

# A method for genome-wide genealogy estimation for thousands of samples

Leo Speidel<sup>1</sup>, Marie Forest<sup>2</sup>, Sinan Shi<sup>1</sup>, Simon R. Myers<sup>1,3,\*</sup>

<sup>1</sup>Department of Statistics, University of Oxford, Oxford, UK

<sup>2</sup>Université du Québec à Montréal, Montréal, Canada

<sup>3</sup>Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

Corresponding author email: [myers@stats.ox.ac.uk](mailto:myers@stats.ox.ac.uk)

## Abstract

Knowledge of genome-wide genealogies for thousands of individuals would simplify most evolutionary analyses for humans and other species, but has remained computationally infeasible. We developed a method, Relate, scaling to > 10,000 sequences while simultaneously estimating branch lengths, mutational ages, and variable historical population sizes, as well as allowing for data errors. Application to 1000 Genomes Project haplotypes produces joint genealogical histories for 26 human populations. Highly diverged lineages are present in all groups, but most frequent in Africa. Outside Africa, these mainly reflect ancient introgression from groups related to Neanderthals and Denisovans, while African signals instead reflect unknown events, unique to that continent. Our approach allows more powerful inferences of natural selection than previously possible. We identify multiple novel regions under strong positive selection, and multi-allelic traits including hair color, body mass index (BMI), and blood pressure, showing strong evidence of directional selection, varying among human groups.

Large-scale genetic variation datasets are now available for many species, including tens of thousands of humans. In principle, all information about a sample's genetic history is captured by their underlying genealogical history, which records the historical coalescence, recombination, and mutation events producing the observed variation patterns. In practice, several key existing approaches (e.g., Refs. [1,2]) leverage an underlying coalescent model, which provides a flexible modelling framework and is the limiting behavior of a variety of finite-population models<sup>3,4</sup>. However, coalescent-based inference is complicated by the structure of the model, and the extremely large space of probabilistically plausible sample histories conditional on observed data<sup>5</sup>. Other approaches<sup>6-11</sup> use more heuristic coalescent approximations, sometimes reducing accuracy: regardless, published existing methods scale to tens or a few hundred samples at most.

These issues have restricted the use of direct genealogy-based approaches to infer recombination, mutational ages, and natural selection to smaller datasets<sup>1,2</sup>, while for larger datasets diverse approaches based on data summaries<sup>12-14</sup> or downsampling<sup>15,16</sup> have predominated. In humans, such tools have detected genetic structure and admixture in good agreement with independent evidence<sup>17,18</sup>, changes in population size<sup>15,19-21</sup> and introgression with archaic groups, including Neanderthals<sup>22</sup>.

We developed a scalable method, Relate, to estimate genome-wide genealogies (**Figure 1; Methods; <https://myersgroup.github.io/relate/>** for implementation). Relate separates two steps; firstly identifying a genealogical framework at each site in the genome, describing ancestral relationships among sequences but not their coalescence times. Secondly, these times are estimated after mutations are mapped to branches of these trees, allowing for variable population sizes simultaneously inferred from the data, to produce complete genealogies. These are then used directly for downstream inferences. Our approach approximates the coalescent model, but performs as well as or better than existing approaches in our simulations, whilst being thousands of times faster.

We demonstrate the utility of a genealogy-based analysis by applying Relate to 4,956 haplotypes of the 1000 Genomes Project (1000GP) dataset<sup>23,24</sup>. We estimate population sizes of all 26 populations in the dataset and their split times using cross-coalescence rates. In agreement with previous work, we identify an increase in the mutation rate of TCC to TTC around 10,000 to 20,000 years ago<sup>25</sup>. The estimated genealogies contain

signals of introgression between Neanderthals and modern humans in Eurasia, and between modern East and South Asians and Denisovans, alongside other signals specific to African groups. Finally, we suggest a test statistic that identifies loci under positive selection by tracking mutation frequencies through time. We demonstrate that for plausible scenarios of selection on complex traits, involving selection dispersed over many loci, this test improves power over the integrated Haplotype Score (iHS)<sup>26</sup>, and identify previously unreported genomic regions under strong positive selection. We find a remarkable enrichment of SNPs identified in genome-wide association studies (GWAS) among targets of selection, and evidence of widespread directional polygenic adaptation.

## Results

### Overview of the Relate approach

At each genomic position, Relate first identifies a non-symmetric distance matrix whose rows estimate the relative order of coalescence events between a particular sequence and the remaining observed sequences. To do this, Relate uses the posterior probabilities output by a hidden Markov model (HMM) similar to that proposed by Li and Stephens<sup>27</sup>, but leveraging knowledge of ancestral and derived status at each single nucleotide polymorphism (SNP) to improve speed and accuracy. This distance matrix is used to construct a rooted binary tree using a bespoke algorithm. Mathematical arguments demonstrate, encouragingly, that if the “infinite-sites” model is satisfied so that each observed mutation occurs exactly once, our approach is guaranteed to generate genealogies exactly producing the observed data, in the limiting cases where either there is no recombination, where the recombination rate is very large, or where all recombination occurs in intense widely spaced hotspots (Supplementary Note). Because the distance matrix is position specific, these binary trees adapt to changes in local genetic ancestry due to recombination. In practice, we save computational time by only rebuilding trees at a subset of sites along the genome (**Methods**, Supplementary Figures 1 and 2).

To estimate branch lengths while allowing for changing population sizes, we first map mutations onto each genealogical tree and then apply an iterative Markov Chain Monte Carlo (MCMC) algorithm to estimate times under a coalescent prior. We simultaneously estimate a stepwise varying effective population size through

time, using the genome-wide collection of estimated genealogies (**Methods**). Our final time estimates account for changes in population size, assuming an unstructured population. We can also explore population stratification within a sample, by leveraging estimated coalescence rates of any pair of sampled sequences. By averaging pairwise coalescence rates within and across groups, we obtain effective population size estimates for sub-populations and cross-coalescence rates between populations. As we show in the next section, this can provide accurate estimates despite the fact that our tree-builder does not account for such population stratification.

## Simulations

We evaluated Relate for its speed, accuracy of inferred trees, robustness, and ability to infer evolutionary parameters, by simulating data under the coalescent with recombination using msprime<sup>28</sup>. We compared performance to ARGweaver<sup>2</sup>, which samples from a time-discretized approximation to the coalescent with recombination, and which we therefore expect to perform well on these simulations. Relate was >4 orders of magnitude faster than ARGweaver, for cases we were able to apply the latter, and also much faster than RENT+<sup>11</sup> (**Figure 2a,b**). Our approach scales linearly in sequence length and quadratically in sample size  $N$ , enabling genealogical inference for e.g. 10,000 human samples genome-wide using a compute cluster.

To evaluate accuracy, we compare, at each locus and for each of the  $\binom{N}{2}$  pairs of haplotypes, the estimated time to their most-recent common ancestor (TMRCA) to the truth (**Figure 2c,d**), observing improved performance relative to both ARGweaver and RENT+. Relate also showed improved robustness to errors in the data, identified misclassified ancestral alleles, and estimated times well in the context of varying population size (Supplementary Figure 3). Other accuracy measures yielded similar improvements (Supplementary Note). Relate identified repeat mutations and variable mutation rates, and is robust to computational rephasing of haplotypes (Supplementary Figure 4). We next compare Relate's inferred population sizes to those from applying two leading specialist approaches, MSMC<sup>20</sup> and SMC++<sup>21</sup>. For multiple previously tested<sup>20,21</sup> scenarios including oscillating population sizes and bottlenecks similar to those observed in out-of-Africa events of modern humans, Relate obtains more accurate estimates, particularly in the recent past (**Figure 2e**, Supplementary Figure 3). While Relate assumes a single population when

estimating branch lengths, when applied to a combined sample from two diverged populations, it still performs well in recovering their distinct population histories and estimating their split time(s) (**Figure 2f,g**).

## Genome-wide human genealogies

We applied Relate to 2,478 1000GP individuals with diverse genetic ancestry and approximately 81 million SNPs (see **Methods** for data pre-processing). Computation time, using up to 300 processors, was ~4 days (Supplementary Table 1). 86% of all SNPs (>96% of SNPs at >0.2% derived-allele frequency (DAF)) map uniquely to trees, falling to 76% for CpG dinucleotides, known to possess strongly elevated mutation rates (Supplementary Figure 5). The number of different trees in a genomic subregion strongly correlates with recombination rate ( $r^2 = 0.63$ ) and the average tree has 3,883 SNPs mapped to it, reflecting block-like structures of human haplotypes between recombination hotspots (Supplementary Figure 5).

We estimated within and across-group coalescence rates for pairs of groups, by first extracting the genealogy for members of a particular subsample of interest embedded within the full genealogy, and then re-estimating coalescence rates for this genealogy. We observe a clear out-of-Africa bottleneck for Eurasian populations (CHB: Chinese in Beijing and GBR: British in England and Scotland shown), and gradual separation from African populations (YRI: Yoruba in Ibadan, Nigeria shown) already visible 200,000 years before present (YBP) and lasting to around 60,000YBP (**Figure 3a,b**). This is consistent with recent studies<sup>15,29</sup> and might reflect several out-of-Africa dispersal events. Asian (CHB shown) and European (GBR shown) populations separate more recently, with a clear and visibly more sudden separation around 30,000 YBP (**Figure 3c**). We also detect, and date, very recent separations <10,000 YBP, such as between CHB-JPT (JPT: Japanese in Tokyo) or FIN-GBR (FIN: Finnish in Finland) (**Figure 3d,e**). Finnish samples exhibit a second bottleneck, around 3000-9,000 YBP following separation from GBR<sup>30,31</sup>, with other population-specific events in e.g. Peruvians and Gujarati individuals (Supplementary Figure 6). The Finnish bottleneck is thought to have caused enrichment of certain disease-causing gene variants, commonly classified as Finnish heritage diseases<sup>30,31</sup>. A strong bottleneck post-dating separation from Eurasian groups is absent in African populations (LWK: Luhya in Webuye, Kenya and YRI shown, **Figure 3f**). All populations show a remarkable increase often to >1,000,000, in the recent past (Supplementary Figure 6), however we note possible inaccuracies due to incomplete power

to detect rare variants<sup>24</sup> leading to underestimation, and computational phasing leading to overestimation (Supplementary Figure 4).

Exploring the relative mutation rate of particular mutation classes through time confirms, as reported previously<sup>25</sup>, a strong elevation in the rate of trinucleotide changes including TCC->TTC in West Eurasian groups, which we date to 5,000-30,000 YBP, but infer to be weak or absent in the present day (**Figure 4a**). Other mutation types show more subtle temporal biases and signatures consistent with GC-biased gene conversion<sup>32</sup> (Supplementary Figure 6). Overall, these results support accuracy of our inferred historical relationships, including the timing of a range of different historical events, identified within a single analysis framework.

## Neanderthal/Denisovan and unexplained introgression events

Introgression from distantly related groups in the past is expected to introduce lineages which forward in time can randomly spread in the tree, and backward in time remain distinct from other lineages, resulting in an excess of long branches associated with particular times. We identified such deep branches (>1 million years (MY) in age and with varying lower end), across human groups (**Figure 4b,c**). It is established that all non-African human groups possess similar levels of Neanderthal introgression, and specific Asian and Australasian groups possess admixture from a group related to Denisovans<sup>22,33</sup>. We therefore label deep branches possessing at least two derived mutations by whether at least one mutation is shared with the sequenced Denisova<sup>33</sup> or Neanderthal<sup>22,34</sup> genomes (**Figure 4b** shows one example of likely introgression from Neanderthals into European GBR, but not African YRI individuals). After classifying deep branches based on their lower-end times, for branches originating within the last 10,000YBP, 85-90% are shared with Neanderthal or Denisovan for most Eurasian groups (**Figure 4c, Methods**). Any lineages from recent introgression events will show a lower-end age younger than the time of introgression, and upper-end older than the split time of the introgressing group, so we expect branches with a younger lower-end to be most enriched with lineages that came from distantly diverged introgressing groups. This suggests that aside from groups closely related to Neanderthals and Denisovans, no strongly diverged hominid has left a major, recent impact in non-African populations studied here. An exception is IBS, which has more long branches shared

with African populations (Supplementary Figure 6). In East and South Asian groups, the data suggest a very recent arrival of Denisovan DNA (mainly <15,000YBP). In non-Africans, Neanderthal sharing remains high for branches with lower-end age younger than ~30,000YBP. These dates are only lower bounds on the introgression time, and an accurate arrival date of Neanderthal DNA would require estimating a joint genealogy which requires further work. Nevertheless, they are consistent with previous estimates based on linkage disequilibrium (LD)<sup>35</sup>, and of direct evidence of hybrids<sup>35,36</sup> around 40,000 YBP. Moreover, elevation in the sharing of quite deep haplotypes with Neanderthals steadily increases for branches with lower-end age of ~100,000 YBP towards the present, which is suggestive of introgression beginning from this time in non-African individuals, although it is important to note that our date estimates for individual events might be over- or under-estimates in some cases.

In contrast to non-African groups, sharing with Neanderthal/Denisovans is lower (<20%, **Figure 4c**) in African populations, and declines towards the present, suggesting minimal recent interactions<sup>22,33</sup>. This is despite the fact that African populations possess far more long branches (on average, on deep branches with lower coalescence age <30,000 YBP; 42,434 vs. 7,012 mutations occur in African vs. non-African populations). Of mutations on long branches found in Africa, 98% are Africa-specific, indicating separate events occurring in non-African and African populations (Supplementary Figure 6). Comparing YRI, GBR, BEB (Bengali in Bangladesh), and CHB to expectations under panmixia, we observe a strong excess of mutations on deep branches with lower coalescence age <40,000 YBP in all cases, which is almost entirely explained by Neanderthals/Denisovans in the non-African populations, but not in YRI (**Figure 4d, Methods**). In panmictic simulations with matched population size histories, we observe no such excess (Supplementary Figure 6). This evidences ancient but uncharacterised population structure within Africa, as suggested elsewhere<sup>37,38</sup>. **Figure 4b** shows one example consistent with an introgression event in YRI, not involving a closely relative of Neanderthals.

## Powerful tree-based approaches to study natural selection

By directly modelling how mutations arise and spread, genealogical trees offer the potential to powerfully investigate different modes of natural selection. For example, a recent method, SDS, indirectly tests for

differences in tree tip branch lengths between carriers and non-carriers using the density of singletons around a focal SNP<sup>39</sup> and a tree-based analogue (trSDS) tests this directly<sup>40</sup>. Here, we propose a class of approaches (Relate Selection Tests) based on estimating the speed of spread of a particular lineage (carrying a particular mutation), relative to other “competing” lineages, over some chosen time range. To test for selection over the entire lifetime of a mutation, we condition on the number of lineages present when it first arises, and use as a test statistic the number of present-day carriers. Assuming no population stratification, the null distribution of this statistic can be calculated analytically and is robust in principle to population size changes (**Methods**).

Simulated data (**Figure 5a**) show a close match in null no-selection scenarios of our p-values  $p_R$  to the expected uniform distribution. Across a range of selective advantages and SNP frequencies (**Figure 5b**, **Supplementary Figure 6**), our approach increases power relative to (tr)SDS, as well as iHS for weaker selection in particular. trSDS is more powerful than SDS, while applying the Relate Selection Test to *true* genealogical trees yields a test that is uniformly more powerful than other approaches (**Figure 5b**), indicating the strength of tree-based approaches. In practice, there is some decrease in power from the need to infer trees via Relate. The power increases for weak selection might be particularly beneficial for testing complex, polygenic traits, where small effect sizes at individual loci are expected to yield small selection coefficients<sup>41</sup>.

Calculating  $p_R$  for SNPs across twenty 1000GP populations (**Methods**) identified 35 distinct (24 novel) stringent signals genome-wide ( $p_R < 5 \times 10^{-8}$  in each of three or more groups) (Supplementary Table 3). These include the LCT region associated with Lactose tolerance in Europeans, and a mutation in the *EDAR* gene in East Asian populations<sup>42,43</sup>, with both likely causal variants strongly associated with our most significant mutation ( $r^2 \geq 0.8$ ). We also observe a previously-detected strong signal of positive selection in the MHC region in GBR<sup>44</sup> (**Figure 5c**). Among new regions, we identify selection evidence at the *EDARADD* gene – which interacts with the *EDAR* gene<sup>45</sup> in the formation of hair follicles, sweat glands, and teeth<sup>43</sup> – in all South Asian populations and Finns, with  $p_R < 10^{-6}$  in all other European populations. In 16 of 35 regions, we identify GWAS catalogue hits (OR=6.44; p=0.01), non-synonymous mutations (OR=2.49; p=0.16), or expression quantitative trait loci (eQTLs; OR=1.74; p=0.1), in LD with the mutation with strongest selection evidence ( $r^2 \geq 0.8$ , **Methods**), suggesting functional effects, reaching statistical significance for the case of



GWAS hits despite the small number of cases tested. Notably, 18 of the 35 regions are found only for African populations.

SNPs in functional parts of the genome are significantly enriched among targets of positive selection (**Figure 5d, Methods**), with strongest enrichment for GWAS hits, across all considered populations. This encouragingly supports a link between evidence of selection and SNPs with detectable influences on phenotypes at the organism level. Multiple previous studies<sup>46–49</sup> have attempted to test polygenic traits for evidence of directional selection, but confounding due to population stratification<sup>50,51</sup> is potentially problematic in practice. To leverage potential power gains, we tested whether derived mutations increasing (or decreasing) a trait show increased selection evidence relative to randomly sampled control mutations of the same frequency (one-sided Wilcoxon test; **Methods**). For each trait, we thin GWAS hits to account for LD and examined only SNPs showing “genome-wide significant” associations ( $p < 5 \times 10^{-8}$ ), because confounding due to population stratification is thought to occur through relatively small - but systematic - biases in effect size estimates<sup>50,51</sup>, but is not expected, in general, to produce genome-wide significant false-positives. At each SNP, we use only the association direction, rather than its strength, to offer additional robustness to potential confounding.

If positive selection influences a trait in a certain direction, e.g. increasing, we would expect positive selection on trait-increasing and negative selection on trait-decreasing mutations. We expect our test to be sensitive mainly to the former, because selection will increase frequencies of such SNPs, and the Relate Selection Test has reduced power to identify selection at rarer markers (**Figure 5b**). However, for traits with a large number of hits and strong selection, it is theoretically possible to observe some selection evidence in both directions<sup>52,53</sup>, because to avoid ascertainment effects we condition on SNP allele frequencies at trait-influencing sites. Therefore, we additionally test for differences in present-day DAFs between trait-increasing and trait-decreasing mutations, which can provide orthogonal evidence of polygenic adaptation, aiding interpretation of results (**Methods**).

As a positive control, we applied our test to GWAS for hair colour within the UK Biobank<sup>54</sup> (**Figure 6a**). As in previous studies<sup>49,55,56</sup>, we find a signal for SNPs associated with blonder hair color among European

populations. We further observe strong selection towards light brown hair color and against black hair color, including more weakly in South Asians, but not in other groups. Testing based on iHS scores decreases significance by up to 4 orders of magnitude (**Figure 6a**), and some signals become non-significant. We applied the same test to test 84 traits: 6 from the UK Biobank, and 78 with at least 10 genome-wide significant GWAS catalogue association signals in each effect direction, in all populations except recently admixed groups. 61 of these (73%) showed nominal selection evidence ( $p < 0.05$ ) in at least one population (**Figure 6b**), with strong geographic clustering. The most significant signal ( $p = 6 \times 10^{-14}$ ) is for SNPs associated with decreased Body Mass Index (BMI) in CEU. The largest number of selection signals are observed for Europeans, likely because many GWAS were conducted in these groups. Interestingly, East Asians have the fewest signals and no enrichment of low p-values (Supplementary Figure 8), possibly explained by their stronger population bottleneck, which would theoretically be expected to weaken selection signals.

Height, BMI, and Schizophrenia have been studied previously and show a large number of association signals<sup>57</sup>. While several studies have reported genetic differentiation between populations<sup>58–60</sup>, evidence for selection remains controversial<sup>40,47–51,58,59,61</sup> and some studies reporting recent selection on increased height in Europeans appear confounded by subtle population stratification<sup>40,50,51</sup>. Our test finds selection evidence for *both* effect directions in each population for height, except in East Asians, using the UK Biobank GWAS. DAFs tend to be larger towards the height-decreasing direction. This complex picture may be a consequence of both negative and positive selection acting on height, as well as pleiotropy. SNPs impacting other traits might also impact height (Supplementary Note). We identify strong evidence of selection favouring BMI-decreasing SNPs across almost all populations, with agreement of DAF shifts, indicative of directional selection. For both traits, we detect little evidence of selection in the smaller GWAS catalogue collection. Decreased risk of Schizophrenia has evidence of selection in Europeans, and some South and East Asian populations, while African populations show selection evidence towards a risk increase.

Among other phenotypes, we see selection evidence for a variety of blood-related phenotypes, with congruent DAF signals. In Europeans and some South Asians, we detect a strong signal favoring SNPs associated with blood pressure increases, contrary to previous studies suggesting the opposite direction<sup>55,62</sup>. We moreover

find evidence in many groups for selection favouring SNPs associated with decreased hemoglobin concentration and related traits, and with increases in platelet-related traits.

## Discussion

We introduced Relate, a scalable method for estimating genealogies genome-wide and demonstrated its accuracy and utility on a diverse set of applications. In many settings, Relate improved on existing state-of-the-art methods, which have previously required separate analyses: by instead obtaining inferences from the same genealogies, comparisons across different applications become straightforward. This approach is highly modular; methods developed for genealogy-based inference should be applicable regardless of the specific algorithm used for estimating marginal trees. Although we have focused on human genomes, Relate should work equally well in other recombining species.

In our 1000GP data analysis, we provide several examples whereby Relate-based trees are able to capture evolutionary processes that are themselves evolving through time: “evolution of evolution”. Temporal changes in mutation rates, population size, migration, and archaic admixture are simultaneously inferred, as are population-specific signals of natural selection. Genealogies provide a powerful, natural way to study these complex, interacting phenomena, and we believe studies of other evolutionarily and temporally dynamic processes – for example of evolution of recombination rates<sup>63,64</sup> - will yield new insights.

