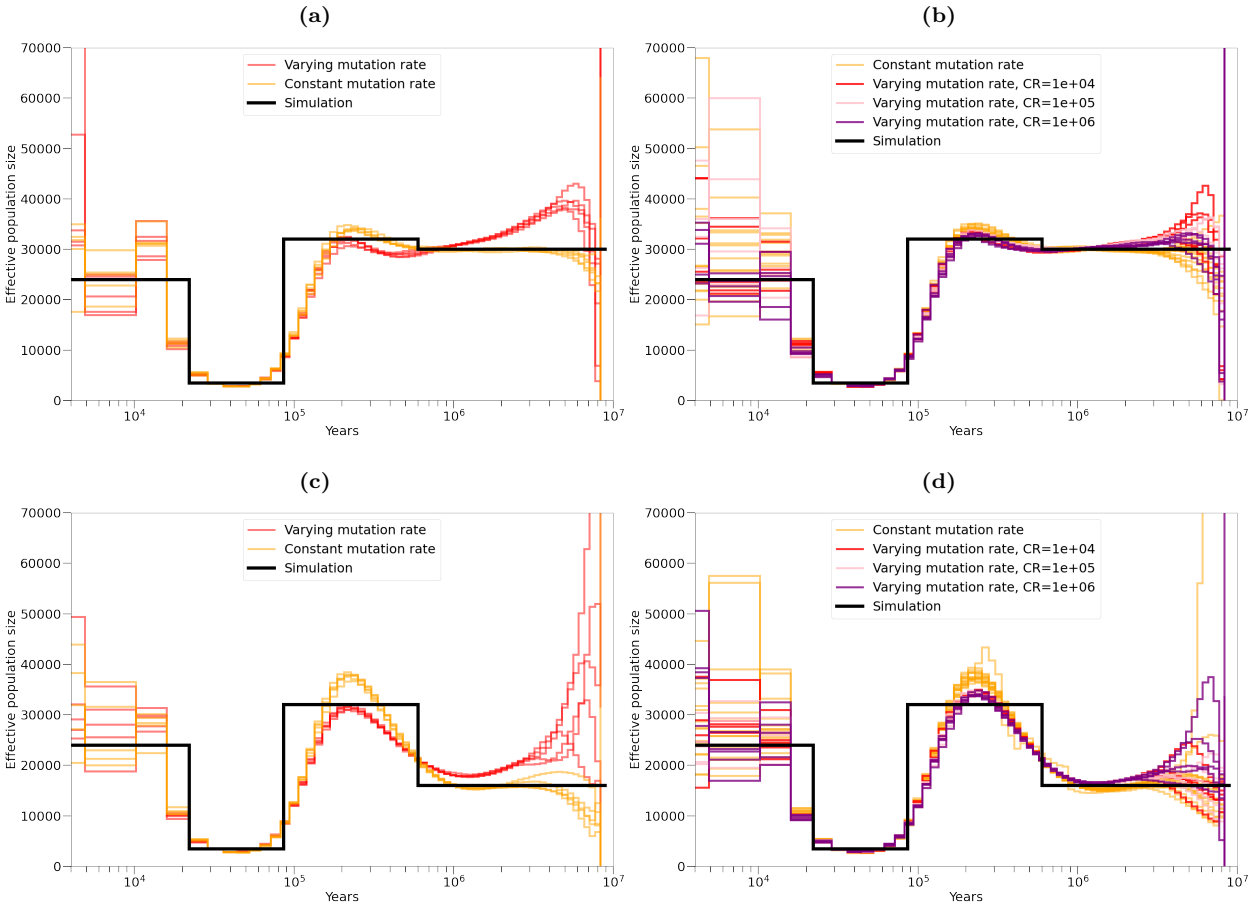

Figure A7: PSMC inference on simulations with a mutation rate that varies across the genome. The mutation rate changes at disjoint intervals, whose length is exponentially distributed with rate L. In each interval, the mutation rate is drawn from gamma distribution with mean 1.25e-08 and a coefficient of variation v. The red, pink and purple lines show inference from PSMC on a simulation where the mutation rate varies, and the gold lines show inference from PSMC on a simulation where the mutation rate is constant. **(a)** and **(c)** $v = 0.25$ and L =10kb. **(b)** and **(d)** $v = 0.15$ and L=10kb (red), 100kb (pink), and 1Mb (purple).

regions with faster mutation rates will have more mutations than otherwise, which PSMC will interpret as coming from a more ancient TMRCA, and vice versa. Consequently, overestimation and underestimation of TMRCAs can occur, which can confound efforts to infer accurate evolutionary histories.

To test this effect, we performed a series of simulations of a spatially varying mutation rate that changes according to a gamma distribution, in line with the models proposed in [73]. We generated a mutation rate map as follows. First, we divided the genome into disjoint intervals, whose length is exponentially distributed according to a rate parameter, L. Then, for each interval, we drew a mutation rate from a gamma distribution whose mean is fixed to be 1.25e-08 [64, 65, 66, 67, 68, 69] with a coefficient of variation, v. We simulated over $L \in \{10\text{kb},100\text{kb},1\text{Mb}\}$ and $v \in \{0.15,0.25\}$. We simulated diploid genomes according to a coarse piecewise population size trajectory using msprime [74], then generated pointwise mutations according to the mutation rate map. We generated 5 replicates of this simulation and then inferred the population size history using PSMC. Finally, as a control, we generated an additional 5 replicates using the same procedure, but with a constant mutation rate of 1.25e-08, and infer $N(t)$ using PSMC as well.

This simulation study shows that if the variation in mutation rate is strong enough (v=0.15), and the rate of change along the genome is fast enough (L=10kb), PSMC detects a spurious peak in ancient time (Figure A7a). However, we note that qualitatively, the false peaks across different replicates seem to vary more in location and height than in the human inference, for which the ancient peaks are well aligned (Figure 1a). Additionally, this simulation used v=0.25, which is larger than mutation rate model fitted by [73] for the datasets in [75, 76] (0.18 and 0.15, respectively) but smaller than in [77] (0.27). In a simulation with v=0.15, we do not consistently observe a second peak (Figure A7b) for any value of L. In a simulation with lower ancient $N(t)$, v=0.25 does generate an artificial inflation of ancient $N(t)$, but not a discernible peak (Figure A7c). A simulation with $v = 0.15$ and L=1Mb again can generate a false peak, though it is not consistent across replicates (purple line, Figure A7d); L=10kb or L=100kb with v=0.15 does not seem to have much of an effect as the inference is similar to the constant mutation rate inference.

If the ancient peak was an artefact caused by variable mutation rates, we would expect to see a correlation between the mutation rate and the inferred TMRCA. To test this, we used de-novo mutation abundance counts obtained from trios [78] to generate a pointwise mutation rate map that depends on the local trinucleotide context (see Methods). The generated mutation map is positively correlated with SNP density as reported in 1000GP's dbSNP (r^2=0.09, p-value<1e-100, averaged across non-overlapping windows of 100bp), indicating positions with higher inferred rates have an elevated probability of mutations. We compared the variable mutation rates with the maximum inferred TMRCA across the 26 individuals in 1000GP, evaluated every 1kb. We observed a statistically insignificant Pearson's correlation (r=0.002, p-value=0.29) and a significant negative Spearman's correlation (rho=-0.04, p-value=3.7e-93) between TMRCA and mutation rate map. This suggests variable mutation rates do not play a significant role in causing the ancient peak.

Finally, we explored the role of CpG islands in affecting $N(t)$ inference. CpG islands are regions of DNA characterised by a high frequency of cytosine-guanine dinucleotides. They are often found near the transcription start site of genes and are associated with gene regulation. Due to their regulatory role, CpG islands have fewer mutations than expected by their high GC content, and therefore may contribute to mutation rate variability along the genome. The fraction of the genome in a CpG island is 0.007, and we