Interpretation of our findings regarding natural selection requires some care. A strength of our selection test is that it provides p-values, which are naturally calibrated, even if population sizes vary through time. In common with previous studies, we find a relatively small (<40) number of clear signals of strong, ongoing selection across multiple human populations. In contrast, we find a much larger collection of phenotypes where – based on published GWAS – there is evidence of an influence of directional selection. These traits include BMI, blood pressure, and white and red blood cell counts, and more generally, we see an enrichment of selection evidence at loci shown to associate with human phenotypes. These findings appear highly consistent with the polygenic nature of most human phenotypes - which are expected to impose very weak selection, but on a large collection of loci<sup>41</sup>. However, temporal changes in selection, overlapping genetic

influences across traits, and the possibility of compensatory evolution in response to other genetic changes or the environment, are among reasons complicating the assignment of selection signals to specific phenotypes (Supplementary Note).

Relate provides age estimates for mutations and other events, and these enable us to construct statistics to understand evolutionary history, including natural selection either on individual mutations or collections of mutations. We regard the selection statistics introduced here as initial approaches along a path towards a richer inference framework, including e.g. background selection, full selective sweeps, or balancing selection. Development of methods to better understand directional migration and ancient admixture is another direction for future work. As one example, our results suggest a large impact of ancient substructure specific to African populations, as has been previously hypothesized<sup>37,38</sup>. More generally, we hope that methods will be developed to perform statistical analyses on a set of trees generated either by Relate or other approaches. Other analyses might use estimated mutational ages obtained here directly (<https://zenodo.org/record/3234689>).

There are several natural extensions to Relate itself, e.g. allowing for increasing sample sizes. A recently developed method, tsinfer<sup>65</sup>, has impressive scaling with sample size and might readily extend to even millions of samples, while Relate currently only handles at most a few tens of thousands of samples genome-wide. While tsinfer currently only infers tree topologies (as part of a full ancestral recombination graph structure), and so cannot infer tree times or model demographic histories, it would be possible to use tsinfer-based tree topologies in our framework, allowing full tree-based inference for huge sample sizes. Incorporation of ancient DNA sequences is another important direction. Such samples may have substantially higher error rates or more missing data than modern-day individuals, potentially requiring an approach that “threads” (ancient) sequences through genealogies that are initially built using modern individuals<sup>2</sup>. This approach might also be useful for efficient statistical phasing and/or imputation of individuals only typed at a subset of markers.

## Acknowledgements

We thank Nick Barton, Daniel Falush, Molly Przeworski, Guy Sella, Jonathan Terhorst, Pier Palamara, Gerton Lunter, Jonathan Marchini, Sile Hu, Christopher B. Cole, Thaddeus Aid, Clare E. West for helpful comments,

ideas, and suggestions. L.S. acknowledges the support provided through the Engineering and Physical Sciences Research Council (EPSRC) [grant number EP/G03706X/1]. M.F. acknowledges the support provided through the Natural Sciences and Engineering Research Council of Canada (NSERC, PGS D) and the Clarendon Scholarship. S.R.M. acknowledges the support provided by the Wellcome Trust Investigator Award [grant number 098387/Z/12/Z and 212284/Z/18/Z]. Computation used the Oxford Biomedical Research Computing (BMRC) facility, a joint development between the Wellcome Centre for Human Genetics and the Big Data Institute supported by Health Data Research UK and the NIHR Oxford Biomedical Research Centre. Financial support was provided by the Wellcome Trust Core Award Grant Number 203141/Z/16/Z. The views expressed are those of the authors and not necessarily those of the NHS, the NIHR or the Department of Health.

## Author contributions

S.R.M. designed the study. L.S. and S.R.M. developed Relate with contributions by M.F. in the development of the algorithm for estimating coalescence rates. L.S. and S.R.M. performed the analysis, S.S. provided supplementary data and L.S. and S.R.M. wrote the manuscript.

## Competing Interests

S.R.M. is a director of GENSCI limited. The remaining authors declare no competing financial interests.

## Main text references

1. Griffiths, R. C. & Marjoram, P. Ancestral inference from samples of DNA sequences with recombination. *J. Comput. Biol.* **3**, 479–502 (1996).
2. Rasmussen, M. D., Hubisz, M. J., Gronau, I. & Siepel, A. Genome-Wide Inference of Ancestral Recombination Graphs. *PLoS Genet.* **10**, (2014).
3. Kingman, J. F. C. On the genealogy of large populations. *J. Appl. Probab.* **19**, 27–43 (1982).
4. Hudson, R. R. Properties of a neutral allele model with intragenic recombination. *Theor. Popul. Biol.* **23**, 183–201 (1983).
5. McVean, G. A. T. & Cardin, N. J. Approximating the coalescent with recombination. *Philos. Trans. R. Soc. London B Biol. Sci.* **360**, 1387–1393 (2005).
6. Hein, J. Reconstructing evolution of sequences subject to recombination using parsimony. *Math. Biosci.* **98**, 185–200 (1990).
7. Song, Y. S. & Hein, J. Constructing minimal ancestral recombination graphs. *J. Comput. Biol.* **12**, 147–169 (2005).
8. Kececioglu, J. & Gusfield, D. Reconstructing a history of recombinations from a set of sequences. *Discret. Appl. Math.* **88**, 239–260 (1998).

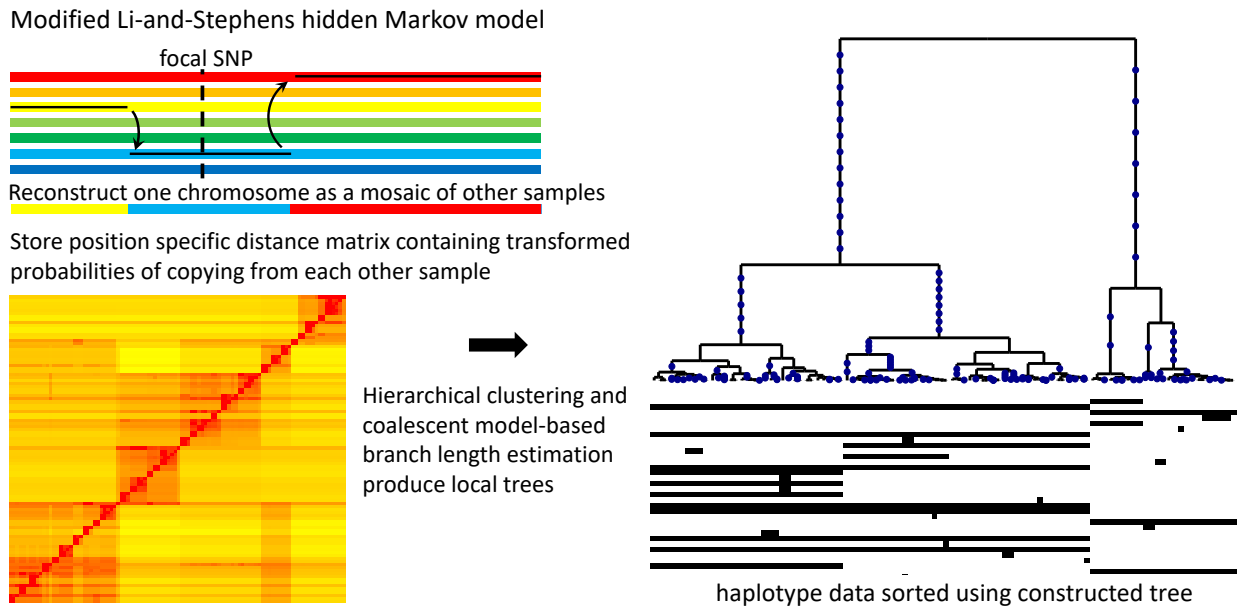
9. Wang, L., Zhang, K. & Zhang, L. Perfect phylogenetic networks with recombination. *J. Comput. Biol.* **8**, 69–78 (2001).
10. Wu, Y. New methods for inference of local tree topologies with recombinant SNP sequences in populations. *IEEE/ACM Trans. Comput. Biol. Bioinforma.* **8**, 182–193 (2011).
11. Mirzaei, S. & Wu, Y. RENT+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics* **33**, 1021–1030 (2017).
12. Menozzi, P., Piazza, A. & Cavalli-Sforza, L. Synthetic maps of human gene frequencies in Europeans. *Science* **201**, 786–792 (1978).
13. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
14. Novembre, J. *et al.* Genes mirror geography within Europe. *Nature* **456**, 98–101 (2008).
15. Li, H. & Durbin, R. Inference of human population history from individual whole-genome sequences. *Nature* **475**, 493–496 (2011).
16. Henderson, D., Zhu, S. (Joe) & Lunter, G. Demographic inference using particle filters for continuous Markov jump processes. *bioRxiv*: 382218 (2018).
17. Pritchard, J. K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
18. Falush, D., Stephens, M. & Pritchard, J. K. Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. *Genetics* **164**, 1567–1587 (2003).
19. Reich, D. D. E. *et al.* Linkage disequilibrium in the human genome. *Nature* **411**, 199–204 (2001).
20. Schiffels, S. & Durbin, R. Inferring human population size and separation history from multiple genome sequences. *Nat. Genet.* **46**, 919–925 (2014).
21. Terhorst, J., Kamm, J. A. & Song, Y. S. Robust and scalable inference of population history from hundreds of unphased whole genomes. *Nat. Genet.* **49**, 303–309 (2017).
22. Green, R. E. *et al.* A draft sequence of the Neandertal genome. *Science* **328**, 710–722 (2010).
23. Sudmant, P. H. *et al.* An integrated map of structural variation in 2,504 human genomes. *Nature* **526**, 75–81 (2015).
24. The 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
25. Harris, K. Evidence for recent, population-specific evolution of the human mutation rate. *Proc. Natl. Acad. Sci. U. S. A.* **112**, 3439–3444 (2015).
26. Voight, B. F., Kudaravalli, S., Wen, X. & Pritchard, J. K. A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
27. Li, N. & Stephens, M. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics* **165**, 2213–2233 (2003).
28. Kelleher, J., Etheridge, A. M. & McVean, G. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Comput. Biol.* **12**, e1004842 (2016).
29. Bae, C. J., Douka, K. & Petraglia, M. D. On the origin of modern humans: Asian perspectives. *Science* **358**, eaai9067 (2017).
30. Liu, X. & Fu, Y.-X. Exploring population size changes using SNP frequency spectra. *Nat. Genet.* **47**, 555–559 (2015).
31. Chheda, H. *et al.* Whole genome view of the consequences of a population bottleneck using 2926 genome sequences from Finland and United Kingdom. *Eur. J. Hum. Genet.* **25**, 477–484 (2017).
32. Duret, L. & Galtier, N. Biased Gene Conversion and the Evolution of Mammalian Genomic Landscapes. *Annu. Rev. Genomics Hum. Genet.* **10**, 285–311 (2009).
33. Meyer, M. *et al.* A high-coverage genome sequence from an archaic Denisovan individual. *Science* **338**, 222–226 (2012).

34. Prüfer, K. *et al.* The complete genome sequence of a Neanderthal from the Altai Mountains. *Nature* **505**, 43–49 (2014).
35. Sankararaman, S., Patterson, N., Li, H., Pääbo, S. & Reich, D. The date of interbreeding between Neandertals and modern humans. *PLoS Genet.* **8**, e1002947 (2012).
36. Fu, Q. *et al.* An early modern human from Romania with a recent Neanderthal ancestor. *Nature* **524**, 216–219 (2015).
37. Hammer, M. F., Woerner, A. E., Mendez, F. L., Watkins, J. C. & Wall, J. D. Genetic evidence for archaic admixture in Africa. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 15123–15128 (2011).
38. Ragsdale, A. P. & Gravel, S. Models of archaic admixture and recent history from two-locus statistics. *bioRxiv*: 489401 (2018).
39. Mathieson, I. *et al.* Genome-wide patterns of selection in 230 ancient Eurasians. *Nature* **528**, 499–503 (2015).
40. Edge, M. & Coop, G. Reconstructing the history of polygenic scores using coalescent trees. *bioRxiv*: 389221 (2018).
41. Simons, Y. B., Bullaughey, K., Hudson, R. R. & Sella, G. A population genetic interpretation of GWAS findings for human quantitative traits. *PLoS Biol.* **16**, e2002985 (2018).
42. Enattah, N. S. *et al.* Identification of a variant associated with adult-type hypolactasia. *Nat. Genet.* **30**, 233–237 (2002).
43. Hardouin, E. *et al.* Positive Selection in East Asians for an EDAR Allele that Enhances NF- $\kappa$ B Activation. *PLoS One* **3**, e2209 (2008).
44. Miretti, M. M. *et al.* A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 634–646 (2005).
45. Sadier, A., Viriot, L., Pantalacci, S. & Laudet, V. The ectodysplasin pathway: from diseases to adaptations. *Trends Genet.* **30**, 24–31 (2014).
46. Pritchard, J. K., Pickrell, J. K. & Coop, G. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr. Biol.* **20**, R208–R215 (2010).
47. Zhang, G., Muglia, L. J., Chakraborty, R., Akey, J. M. & Williams, S. M. Signatures of natural selection on genetic variants affecting complex human traits. *Appl. Transl. Genomics* **2**, 78–94 (2013).
48. Berg, J. J. & Coop, G. A population genetic signal of polygenic adaptation. *PLoS Genet.* **10**, e1004412 (2014).
49. Field, Y. *et al.* Detection of human adaptation during the past 2000 years. *Science* **354**, 760–764 (2016).
50. Sohail, M. *et al.* Signals of polygenic adaptation on height have been overestimated due to uncorrected population structure in genome-wide association studies. *bioRxiv*: 355057 (2018).
51. Berg, J. J. *et al.* Reduced signal for polygenic adaptation of height in UK Biobank. *bioRxiv*: 354951 (2018).
52. Maruyama, T. The age of an allele in a finite population. *Genet. Res.* **23**, 137 (1974).
53. Kiezun, A. *et al.* Deleterious Alleles in the Human Genome Are on Average Younger Than Neutral Alleles of the Same Frequency. *PLoS Genet.* **9**, e1003301 (2013).
54. Bycroft, C. *et al.* The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
55. Casto, A. M. & Feldman, M. W. Genome-Wide Association Study SNPs in the Human Genome Diversity Project Populations: Does selection affect unlinked SNPs with shared trait associations? *PLoS Genet.* **7**, e1001266 (2011).
56. Wilde, S. *et al.* Direct evidence for positive selection of skin, hair, and eye pigmentation in Europeans during the last 5,000 y. *Proc. Natl. Acad. Sci. U. S. A.* **111**, 4832–4837 (2014).
57. Bulik-Sullivan, B. K. *et al.* LD score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat. Genet.* **47**, 291–295 (2015).
58. Turchin, M. C. *et al.* Evidence of widespread selection on standing variation in Europe at height-associated SNPs. *Nat. Genet.* **44**, 1015–1019 (2012).

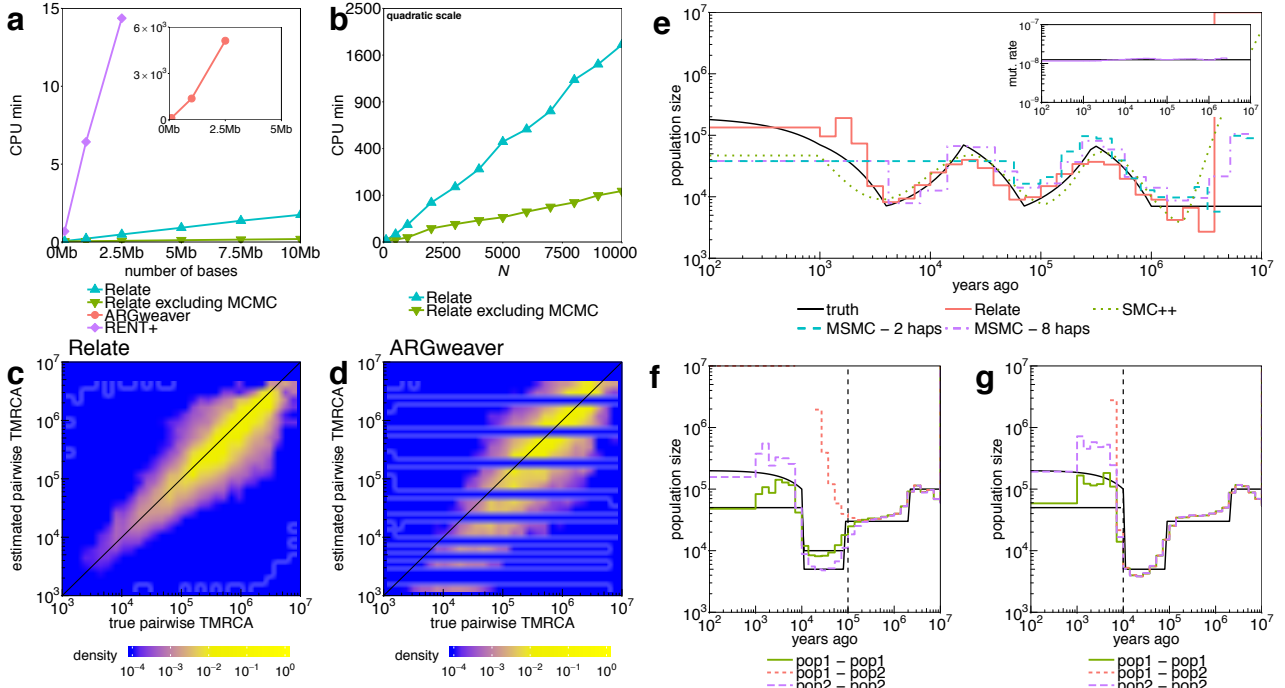
59. Robinson, M. R. *et al.* Population genetic differentiation of height and body mass index across Europe. *Nat. Genet.* **47**, 1357–1362 (2015).
60. Novick, D., Montgomery, W., Treuer, T., Moneta, M. V. & Haro, J. M. Sex differences in the course of schizophrenia across diverse regions of the world. *Neuropsychiatr. Dis. Treat.* **12**, 2927–2939 (2016).
61. Crespi, B., Summers, K. & Dorus, S. Adaptive evolution of genes underlying schizophrenia. *Proc. R. Soc. B Biol. Sci.* **274**, 2801–2810 (2007).
62. Young, J. H. *et al.* Differential susceptibility to hypertension is due to selection during the out-of-Africa expansion. *PLoS Genet.* **1**, e82 (2005).
63. Hinch, A. G. *et al.* The landscape of recombination in African Americans. *Nature* **476**, 170–5 (2011).
64. Fledel-Alon, A. *et al.* Variation in human recombination rates and its genetic determinants. *PLoS One* **6**, e20321 (2011).
65. Kelleher, J., Wong, Y., Albers, P., Wohns, A. W. & McVean, G. Inferring the ancestry of everyone. *bioRxiv*: 458067 (2018).