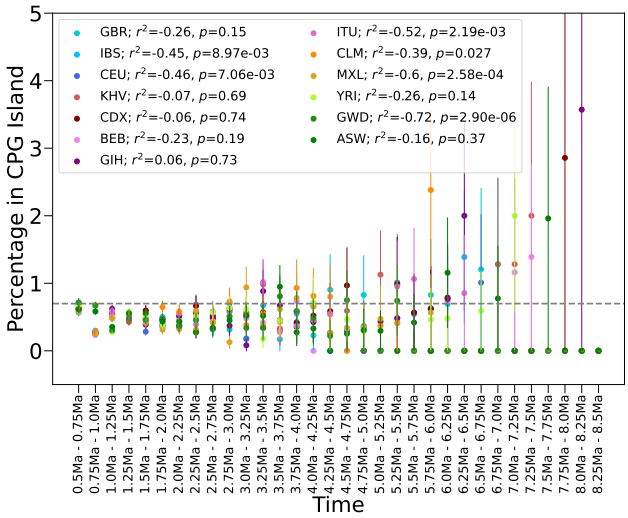

Figure A8: The correlation between the posterior mean TMRCA and the fraction of regions in a CpG island, on chromosome 1 for one individual from GBR, JPT, and ESN in the 1000GP. The Pearson r^2 and p-value are shown in the legend, although the relationship appears to be non-linear, and dominated by noise in ancient TMRCA bins. The dashed line indicates the genome-wide fraction of regions in a CpG island.

stratify this by inferred TMRCA in Figure A8. The relationship between the two variables is unclear, so to test the effect on inference we masked out CpG islands and inferred $N(t)$. We still observed an ancient peak (Figure A6c), and thus concluded that CpG islands are not its cause.

6.2.3 Variable mutation rate through time

The PSMC model also assumes a constant mutation rate across past generations. While this is a sufficient approximation to allow inference in the last ∼50kya [26], it may not hold in more ancient times. In [26], the authors show that by using a yearly mutation rate of 0.38e-09 to 0.49e-09, the PSMC curves of modern humans and a 45,000-year-old individual from Siberia are well aligned. Assuming a generation time of 30 years, this indicates that a per generation rate of 1.25e-08 is relatively stable in the last ∼50kya. Indeed, it has been suggested that the mutation rate has changed over the course of human evolution, also referred to as the "hominoid rate slowdown" [36, 37, 38, 39]. This is based on the observation that the yearly mutation rate estimated from human pedigree studies is almost half the rate inferred by considering observed differences between human and chimpanzee genomes. The effect of a possible slowdown on PSMC estimates is unclear.

To this end, we simulated 10 replicates of a diploid genome, arising from a constant population size, and a mutation rate that is fixed at 2.5e-08 at 10Mya then decreases to 1.25e-8 at present time (Figure A9a). To simulate a temporally changing mutation rate, we discretised time into D=64 segments with 65 time interval boundaries, $\tau = [\tau_1, \ldots, \tau_{64}]$, where

$$\tau_i = \omega exp\left(\frac{i}{D} log\left(1 + \frac{T_{max}}{\omega}\right) - 1\right) \tag{1}$$

and the changes in mutation rate are piecewise constant in these intervals, $\mu = [\mu_1, \ldots, \mu_{64}]$. We simulate the coalescent process using msprime [74] with Hudson's model of recombination [79], then utilise the memoryless