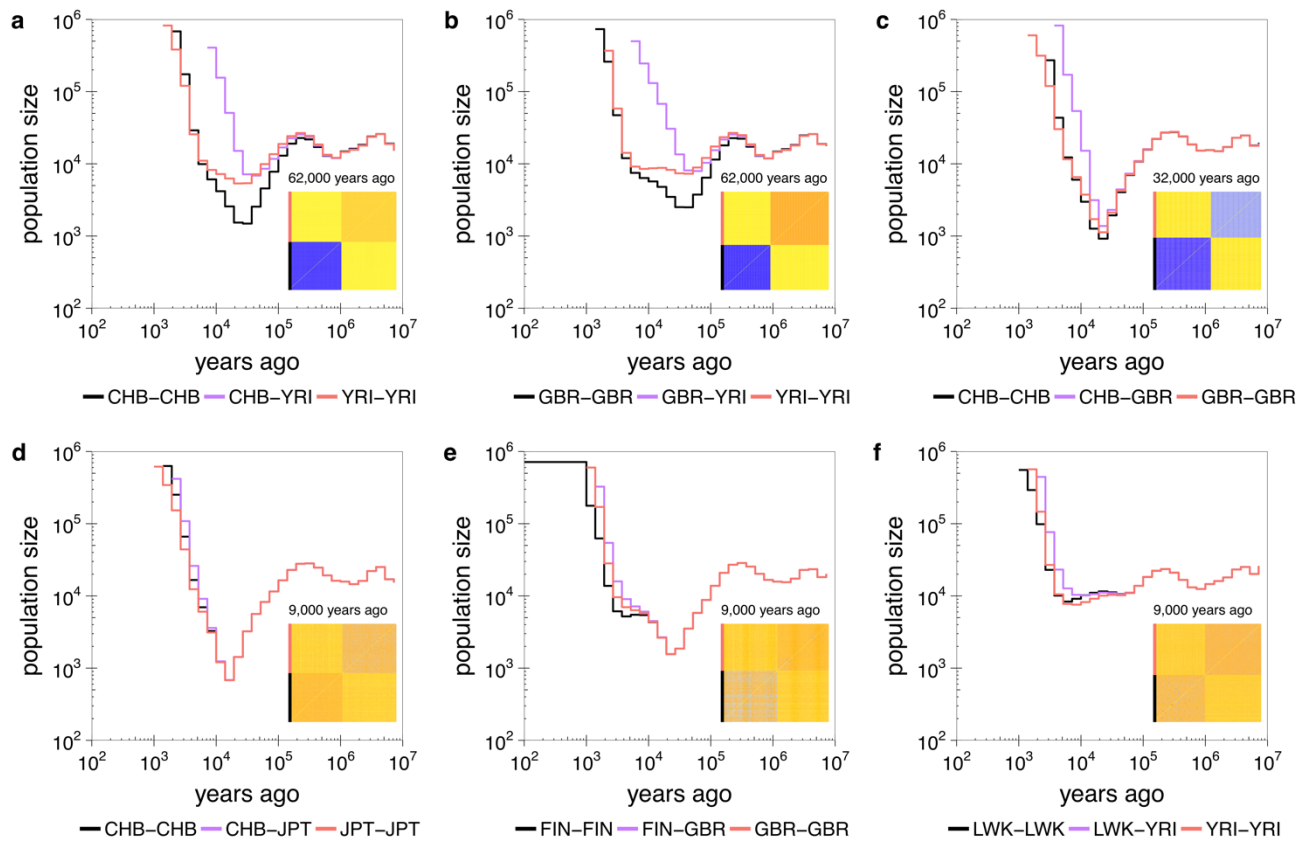
## Figures



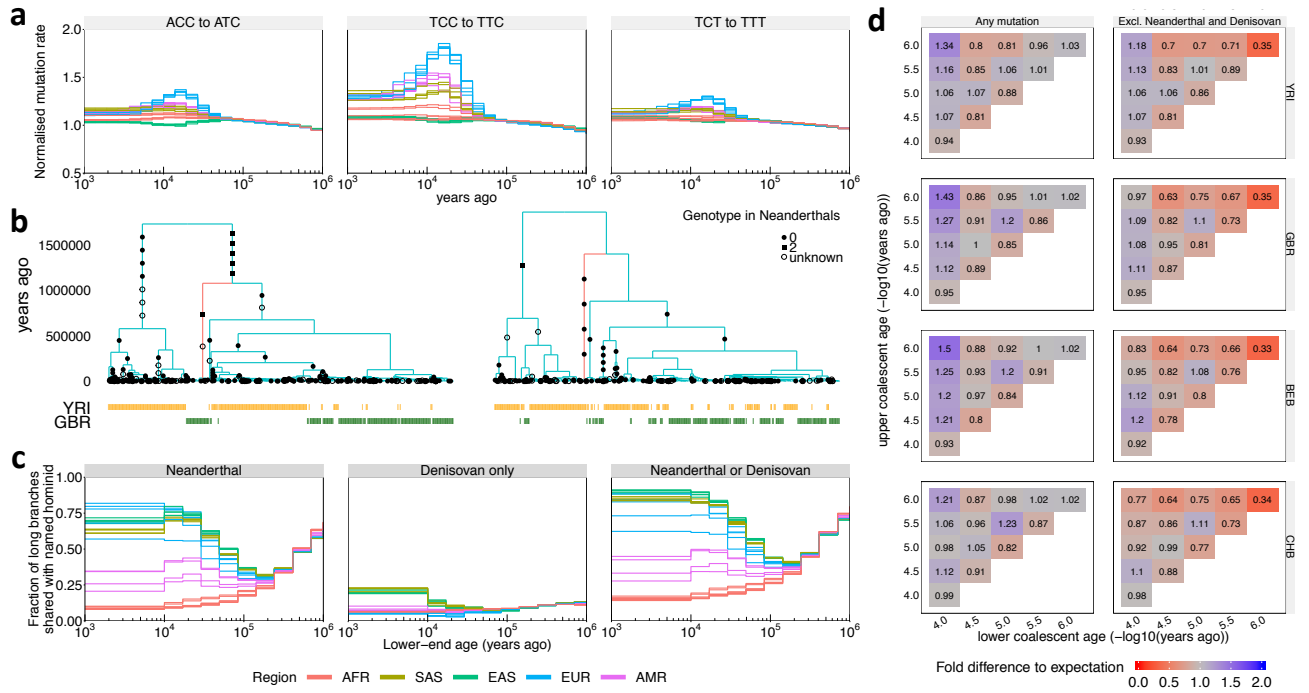
**Figure 1 Relate method overview:** Our method applies a version of the Li-and-Stephens hidden Markov model<sup>15</sup>, modified to take ancestral and derived states into account, to calculate at a focal SNP (dotted vertical line) a position-specific distance matrix  $d$  (bottom left). Each entry  $d_{ij}$  of this matrix stores the rescaled log-likelihood of generating haplotype  $i$  by copying from haplotype  $j$ , which can be interpreted as the number of mutations carried by  $i$ , but not by  $j$ , locally around the focal SNP. (**Methods**; Supplementary Note). Our tree builder uses the resulting inferred distance matrix to coalesce haplotypes (right-hand side). After mapping mutations to their corresponding branches, we estimate branch lengths using an MCMC algorithm that employs a coalescent prior model.



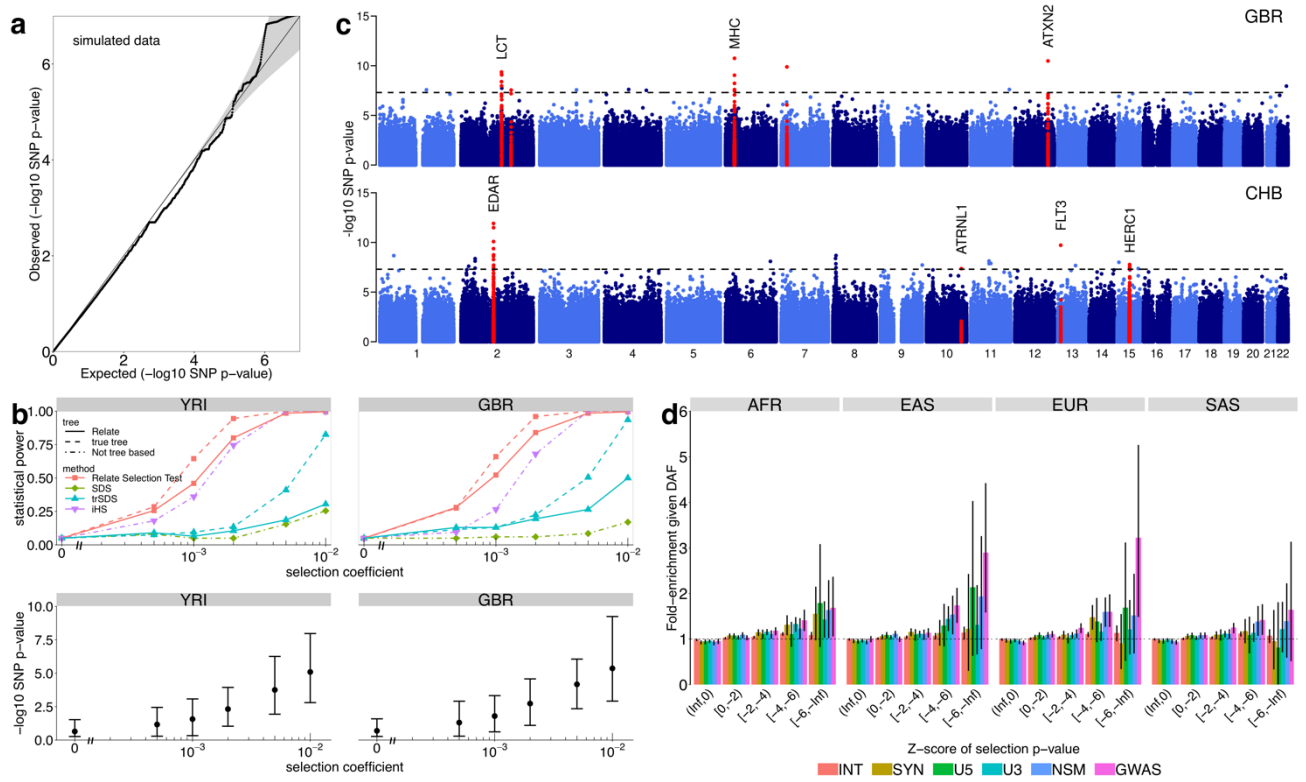
**Figure 2 Simulated data:** **a**, Runtimes of Relate, RENT+ and ARGweaver in CPU minutes as a function of the number of bases simulated with  $N = 200$ ,  $\mu = 1.25 \times 10^{-8}$ ,  $2N_e = 30,000$ , and recombination rates taken from human chromosome 1. We also show the runtime of Relate excluding the estimation of branch lengths. **b**, Runtime of Relate in minutes as a function of sample size  $N$ , where we simulate 2.5Mb for each data point. Other parameters are the same as in **a**; y-axis is on a quadratic scale. **c**, True TMRCA for pairs of haplotypes (x-axis) versus those estimated by Relate (y-axis). **d**, As **c**, except showing results for ARGweaver. **e**, Comparison of population size estimates across methods for a simulation of 200 haplotypes, 200Mb and an oscillating population size<sup>20</sup>. Relate and SMC++ estimates are based on 200 haplotypes, MSMC estimates are obtained from 2 or 8 haplotypes. Inset shows the mutation rate over time estimated by Relate. **f**, **g**, Population-specific estimates of population size and cross-population coalescence rates for a simulation with a discrete bottleneck for two populations that separated 80,000 YBP (**f**; vertical dashed line), or 10,000YBP (**g**). TMRCA, time to most recent common ancestor.



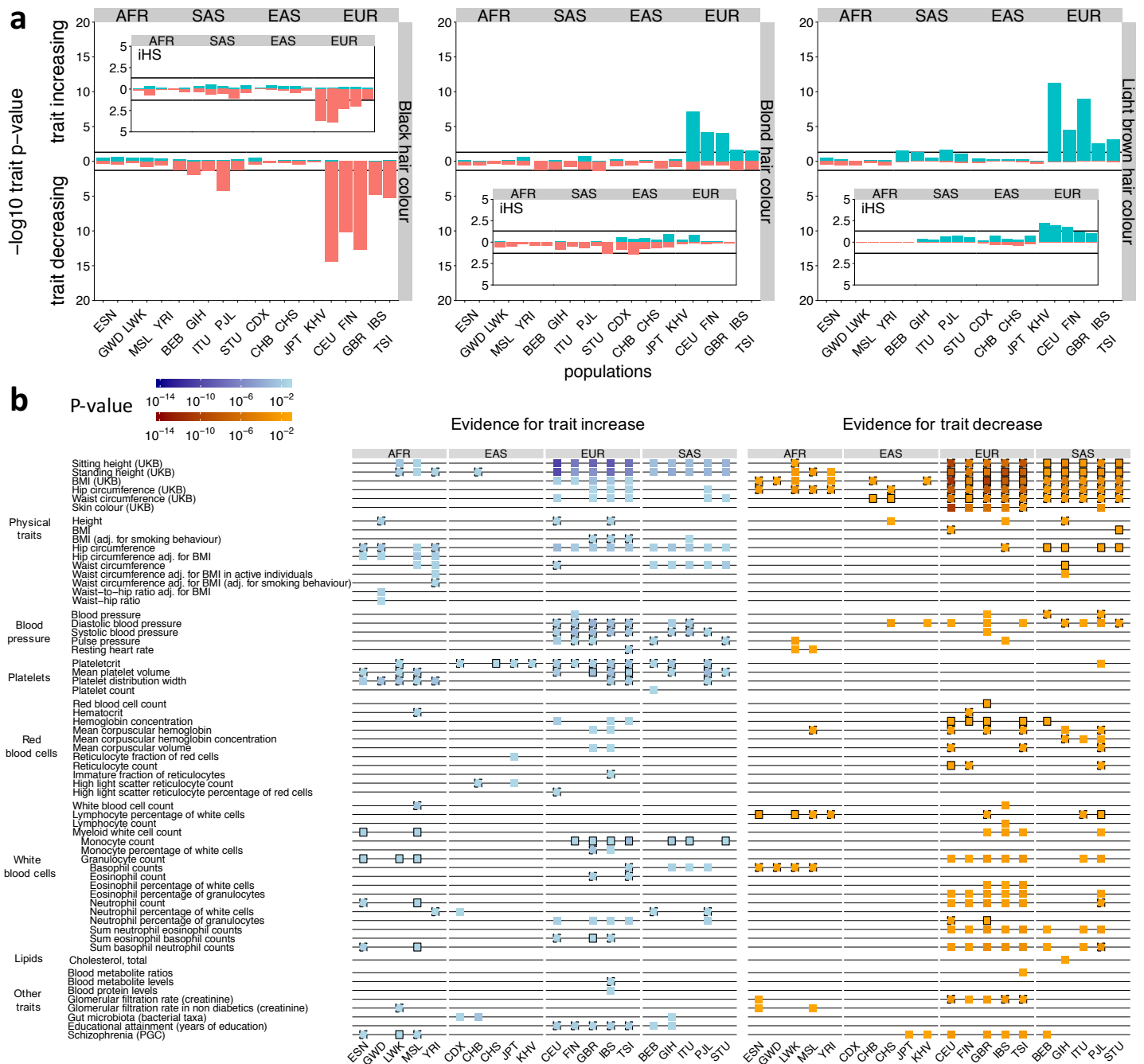
**Figure 3 Population sizes and split times in 1000 GP:** Relate-based population-specific estimates of population size and cross-population coalescence rates using genome-wide genealogies for CHB and YRI (a), GBR and YRI (b), CHB and GBR (c), CHB and JPT (d), FIN and GBR (e), and LWK and YRI (f). Insets show the matrices of coalescence rates between pairs of haplotypes at the indicated time. Rows and columns are sorted by population labels of haplotypes, as indicated by the colour on the left of each matrix.



**Figure 4 Evolution of human mutation rates and evidence for introgression:** **a**, Evolution of mutation rates for three triplet mutations ACC to ATC, TCC to TTC, and TCT to TTT (see Supplementary Fig. 7 for all 96 triplet mutations, **Methods** for normalization). **b**, Marginal trees for a subregion on chromosome 14 (left) and chromosome 11 (right). The tree on the left contains a long branch with descendants only in GBR (red) consistent with Neanderthal introgression into GBR. The tree on the right contains a long branch with descendants only in YRI (red) consistent with introgression in YRI involving a hominid not closely related to Neanderthals. **c**, Fraction of branches with an upper-end age >1M YBP that are shared with Neanderthals (left), Denisovans and not Neanderthals (center), or Neanderthals or Denisovans (right) (**Methods**). In **a** and **c**, colours encode geographic regions (AFR: Africa, EAS: East Asia, EUR: Europe, SAS: South Asia, AMR: Americas). **d**, Number of mutations binned by age of upper and lower coalescent event, relative to the expected number of mutations when randomising topology while fixing ages of coalescence events (**Methods**). Right column shows mutations not present in Neanderthal or Denisovan samples.



**Figure 5 Natural selection:** **a**, QQ-plot of p-values  $p_R$  for selection evidence of SNPs. We simulated 250Mb for  $N = 1000$  haplotypes using the recombination map of chromosome 1 and a bottleneck population size resembling that of non-African populations. **b**, Power simulations using  $N = 1000$  haplotypes. We use historical population sizes estimated by Relate for YRI (left) and GBR (right). Top row shows statistical power and bottom row shows  $p_R$  with the mean indicated by circles and the 5<sup>th</sup> and 95<sup>th</sup> percentiles indicated by the error bars. We performed 500 simulations for the neutral case and 200 simulations for  $s > 0$ . **c**, Manhattan plot showing  $p_R$  of SNPs, for GBR and CHB. We highlight regions containing a SNP with  $p_R < 5 \times 10^{-8}$  in at least three populations (see Supplementary Table 3 for a full list), as well as the MHC region in GBR. **d**, Mean enrichment of functional annotation among targets of selection, conditional on allele frequency. Error bars show 95% confidence intervals estimated from 1000 iterations of a block bootstrap resampling (**Methods**). We group SNPs by mean regional Z-score corresponding to the log p-value for selection evidence, where a smaller Z-score indicates stronger selection evidence. SNPs are binned by partially overlapping functional annotations: intronic mutations (INT), synonymous mutations (SYN), mutations at the 5' end and 3' end of a gene (U5, U3), non-synonymous mutations (NSM), and GWAS hits (GWAS).



**Figure 6 Evidence of selection on traits: a**, P-values for evidence of directional selection of black, blond, and light brown hair color (see **Methods** for calculation of p-values). Insets show p-values for the same test but using iHS scores instead, where iHS scores are calculated for each population separately for any variant with a minor allele frequency >5% in that population. **b**, Evidence for directional or bidirectional selection on multi-allelic traits. Each trait is associated with at least 10 SNPs in both effect directions in each of the considered populations. We show evidence for a trait increasing over time (left) and evidence for a trait decreasing over time (right) if  $p \leq 0.05$ . Black boundaries indicate consistency with an additional test that tests for shifts in the DAFs (solid:  $p \leq 0.05$ , dashed:  $p \leq 0.5$ , **Methods**).

## Online Methods

### Relate overview

We estimate genealogies as a sequence of rooted binary trees, where each tree captures the genealogy for a subregion of the genome. This representation serves as an approximation of an Ancestral Recombination Graph (ARG)<sup>4</sup>. We estimate local ancestry without global constraints on tree topology, thereby transforming genealogy reconstruction into a feasible and highly parallelizable problem.

Our approach can be divided roughly into three steps, which we detail below (also see **Figure 1**, Supplementary Figure 1, and Supplementary Note).

### Calculating position specific distance matrices

While trees vary along the genome, our method heavily utilizes ancestry information from nearby SNPs to reconstruct the tree at a specific position. We achieve this by using a HMM similar to that first proposed by Li and Stephens<sup>27</sup> (see Supplementary Figure 2 for parameter choices). Intuitively, this HMM reconstructs a haplotype as a mosaic of other sample haplotypes along the genome (Supplementary Figure 1), allowing for mismatching in the copying process, and viewing changes in haplotype as recombination events. After applying the HMM, at a focal SNP  $\ell$  each of the other haplotypes  $j$  therefore has some probability  $p_{ij\ell}$  of being copied from, to generate haplotype  $i$ . After rescaling  $\log p_{ij\ell}$  appropriately (Supplementary Note), we obtain a position-specific distance matrix  $d$  whose entry  $(i, j)$  converges to the number of mutations derived in  $i$  and ancestral in  $j$  in the limit of no recombinations. In the presence of recombination, this  $d$  can be interpreted to store a local number of derived mutations, where more closely related haplotypes tend to have fewer mismatches over longer stretches, therefore receiving a smaller distance in this matrix.

We modified the Li-and-Stephens HMM to account for ancestral and derived states, a modification that guarantees our approach will construct the correct tree topology under the infinite-sites assumption with no recombination, while simultaneously speeding up the calculation of posterior copying probabilities.

## Tree builder

The distance matrix is turned into a binary tree using a hierarchical clustering algorithm. This hierarchical clustering algorithm is motivated by the observation that each row of the distance matrix should indicate the order in which this haplotype coalesced with other haplotypes of the dataset. This can be shown mathematically in some limit conditions, such as the case with no recombination (Supplementary Note).

Our algorithm iteratively merges clades of haplotypes, corresponding to past coalescences. After merging clades, we update the distance matrix by combining the corresponding rows and columns using a weighted sum, with weights determined by the size of clades. In each step of the algorithm, we merge the pair of clades that coalesce with each other before coalescing with any other clade, as determined using rows of the distance matrix. If multiple pairs of clades satisfy this condition, we choose the pair with minimum symmetrized score in the distance matrix. If the data are consistent with a binary tree under the infinite-sites model, such a pair always exists. In practise, errors in the data, complex recombination histories, or violations of assumptions made by our model, may result in a distance matrix that is inconsistent with a binary tree. To be robust to such cases, we relax the conditions for identifying pairs of clades to coalesce.

## Mapping mutations to branches and estimating branch lengths

Once tree topology is estimated as above, where possible we map mutations to the (unique) branch that has the identical descendants as the carriers of the derived allele in the data. To be robust to errors, where necessary we use an approximate rule for such mapping; however some mutations, e.g. repeat mutations or error-prone loci, may still not map to a unique branch. For these loci, we determine the smallest collection of branches, such that the data can be fully recovered. If a mutation maps to the tree only after reinterpreting the derived allele as the ancestral allele (and vice versa), we “flip” ancestral and derived alleles at this locus. For computation efficiency, to avoid having to construct a new tree at every locus we construct trees starting at the 5' end of a region or chromosome, and move along the region constructing a new tree only when a SNP is flipped or cannot be mapped to a unique branch. Finally, after identifying equivalent branches in adjacent trees along the genome, we apply a Metropolis-Hastings type Markov Chain Monte Carlo (MCMC) algorithm



to estimate branch lengths. The MCMC algorithm has a coalescent prior assuming a single panmictic population<sup>3</sup>.

## Estimating coalescence rates through time

We estimate the effective population size, defined as the inverse of the coalescence rate, by applying the following iterative algorithm. We initially estimate branch lengths using a constant effective population size. We then calculate a maximum-likelihood estimate of the coalescence rates between pairs of haplotypes given the branch lengths (Supplementary Note). By averaging coalescence rates over all pairs of haplotypes and taking the inverse, we obtain a population-wide estimate of the effective population size. We then use this population size estimate to re-estimate branch lengths, which requires only the final MCMC step of the branch-length estimation. By repeating these two steps until convergence (in practice, we use only 5 iterations as this provides good performance), we obtain a self-contained algorithm for jointly estimating branch lengths and the effective population size. We can average pairwise coalescence rates in different ways to obtain rates for sub-populations and cross-coalescence rates between populations.

## Estimating relative mutation rates through time

We estimate the mutation rate through time for all 96 triplet mutations (**Figure 4a**, Supplementary Figure 6). To estimate mutation rates for a mutation category of interest, we calculate, for each epoch, the quotient of the number of mutations in that category by the total branch length over bases at which such a mutation may have occurred. In our model, we fix the average mutation rate to a constant value through time, such that any change in average mutation rate should in theory be absorbed in our population size estimate. We therefore first eliminate any remaining temporal trends in the average mutation rate by dividing by the average mutation rate in each epoch. For each population, we then normalise the mutation rates such that the average rate over time equals 1. In simulations (Supplementary Figure 4), we show that variable mutation rates among categories can be detected by this approach, and approximately dated.

## Pre-processing of the 1000 Genomes Project dataset

The 1000 Genomes Project dataset comprises 2504 individuals, from 26 populations. We obtained a phased version of the dataset (see **Data availability**). We next excluded multi-allelic SNPs, and we exclude one individual (two haplotypes) from each population for future applications, and analyzed the remaining 2,478 individuals (Supplementary Table 2). We use a genomic mask provided with the 1000 Genomes Project dataset (see **Data availability**) to exclude regions in the marked as "not passing" in the pilot mask, to remove loci with low certainty of genotypes. We also exclude any base for which the fraction of "not passing" bases within 1,000 bases to either side exceeds 0.9. To account for this filtering, we readjust the number of bases between SNPs at which we could have potentially observed a SNP. We use an estimate of the human ancestral genome (see **Data availability**) to identify the most likely ancestral allele for each SNP.