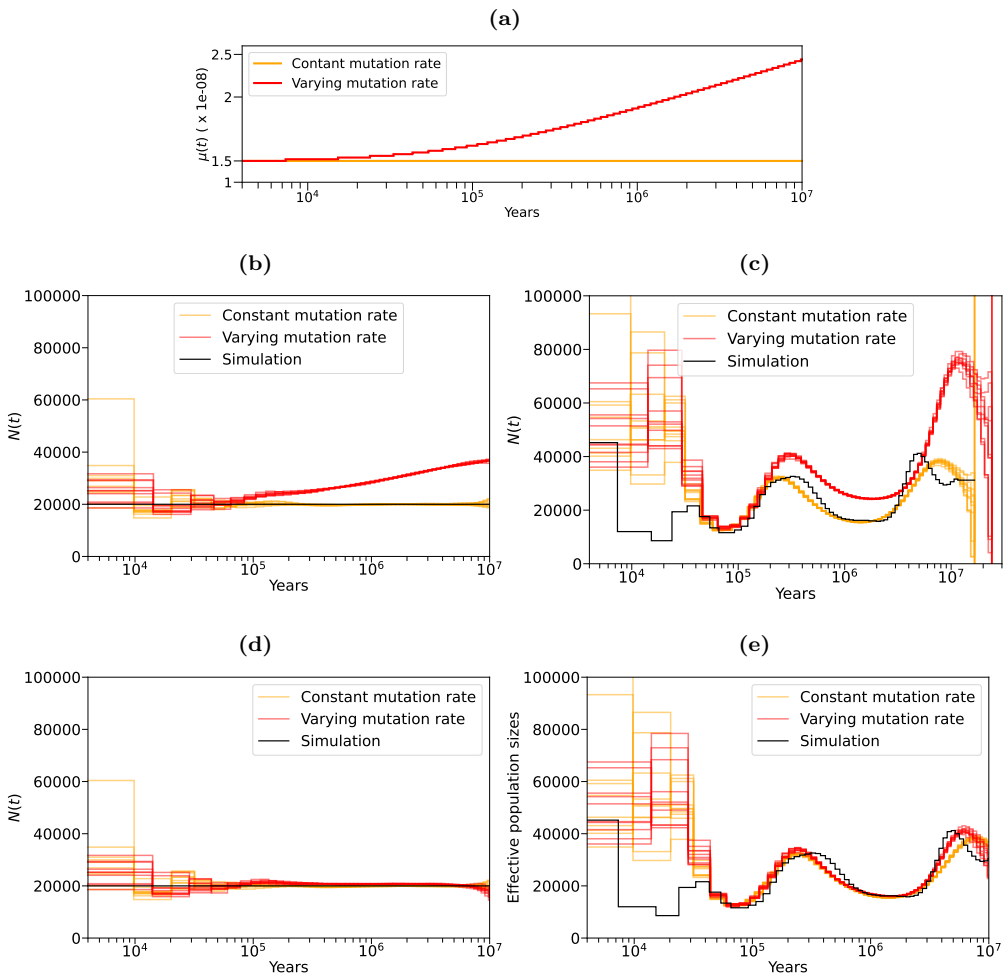

Figure A9: PSMC inference on simulations with a mutation rate that varies through time. **(a)** The mutation rate model, $\mu(t)$: at 10Mya the mutation rate is 2.5e-08, then it slows down to 1.25e-08 at present. **(b)** Inference from PSMC on a simulation with constant population size, where the mutation rate changes through time according to the model in (a) in red, and a constant mutation rate (1.25e-8) in gold. **(c)** The same simulation as in (b) but with a population undergoing size changes. **(d)** and **(e)** are the same as (b) and (c), respectively, although the inferred $N(t)$ on the simulation with a varying mutation rate (red) has been multiplied through by $\mu(t)$ (red line in (a)).

property of the exponential distribution to add mutations at each position. With a coalescence time of t where $\tau_i \leq t < \tau_{i+1}$, the probability of a mutation arising on either lineage is:

$$P(\text{Mutation arises}|t) = 1 - exp\left(-4N\left(\sum_{j=1}^{i-1}\mu_j\Delta_j - \mu_i(t-\tau_i)\right)\right) \tag{2}$$

where $\Delta_j = \tau_{j+1} - \tau_j$.

We used PSMC to infer an $N(t)$ curve, and compared the results to $N(t)$ curves inferred from simulations with a fixed mutation rate of 1.25e-8. We observed that PSMC infers an increasingly inflated $N(t)$ in ancient times (Figure A9b). With the same mutation model, we also simulated changes in $N(t)$ similar to that as