## Identifying branches indicative of Neanderthal and Denisovan introgression

We use genome sequences of the Vindija<sup>22</sup> and Altai<sup>34</sup> Neanderthals (NEA), and a Denisovan (DEN)<sup>33</sup> to identify branches indicative of Neanderthal and Denisovan introgression into non-African populations. To identify branches that remain segregated from other human lineages for a long time, we use the world-wide genealogy of 2,487 samples. To identify whether a branch is shared with NEA or DEN, at least one mutation needs to be mapped to that branch. We therefore exclude any mutation that has not been genotyped (or does not pass the genomic masks) in these ancient genomes. We further restrict our analysis to branches with at least two mutations mapped to them, as well as having an upper end that is older than 1M YBP. Of any such branches, we calculate the fraction of branches with at least one NEA or DEN mutation. In **Figure 4c**, we plot these fractions as functions of the lower-end age of the branch. Because the same branch may persist over multiple trees, we identify equivalent branches (Supplementary Note) and average ages of lower and upper ends across these equivalent branches. We assign a branch to a population if at least one descendant of that branch is in the population.

In **Figure 4d**, we observe an enrichment of branches indicative of introgression. This enrichment is identified by comparing the observed number of mutations in bins divided by upper and lower coalescence age to that

expected in a panmictic history. To calculate the expected number of mutations in each bin, we fix the ages of coalescence events in each tree but randomise the topology assuming a panmictic population. The probability of upper and lower coalescence ages falling into bins  $s$  and  $r$ , conditional on the mutation arising while  $k$  lineages remain, is given by  $P(r, s|k) = \sum_{\ell \geq k, h < k} I_{t_\ell \in s} I_{t_h \in r} \frac{2h}{\ell(\ell-1)} \frac{1}{k}$ , where  $I$  denotes the indicator function. Assuming neutrality, a mutation is equally likely to have arisen anywhere on the branch it maps to. We therefore calculate the weighted average  $\sum_{k=2}^N w_k P(r, s|k)$ , with weights  $w_k$  defined as the proportions of a branch while  $k$  lineages remain. Summing this over all SNPs yields the expected number of mutations with upper and lower coalescence age falling, respectively, into bins  $s$  and  $r$ . In **Figure 4d**, log10 age bins are defined by  $[-\infty, 4.25), [4.25, 4.75), [4.75, 5.25), [5.25, 5.75), [5.75, \infty)$ .

## Tree-based statistic for detecting positive selection

Positive selection is expected to result in favourable mutations spreading rapidly in a population. One approach to capture this is via the number of lineages ultimately descending from the potentially favourable mutation(s): although we note that this is not the maximum likelihood approach, it has the benefit of making calculations straightforward. Under a null model of the standard coalescent model without selection, it is known that while  $k$  lineages remain, the joint distribution of the number of descendants of these  $k$  lineages is uniform in the partitions of  $N$  haplotypes to  $k$  lineages (see e.g., Ref. [66]). Using this property, we analytically calculate the marginal distribution that two of  $k$  lineages have more than  $f_N$  descendants, where  $f_N$  is the present-day DAF of the mutation. Here, we choose  $k$  to be the number of lineages remaining when the mutation of interest increased from frequency 1 to 2 (see Supplementary Note for the mathematical details).

To remove false-positive selection hits due to poorly inferred genealogies, our analysis for the 1000 Genomes Project dataset is based on a subset of all SNPs mapping to trees, and present in 3 or more copies in the dataset. Specifically, we remove SNPs failing any of the following filters: (i) the number of mutations mapping to that SNP's tree is in the bottom 5<sup>th</sup> percentile, or (ii) the fraction of tree branches having at least one SNP is in the bottom 5<sup>th</sup> percentile. This excludes approximately 7% of SNPs.

## Simulation of positive selection

To simulate positive natural selection, we adopt the pipeline outlined in Ref. [49]. We first simulate the trajectory of the DAF using simuPOP<sup>67</sup>. We vary the selection coefficient between  $s = 0.001$  and  $s = 0.05$  and assume that the selected allele is beneficial throughout its history. We fix the present-day DAF to 0.7 (see Supplementary Figure 7 for other present-day DAFs). We then use mbs2<sup>68</sup> (mutation rate  $\mu = 1.25 \times 10^{-8}$ , constant recombination rate  $\rho = 5 \times 10^{-9}$ ) to simulate a region of 20 Mb, given the DAF trajectory for the central selected SNP. For each non-zero selection coefficient, we perform 200 simulations, and we perform 500 simulations for the neutral case. We assume a population size history as for our estimates for YRI and GBR, in separate simulations.

We compare to iHS, SDS, and a tree-based variant of SDS (trSDS) proposed in Ref. [40]. For iHS, SDS, and trSDS, we standardise scores using the mean and standard deviation in the neutral case, which is an idealised setting that should favour the power estimates of these methods. We then determine a critical standardized score that corresponds to a given type I error rate in the neutral case to estimate the statistical power. For Relate, we use frequency-conditioned p-values, by calculating a critical p-value that yields the desired false-positive rate in the neutral case (for the statistical power using raw p-values, see Supplementary Figure 7).

## Enrichment of SNPs with functional annotation among targets of positive selection

We merge selection evidence for SNPs by region (AFR: Africans, EAS: East Asians, EUR: Europeans, SAS: South Asians) by first calculating Z-scores of the logarithm of selection p-values within populations, and then averaging these Z-scores across populations. We exclude groups expected to be highly admixed<sup>69</sup> (ACB, ASW, CLM, MXL, PEL, PUR (Supplementary Table 2)), because recent admixture may confound selection signals. We further exclude SNPs with a DAF <5% in the region of interest.

To assess statistical significance for the observed enrichment of GWAS hits and functional mutations in groups of SNPs showing evidence of selection, we used a block bootstrap with a block size of 1 Mb. This will account for LD at scales below this threshold. In each bootstrap iteration, we resample blocks containing SNPs with a selection Z-score within the range of interest, with replacement, and calculate the ratio of the number of SNPs

with functional annotation obtained using the HaploReg database<sup>70</sup> (see **Data availability**) and the GWAS catalogue to the expected number of such SNPs, conditional on DAF. We condition on frequency, to account for the possibility that skewed frequency spectra in functional SNPs could be driving the signal.

## Pre-processing of GWAS

We use SNP-trait associations documented in the GWAS catalogue<sup>71</sup> (see **Data availability**) to study polygenic adaptation. We use only association signals whose GWAS p-value is smaller than  $5 \times 10^{-8}$ . For each trait, we also remove any duplicate SNPs.

For every combination of population, trait, and effect direction, we compile a set of approximately independent GWAS signals as follows.

For each pair of population and trait, we remove associations that are in close physical proximity and may therefore be in LD. For this, we first group SNPs into approximately independent blocks, such that any two GWAS hits in separate blocks are separated by at least 100 kb and there are no intervals larger than 100 kb with no GWAS hit inside a block. We then choose one GWAS hit from each block uniformly at random. We remove any SNP with a DAF <5%. To determine the effect direction of a SNP, we use the annotation in column “95% CI (TEXT)” combined with the indicated risk allele. We then realign the effect direction to the derived allele. We only consider SNPs for which an effect direction can be determined with this procedure. As described in the main text, we only analyze traits with at least 10 independent hits in both effect directions in all populations. This results in 76 traits and a total of 7,302 GWAS hits (before filtering for SNPs in close proximity in each population).

For Schizophrenia, we are unable to obtain an effect direction using the procedure described above. Instead, we downloaded results for a large-scale GWAS conducted by the Psychiatric Genomics Consortium<sup>72</sup>. We considered SNPs reaching a GWAS p-value of  $5 \times 10^{-8}$  of which there were 9,138. We intersected this set of SNPs with SNPs segregating in each of the considered populations. As for the GWAS catalogue, we identified approximately independent blocks. We then chose the SNP with lowest GWAS p-value in each block, resulting in 81 to 89 hits per population.

In addition, we use GWAS conducted as part of the UK Biobank<sup>54</sup>, focussing on highly polygenic physical traits. Our pre-processing protocol is analogous to that for schizophrenia detailed above. The number of approximately independent hits per population range from 272 hits for waist circumference to 989 hits for standing height.

## Trait selection test

For every combination of population, trait, and effect direction, we test whether p-values are smaller than expected. For this test, we first sample SNPs that we use for comparison. For each SNP associated with the population, trait, and effect direction tuple of interest, we sample 20 SNPs uniformly at random with replacement from SNPs, with the same present-day DAF in the population of interest. We then use a one-sided Wilcoxon rank-sum test to test whether the p-values of SNPs associated with the tuple of interest tend to be smaller than those for the frequency-matched set of SNPs. We repeat this test 20 times and report the mean p-value of the Wilcoxon rank-sum test.

Our primary test identifies selection evidence conditional on DAF. However, shifts in DAF can themselves serve as orthogonal evidence of polygenic adaptation, complementing our inferences. Therefore, we conducted a one-sided Wilcoxon rank-sum test to test whether DAFs of SNPs associated with the effect direction with selection evidence tend to exceed those associated with the opposing effect direction, and compared to our results conditional on SNP frequency. We note that we expect to lack power to reliably detect selection with this test, given that there are typically only tens of SNPs independently associating with each trait. In addition, the relationship between selection and SNP frequencies can be complex if selection strength varies through time and/or geographic locations.

## Data availability

Relevant-estimated coalescence rates, allele ages, and selection p-values for the 1000 Genomes Project can be downloaded from <https://zenodo.org/record/3234689>.

**Datasets used in the current study were obtained from the following URLs:**

1000 Genomes Project phased dataset, [https://mathgen.stats.ox.ac.uk/impute/1000GP\\_Phase3.html](https://mathgen.stats.ox.ac.uk/impute/1000GP_Phase3.html) (13 Jan 2017); Genomic mask, [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible\\_genome\\_masks/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/supporting/accessible_genome_masks/) (20

Jul 2017); Human ancestral genome, [ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis\\_results/supporting/ancestral\\_alignments/](ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/) (20 Jul 2017); GWAS catalogue, <https://www.ebi.ac.uk/gwas/api/search/downloads/full> (9 Nov 2017); PGC GWAS study, <https://www.med.unc.edu/pgc/results-and-downloads> (23 Nov 2018); HaploReg, [http://archive.broadinstitute.org/mammals/haploreg/data/haploreg\\_v4.0\\_20151021.vcf.gz](http://archive.broadinstitute.org/mammals/haploreg/data/haploreg_v4.0_20151021.vcf.gz) (21 Oct 2017); GTEx eQTL [https://storage.googleapis.com/gtex\\_analysis\\_v7/single\\_tissue\\_eqtl\\_data/GTEx\\_Analysis\\_v7\\_eQTL.tar.gz](https://storage.googleapis.com/gtex_analysis_v7/single_tissue_eqtl_data/GTEx_Analysis_v7_eQTL.tar.gz) (13 Jan 2019); UK Biobank GWAS summary statistics, <http://www.nealelab.is/uk-biobank> (4 Oct 2018); PopHumanScan, <https://pophumanscan.uab.cat> (13 Jan 2019)

## Code availability

The software Relate can be downloaded from <https://myersgroup.github.io/relate> under an Academic Use Licence.

### External software used in the current study were downloaded from the following URLs:

ARGweaver, <https://github.com/mdrasmus/argweaver> (24 Jan 2017); RENT+, <https://github.com/SajadMirzaei/RentPlus> (2 Oct 2017); msprime, <https://github.com/tskit-dev/msprime> (22 Jul 2017); msmc, <https://github.com/stschiff/msmc2> (14 Oct 2017); SMC++, <https://github.com/popgenmethods/smcpp> (14 Oct 2017); simuPOP, <http://simupop.sourceforge.net/> (27 Jun 2018); mbs, <http://www.sendou.soken.ac.jp/esb/innan/InnanLab/> (27 Jun 2018); SDS, <https://github.com/yairf/SDS> (27 Jun 2018), selscan, <https://github.com/szpiech/selscan> (31 Jul 2018); hapbin, <https://github.com/evotools/hapbin> (11 Dec 2018)

## Online methods references

66. Griffiths, R. C. & Tavaré, S. The age of a mutation in a general coalescent tree. *Stoch. Model.* **14**, 273–295 (1998).
67. Peng, B. & Kimmel, M. simuPOP: a forward-time population genetics simulation environment. *Bioinformatics* **21**, 3686–3687 (2005).
68. Teshima, K. M. & Innan, H. mbs: modifying Hudson’s ms software to generate samples of DNA sequences with a biallelic site under selection. *BMC Bioinformatics* **10**, 166 (2009).
69. Ruiz-Linares, A. *et al.* Admixture in Latin America: Geographic Structure, Phenotypic Diversity and Self-Perception of Ancestry Based on 7,342 Individuals. *PLoS Genet.* **10**, e1004572 (2014).
70. Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res.* **40**, D930–D934 (2011).
71. MacArthur, J. *et al.* The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Res.* **45**, D896–D901 (2016).
72. Ruderfer, D. M. *et al.* Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell* **173**, 1705–1715.e16 (2018).

Supplementary Information:  
A method for genome-wide genealogy estimation  
for thousands of samples

Leo Speidel<sup>1</sup>, Marie Forest<sup>2</sup>, Sinan Shi<sup>1</sup>, and Simon R. Myers<sup>1,3</sup>

<sup>1</sup>Department of Statistics, University of Oxford, Oxford, UK

<sup>2</sup>Université du Québec à Montréal, Montréal, Canada

<sup>3</sup>Wellcome Centre for Human Genetics, University of Oxford, Oxford, UK

July 16, 2019



# Contents

<b>Supplementary Note:</b>	
<b>Method details</b>	<b>2</b>
1 Overview of Relate	2
2 Tree builder	3
3 Calculating distance matrices	6
4 Estimating branch lengths	11
5 Estimating coalescence and mutation rates through time	14
6 A tree-based statistic for detecting positive selection	16
<b>Supplementary Note:</b>	
<b>Simulations</b>	<b>18</b>
1 Accuracy of TMRCA and mutation ages	18
2 Accuracy measured using the Robinson-Foulds metric and a TMRCA metric	19
3 Estimating coalescence and mutation rates	19
4 Simulations with perturbations from infinite-sites, constant mutation rates, or perfect phase	21
5 Positive natural selection	22
<b>Supplementary Note:</b>	
<b>1000 Genomes Project data set</b>	<b>24</b>
1 Runtime	24
2 Number of trees built	24
3 CpG mutations map less frequently than other mutations	24
4 Historical population sizes of 26 populations	25
5 Evolution of triplet mutation rates	25
6 Positive selection	25
<b>Supplementary Note:</b>	
<b>Interpretation of polygenic selection</b>	<b>27</b>

# Supplementary Note:

## Method details

### 1 Overview of Relate

We first describe Relate in the absence of recombination. In this case, a single tree describes the genealogy of the whole genome. We define the number of *derived* mutations  $d(i, j)$  as the number of mutations carried by haplotype  $i$  and not by haplotype  $j$ . Notice that  $d(i, j) \neq d(j, i)$ . Assuming that every mutation happened exactly once, we can determine the order in which haplotype  $i$  coalesced with other haplotypes by ordering them in ascending order of derived mutations  $d(i, j)$  ( $j = 1, \dots, i-1, i+1, \dots, N$ ). Once we know the relative order of coalescences, we reconstruct the tree topology using the hierarchical clustering algorithm described in Section 2. The resulting tree topology is guaranteed to be consistent with the truth in the sense that the constructed tree is a subtree of the gene tree describing the data (see Section 2.3).

In the presence of recombination, the relative order of coalescences changes along the genome. We apply a modified version of a Li-and-Stephens type hidden Markov model (HMM) 1 to calculate a local number of derived mutations  $d(i, j; \ell)$  at every SNP  $\ell$  (see Section 3). We modified the Li-and-Stephens model to take ancestral and derived states into account, which is necessary for  $d(i, j; \ell)$  to converge to  $d(i, j)$  in the limit of no recombination. We use  $d(i, j; \ell)$  to reorder haplotypes at every SNP and apply the tree building algorithm to reestimate tree topology. Our method builds trees that are consistent with the truth if  $d(i, j; \ell)$  orders haplotypes correctly. This is guaranteed for a recombination map consisting of zero and infinite recombination rates, which can be seen as a limit case of a hotspot recombination map (Section 2.3).

It is computationally inefficient, but possible, to reestimate tree topology at every SNP, because trees are unchanged if no recombination event occurred between SNPs. Instead, Relate initially estimates the tree topology at the first SNP of the 5' end of a chromosome. It then only reestimates the tree topology if a mutation cannot be uniquely *mapped* to a branch of the tree describing the previous SNP or is potentially *flipped*. A mutation is mapped to the branch for which the descendants coincide with the carriers of the alternative allele. Such a branch exists as long as the mutation occurred exactly once in human history and the estimated tree topology is correct. To be robust to errors in the data and the inferred tree, we relax this requirement such that the descendants of the branch only have to approximately coincide with the carriers of the mutation (see Section 2.2 for details). A mutation is potentially flipped, if it maps to a branch only after reinterpreting non-carriers as carriers and vice versa. While this introduces a small bias in the placement of recombination points, we note that this bias does not propagate along the genome because marginal trees are constructed using the distance matrix for the genomic position at which tree topology is reestimated.

Finally, once tree topologies are estimated, we estimate the branch lengths of each tree using an MCMC approach with a coalescent prior (see Section 4). We developed an algorithm that jointly estimates coalescence rates and branch lengths if historical coalescence rates are unknown (see Section 5).

## 2 Tree builder

### 2.1 Hierarchical Clustering

We begin by describing how we construct a tree using a distance matrix  $d = (d(i, j))_{1 \leq i, j \leq N}$  as input. We note that in general  $d$  does not provide a distance metric, and  $d(i, j) \neq d(j, i)$ . A schematic is depicted in Supplementary Fig. 1a.

The tree builder is initialised by placing each haplotype in a separate cluster. The algorithm proceeds by finding pairs of clusters that coalesce with each other before coalescing with any other haplotype. Such a pair satisfies

$$\begin{aligned}\mathcal{A} &= \arg \min_{\mathcal{A}'} d(\mathcal{B}, \mathcal{A}') \\ \mathcal{B} &= \arg \min_{\mathcal{B}'} d(\mathcal{A}, \mathcal{B}'),\end{aligned}\tag{1}$$

where the distance  $d(\mathcal{A}', \mathcal{B}')$  between two clusters  $\mathcal{A}'$  and  $\mathcal{B}'$  of cardinality  $|\mathcal{A}'|$  and  $|\mathcal{B}'|$ , respectively, is given by

$$d(\mathcal{A}', \mathcal{B}') = \frac{1}{|\mathcal{A}'||\mathcal{B}'|} \sum_{x \in \mathcal{A}'} \sum_{y \in \mathcal{B}'} d(x, y).\tag{2}$$

There might be more than one pair satisfying Eq. (1), in which case we choose the pair with the smallest symmetrised distance  $d(\mathcal{A}, \mathcal{B}) + d(\mathcal{B}, \mathcal{A})$ . The chosen pair is then combined to a new cluster comprising all haplotypes of both clusters. The algorithm is terminated when all haplotypes are in one cluster.

A pair satisfying Eq. (1) is guaranteed to exist as long as there exists a tree consistent with the order of coalescence events implied by matrix  $d$ . Sometimes such a tree cannot be constructed. To make our algorithm robust to such situations, we replace Eq. (1) by

$$\begin{aligned}\mathcal{A} &\in \{\mathcal{A}' : |d(\mathcal{B}, \mathcal{A}') - \min_C d(\mathcal{B}, C)| < \varepsilon\} \\ \mathcal{B} &\in \{\mathcal{B}' : |d(\mathcal{B}', \mathcal{A}) - \min_C d(C, \mathcal{A})| < \varepsilon\},\end{aligned}\tag{3}$$

which allows for a tolerance  $\varepsilon > 0$  in finding feasible pairs. In our implementation, we set  $\varepsilon = 0.2$ . In addition, in case pairs satisfying Eq. (3) cannot be found, we choose the pair with the smallest symmetrised distance.

### 2.2 Deciding when to build a new tree

It is computationally inefficient, but possible, to reestimate tree topology at every SNP, because trees are unchanged if no recombination event occurred between SNPs. Instead, Relate initially estimates the tree topology at the first SNP of the 5' end of a chromosome. It then only reestimates the tree topology if a mutation cannot be uniquely *mapped* to a branch of the tree describing the previous SNP (or the mutation is potentially *flipped*, see below). While this introduces a small bias in the placement of recombination points, we note that this bias does not propagate along the genome because marginal trees are constructed using the distance matrix for the genomic position at which tree topology is reestimated.

We use present-day genome data of outgroups, such as chimpanzees and other primates for humans, to determine the ancestral and derived alleles. Occasionally, the ancestral allele can be confused with the alternative allele due to repeat mutations between species or sequencing errors. We can infer such cases if a SNP maps onto the tree only after non-carriers of the mutation are reinterpreted as carriers and vice versa. We refer to these SNPs as *flipped* SNPs and reinfer tree topology whenever we detect a potentially flipped SNP.

A mutation is mapped to the branch for which the descendants coincide with the carriers of the derived allele. Such a branch exists as long as the mutation occurred exactly once in human history

and the estimated tree topology is correct. To be robust to errors in the data and the inferred tree, we relax this requirement as follows.

By placing a mutation on a branch, we indicate that all descendants below that branch carry the mutation. Let us denote the set of haplotypes that carry the mutation by  $C_t$  and the set of haplotypes that do not carry the mutation by  $N_t$ . Similarly, let us denote the set of haplotypes that (do not) carry the mutation in the data set by  $C_d$  and  $N_d$ . For a mutation to map to a branch, it needs to satisfy

$$\begin{aligned} \frac{|C_t \cap C_d|}{\max\{|C_t|, |C_d|\}} &> 0.7 \\ \frac{|N_t \cap N_d|}{\max\{|N_t|, |N_d|\}} &> 0.7. \end{aligned} \quad (4)$$

These conditions should identify suitable candidate branches and should prevent mapping infrequent mutations, such as doubletons, to a unique branch, when in fact they cannot have arisen by a single mutation. Out of all remaining candidate branches, we calculate the fraction of missclassified haplotypes given by

$$\frac{|N_t \cap C_d| + |C_t \cap N_d|}{N}. \quad (5)$$

We also calculate the same quantity for branches that satisfy Eq. (4) after reinterpreting carriers as non-carriers and vice-versa. We then accept the branch with the minimum score given by Eq. (5) if it is also less than 0.03. We only flip a SNP if this leads to a smaller score (i.e., in case of a tie, we do not flip the SNP). These rules, though heuristic, allow approximate mapping for mutations in 4 or more copies in the data set (Supplementary Fig. 2b).