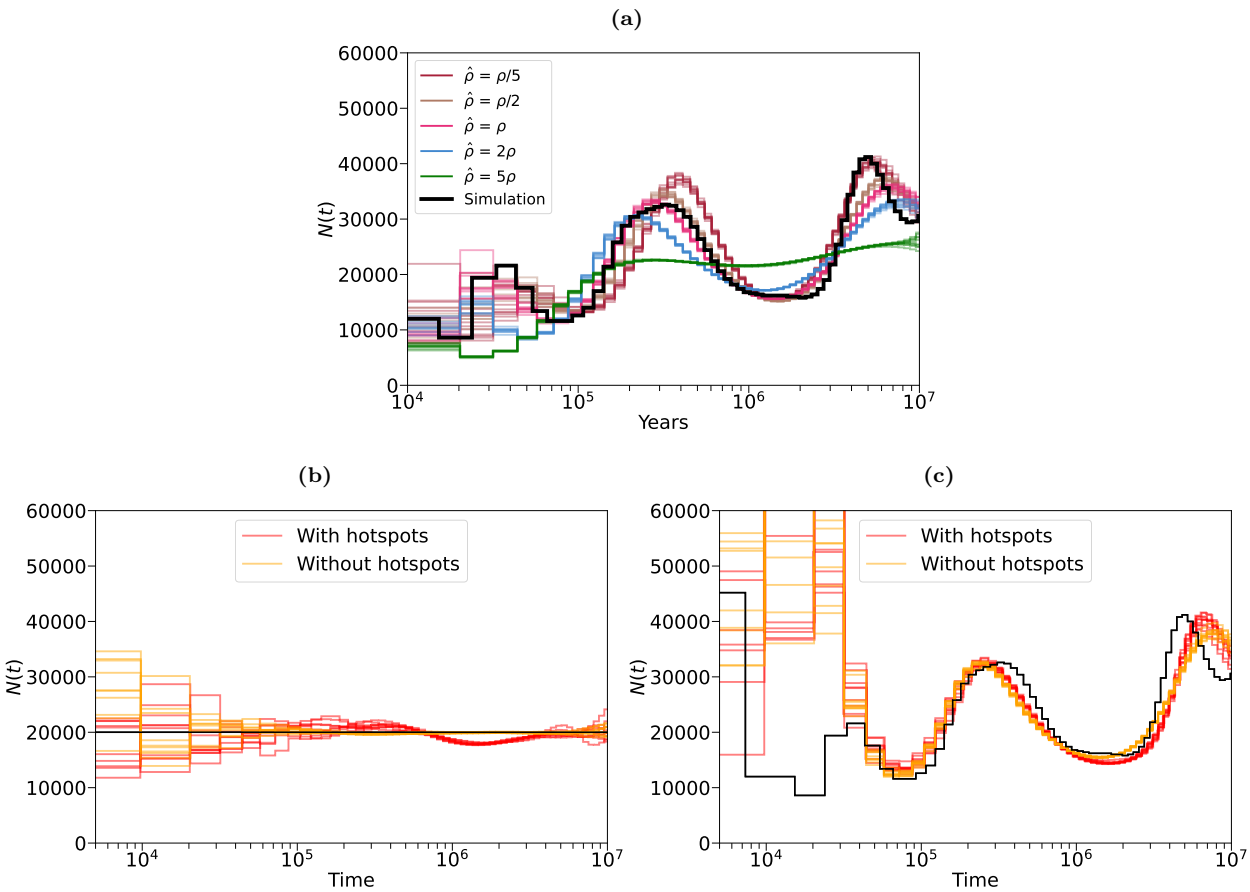

Figure A10: Inference from PSMC when its model of recombination is violated. **(a)** In a simulation with constant recombination rate ρ, we deliberately fixed the recombination rate used by the PSMC algorithm $\hat{\rho}$ at an incorrect value. The black line indicated the simulated $N(t)$ and the coloured lines indicate the various values of $\hat{\rho}$, which are expressed relative to the simulated value. **(b)** and **(c)** Inference from PSMC on simulations with changes in the recombination rate according to HapMap (red lines) and a constant recombination rate (gold lines).

inferred in the ESN (Figure 1). Again, we observe the inference of $N(t)$ from PSMC is increasingly inflated in ancient times, though the general shape of the trajectory is recovered (Figure A9c). We note that if the changes in mutation rate are known, we can simply scale the inference appropriately to correct the error (Figure A9d and e). Even though changes in the mutation rate over time can affect the estimation of $N(t)$, it is unlikely that the observed ancient peak in humans is due to this, as producing a peak as an artefact of changing mutation rates would require several rapid and severe mutation rate fluctuations, which seems unlikely. We conclude that mutation rate variation through time is not a likely cause of an ancient peak, and note that this may explain the differences in ancient $N(t)$ magnitudes in humans and chimpanzees (Figures 1a and b).

6.2.4 Variable recombination rate across the genome

The PSMC model also assumes a constant recombination rate across the genome, and uniformly through all past generations. However, recombination rate varies along the genome, with high rates at recombination hotspots and lower rates in different regions [80, 81, 42, 43, 44]. Looking back in time, the landscape of

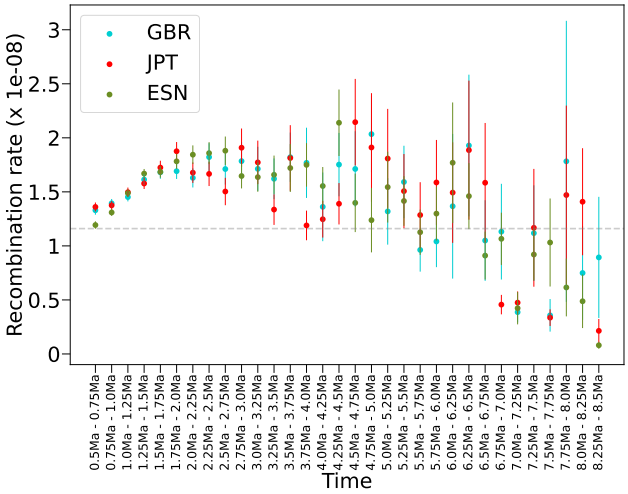

Figure A11: The relationship between the posterior mean TMRCA and the HapMap recombination rate map, on chromosome 1 for one individual from GBR, JPT, and ESN in the 1000GP. The reported Pearson r^2 for each population is 0.45, 0.41, and 0.50, with p-value<0.02 for each. However, clearly the relationship is non-linear; it appears that recombination rate increases as does TMRCA up until ∼2Mya, after which they become negatively correlated. This may come from model violations, in that PSMC assumes a uniform rate of recombination, or that the recombination landscape changes over time [82]. The grey, dashed line indicates the chromosome wide average recombination rate.

recombination is known to transform every few hundred thousand years [82, 83, 84, 85]. It is thus unclear how recombination rate variation affects $N(t)$ inference.

Previous work [2] has demonstrated that $N(t)$ inference in PSMC is robust even under simulations including recombination hotspots. As additional confirmation, we tested the effect of mis-specifying the recombination rate. Denote the simulated scaled recombination rate as ρ; we fix PSMC's recombination rate $\hat{\rho}$ as $0.2\rho, 0.5\rho, 1\rho, 2\rho$, or 5ρ, and infer $N(t)$ (Figure A10a). We observed that PSMC is able to recover $N(t)$ relatively accurately for all values of the recombination rate except $\hat{\rho}/\rho$ equal to 5. Finally, we saw a significant correlation between the local recombination probability (genetic map taken from HapMap [86, 87]; downloaded from `https://alkesgroup.broadinstitute.org/Eagle/`) and inferred TMRCAs (Figure A11), though the relationship appears to be non-linear. We investigated this by simulating from this genetic map, and inferring an $N(t)$ curve assuming a constant recombination rate. We observed that this type of model mis-specification does not significantly alter inference (Figure A10b and A10c). Given these observations, we believe it is unlikely that a spatially varying recombination rate could generate a fake ancient peak.

6.2.5 Recurrent mutations

PSMC assumes an infinite sites model, in which a site can only experience one mutation. In reality, some ancient sites will have experienced recurrent mutations. Under a Jukes-Cantor model of mutation [88], which assumes that each base pair mutates with uniform probability to another, with probability 1/3 a recurrent mutation will be revert to its ancestral state. These will be observed as a homozygote and therefore may be inferred as a lower TMRCA by PSMC. With probability 2/3 a recurrent mutation induces a distinct biallelic SNP, which is observed as a heterozygote. These biallelic SNPs will appear to PSMC as a younger TMRCA,