If such unique branch cannot be found, we map the mutation to more than one branch. In this case, we find the smallest set of branches, such that all carriers of the mutation  $C_d$  are below one chosen branch and such that the summed score given by Eq. (5) equals zero. We do the same after flipping the SNP. We choose to flip the SNP only if this leads to a smaller set of branches. We then add one over the number of branches chosen to the number of mutations on that branch. For many analyses, we only consider mutations mapping to a unique branch, however we note that e.g., CpG mutations often occur on multiple branches.

### 2.3 Consistency of estimated tree topology in the absence of recombination

We prove that the use of the number of derived mutations  $d(i, j)$  as a distance matrix always guarantees estimation of a tree topology consistent with the truth, assuming that every mutation is unique in history, and no recombination. We recall that we defined the number of derived mutations  $d(i, j)$  as the number of mutations carried by haplotype  $i$  and not by haplotype  $j$ . We assume that there is a gene tree that is consistent with SNPs in the data in the sense that every SNP can be mapped to a unique branch of the gene tree [2]. Such a gene tree contains polytomies reflecting branches unresolved by observed mutations, and always exists and is unique assuming the infinite-sites model. We say, that a binary coalescence (sub-)tree, is consistent with a gene tree if all carriers of any mutation coalesce before non-carriers of a mutation.

We prove that our tree builder described in Section 2 constructs a tree that is consistent with the gene tree. The input matrix is the matrix of derived mutations  $(d(i, j))_{i,j=1,\dots,N}$  and we set  $\varepsilon = 0$  in the tree builder. The tree builder therefore coalesces haplotypes according to Eq. (1).

**Proposition 2.1.** *The tree builder cannot coalesce carriers and non-carriers before it has coalesced all carriers of any SNP.*

*Proof.* Assume without loss of generality (w.l.o.g.) that  $1, \dots, k$  are carriers of a SNP and  $k+1, \dots, N$  are non-carriers of the same SNP. We first observe that if a branch in the gene tree has a mutation, all

descendants of that branch carry at least one more derived mutation to any non-carrier of the mutation than to a carrier of the mutation. Using this observation, we obtain

$$d(i, j) < d(i, h) \text{ for } i, j \in \{1, \dots, k\} \text{ and } h \in \{k+1, \dots, N\}. \quad (6)$$

This property is why it is important to distinguish derived mutations in determining distances; the equivalent to Eq. (6) does not hold if  $d(i, j)$  is determined as the number of differences between sequences  $i$  and  $j$ .

It follows that coalescing a carrier and a non-carrier in the first step of the algorithm is not feasible in Eq. (1). Let us assume that we have not coalesced carriers and non-carriers until the  $x$ 'th step of the algorithm and that there still exist more than one cluster of carriers. We prove that it is still not feasible to coalesce a cluster of carriers and non-carriers in the  $x+1$ st step. Let us denote by  $\mathcal{A}$  and  $\mathcal{B}$  clusters containing only carriers and by  $\mathcal{C}$  a cluster containing only non-carriers. We obtain from Eq. (6),

$$d(\mathcal{A}, \mathcal{B}) < d(\mathcal{A}, \mathcal{C}), \quad (7)$$

because an average of  $d(i, j)$  with  $i, j \in \{1, \dots, k\}$  is always smaller than an average of  $d(i, h)$  with  $i \in \{1, \dots, k\}$  and  $h \in \{k+1, \dots, N\}$ . Therefore coalescing  $\mathcal{A}$  and  $\mathcal{C}$  does not satisfy Eq. (1).  $\square$

With Proposition 2.1, we know that the tree builder never violates the gene tree and therefore, if the algorithm successfully coalesces all haplotypes, we obtain a tree that is consistent with the gene tree. It remains to prove that the tree builder can always find a next coalescence event and therefore, by induction, terminates with all haplotypes in one cluster.

**Proposition 2.2.** *Assuming there is a gene tree consistent with the data, the tree builder terminates with all haplotypes in one cluster.*

*Proof.* Assume that we are in the  $x$ th step of the tree builder. We prove that we can coalesce a pair of clusters to proceed to the  $x+1$ st step of the tree builder. We have already proved that the tree constructed until the  $x$ th step is consistent with the gene tree. We can therefore find two clusters  $\mathcal{A}$  and  $\mathcal{B}$ , such that coalescing these clusters is consistent with the gene tree.

Let  $\mathcal{C}$  be a third cluster distinct from  $\mathcal{A}$  and  $\mathcal{B}$ . For  $\mathcal{A}$  and  $\mathcal{B}$  to satisfy Eq. (1), we require

$$\begin{aligned} d(\mathcal{A}, \mathcal{B}) &\leq d(\mathcal{A}, \mathcal{C}) \text{ and} \\ d(\mathcal{B}, \mathcal{A}) &\leq d(\mathcal{B}, \mathcal{C}). \end{aligned} \quad (8)$$

We show  $d(\mathcal{A}, \mathcal{B}) \leq d(\mathcal{A}, \mathcal{C})$  and note that  $d(\mathcal{B}, \mathcal{A}) \leq d(\mathcal{B}, \mathcal{C})$  can be shown analogously.

Let us define by  $d_\ell(\mathcal{A}, \mathcal{B})$  the distance considering only SNP  $\ell$ , such that  $d(\mathcal{A}, \mathcal{B}) = \sum_{\ell=1}^L d_\ell(\mathcal{A}, \mathcal{B})$  with  $L$  denoting the number of SNPs. We note that any SNP with no derived sequences in  $\mathcal{A}$  yields  $d_\ell(\mathcal{A}, \mathcal{B}) = d_\ell(\mathcal{A}, \mathcal{C}) = 0$ . Let us therefore consider the possible types of SNPs with at least one derived mutation in  $\mathcal{A}$ . We note that we cannot have SNPs with both carriers and non-carriers in more than one cluster, because we would have violated Proposition 2.1. Therefore, the possible SNPs are as follows.

- If a SNP is only derived in sequences in  $\mathcal{A}$ , then  $d_\ell(\mathcal{A}, \mathcal{B}) = d_\ell(\mathcal{A}, \mathcal{C})$ .
- If a SNP is derived in all sequences of  $\mathcal{A}$  and  $\mathcal{B}$ , but no sequences in  $\mathcal{C}$ , then

$$d_\ell(\mathcal{A}, \mathcal{B}) = 0 < 1 = d_\ell(\mathcal{A}, \mathcal{C}). \quad (9)$$

- If a SNP is derived in all sequences of  $\mathcal{A}$  and  $\mathcal{C}$ , but no sequences of  $\mathcal{B}$ , then this SNP violates the assumption that coalescing  $\mathcal{A}$  and  $\mathcal{B}$  is consistent with the gene tree.
- If a SNP is derived in all sequences of  $\mathcal{A}$ ,  $\mathcal{B}$ , and  $\mathcal{C}$ , then  $d_\ell(\mathcal{A}, \mathcal{B}) = d_\ell(\mathcal{A}, \mathcal{C}) = 0$ .

By summing over all SNPs, we obtain

$$d(\mathcal{A}, \mathcal{B}) = \sum_{\ell=1}^L d_{\ell}(\mathcal{A}, \mathcal{B}) \leq \sum_{\ell=1}^L d_{\ell}(\mathcal{A}, \mathcal{C}) = \sum_{\ell=1}^L d(\mathcal{A}, \mathcal{C}), \quad (10)$$

as required.  $\square$

We conclude with a few observations. First, we notice that all branches that have a mutation in the gene tree are guaranteed to exist in the tree built by our tree builder. In particular, if all branches in the gene tree have a mutation, the tree builder is guaranteed to build the correct tree. The tree builder therefore constructs the correct tree in the limit of an infinite mutation rate (and under the infinite-sites assumption). Second, we notice that if recombination rates are only either infinite or zero, we can divide the genome into regions of zero recombination rates. In this case, the tree builder constructs trees consistent with the gene trees of these regions. Finally, we notice that our proof did not depend on the absolute values of the input distance matrix, and that the only requirement was that rows of the distance matrix are perfectly correlated to the order of coalescences of a haplotype with other haplotypes. Therefore, we expect our tree builder to construct an accurate tree as long as rows of the distance matrix obtained from the modified Li-and-Stephens HMM (Section 3) is well correlated with the order in which a haplotype coalescences with other haplotypes.

### 3 Calculating distance matrices

#### 3.1 Assumptions about the input data

We apply a version of the Li-and-Stephens HMM to calculate distance matrices. To apply this algorithm, we assume haplotype SNP data as input, which can be inferred by phasing genotype data [3, 4]. We assume a high coverage of SNPs along the genome and no bias with respect to the frequency of a mutation in the population. We also assume that at most one mutation has occurred at any given position along the genome since the divergence of humans from chimpanzees and other primates. This assumption, known as the infinite sites model, is justified by a small average mutation rate in humans [5].

Additionally, we require knowledge of the ancestral allele for every recorded mutation. The ancestral allele can be determined by aligning the human genome to present day genomes of other primates [6]. Assuming that it is unlikely to observe a mutation at the same genomic position in humans and other primates, the allele carried by other primates is declared to be the ancestral allele. For humans, we use the ancestral genome that was inferred as part of the 1000 Genomes Project (see URLs in main text).

For the robustness of Relate to genotype errors and occasional confusion of ancestral and alternative alleles, please refer to the Supplementary Note: Simulations.

Under these assumptions, we can represent SNP data as a binary matrix  $D$  with each row corresponding to one haplotype. The matrix  $D$  therefore has dimensions  $N \times L$ , where  $N$  is the number of haplotypes and  $L$  is the number of SNPs. In this representation, a haplotype is a binary vector, where the ancestral allele is denoted by 0 and the alternative allele is denoted by 1. We denote row  $i$  of  $D$  by  $D^{(i)}$ .

#### 3.2 Modified Li-and-Stephens HMM

A recombination event may change the order of coalescence events between haplotypes. Therefore, mutations close to the SNP of consideration are more informative than SNPs further away. To capture this, we apply a Li-and-Stephens type HMM [1]. We modified the original HMM by changing emission probabilities such that they take ancestral and derived states into account. A schematic of the HMM is depicted in Supplementary Fig. 1c.

The HMM can be interpreted as a generative model for haplotype  $D^{(i)}$  using all other haplotypes as inputs. To generate the allele  $D_\ell^{(i)}$  at SNP  $\ell$ , we first choose one reference haplotype  $H_\ell$  from all other haplotypes. This reference haplotype is the hidden state of the HMM. In Ref. [1], the emission probabilities are defined by

$$P_c(D_\ell^{(i)}|H_\ell = j) = \begin{cases} p & \text{if } D_\ell^{(i)} \neq D_\ell^{(j)}, \\ 1 - p & \text{if } D_\ell^{(i)} = D_\ell^{(j)}. \end{cases} \quad (11)$$

Here,  $p$  is the mismatch probability. In this conventional definition of the emission probabilities, a mutation may have occurred on the branch from  $i$  to the MRCA with  $j$ , or on the branch from  $j$  to the MRCA with  $i$ . Therefore, information about on which branch the mutation occurred is lost. To preserve this information, we change the emission probabilities to

$$P_m(D_\ell^{(i)}|H_\ell = j) = \begin{cases} p & \text{if } D_\ell^{(i)} = 1, D_\ell^{(j)} = 0, \\ 1 - p & \text{if } D_\ell^{(i)} = 0, D_\ell^{(j)} = 0, \\ 1 - p & \text{if } D_\ell^{(i)} = 1, D_\ell^{(j)} = 1, \\ 1 - p & \text{if } D_\ell^{(i)} = 0, D_\ell^{(j)} = 1. \end{cases} \quad (12)$$

We can interpret  $p$  as the probability of a mutation since the MRCA. The emission probability equals  $p$ , if haplotype  $i$  carries a mutation at site  $\ell$  which is not carried by the reference haplotype  $j$ , such that the mutation must have occurred on the branch from  $i$  to the MRCA with  $j$ . Otherwise, no mutation occurred on this branch and the emission probability equals  $1 - p$  (see Supplementary Fig. 1 c). The hidden state may change between neighbouring SNPs according to transition probabilities  $r_\ell$ . The transition probabilities are proportional to the recombination probabilities obtained from a recombination map. We describe how to choose  $p$  and  $r_\ell$  in Section 3.3.

Using the modified Li-and-Stephens HMM, we can calculate a distance matrix at every SNP which we will use as input for the tree builder described in Section 2. We first derive a distance matrix for the case when transition probabilities are set to zero, which corresponds to no recombination. In this case, the reference haplotype remains the same along the genome. The likelihood of observing  $D^{(i)}$  given reference haplotype  $H_\ell = j$  is given by

$$P_m(D^{(i)}|H_\ell = j) = p^{d(i,j)}(1 - p)^{L-d(i,j)}, \quad (13)$$

where  $d(i, j)$  is the number of derived mutations defined in Section 1. By taking logarithms on both sides of Eq. (13), we obtain

$$\log P_m(D^{(i)}|H_\ell = j) = d(i, j) \log \left( \frac{p}{1 - p} \right) + L \log(1 - p). \quad (14)$$

By rearranging Eq. (14), we obtain

$$d(i, j) = \frac{\log P_m(D^{(i)}|H_\ell = j) - L \log(1 - p)}{\log \left( \frac{p}{1 - p} \right)}. \quad (15)$$

We can generalise Eq. (15) to the case of non-zero recombination rates to define, for each SNP  $\ell$ , the *local* number of derived mutations

$$d(i, j; \ell) = \frac{\log P_m(D^{(i)}|H_\ell = j) - L \log(1 - p)}{\log \left( \frac{p}{1 - p} \right)}. \quad (16)$$

Equation (16), corresponding to a local number of derived mutations, orders coalescence events locally at every SNP. In practice, we use  $d(i, j; \ell) - \min_{j \neq i} d(i, j; \ell)$  as our distance matrix, which only affects

our tie-breaking heuristic in case no pair of haplotypes (or clusters) satisfy Eq. (3) and we choose a pair with minimum entry in the symmetrised matrix. By subtracting the minimum entry from each row, we remove residuals interpretable as the number of derived mutations on the tip branches, which could confound the symmetrised distance if some haplotypes have many more mutations at their tips than other haplotypes. The quantity  $\log P_m(D^{(i)}|H_\ell = j)$  in Eq. (16) can be calculated using the forward-backward algorithm. We describe how to efficiently implement the Li-and-Stephens HMM in Section 3.4.

We notice that with the conventional definition of the emission probabilities (Eq. (11)) and no recombination, we obtain

$$d(i, j) + d(j, i) = \frac{\log P_c(D^{(i)}|H_\ell = j) - L \log(1 - p)}{\log\left(\frac{p}{1-p}\right)}. \quad (17)$$

In this case, we obtain a symmetric matrix that stores the number of mutations that differ between two haplotypes. Using the matrix  $d(i, j) + d(j, i)$  as input to our tree builder is equivalent to applying the UPGMA algorithm [7], which is an alternative hierarchical clustering algorithm, to the matrix  $d(i, j) + d(j, i)$ . We show in Supplementary Fig. 1a, how this can lead to estimation of a tree topology that is inconsistent with the data. Intuitively, this is because  $d(i, j) + d(j, i)$  combines the number of mutations of two branches, such that information about the number of mutations on each branch is lost. Therefore,  $d(i, j)$  preserves more information about the tree topology than  $d(i, j) + d(j, i)$ .

### 3.3 Choosing parameters for the modified Li-and-Stephens HMM

The transition probabilities in the HMM are determined by the recombination map, multiplied by a constant factor  $R$ . The mutation probability  $p$  should reflect the true biological mutation rate and errors in the data set. In practice, we find that the tree topology constructed using distance matrix  $d(i, j; \ell)$  (Eq. (16)) is very robust with respect to choice of  $p$  and  $R$ . To illustrate this, we evaluate the performance of our algorithm for different choices of  $p$  and  $R$ . In our implementation, we have fixed  $p = 0.025$  and  $R = 2500$ .

To evaluate how well a pair  $(p, R)$  captures ancestry information in a subregion of the genome, we sample subregions of lengths 1200 SNPs at random. We first apply the modified Li-and-Stephens HMM on the chosen subregion. For every SNP  $400 < \ell < 800$ , we then do the following. We first calculate the distance matrix as described in Section 3. We modify the distance matrix by hiding the focal SNP  $\ell$  which can be done by subtracting 1 from any entry  $(i, j)$  where  $i$  is a carrier and  $j$  is not a carrier of the mutation. We then build the tree with this modified distance matrix using the hierarchical clustering algorithm described in Section 1. Finally, we attempt to place the focal SNP on branches of the tree and record whether a branch exists such that the descendants of that branch coincide with carriers of the SNP.

We therefore evaluate a pair  $(p, R)$  by how well we can build trees at focal SNPs using information from surrounding SNPs only. As  $R$  becomes larger and  $p$  becomes smaller, only SNPs close to the focal SNP will influence the distance matrix. As  $R$  becomes smaller and  $p$  becomes larger, SNPs further away will influence the distance matrix. We should find an optimal pair  $(p, R)$  such that we include enough SNPs to be able to build a tree onto which the focal SNP can be mapped and such that we prevent SNPs that map onto different trees from distorting the signal.

We applied this method to 50 subregions of chromosome 20 of the 1000 Genomes Project data set. In Supplementary Fig. 2c and d, we show that the number of non-mapping SNPs is relatively robust to the choice of  $(p, R)$ , particularly along the axis  $p/R = 10^{-6}$ . We therefore fix  $p = 0.025$  and  $R = 2500$  in our implementation.



### 3.4 Speed-up and approximation of the modified Li-and-Stephens HMM

To calculate  $d(i, j; \ell)$  defined in Eq. (16), we need to calculate  $P_m(D^{(i)}|H_\ell = j)$ . We apply Bayes' theorem and obtain

$$P_m(D^{(i)}|H_\ell = j) = \frac{P_m(H_\ell = j|D^{(i)})P_m(D^{(i)})}{P_m(H_\ell = j)}. \quad (18)$$

We assume the prior probability of copying from any haplotype to be identical, such that  $P_m(H_\ell = j) = 1/(N-1)$ , and calculate  $P_m(H_\ell = j|D^{(i)})$  using a forward-backward algorithm. The forward algorithm calculates  $\alpha_j(\ell) = P_m(H_\ell = j, D_{1:\ell}^{(i)})$  and the backward algorithm calculates  $\beta_j(\ell) = P(D_{(\ell+1):L}^{(i)}|H_\ell = j)$  such that

$$P_m(H_\ell = j|D^{(i)}) = \frac{\alpha_j(\ell)\beta_j(\ell)}{P_m(D^{(i)})}. \quad (19)$$

By substituting Eq. (19) in Eq. (18) and taking logarithms, we obtain

$$\log P_m(D^{(i)}|H_\ell = j) = \log(\alpha_j(\ell)\beta_j(\ell)) + \log(N-1). \quad (20)$$

By substituting Eq. (20) in Eq. (16), we obtain

$$d(i, j; \ell) = \frac{\log(\alpha_j(\ell)\beta_j(\ell)) + \log(N-1) - L \log(1-p)}{\log\left(\frac{p}{1-p}\right)}. \quad (21)$$

To speed-up the calculation of  $\alpha_j(\ell)$  and  $\beta_j(\ell)$  in Eq. (21), we calculate  $P_m(H_\ell = j|D^{(i)}) \propto \alpha_j(\ell)\beta_j(\ell)$  only for SNPs  $\ell$  at which haplotype  $i$  is a carrier of the derived allele, i.e.  $D_\ell^{(i)} = 1$ . This implies that the forward-backward algorithm is applied to a different set of SNPs depending on the haplotype  $i$ .

The calculated  $P_m(H_\ell = j|D^{(i)})$  at SNPs  $\ell$  with  $D_\ell^{(i)} = 1$  are exact up to a multiplicative constant  $(1-p)^{L-L^{(i)}}$ , where  $L^{(i)} = \sum_{\ell=1}^L D_\ell^{(i)}$  is the number of derived mutations carried by  $i$ . We therefore calculate Eq. (21) as

$$d(i, j; \ell) = \frac{\log(\alpha_j^d(\ell)\beta_j^d(\ell)) + \log(N-1) - L^{(i)} \log(1-p)}{\log\left(\frac{p}{1-p}\right)}, \quad (22)$$

where  $\alpha_j^d(\ell)$  and  $\beta_j^d(\ell)$  are the forward and backward probabilities calculated for SNPs  $\ell$  at which  $D_\ell^{(i)} = 1$ .