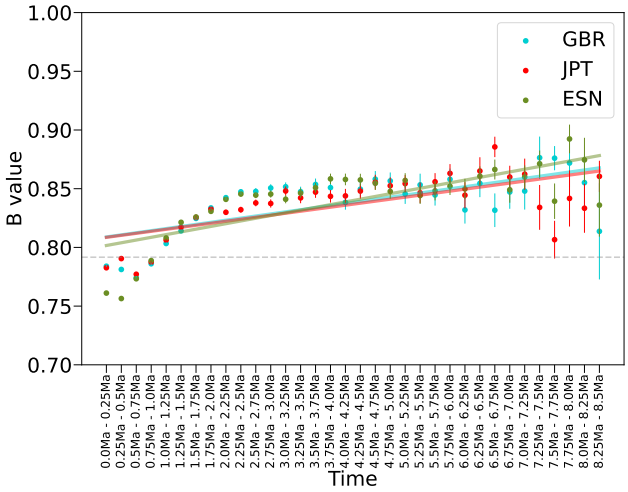

Figure A12: The relationship between the posterior mean TMRCA and a high resolution map of background selection [46], on chromosome 1 for one individual from GBR, JPT, and ESN in the 1000GP. The reported Pearson r^2 for each population is 0.67, 0.67, and 0.75, with p-value<1.3e-5 for each. In general this suggests that B value increases with TMRCA (a linear line of best fit has been added to show this), though the relationship appears to be non-linear. The grey, dashed line indicates the chromosome wide average B value.

so it is unlikely that these will significantly affect inference of ancient $N(t)$. Moreover, it is generally a whole region of a chromosome that is informative about local TMRCA, rather than a single base pair. If a segment is old enough to experience a recurrent mutation, then there are likely many other neighbouring SNPs from which the TMRCA can be accurately inferred.

Indeed, in almost all the simulations previously presented a Jukes-Cantor model [88] was used to generate mutations, in which recurrent mutations are allowed to occur. Generally, in simulations with uniform μ in time and space, we observe that inference of ancient $N(t)$ is reasonably accurate in ancient time (Figures A7, A9, and A10).

6.2.6 Background selection

Background selection (BGS) is a form of linked selection, where the removal of deleterious mutations reduces genetic diversity in the surrounding regions due to linkage [89, 90]. It has been demonstrated that BGS is pervasive throughout the human genome, and that this explains roughly 60% of the variance in diversity at the megabase scale [45, 46]. Many inference methods, however, assume that the genome evolves neutrally. This is problematic, as it has been shown that wrongly assuming the absence of selection means that inference methods are not able to accurately reconstruct the demographic history [47, 49, 48].

We correlated the inferred TMRCAs against a high resolution map that describes the strength of BGS across the human genome [46]. We saw a significant positive correlation (Pearson $r^2 \approx 0.7$, p-value<1e-5; Figure A12), which is consistent with models of BGS that model regions under strong linked selection as having lower effective population size [90]. To explore the effect of BGS on inference of $N(t)$ in real data, the authors in [58] binned the genome of 10 YRI individuals into quintiles according the strength of BGS and ran PSMC in each (reproduced in Figure A13). In all quintiles except the one with strongest BGS, a clear second peak is seen in ancient time. The peak is not seen in the strongest BGS quintile likely because

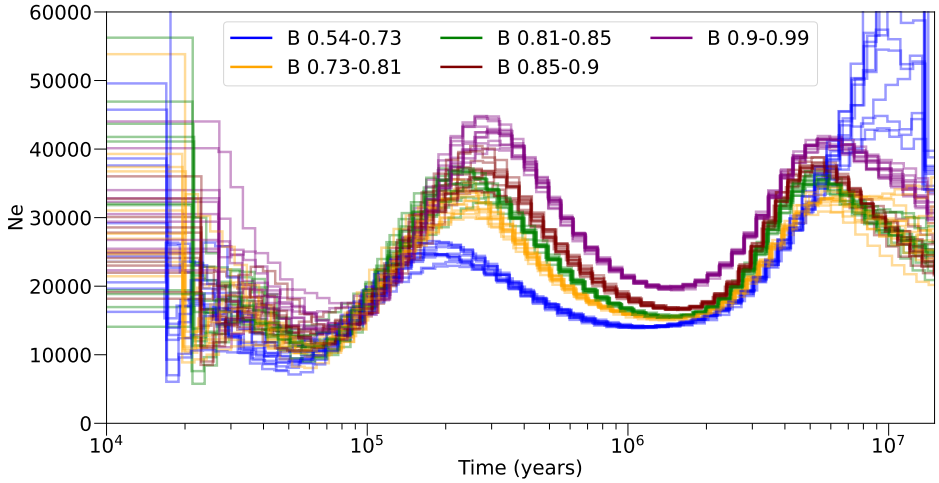

Figure A13: PSMC inference of $N(t)$ in 10 YRI individuals from the 1000GP project, stratified into quintiles according to the strength of BGS (as inferred by [46]). B values indicate the local strength of BGS, with 1 being no reduction in genetic diversity due to selection and 0 being full removal of genetic diversity. This figure is reproduced with permission from [58].

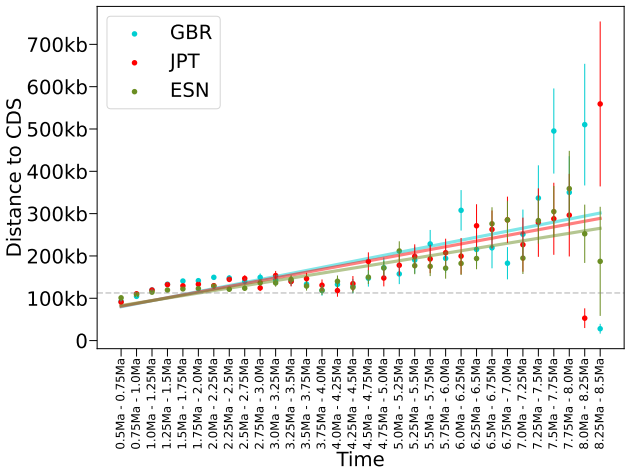

Figure A14: The relationship between the posterior mean TMRCA and mean distance to coding sequence (CDS), on chromosome 1 for one individual from GBR, JPT, and ESN in the 1000GP. The reported Pearson r^2 for each population is 0.63, 0.67, and 0.83, with p-value<0.0001 for each. In general this suggests that dcCDS increases with TMRCA (a linear line of best fit has been added to show this), although the relationship appears reversed from ∼3Mya to ∼4.5Ma. The grey, dashed line indicates the chromosome wide average distance to CDS.

all of the input sequence has already coalesced. In none of the simulations with realistic parameters of BGS, as shown in [48, 91, 58, 92], does PSMC infer infer a false ancient peak. We thus find it implausible that widespread BGS could generate an ancient peak as observed in humans.

6.2.7 Balancing selection

Balancing selection (BLS) is a type of natural selection that maintains genetic diversity in a population by favouring the maintenance of multiple alleles. This type of selection can occur through various mechanisms, such as heterozygote advantage, frequency-dependent selection, or spatially varying selection [93]. BLS oper-

ating for a long time period will maintain advantageous polymorphism and result in an older TMRCA than expected under neutrality. Therefore, if BLS were sufficiently prevalent in the genome, this would manifest as enrichment of older coalescence times, which would increase inferred $N(t)$ in the past.

The prevalence of BLS in the human genome is unclear. Initially considered a rarity [94, 95] and overlooked, balanced polymorphisms have recently received renewed attention with several lines of evidence showing their relevance in human evolution [96, 97, 98]. Recently, hundreds of loci were implicated in possible trans-species BLS, maintained since earlier than human-chimpanzee speciation [99]. By finding regions of ancient shared ancestry across 54 individuals from Complete Genomics [100], in [101] the authors suggest numerous regions that are under BLS. We calculated the correlation between inferred TMRCAs and distance to coding sequence (CDS), and found that ancient TMRCAs tend to be increasingly far away from CDS (Figure A14; Pearson $r^2 \approx 0.71$, p-value<0.0001). As BLS usually acts on or near functionally important parts of the genome, this makes it unlikely that the ancient TMRCAs underlying our ancient peak are driven by BLS and that this is the cause of the ancient peak.

6.3 Methods

6.3.1 Correlations with annotations

We downloaded the the tracks for repeats, segmental duplications, and CpG islands from the UCSC Genome Browser `https://genome.ucsc.edu/cgi-bin/hgTables` [102]. The positions of coding sequence were obtained from Gencode `ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_45/gencode.v45.chr_patch_hapl_scaff.basic.annotation.gff3.gz`. We used the B-map as inferred by Murphy et al. [46] and lifted over from GRCh37 to GRCh38 [102].

We took the posterior mean TMRCA across chromosome 1 for the 26 1000GP samples. We stratified the TMRCAs into windows of 250kya and analysed how various functional annotations correspond to these TMRCA windows, and ensured that positions in the analysis passed the strict mappability mask.