For a site  $\ell$  at which haplotype  $i$  is not a carrier of the derived allele, we approximate  $P_m(H_\ell = j|D^{(i)}) \propto \alpha_j(\ell)\beta_j(\ell)$  using  $P_m(H_{\ell_{\text{left}}} = j|D^{(i)})$  and  $P_m(H_{\ell_{\text{right}}} = j|D^{(i)})$ , where  $\ell_{\text{left}}$ ,  $\ell_{\text{right}}$  are the nearest derived sites to either side of  $\ell$ , i.e.,  $D_{\ell_{\text{left}}}^{(i)} = D_{\ell_{\text{right}}}^{(i)} = 1$  and  $D_\ell^{(i)} = 0$  for  $\ell_{\text{left}} < \ell < \ell_{\text{right}}$ . For such  $\ell$ , we approximate  $P_m(H_\ell = j|D^{(i)})$  as a weighted average of  $P_m(H_{\ell_{\text{left}}} = j|D^{(i)})$  and  $P_m(H_{\ell_{\text{right}}} = j|D^{(i)})$ . This approximation is valid if the probability for two or more recombinations between  $\ell_{\text{left}}$  and  $\ell_{\text{right}}$  is sufficiently small. We expand

$$\begin{aligned} P_m(H_\ell = j|D^{(i)}) &= P_m(H_\ell = j, 0 \text{ recombinations in } [\ell_{\text{left}}, \ell_{\text{right}}]|D^{(i)}) \\ &\quad + P_m(H_\ell = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}]|D^{(i)}) \\ &\quad + P_m(H_\ell = j, \text{ more than 1 rec. in } [\ell_{\text{left}}, \ell_{\text{right}}]|D^{(i)}). \end{aligned} \quad (23)$$

For the third term on the right hand side of Eq. (23), we obtain

$$P_m(H_\ell = j, \text{ more than 1 rec. in } [\ell_{\text{left}}, \ell_{\text{right}}]|D^{(i)}) = O((r_{\text{left}} + r_{\text{right}})^2). \quad (24)$$

For the second term on the right hand side of Eq. (23), we obtain

$$\begin{aligned}
& P_m(H_\ell = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}) \\
&= P_m(H_{\ell_{\text{right}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell], 0 \text{ rec. in } [\ell, \ell_{\text{right}}] | D^{(i)}) + \\
& P_m(H_{\ell_{\text{left}}} = j, 0 \text{ rec. in } [\ell_{\text{left}}, \ell], 1 \text{ rec. in } [\ell, \ell_{\text{right}}] | D^{(i)}).
\end{aligned} \tag{25}$$

We notice that the emission probability at any site  $\ell_{\text{left}} < \ell < \ell_{\text{right}}$  equals  $1 - p$  regardless of the hidden state  $H_\ell$ , because  $D_\ell^{(i)} = 0$ . Therefore, we find that the probability of a recombination (i.e., switch of hidden state), given the data  $D^{(i)}$ , is the same at any position between  $\ell_{\text{left}}$  and  $\ell_{\text{right}}$ . We denote the recombination distance from  $\ell_{\text{left}}$  to  $\ell$  by  $r_{\text{left}}$  and the recombination distance from  $\ell$  to  $\ell_{\text{right}}$  by  $r_{\text{right}}$ . These recombination distances correspond to the probability of a switch of hidden states between the two SNPs. We therefore obtain

$$\begin{aligned}
& P_m(H_{\ell_{\text{right}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell], 0 \text{ rec. in } [\ell, \ell_{\text{right}}] | D^{(i)}) \\
&= P_m(1 \text{ rec. in } [\ell_{\text{left}}, \ell] | 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}]) P_m(H_{\ell_{\text{right}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}) \\
&= \frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{right}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}).
\end{aligned} \tag{26}$$

Analogously, we obtain

$$\begin{aligned}
& P_m(H_{\ell_{\text{left}}} = j, 0 \text{ rec. in } [\ell_{\text{left}}, \ell], 1 \text{ rec. in } [\ell, \ell_{\text{right}}] | D^{(i)}) \\
&= \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{left}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}).
\end{aligned} \tag{27}$$

We substitute Eqs. (26) and (27) in Eq. (25) and obtain

$$\begin{aligned}
& P_m(H_\ell = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}) \\
&= \frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{right}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}) \\
&+ \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{left}}} = j, 1 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}).
\end{aligned} \tag{28}$$

For the first term on the right hand side of Eq. (23), we use that

$$\frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} + \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} = 1, \tag{29}$$

to obtain

$$\begin{aligned}
& P_m(H_\ell = j, 0 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}) \\
&= \frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{right}}} = j, 0 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}) \\
&+ \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{left}}} = j, 0 \text{ rec. in } [\ell_{\text{left}}, \ell_{\text{right}}] | D^{(i)}).
\end{aligned} \tag{30}$$

We substitute Eqs. (24), (28), and (30) in Eq. (23) and obtain

$$\begin{aligned}
P_m(H_\ell = j | D^{(i)}) &= \frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{right}}} = j | D^{(i)}) \\
&+ \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} P_m(H_{\ell_{\text{left}}} = j | D^{(i)}) + O((r_{\text{left}} + r_{\text{right}})^2).
\end{aligned} \tag{31}$$

By multiplying both sides of Eq. (31) by  $P_m(D^{(i)})$ , we obtain

$$\begin{aligned}
\alpha_j^d(\ell) \beta_j^d(\ell) &= \frac{r_{\text{left}}}{r_{\text{left}} + r_{\text{right}}} \alpha_j^d(\ell_{\text{left}}) \beta_j^d(\ell_{\text{left}}) \\
&+ \frac{r_{\text{right}}}{r_{\text{left}} + r_{\text{right}}} \alpha_j^d(\ell_{\text{right}}) \beta_j^d(\ell_{\text{right}}) + O((r_{\text{left}} + r_{\text{right}})^2).
\end{aligned} \tag{32}$$

By substituting Eq. (32) in Eq. (22) and dropping terms  $O((r_{\text{left}} + r_{\text{right}})^2)$ , we obtain an approximation for calculating  $d(i, j; \ell)$ .

The complexity of the modified Li-and-Stephens HMM is reduced from  $N^2 L$  to  $N^2 L^{(i)}$ . We expect  $L^{(i)}$  to be independent of the sample size  $N$  for large  $N$ . This is because in the standard coalescent, the expected TMRCA is, in the limit  $N \rightarrow \infty$ , given by  $4N_e$ , such that the expected number of derived mutations carried by haplotype  $i$  since the population's TMRCA is  $E[L^{(i)}] = 4N_e\mu$ . In practice, this can make the application of the algorithm  $> 10$  times faster even for modest samples of a few thousand individuals.

## 4 Estimating branch lengths

We have now estimated the genetic ancestry of each SNP in form of a rooted binary tree. To estimate the branch lengths  $t_b$  ( $b = 0, \dots, 2N - 2$ ) of these trees, we use a Metropolis-Hastings type MCMC algorithm. We note that while MCMC sampling is a computationally expensive approach, it is a flexible approach allowing, for instance, the specification of demographic histories. By sampling posterior branch lengths, we estimate the mean age of a coalescent event and use these to calculate branch lengths. This approach will yield branch lengths that reflect the coalescent prior for trees (or subtrees) with little information about branch lengths, which is desirable for many applications.

Before we can apply the MCMC algorithm, we notice that a recombination event changes only a few branches in adjacent trees along the genome. Some branches can persist over multiple trees. We therefore identify equivalent branches in adjacent trees along the genome (see Section 4.1). We then count the number of mutations across equivalent branches and calculate a cumulative mutation rate for each branch. This information is fed into the MCMC algorithm (see Section 4.2). We assume a coalescent prior given by the standard coalescent [8]. The effective population size is predetermined and assumed to be constant. We initialise the order of coalescence events to a random order, but such that no topological constraints are violated (see Section 4.3). We initialise the times while  $k$  ancestors remain using an Expectation-maximization (EM) algorithm that calculates the maximum-likelihood estimates (MLEs) of the times while  $k$  ancestors remain, where the order of coalescence events is kept fixed (see Section 4.4).

Once the constant population size model has been fitted, we can jointly infer piecewise-constant historical population sizes and branch lengths under a coalescent prior with variable historical population sizes (see Section 5).

### 4.1 Identifying equivalent branches in neighbouring trees

Let us take a branch  $b_1 = (c_1, d_1)$  from one tree and a branch  $b_2 = (c_2, d_2)$  from another tree, where  $c_i$  and  $d_i$  ( $i = 1, 2$ ) denote coalescence events. We then say that branches  $b_1$  and  $b_2$  are equivalent if the descendants of  $c_1$  coincide with those of  $c_2$  and the descendants of  $d_1$  coincide with those of  $d_2$ . To be robust to errors, we slightly relax this requirement as follows.

For every coalescence event, we store a vector of length  $N$  containing its present-day descendants, where the  $m$ 'th entry of the vector equals 1 if haplotype  $m$  is below the coalescence event, and 0 otherwise. Then, for every pair of branches, with branches coming from different trees, we calculate the correlation coefficient of these vectors for the coalescence events at the lower ends of the branches and the upper ends of the branches.

Two branches are exactly equivalent if the correlation coefficient on both ends of the branches equal one. We first identify exactly equivalent pairs of branches. For the remaining branches, we require that a pair of branches is equivalent if the correlation coefficients on both ends are greater than 0.9. Notice that a branch can satisfy this condition for more than one branch in the other tree. To guarantee that each branch is associated with at most one other branch, we sort pairs of candidate branches in descending order of the correlation coefficient at the lower end of the branches. We then associate

branches in the order at which they appear in this list, where we delete any entry for which one of the branches has already been associated with a different branch.

Once we identified equivalent branches, we calculate the number of mutations  $m_b$  on a branch  $b$  by adding the number of mutations across all equivalent branches. Next, we calculate the cumulative mutation rate for a branch. We assume that mutations occur at a constant rate of  $\theta/2$  per base in coalescence time, where  $\theta = 4\mu N_e$  and  $\mu$  is the per-generation mutation rate. For each branch  $b$ , we denote the mutation rate summed over bases for which  $b$  persists by  $\theta_b/2$ .

## 4.2 Metropolis-Hastings type MCMC to estimate branch lengths in a constant population size

We use a Metropolis-Hastings type MCMC algorithm to estimate branch lengths. The likelihood of observing branch lengths  $\mathbf{t} = \{t_b\}_{b=0,\dots,2N-2}$ , conditional on the number of mutations on branches  $\mathbf{m} = \{m_b\}_{b=0,\dots,2N-2}$ , is given by

$$P(\mathbf{t}|\mathbf{m}) \propto P(\mathbf{t})P(\mathbf{m}|\mathbf{t}) = P(\mathbf{t}) \prod_{b=0}^{2N-2} P(m_b|t_b), \quad (33)$$

where  $P(m_b|t_b)$  is Poisson distributed with mean  $\theta_b t_b/2$  and the prior  $P(\mathbf{t})$  is given by the standard coalescent.

We can now define a reversible Markov-Chain with a unique stationary distribution which is the target distribution  $P(\mathbf{t}|\mathbf{m})$ . For this, we assign a label  $\{1, \dots, N-1\}$  to every coalescence event. The coalescence event that decreases the number of lineages from  $k$  to  $k-1$  is stored in variable  $n_k$  ( $k = 2, \dots, N$ ). We denote the time while  $k$  ancestors exist by  $\tau_k$  ( $k = 2, \dots, N$ ). Notice that the branch lengths are uniquely determined by  $\mathbf{n} = \{n_k\}_{k=2,\dots,N}$  and  $\boldsymbol{\tau} = \{\tau_k\}_{k=2,\dots,N}$ . In particular, the coalescent prior is given by

$$P(\mathbf{t}) = P(\mathbf{n})P(\boldsymbol{\tau}) = P(\mathbf{n}) \prod_{k=2}^N P(\tau_k), \quad (34)$$

where  $\mathbf{n}$  is uniformly distributed over all possible orders of coalescence events and  $\tau_k$  is exponentially distributed with rate  $\binom{k}{2}$ . We initialise  $\mathbf{n}$  and  $\boldsymbol{\tau}$  as described in Sections 4.3 and 4.4. In every step of the Metropolis-Hastings algorithm, we propose a change in  $\mathbf{n}$  with probability  $q$  and a change in  $\boldsymbol{\tau}$  with probability  $1-q$ . In our implementation, we have chosen  $q = 0.8$ .

For a change in  $\mathbf{n}$ , we first choose one coalescence event  $e_1 = n_k$  uniformly at random. This is the event that decreases the number of lineages from  $k$  to  $k-1$ . We then propose to swap  $e_1$  with another coalescence event  $e_2 = n_h$  chosen uniformly at random from any event with age between  $e_1$ 's parent event and  $e_1$ 's daughter events. If the proposal is accepted, event  $e_1$  now decreases the number of lineages from  $h$  to  $h-1$  and  $e_2$  decreases the number of lineages from  $k$  to  $k-1$ . We discard any proposal that would violate topological constraints of the tree if all other events retain their times. In a swap of the order of coalescence events, we keep  $\boldsymbol{\tau}$  fixed. Such a swap therefore changes the lengths of six branches, but keeps all other branch lengths fixed. Denote these six branches by  $b_1, \dots, b_6$ . By using Eqs. (33) and (34) and noticing that the proposal distribution is symmetric, the acceptance probability is given by

$$\min \left( 1, \frac{P(\tilde{\mathbf{t}}|\mathbf{m})}{P(\mathbf{t}|\mathbf{m})} \right) = \min \left( 1, \prod_{\ell=1}^6 \frac{P(m_{b_\ell}|\tilde{t}_{b_\ell})}{P(m_{b_\ell}|t_{b_\ell})} \right), \quad (35)$$

where  $\tilde{\mathbf{t}} = \{\tilde{t}_b\}_{b=0,\dots,2N-2}$  are the proposed branch lengths.

For a change in  $\boldsymbol{\tau}$ , we choose one  $\tau_k$  uniformly at random. We propose  $\tilde{\tau}_k$  from an exponential distribution with expectation  $\tau_k$ . A change in  $\tau_k$  changes the length of  $k$  branches that are present at the time while  $k$  ancestors remain. Denote these  $k$  branches by  $b_1, \dots, b_k$ . The remaining branch

lengths remain unchanged. By using Eqs. (33) and (34), the acceptance probability is given by

$$\min \left( 1, \frac{\tilde{\tau}_k}{\tau_k} \exp \left[ -\frac{\tilde{\tau}_k}{\tau_k} + \frac{\tau_k}{\tilde{\tau}_k} - \binom{k}{2} (\tilde{\tau}_k - \tau_k) \right] \prod_{\ell=1}^k \frac{P(m_{b_\ell} | \tilde{t}_{b_\ell})}{P(m_{b_\ell} | t_{b_\ell})} \right). \quad (36)$$

Using this MCMC algorithm, we calculate the mean age of every coalescence event. We then calculate the branch lengths as the differences in the mean ages of coalescence events. This guarantees that the time from any tip to the root is equal, a property also known as ultrametric.

In our implementation, we initially perform  $\max\{10N, 1000\}$  burn-in iterations. We then apply the algorithm until every  $\tau_k$  ( $k = 2, \dots, N$ ) has had at least 20 proposals and then terminate the MCMC algorithm, conditional on all branch lengths being positive, and continue until the latter condition is satisfied otherwise.

### 4.3 Initialising the order of coalescence events

We initialise the order of coalescence events  $\mathbf{n}$  by applying a simple MCMC algorithm. In the standard coalescent, any order of coalescence events is equally likely, provided that the order does not contradict the topological constraints of the tree. We therefore propose the following swap moves. We choose a coalescence event  $n$  uniformly at random and propose a swap with another coalescence event  $n'$ . To ensure that we do not contradict tree topology after swapping the two events, we choose  $n'$  from coalescence events that are between the parent and the children of  $n$ . We then assert whether  $n$  is between the parent and children of  $n'$ . If this is the case, we accept the proposal. We accept a proposal with probability 1 because the transition probabilities are symmetric and any order of coalescence events is equally likely. We initialise the order of coalescence events by the order obtained after proposing  $N^2$  swap moves.

### 4.4 Initialising the time while $k$ ancestors remain

After initialising  $\mathbf{n}$ , we initialise the times  $\tau_k$  while  $k$  ancestors remain using the MLE of  $\tau_k$  conditional on a fixed order of coalescence events. For each  $k$ , let  $b_1, \dots, b_k$  be the  $k$  branches while there are  $k$  lineages remaining in the tree and define  $\tilde{\mathbf{m}}_k = \{\tilde{m}_{b_1,k}, \dots, \tilde{m}_{b_k,k}\}$ , where  $\tilde{m}_{b,k}$  is the number of mutations that are on branch  $b$  in the interval while there are  $k$  lineages remaining. We set up an EM algorithm to estimate the MLE of the parameters  $\boldsymbol{\tau}$ , given the data  $\mathbf{m}$  and the unobserved variables  $\tilde{\mathbf{m}}_k = \{\tilde{m}_{b_i,k}\}_{i=1,\dots,k}$  ( $k = 2, \dots, N$ ) corresponding to the number of mutations on branches  $b_i$  while  $k$  ancestors remain. We denote by  $\hat{\boldsymbol{\tau}}^{(s)} = \{\hat{\tau}_k^{(s)}\}_{k=N,\dots,2}$  the estimate of the MLE of  $\tau_k$  after  $s$  iterations. We initialise  $\hat{\tau}_k^{(0)} = \binom{k}{2}^{-1}$ .

We fix  $k$ . The update rule for the EM algorithm is given by the expectation of the log likelihood function  $\log P(\tau_k | \tilde{\mathbf{m}}_k)$ , where the expectation is taken conditional on the data and parameters from the previous iteration. We obtain,

$$\hat{\tau}_k^{(s+1)} = \arg \max_{\tau_k} E \left[ \log P(\tau_k | \tilde{\mathbf{m}}_k) \mid \mathbf{m}, \hat{\boldsymbol{\tau}}^{(s)} \right]. \quad (37)$$

We first calculate  $\log P(\tau_k | \tilde{\mathbf{m}}_k)$ . By using Bayes' theorem, we obtain

$$\begin{aligned} P(\tau_k | \tilde{\mathbf{m}}_k) &\propto P(\tilde{\mathbf{m}}_k | \tau_k) P(\tau_k) \\ &= P(\tau_k) \prod_{\ell=1}^k P(\tilde{m}_{b_\ell,k} | \tau_k), \end{aligned} \quad (38)$$

where  $P(\tau_k)$  is an exponential distribution with rate  $\binom{k}{2}$  and  $P(\tilde{m}_{b_\ell,k} | \tau_k)$  is a Poisson distribution with mean  $\theta_{b_\ell} \tau_k / 2$ . By using this together with Eq. (38), we obtain

$$P(\tau_k | \tilde{\mathbf{m}}_k) \propto \tau_k^{\sum_{\ell=1}^k \tilde{m}_{b_\ell,k}} \exp \left[ - \left( \sum_{\ell=1}^k \frac{\theta_{b_\ell}}{2} + \binom{k}{2} \right) \tau_k \right]. \quad (39)$$

By taking logarithms on both sides of Eq. (39), we obtain

$$\log P(\tau_k | \tilde{\mathbf{m}}_k) = \left( \sum_{\ell=1}^k \tilde{m}_{b_\ell, k} \right) \log(\tau_k) - \left( \sum_{\ell=1}^k \frac{\theta_{b_\ell}}{2} + \binom{k}{2} \right) \tau_k + \text{const.} \quad (40)$$

We substitute Eq. (40) into Eq. (37) and obtain

$$\hat{\tau}_k^{(s+1)} = \arg \max_{\tau_k} \sum_{\ell=1}^k E \left[ \tilde{m}_{b_\ell, k} | \mathbf{m}, \hat{\tau}^{(s)} \right] \log(\tau_k) - \left( \sum_{\ell=1}^k \frac{\theta_{b_\ell}}{2} + \binom{k}{2} \right) \tau_k. \quad (41)$$

To find the  $\tau_k$  maximising Eq. (41), we take the derivative with respect to  $\tau_k$ , we obtain the condition

$$\sum_{\ell=1}^k E \left[ \tilde{m}_{b_\ell, k} | \mathbf{m}, \hat{\tau}^{(s)} \right] \frac{1}{\hat{\tau}_k^{(s+1)}} - \left( \sum_{\ell=1}^k \frac{\theta_{b_\ell}}{2} + \binom{k}{2} \right) = 0. \quad (42)$$

After reorganizing the terms, we obtain

$$\hat{\tau}_k^{(s+1)} = \frac{\sum_{\ell=1}^k E \left[ \tilde{m}_{b_\ell, k} | \mathbf{m}, \hat{\tau}^{(s)} \right]}{\sum_{\ell=1}^k \frac{\theta_{b_\ell}}{2} + \binom{k}{2}}. \quad (43)$$

It therefore remains to calculate  $E \left[ \tilde{m}_{b_\ell, k} | \mathbf{m}, \hat{\tau}^{(s)} \right]$  for  $\ell = 1, \dots, k$ . In the standard coalescent, mutation events are uniformly distributed on each branch. Therefore, the likelihood that an event on branch  $b$  falls within the interval while  $k$  lineages are remaining is given by  $\hat{\tau}_k^{(s)} / t_b^{(s)}$ . Therefore, the likelihood that  $\tilde{m}_{b, k}$  mutations fall into that interval is given by the binomial distribution

$$P(\tilde{m}_{b, k} | m_b, \hat{\tau}_k^{(s)}) = \binom{m_b}{\tilde{m}_{b, k}} \left( \frac{\hat{\tau}_k^{(s)}}{t_b^{(s)}} \right)^{\tilde{m}_{b, k}} \left( 1 - \frac{\hat{\tau}_k^{(s)}}{t_b^{(s)}} \right)^{m_b - \tilde{m}_{b, k}}. \quad (44)$$

It follows that

$$E \left[ \sum_{\ell=1}^k \tilde{m}_{b_\ell, k} | \mathbf{m}, \hat{\tau}^{(s)} \right] = \sum_{\ell=1}^k \frac{m_{b_\ell} \hat{\tau}_k^{(s)}}{t_{b_\ell}^{(s)}}. \quad (45)$$

By substituting Eq. (45) into Eq. (43), we obtain

$$\hat{\tau}_k^{(s+1)} = \frac{\sum_{\ell=1}^k m_{b_\ell} \frac{\hat{\tau}_k^{(s)}}{t_{b_\ell}^{(s)}}}{\sum_{\ell=1}^k \frac{\theta_{b_\ell}}{2} + \binom{k}{2}}. \quad (46)$$

We iterate  $\hat{\tau}^{(s)}$  until convergence using Eq. (46).

## 5 Estimating coalescence and mutation rates through time

We have developed a method for jointly estimating (cross-)coalescence rates and mutation rates through time, as well as branch lengths reflecting these estimated coalescence and mutation rates. This yields a self-contained method for inferring the demographic history and branch lengths, and should, for instance, improve age estimates of mutations. It can also be used to infer separation histories between diverged populations or track fine-scale population structure through time (see Section 5.1).

When estimating genealogies, we fit a constant population size model as described in Sections 4.2 to 4.4. In practice, we use  $2N_e = 30,000$  for humans. We then use these branch lengths as the initial

state for an iterative algorithm in which we repeatedly estimate coalescence rates and update branch lengths. This iterative algorithm proceeds as follows.

We first estimate a population-wide coalescence rate (see Section 5.1). Second, we estimate a population-wide mutation rate over time, where we divide time into epochs and calculate the quotient of the number of mutations that occurred in an epoch and the total branch length in that epoch. We use this estimated mutation rate in a heuristic step intended to speed-up convergence. In this step, we multiply the population-wide coalescence rate by the quotient calculated by dividing a predetermined constant mutation rate  $\mu$  ( $\mu = 1.25 \times 10^{-8}$  for humans) by the estimated mutation rate over time, reflecting the assumption of a constant population-wide mutation rate through time. Finally, using this rescaled coalescence rate estimate as input, we reestimate branch lengths, now under a variable population size model (see Section 5.2). We then return to the first step.

We terminate this algorithm after five iterations. To speed up convergence and computation time, we apply this algorithm only to trees with sufficiently many mutations, which in our implementation is  $N$  mutations per tree, where  $N$  is the number of haplotypes. Using the output coalescence rates, we then calculate a final estimate of the branch lengths by reestimating branch lengths for all trees for a final time.

## 5.1 Estimating the coalescence rate for a pair of haplotypes

Here, we derive an MLE for the historical coalescence rates for a pair of haplotypes, given a genealogy with estimated branch lengths. To obtain a population-wide estimate of the coalescence rate, we take the mean over all pairs of haplotypes in a population. We note that this is not the MLE for the population-wide coalescence rate assuming a panmictic population. In practice, this approach, though heuristic, allows us to avoid having to assume panmixia in this step of the algorithm. It also enables us to calculate coalescence rates for any subset of haplotypes, cross-coalescence rates between subpopulations, or track fine-scale population structure through time.

To estimate pairwise coalescence rates, we divide time into epochs. Within an epoch, we assume that the coalescence rate remains constant. For every pair of haplotypes, we estimate these piecewise constant coalescence rates using an MLE. Denote epochs by  $e = 0, \dots, E$ , where epoch  $e$  begins at time  $T_e$  and ends at time  $T_{e+1}$ . We denote the coalescence rate in epoch  $e$  by  $\gamma(e)$ .

We denote the time at which the two haplotypes coalesce in tree  $z$  by  $t_z$ . Also, we denote by  $e_z$  the index of the epoch in which the haplotypes coalesce. We therefore have

$$T_{e_z} \leq t_z < T_{e_z+1}. \quad (47)$$

Conditioning on tree topology and branch lengths, the likelihood that the two haplotypes coalesce at time  $t_z$  is given by

$$P(t_z) = \gamma(e_z) \exp[-\gamma(e_z)(t_z - T_{e_z})] \left[ \prod_{e=1}^{e_z} \exp[-\gamma(e-1)(T_e - T_{e-1})] \right]. \quad (48)$$

By taking logarithms on both sides of Eq. (48), we obtain

$$\log P(t_z) = \log \gamma(e_z) - \gamma(e_z)(t_z - T_{e_z}) - \sum_{e=1}^{e_z} \gamma(e-1)(T_e - T_{e-1}). \quad (49)$$

Assuming independence across trees, the log-likelihood for the whole genome is given by  $\sum_{z=0}^M \log P(t_z)$ , where  $M$  is the number of trees built. By differentiating with respect to  $\gamma(e)$ , we obtain an MLE given by

$$\hat{\gamma}(e) = \frac{n_e}{\sum_{z:e=e_z} (t_z - T_e) + \sum_{z:e < e_z} (T_e - T_{e-1})}, \quad (50)$$

where  $n_e$  denotes the number of trees for which the two haplotypes coalesce in epoch  $e$ .

## 5.2 Reestimating branch lengths using a coalescent prior with variable population sizes

We reestimate branch lengths using an MCMC sampling identical to that described in Section 4.2 but with a modified Eq. (36) as shown below, reflecting a coalescent prior that incorporates piecewise-constant coalescence rates. The MCMC sampler is initialised using the branch lengths of the input genealogies.

Whenever  $\tau_k$  is updated in an MCMC iteration by a proposed value  $\tilde{\tau}_k$ , all coalescence events older than this event are updated by  $\Delta\tau = \tilde{\tau}_k - \tau_k$ . Therefore, older events may now coalesce in a different time epoch to before, which we need to reflect in the acceptance probability of  $\tilde{\tau}_k$ . We modify Eq. (36), which states the acceptance probability of a proposed update  $\tilde{\tau}_k$  of  $\tau_k$ , to reflect a variable population size and obtain

$$\min \left( 1, \frac{\tilde{\tau}_k}{\tau_k} \exp \left[ -\frac{\tilde{\tau}_k}{\tau_k} + \frac{\tau_k}{\tilde{\tau}_k} \right] \prod_{\ell=1}^k \frac{P(m_{b_\ell}|\tilde{t}_{b_\ell})}{P(m_{b_\ell}|t_{b_\ell})} \prod_{m=2}^k c_m \right), \quad (51)$$

for  $c_m$  ( $m = 2, \dots, k$ ) which we will derive below.

Let us define a function  $\eta(t) \in \{0, \dots, E\}$  mapping time  $t$  to its corresponding epoch. For a piecewise constant coalescence rate as defined in Section 5.1, the time while  $\tau_k$  ancestors remain, conditional on  $(\tau_\ell)_{\ell=k+1, \dots, N}$  has density

$$f_{\tau_k|\tau_{k+1}, \dots, \tau_N} = \binom{k}{2} \gamma \left( \eta \left( \sum_{\ell=k}^N \tau_\ell \right) \right) \exp \left[ -\binom{k}{2} \int_{\sum_{\ell=k+1}^N \tau_\ell}^{\sum_{\ell=k}^N \tau_\ell} \gamma(\eta(\tau)) d\tau \right]. \quad (52)$$

It follows that the ratio of the prior probabilities of  $\tilde{\tau}_k$  and  $\tau_k$ , conditional on  $\tau_{k+1}, \dots, \tau_N$ , is given by

$$c_k = \frac{\gamma \left( \eta \left( \Delta\tau + \sum_{\ell=k}^N \tau_\ell \right) \right)}{\gamma \left( \eta \left( \sum_{\ell=k}^N \tau_\ell \right) \right)} \exp \left[ -\binom{k}{2} \int_{\sum_{\ell=k}^N \tau_\ell}^{\Delta\tau + \sum_{\ell=k}^N \tau_\ell} \gamma(\eta(\tau)) d\tau \right]. \quad (53)$$

Because we also update the times of all events older than event  $k$ , we need to calculate the ratio of prior probabilities for these events, and we obtain for  $m < k$ ,

$$c_m = \frac{\gamma \left( \eta \left( \Delta\tau + \sum_{\ell=m}^N \tau_\ell \right) \right)}{\gamma \left( \eta \left( \sum_{\ell=m}^N \tau_\ell \right) \right)} \exp \left[ -\binom{m}{2} \int_{\sum_{\ell=m}^N \tau_\ell}^{\sum_{\ell=m}^N \tau_\ell + \Delta\tau} [\gamma(\eta(\tau + \Delta\tau)) - \gamma(\eta(\tau))] d\tau \right]. \quad (54)$$

Substituting Eqs. (53) and (54) in Eq. (52), we obtain the acceptance probability of a proposed change  $\Delta\tau$  to the time while  $k$  ancestors remain. In particular, we note that if  $\gamma(e) \equiv 1$  for all epochs  $e$ , we obtain  $c_k = \exp \left[ -\binom{k}{2} \Delta\tau \right]$  and  $c_m = 1$  ( $m < k$ ), reducing Eq. (52) to Eq. (36).

## 6 A tree-based statistic for detecting positive selection

We define a tree-based statistic for detecting positive selection. Let  $f_N$  be the number of carriers of a mutation today, and  $k$  be the number of lineages when the mutation increased from frequency 1 to frequency 2. Then, under the standard coalescent model, the probability that a mutation spreads to  $f_N$  haplotypes is given by (see e.g., Ref. [9])

$$P(f_N) = \frac{(f_N - 1) \binom{N - f_N - 1}{k - 3}}{\binom{N - 1}{k - 1}}. \quad (55)$$

It follows that the probability that the mutation spreads to *at least*  $f_N$  haplotypes is given by

$$p_R = \sum_{f=f_N}^{N-k+2} P(f) = \sum_{f=f_N}^{N-k+2} \frac{(f - 1) \binom{N - f - 1}{k - 3}}{\binom{N - 1}{k - 1}}. \quad (56)$$



We reject the null hypothesis that the frequency change happened under random drift (and hence no selective pressures) if this p-value is sufficiently small. We note that this test statistic does not use branch length estimates, and relies only on the order of coalescences. For this reason, we expect this test to be robust to misspecification of population size histories.

## Bibliography

- [1] N. Li and M. Stephens. Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213–2233, 2003.
- [2] M. Bahlo and R. C. Griffiths. Inference from gene trees in a subdivided population. *Theoretical Population Biology*, 57:79–95, 2000.
- [3] M. Stephens, N. J. Smith, and P. Donnelly. A new statistical method for haplotype reconstruction from population data. *The American Journal of Human Genetics*, 68:978–989, 2001.
- [4] S. R. Browning and B. L. Browning. Haplotype phasing: existing methods and new developments. *Nature Reviews Genetics*, 12:703–714, 2011.
- [5] M. Kimura. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics*, 61:893–903, 1969.
- [6] J. G. Hacia et al. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nature Genetics*, 22:164–167, 1999.
- [7] R. Sokal and C. Michener. A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38:1409–1438, 1958.
- [8] J. F. C. Kingman. On the genealogy of large populations. *Journal of Applied Probability*, 19:27–43, 1982.
- [9] R. C. Griffiths and S. Tavaré. The age of a mutation in a general coalescent tree. *Stochastic Models*, 14:273–295, 1998.

# Supplementary Note: Simulations

We test Relate on data simulated using msprime [1] which simulates the standard coalescent with recombination [2]. In the main text, we compared accuracy attained by Relate to that of ARGweaver [3]. Here, we additionally compare to accuracy of RENT+ [4] which is a recent non-parametric method that estimates genealogies more efficiently than ARGweaver. For all methods, we use default parameters and the true mutation rates, as well as effective population size. We provide the true recombination rates for ARGweaver, which are not required for RENT+.

Whenever comparisons across methods are made, we use small sample sizes ( $N = 50$ ,  $N = 200$ ). We note that ARGweaver is computationally infeasible for larger data sets. RENT+ is faster than ARGweaver but about 30 times slower than Relate on small data sets. We tested RENT+ on a data set of  $N = 1000$  and 2.5Mb, where it ran out of memory after approximately 400 CPU hours on a computing server with 100GB of RAM.

## 1 Accuracy of TMRCAs and mutation ages

We compare the TMRCA of pairs of haplotypes to the truth (Supplementary Fig. 3a). We find that estimates using Relate appear to be unbiased, whereas ARGweaver underestimates the TMRCA for recent coalescences and RENT+ exhibits a non-linear relationship to the true TMRCA. Next, we map mutations to trees to estimate mutation ages (Supplementary Fig. 3b). We applied our code to do the same for ARGweaver and RENT+ because the output files of these methods do not specify on which branches the mutations occurred. We find that across all methods, age estimates of mutations appear to have a smaller deviation from the truth than TMRCAs. This may indicate that branches with at least one mutation have less uncertainty in their age estimates. In terms of accuracy, we observe similar trends as for TMRCAs, where our method appears to have the least bias in terms of dating mutations.

Additionally, we test Relate on a simulated data set with a variable population size history ( $N = 200$ ). We simulate 200Mb using the population size inferred for GBR using Relate. In Supplementary Fig. 3c, we estimate branch lengths using a constant population size of  $2N_e = 30,000$ . Here, we deliberately misspecified the population size history and we observe a clear bias in the estimated TMRCAs, e.g., a coalescence event dated at approximately  $10^5$  years before present may have occurred between  $10^4$  to  $2 \times 10^6$  years before present in the true trees. In Supplementary Fig. 3d, we jointly infer branch lengths and population sizes for the same data set. We observe that this corrects for most of this bias, highlighting the importance of accounting for the demographic history of the sample.

## 2 Accuracy measured using the Robinson-Foulds metric and a TMRCA metric

In addition to measuring accuracy using pairwise TMRCAs, we use two distance metrics between trees. For each distance metric, we calculate a genome-wide mean score between an estimate of the genealogy and the truth.

We compare tree topologies using the Robinson-Foulds metric  $d_{\text{RF}}$  adapted for rooted binary trees [5]. For each coalescence event, we find the set of present-day descendants, which we call a clade. We count the number of clades that exist in one tree but not the other. We then divide this number by  $4N - 2$  such that  $d_{\text{RF}} = 1$  if trees are entirely different and  $d_{\text{RF}} = 0$  if they are exactly the same. Notice that this metric is independent of branch lengths. For two uncorrelated trees drawn from the standard coalescent, the Robinson-Foulds metric equals 1 with probability 1 as  $N \rightarrow \infty$ .

To also compare the accuracy of estimated branch lengths, we define a second metric  $d_{\text{PTMRCA}}$ , in which we compare the time to the MRCA (TMRCA) of every pair of haplotypes. For each tree, we calculate a vector of lengths  $\binom{N}{2}$  containing the TMRCAs between every pair of tips. We then calculate, at every SNP, the mean squared difference between the vectors corresponding to the estimated and true trees and divide the result by the diploid effective population size  $N_e$ . We note that  $d_{\text{PTMRCA}}$  inherits its metric properties from the mean square difference together with the fact that for trees  $T_1$  and  $T_2$ , we have  $d_{\text{PTMRCA}}(T_1, T_2) = 0$  if and only if  $T_1 = T_2$ . In this metric, the expected score for two uncorrelated trees drawn at random from the standard coalescent model equals 1.

### 2.1 Impact of errors in the data set

We evaluate Relate at varying levels of errors introduced to the data. We simulate 2.5Mb and  $N = 1000$  haplotypes with  $2N_e = 30,000$ ,  $\mu = 1.25 \times 10^{-8}$ , and recombination rates taken from the 1000 Genomes Project map for chromosome 1. We then subset this dataset to  $N = 50$  haplotypes. We estimate genealogies for the same 50 haplotypes using Relate, RENT+, and ARGweaver. In addition, we estimate the genealogy for 1000 haplotypes using Relate, and extract the embedded genealogy corresponding to the same 50 haplotypes.

We introduce errors to the 50 haplotypes by first choosing a SNP and a haplotype uniformly at random and then, changing this haplotype. We find that in the absence of errors, accuracy across the three methods is similar by the  $d_{\text{RF}}$  metric, while Relate offers improvements for the  $d_{\text{PTMRCA}}$  metric. However when we introduce errors, Relate outperforms ARGweaver and RENT+ in both the Robinson-Foulds and PTMRCA metrics (Supplementary Figs. 3e and f).

### 2.2 Impact of incorrect ancestral allele estimates

We evaluate the robustness of Relate with respect to incorrect identification of ancestral and alternative alleles. We simulate  $N = 200$  haplotypes, with other parameters identical to before. We find that over 99% of SNPs can be mapped to a unique branch (Supplementary Fig. 3g). The fraction of correctly unflipped SNPs remains above 98% regardless of the fraction of flips introduced to the data. The fraction of correctly flipped SNPs decreases slightly but stays at around 85%. We note that for SNPs that map to a branch connected to the root of the tree, we cannot identify whether it is flipped or unflipped. We therefore excluded such SNPs from this analysis. In addition, we also excluded singleton mutations because these can always be mapped to a unique branch.

## 3 Estimating coalescence and mutation rates

We compare our algorithm for estimating historical coalescence and mutation rates to MSMC [6] and SMC++ [7] across different simulated population size histories. We simulate 200Mb for  $N = 200$

haplotypes and a constant mutation rate of  $1.25 \times 10^{-8}$ . The recombination rates are taken from the 1000 Genomes Project map for chromosome 1.

The command line used for MSMC is

```
msmc_2.0.0 \
  -t 6 \
  -p 1*2+15*1+1*2 \
  -o msprime.msmc2 \
  msprime.multihetsep.txt
```

where “msprime.multihetsep.txt” denotes the input filename and “msprime.msmc2” denotes the output filename. The command line used for SMC++ is

```
smc++ estimate \
  --regularization_penalty 5.0 \
  --knots 16 \
  --timepoints 35,100000 \
  1.25e-8 \
  -o analysis/ \
  out/msprime.smc.gz
```

where “analysis/” denotes the directory into which the output of SMC++ is saved and “out/msprime.smc.gz” denotes the input filename. We noticed that the accuracy of SMC++ is sensitive to the choice of parameters “regularization\_penalty” and “knots”. Our choice of these parameters is based on advice (personal communication) from the authors of Ref. [7].

We simulate three scenarios: a discrete bottleneck, with a ten-fold change in population size ranging from  $2N_e = 7,000$  to  $2N_e = 70,000$ , an increasing trend, and a decreasing trend (Supplementary Fig. 3h). We find that Relate estimates the population size with high accuracy in all three scenarios. The accuracy of MSMC depends on the number of haplotypes used, where its population size estimate is accurate up until 100,000 years before present with 2 haplotypes and 10,000 with 8 haplotypes. SMC++ has a comparable accuracy to MSMC, but detects trends in the more recent past as well.

A consequence of correctly estimated population sizes is a mutation rate that remains constant and close to  $\mu = 1.25 \times 10^{-8}$  through time. The insets of Supplementary Fig. 3h show that indeed the mutation rate stays mostly constant and close to the truth.

We note that Relate, like other previous methods, estimates *effective* population sizes which may be influenced by unmodeled complexities not accounted for the inference model [8]. To assess how well simulated data that assume panmixia and estimated effective population sizes emulate real data, we followed the analysis of Ref. [8] and compared three statistics capturing distinct aspects of variation patterns. In Supplementary Fig. 4a, we compare the expected heterozygosity, defined by

$$\pi = \frac{N}{N-1} \frac{\sum_{i=1}^L 2p_i(1-p_i)}{L}, \quad (57)$$

where  $p_i$  is the frequency of an allele,  $L$  is the total number of callable sites in the window (passing the Pilot mask for the 1000 Genomes Project data set), and  $N$  is the number of sampled haplotypes ( $N = 20$ ). We calculated  $\pi$  for 20,000 randomly chosen 100kb windows using chromosome 1 of 10 individuals (20 haplotypes). We find that panmictic simulations closely match the data, with better or equal accuracy to all methods included in Ref. [8]. In Supplementary Fig. 4b, we compare the site-frequency spectrum across all biallelic variants passing the Pilot mask on chromosome 1. We observe some discrepancies, particularly for rare variants. In Supplementary Fig. 4c, we compare how LD patterns decay with physical distance from a focal SNP. While none of the methods closely matched the data in Ref. [8], we find that our demographic estimates appears to perform reasonably well at capturing the LD decay pattern in CEU. For CHB and YRI, we appear to match the data less well, although better than methods in Ref. [8], possibly because recombination maps are biased towards European hotspots.

## 4 Simulations with perturbations from infinite-sites, constant mutation rates, or perfect phase

We conduct additional simulations to assess the robustness of Relate to perturbations from the infinite-sites assumption, constant mutation rates, as well as perfect phasing of haplotypes, which are likely to be present in real data.

Our base-line simulation scenario, to which we will add these perturbations, is a simulation of 3000Mb for 1000 haplotypes with a population size history estimated by Relate for the GBR samples in the 1000 Genomes Project data set. We use human chromosome 1 recombination rates, and assume a constant mutation rate of  $1.25 \times 10^{-8}$ . We subset 200 haplotypes used for inference and retain the remaining 800 haplotypes as a reference panel for rephasing (see below). This simulation assumes infinite-sites, such that every mutation occurs at a new base-pair position. We then add perturbations to this simulation scenario as follows.

First, we emulate the variable mutation rate observed for the triplet mutation TCC to TTC in Europeans (Figure 4). We assume that with a probability of 1/96, a base-pair position mutates with a varying rate through time and otherwise, it mutates with a constant rate of  $1.25 \times 10^{-8}$ , independently of other mutations. To achieve this, we first classify a mutation in the base-line simulation as belonging to the variable mutation rate category with probability 1/96, independently of other mutations. We then add novel mutations occurring at rate  $1/96 \times 1.25 \times 10^{-8}$  between 10,000 and 50,000 YBP. This is equivalent to assigning a mutation rate of  $2.5 \times 10^{-8}$  for this mutation category between 10,000 and 50,000 YBP, and a mutation rate of  $1.25 \times 10^{-8}$  otherwise. We find that age estimates of mutations, as well as population size estimates remain accurate (Supplementary Fig. 4 d, e). We calculate the normalised mutation rate, where we first eliminate any remaining temporal trends in the average mutation rate by dividing by the average mutation rate in each epoch. For each mutation category, we then normalise the mutation rates such that the average rate over time equals 1. The elevation in mutation rate is detected with reasonable accuracy; however, we slightly underestimate the absolute elevation in mutation rate and the activity period appears longer than the truth (Supplementary Fig. 4 f).

Second, we consider a scenario in which 1% of all base-pair positions have a 20 times higher mutation rate, emulating CpG dinucleotides in the human genome [9]. We choose these CpG-like base-pair positions uniformly at random and remove any mutation in the base-line simulation that occurred at such a CpG-like position. At these CpG-like positions, multiple mutations may occur, where the mutation rate returns to its usual rate ( $1.25 \times 10^{-8}$ ) on any lineages below the first mutation. To account for the elevated average mutation rate caused by the introduction of CpG-like mutations, we specified a mutation rate of  $1.49 \times 10^{-8}$  in Relate. We find that age estimates of mutations, as well as population size estimates remain accurate (Supplementary Fig. 4d, e). At CpG-like sites that only mutated once, > 99.5% of mutations map to a unique branch, which is identical to the mapping rate in the base-line simulation (Supplementary Fig. 4g). However, at CpG-like sites with more than one mutation, we observe a substantially reduced mapping rate of 82.4%. In particular, of CpG doubletons at sites with more than one mutation, >97% did not map to a unique branch. We note that partial mapping of doubly mutated sites is expected. For example, in a case where two mutations have 1 and 20 descendants, respectively, it is possible to map the mutation to a branch with 20 descendants from the second mutation, if such a branch exists, because our mapping of mutations to trees allows for some noise/error.

Finally, we evaluate Relate on haplotypes that have been rephased using SHAPEIT2 [10]. We rephase 200 haplotypes (100 genotypes) using the remaining 800 haplotypes as a reference panel. Switch errors complicate the matching of rephased haplotypes to true haplotypes and make comparisons of estimated and true genealogies difficult. We therefore evaluate the accuracy of Relate using the accuracy of age estimates and population size histories. Both are almost unchanged in accuracy, with a slight elevation in recent population sizes visible after rephasing (Supplementary Fig. 4 d, e). This elevation in recent population sizes can, at least in part, be explained by the random phase assigned

to singletons, since singletons are on average moved from longer lineages to shorter lineages.

## 5 Positive natural selection

We simulate positive natural selection acting on a single mutation using the pipeline outlined in Ref. [11] (Methods). To assess the accuracy of genealogies estimated by Relate at loci under selection, we calculate the Robinson-Foulds distance ( $d_{\text{RF}}$ ) between the true and estimated tree at the selected locus, as well as the ratio of estimated and true lower-end age of the branch onto which a mutation under selection maps (Supplementary Fig. 7a). We find that tree topology ( $d_{\text{RF}}$ ) remains accurate, but decreases slightly, potentially due to fewer mutations mapping underneath the selected mutation, such that tree topology is less accurate in this part of the tree. The age of a mutation under selection, approximated by the lower-end of the corresponding branch, is over-estimated when the mutation is selected, which is expected because our model assumes neutrality.

In Supplementary Fig. 7b, we plot p-values calculated with the true trees against p-values calculated using estimated trees for selection coefficients equalling 0.00 and 0.001. We observe a high correlation and a clear shift in p-values when selection is turned on.

In Supplementary Fig. 7c and d, we estimate the power of our selection test for varying present-day derived allele frequencies. We compare our method to the integrated haplotype score (iHS) [12] and singleton density score (SDS) [11]. We use selscan [13] and hapbin [14] to calculate iHS scores. In addition, we define a tree-based SDS score (trSDS) as proposed in Ref. [15]. For iHS, SDS, and trSDS, we standardise the raw scores using the frequency specific empirical mean and standard deviation for the neutral case, which is an idealised setting that should favour power estimates of both methods. We find that the statistical power of our test outperforms iHS, SDS, and trSDS when using the true trees in all considered scenarios. Using trees estimated by Relate, we attain a higher power than iHS for weak selection and a similar or slightly lower power than iHS for strong selection. While iHS is similarly powered for all present-day derived allele frequencies, the power of our test statistic increases with larger derived allele frequency. Both methods outperform SDS and trSDS, which have been designed to perform better for very recent selection on standing variation. In most cases, trSDS outperforms SDS. This is an encouraging result, suggesting that adapting existing ideas for trees can improve power. We postpone a more thorough study in this direction, including simulating scenarios that are more suited to SDS and trSDS, to future work.

## Bibliography

- [1] J. Kelleher, A. M. Etheridge, and G. McVean. Efficient coalescent simulation and genealogical analysis for large sample sizes. *PLoS Computational Biology*, 12:e1004842, 2016.
- [2] R. R. Hudson. Properties of a neutral allele model with intragenic recombination. *Theoretical Population Biology*, 23:183–201, 1983.
- [3] M. D. Rasmussen, M. J. Hubisz, I. Gronau, and A. Siepel. Genome-wide inference of ancestral recombination graphs. *PLoS Genetics*, 10:e1004342, 2014.
- [4] S. Mirzaei and Y. Wu. Rent+: an improved method for inferring local genealogical trees from haplotypes with recombination. *Bioinformatics*, 33:1021–1030, 2016.
- [5] D. F. Robinson and L. R. Foulds. Comparison of phylogenetic trees. *Mathematical Biosciences*, 53:131–147, 1981.
- [6] S. Schiffels and R. Durbin. Inferring human population size and separation history from multiple genome sequences. *Nature Genetics*, 46:919–925, 2014.

- [7] J. Terhorst, J. A. Kamm, and Y. S. Song. Robust and scalable inference of population history from hundreds of unphased whole-genomes. *Nature Genetics*, 49:303309, 2017.
- [8] A. C. Beichman, T. N. Phung, and K. E. Lohmueller. Comparison of single genome and allele frequency data reveals discordant demographic histories. *G3: Genes, Genomes, Genetics*, 7: 3605–3620, 2017.
- [9] A. Bird. DNA methylation patterns and epigenetic memory. *Genes & development*, 16:6–21, 2002.
- [10] O. Delaneau, J. Marchini, and J.-F. Zagury. A linear complexity phasing method for thousands of genomes. *Nature Methods*, 9:179–181, 2012.
- [11] Y. Field et al. Detection of human adaptation during the past 2000 years. *Science*, 354:760–764, 2016.
- [12] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4:e72, 2006.
- [13] Z. A. Szpiech and R. D. Hernandez. selscan: An efficient multithreaded program to perform eh-h-based scans for positive selection. *Molecular Biology and Evolution*, 31:2824–2827, 2014.
- [14] C. A. Maclean, Neil P. Chue H., and J. G. D. Prendergast. hapbin: An efficient program for performing haplotype-based scans for positive selection in large genomic datasets. *Molecular Biology and Evolution*, 32:3027–3029, 2015.
- [15] M. Edge and G. Coop. Reconstructing the history of polygenic scores using coalescent trees. *bioRxiv*: 389221, 2018.

# Supplementary Note:

## 1000 Genomes Project data set

### 1 Runtime

Relate terminated after less than 2 CPU years, or 4 days when run on a high-performance cluster, using up to 300 cores (see Supplementary Table [1](#)). Each core had a maximum memory allowance of 16GB and was equipped with an Intel Ivybridge 2.4 GHz or Intel Haswell 2.6 GHz processor.

### 2 Number of trees built

As a first indication of good accuracy of the inferred genealogy, we find a high correlation between the number of trees built and the recombination distance in bins of  $10^5$  SNPs (Supplementary Fig. 5a). In a low recombination rate region, we estimate that we construct approximately one tree every 125 SNPs. The number of trees we build is slightly inflated by the fact that in our implementation of Relate, we rebuild a tree after 200 - 1000 SNPs for computational reasons. Consequently, not every new tree represents a recombination event. We also find that the mean number of SNPs mapping onto a tree increases with reducing local recombination distance (Supplementary Fig. 5b).

### 3 CpG mutations map less frequently than other mutations

We next investigate SNPs that could not be mapped to a unique branch. Singleton mutations always map to a unique branch, such that we exclude singletons from the following analysis. For any non-singleton, 14.3 % of SNPs do not map to a unique branch (Supplementary Fig. 5c). SNPs with a small derived allele frequency are particularly susceptible to errors in the data. How the derived allele frequency of a mutation affects the fraction of non-mapping SNPs is shown in Supplementary Fig. 5f. We can see that the fraction of non-mapping SNPs is relatively high for rare SNPs and rapidly decreases to around 5% for more frequent SNPs. If we exclude any SNP with a derived allele frequency of less or equal to 10, only 4 % of SNPs do not map to a unique branch (Supplementary Fig. 5c).

We further notice that CpG sites, which are sites at which a C nucleotide is followed by a G nucleotide, are known to have a significantly higher rate of mutations  $CG \rightarrow TG$  [1](#). We should therefore expect a higher probability of observing two or more mutations at the same genomic position. Such SNPs usually cannot be mapped to a unique branch. Indeed, we observe that the fraction of non-mapping CpG transitions is 24.3 % which is two-fold higher than the overall average.

To further study the effect of adjacent nucleotides, we group mutations by their two neighbouring nucleotides in sequence. This yields 96 categories after accounting for equivalent mutations due to symmetry on the complementary strands. We observe that the fraction of non-mapping SNPs is highly variable with respect to the mutation category (Supplementary Fig. 5e). In some categories, we observe a higher fraction of non-mapping SNPs resembling those of CpG transitions. We observe a clear elevation in the fraction of non-mapping mutation for  $GT \rightarrow GG$  mutations, which have recently



been suggested to be prone to sequencing artefacts in a subset of 1000 Genomes Project samples [2]. Currently, we cannot explain spikes in other categories, such as ATA → AAA, or TTA → TAA.

Mutations that occur at the same genomic position both in humans and other primates can cause confusion of the ancestral and derived alleles which we can detect as flipped SNPs. The fraction of flipped SNPs appears to be less dependent on the mutation category (Supplementary Fig. 5e). However, we again find a clear signal for CpG transitions  $CG \rightarrow TG$ . This is likely to be caused by a mutation at the same genomic position in humans and other primates, leading to an error in estimating the ancestral allele. For other categories, the fraction of flipped SNPs is approximately 1%. The fraction of flipped SNPs is moderately dependent on the frequency of the alternative allele in the sample (Supplementary Fig. 5g).

## 4 Historical population sizes of 26 populations

We estimate historical population sizes for all 26 populations in the 1000 Genomes Project data set (Supplementary Fig. 6a). All non-African groups show a severe bottleneck following their out-of-Africa migration. As already discussed in the main text, we observe a second bottleneck in the Finnish (FIN) population around 3,000 to 9,000 YBP. Indications of a second bottleneck can also be observed in Gujarati individuals (GIH) and a severe second bottleneck can be observed in the Peruvian population (PEL).

## 5 Evolution of triplet mutation rates

We estimate the evolution of all 96 triplet mutation rates (see **Methods**, Supplementary Fig. 6d) and uncover continent-level differences. The strongest signal is observed for the TCC to TTC category (Figure 4a), with similar but weaker signatures for ACC to ATC and TCT to TTT. Other notable trends include a recent increased mutation rate for African populations of GCA to GGA and GCT to GGT. We note that the apparent decrease in mutation rate for categories involving a CpG dinucleotide is likely an artefact, caused by many rare mutations at these sites being not mappable to the tree due to repeat mutations.

GC-biased gene conversion biases the conversion rate of heterozygotes during repair of double-strand breaks towards C or G nucleotides. This effect leads to a faster-than-expected spread of mutations towards C and G. A consequence is that mutations towards C or G are overrepresented, whereas mutations from C and G are underrepresented in the past. Consistent with this idea, we observe that mutation rates decrease towards the present in T to C and T to G mutations, whereas we observe a small increasing trend in C to A and C to T mutations.

## 6 Positive selection

### 6.1 Genome-wide significant hits for positive selection

Because there is no guarantee that the mutation with the lowest p-value is the causal variant, we determine genomic regions that should contain the causal variant. In Supplementary Table 3, we list all regions in the genome that contain genome-wide significant hits ( $p < 5 \times 10^{-8}$ ) in at least three populations.

To determine genomic regions under selection, we cluster genome-wide significant hits into blocks, such that any two SNPs in different blocks have a Pearson correlation coefficient  $r^2 < 0.5$ , and for any SNP, there is always another SNP in the same block with  $r^2 \geq 0.5$ . We then extend these identified genomic regions, by including any SNP within 2Mb that have an  $r^2 \geq 0.5$  to any of the genome-wide significant hits in a block.

We annotate each region using genes, eQTLs, GWAS hits, and non-synonymous substitutions at SNPs with highest  $r^2$  to the listed mutation. We combined the genome-wide significant GWAS hits ( $p < 5 \times 10^{-8}$ ) of the GWAS catalogue [3] and the UK Biobank project [4]. We used the eQTL annotation provided by GTEx [5] (see URLs).

In Supplementary Table 3, we list all identified regions, where we choose the SNP with the lowest p-value in that genomic region for each population. We observe that all considered geographic regions (AFR, EUR, SAS, EAS) are represented in this table. With the exception of 6 identified genomic regions, populations come from same geographic region in each identified genomic region. To the best of our knowledge, 13 out of 35 regions have been reported previously. We additionally record the statistic that attains a value in the 0.05% tail of its empirical distribution in the 1000 Genomes Project dataset according to the PopHumanScan resource [6] (URLs), whenever the region reported in PopHumanScan overlaps the region listed in this table and is attributed to a population listed in this table. Including such regions, our method detects 18 previously reported regions. Of all unreported regions, 12 are attributed only to African populations, and 5 unreported regions are attributed to non-Africans.

In 9 regions, we find an eQTL and in 8 regions, we find a GWAS hit within  $r^2 \geq 0.8$  of the SNP with the lowest p-value for selection evidence. As discussed in the main text, two well known regions of positive selection, EDAR and LCT, each fall into one identified genomic regions. While the causal variants are not those with the lowest p-value for selection evidence, they are in  $r^2 \geq 0.8$  in both cases. Moreover, we identify a previously unreported hit in the EDARADD gene, which is known to directly interact with the EDAR protein [7], to be selected in all Southern Asian populations, as well as the Finnish population. This mutation achieves a selection p-value of less than  $10^{-6}$  in all European populations.

## Bibliography

- [1] J. Sved and A. Bird. The expected equilibrium of the CpG dinucleotide in vertebrate genomes under a mutation model. *Proceedings of the National Academy of Sciences of the United States of America*, 87:4692–4696, 1990.
- [2] L. Anderson-Trocmé, R. Farouni, M. Bourgey, Y. Kamatani, K. Higasa, J. Seo, C. Kim, F. Matsuda, and S. Gravel. Legacy data confounds genomics studies. *bioRxiv:624908*, 2019.
- [3] J. MacArthur et al. The new NHGRI-EBI Catalog of published genome-wide association studies (GWAS Catalog). *Nucleic Acids Research*, 45:D896–D901, 2016.
- [4] C. Bycroft et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature*, 562: 203–209, 2018.
- [5] J. Lonsdale et al. The genotype-tissue expression (GTEx) project. *Nature Genetics*, 45:580–585, 2013.
- [6] A. Bodeln, J. Murga-Moreno, M. Coronado-Zamora, A. Barbadilla, and S. Casillas. PopHumanScan: the online catalog of human genome adaptation. *Nucleic Acids Research*, 47:D1080–D1089, 2018.
- [7] A. Sadier, L. Viriot, S. Pantalacci, and V. Laudet. The ectodysplasin pathway: from diseases to adaptations. *Trends in Genetics*, 30:24–31, 2014.

# Supplementary Note:

## Interpretation of polygenic selection

Identifying the direction of selection is sometimes challenging, with a number of traits showing selection evidence in both effect directions within the same group (**Figure 6**). To aid interpretability of these results, we additionally tested in which direction, if any, DAFs of associated SNPs are increased. For many traits with selection evidence in one effect direction only, such as platelet-related traits, the direction of selection inferred using our polygenic selection test and the direction of DAF increase align.

Interestingly, for some traits related to white blood cells (WBC), we observe differences between the frequency conditioned selection signals and shift in DAF (**Figure 6b**). For instance, we detect a signal towards increased granulocyte counts in African populations, but decreasing counts in some European and South Asian populations. While DAFs are strongly different ( $p < 0.003$ ; one-sided Wilcoxon test) and in agreement with the inferred direction of selection in African groups, DAFs remain slightly lower for SNPs associated with decreasing granulocyte counts even in Europe and South Asia. ( $p < 0.09$  for EUR,  $p < 0.12$  for SAS; one-sided Wilcoxon test). However, these DAFs for granulocyte-decreasing variants are increased relative to those in African groups ( $p < 0.0013$ ; one-sided paired Wilcoxon test), while DAFs for granulocyte-increasing variants were not significantly different ( $p > 0.4$ ; two-sided paired Wilcoxon test), so a possible resolution is that this trait (or a related trait) was selected to increase in the past, and has more recently been selected to decrease in some non-African groups.

Another important issue is assigning selection to specific phenotypes. This remains challenging for multiple reasons (see e.g., Ref. [1]). For example, a directional signal might be partly driven by selection on other phenotypes correlated to those studied. Moreover, even if mutations e.g. increasing WBC counts have been generally favoured in a group, this does not imply that WBC count itself has increased evolutionarily; if for example a selective sweep has fixed a single SNP of major effect on this phenotype (such as Duffy negativity in Africa, associated both with malaria resistance and decreased WBC count [2]), then selection might be acting on other SNPs to compensate this change. Environmental influences might have similar impacts.

Differences between populations must also be interpreted carefully: aside from impacts of demographic history, most human GWASs to date have been conducted in European populations, so that recently arisen phenotype-influencing mutations in other groups might not have been observed, reducing power in those populations.

Finally, we note that we only utilise the direction of association signals in testing for selection evidence, and test derived mutations, in order to increase robustness to residual population stratification still present in a GWAS, even after attempts to correct for such stratification. We believe that this is likely to resolve the most serious known issues, except in a setting where residual stratification (which can correlate with selection evidence [3]) improves power to observe effects that are genome-wide significant in one direction vs. another. Implicit in our approach is the idea that stratification issues are relatively far weaker for potentially genome-wide significant SNPs (of relatively large effect size) compared to directly using effect size estimates - which may be comparable to the strength of bias across many or all SNPs genome-wide.

## Bibliography

- [1] J. Novembre and N. H. Barton. Tread lightly interpreting polygenic tests of selection. *Genetics*, 208:1351–1355, 2018.
- [2] R. E. Howes et al. The global distribution of the duffy blood group. *Nature Communications*, 2:266, 2011.
- [3] A. L. Price, N. A. Zaitlen, D. Reich, and N. Patterson. New approaches to population stratification in genome-wide association studies. *Nature Reviews Genetics*, 11:459463, 2010.

**Supplementary Table 1: Runtime of Relate on the 1000 Genomes Project data set.**  
CPU time spent in days (second column) and maximum memory usage in Mb (third column) in each stage of the algorithm applied to 4956 haplotypes of the 1000 Genomes Project dataset.

	CPU days	Max. memory usage (Mb)
Modified Li-and-Stephens HMM	6.9	2,802
Tree building	132.2	14,632
Finding Equivalent branches	1.0	7,435
Estimating branch lengths	482.2	83
Other	0.2	6310
Total	622.4	

Supplementary Table 2: Number of 1000 Genomes Project samples used in our analysis by population label.

ACB	ASW	BEB	CDX	CEU	CHB	CHS	CLM	ESN	FIN	GBR	GIH	GWD
95	60	85	92	98	102	104	93	98	98	90	102	112
IBS	ITU	JPT	KHV	LWK	MSL	MXL	PEL	PJL	PUR	STU	TSI	YRI
106	101	103	98	98	84	63	84	95	103	101	106	107

AMR	(Americas)	EAS	(East Asians)	AFR	(Africans)
CLM	Colombian in Medellin, Colombia	CDX	Chinese Dai in Xishuangbanna, China	ACB	African Caribbean in Barbados
MXL	Mexican Ancestry in Los Angeles, CA, USA	CHB	Han Chinese in Beijing, China	ASW	African Ancestry in Southwest US
PEL	Peruvian in Lima, Peru	CHS	Southern Han Chinese, China	ESN	Esan in Nigeria
PUR	Puerto Rican in Puerto Rico	JPT	Japanese in Tokyo, Japan	GWD	Gambian in Western Division, The Gambia
		KHV	Kinh in Ho Chi Minh City, Vietnam	LWK	Luhya in Webuye, Kenya
				MSL	Mende in Sierra Leone
SAS	(Southern Asians)	EUR	(Europeans)	YRI	Yoruba in Ibadan, Nigeria
BEB	Bengali in Bangladesh	CEU	Utah residents with Northern and Western European ancestry		
GIH	Gujarati Indian in Houston, TX, USA	IBS	Iberian populations in Spain		
ITU	Indian Telugu in the UK	FIN	Finnish in Finland		
PJL	Punjabi in Lahore, Pakistan	GBR	British in England and Scotland		
STU	Sri Lankan Tamil in the UK	TSI	Toscani in Italy		

**Supplementary Table 3: Genome-wide significant hits for positive selection.**

Regions containing a SNP with p-value for selection evidence ( $p_R$ ) of less than  $5 \times 10^{-8}$  in at least three populations. We list genes, eQTLs, and GWAS at mutations with  $r^2 \geq 0.5$ , where a bold font corresponds to  $r^2 = 1.0$ , a plain font corresponds to  $r^2 \geq 0.8$ , and brackets correspond to  $r^2 \geq 0.5$ . If a non-synonymous mutation falls within  $r^2 \geq 0.5$ , we indicate this similarly in the NSM column. BP denotes the base-pair position of the SNP (GRCh37). In the notes column, we indicate whether this region has been highlighted in a previous study. Additionally, we record the statistic that attains a significant value according to the PopHumanScan resource, whenever the region reported in PopHumanScan overlaps the region listed in this table and is attributed to a population listed in this table. See Supplementary Note: 1000 Genomes Project data set, Section [6.1](#) for details

**GWAS catalogue phenotypes:**

- a Helix rolling
- b Cholesterol, total+;Blood metabolite levels-
- c Nonsyndromic cleft lip with cleft palate
- d Nonsyndromic cleft lip with cleft palate
- e Glaucoma (primary open-angle)

**UK BIOBANK phenotypes: BIOBANK phenotypes:**

1. Forced vital capacity (FVC)-;Monocyte percentage+;Place of birth in UK - east co-ordinate-;Place of birth in UK - north co-ordinate+
2. Arm fat mass (left)+;Arm fat mass (right)+;Body fat percentage+;Forced vital capacity (FVC)-;Leg fat mass (left)+;Leg fat mass (right)+;Monocyte percentage+;Place of birth in UK - east co-ordinate-;Place of birth in UK - north co-ordinate+;Trunk fat mass+;Trunk fat percentage+;Whole body fat mass+
3. General happiness with own health+;High light scatter reticulocyte count-;High light scatter reticulocyte percentage-;Immature reticulocyte fraction-;Impedance of arm (right)+;Impedance of leg (left)+;Impedance of leg (right)+;Impedance of whole body+;Reticulocyte count-;Reticulocyte percentage-;Standing height+
4. Standing height-
5. White blood cell (leukocyte) count-
6. Impedance of leg (right)-;Impedance of whole body-;Red blood cell (erythrocyte) distribution width+
7. Leg fat-free mass (right)+;Leg predicted mass (right)+;Red blood cell (erythrocyte) distribution width+;Trunk fat-free mass+;Trunk predicted mass+;Whole body fat-free mass+;Whole body water mass+
8. Birth weight-;High light scatter reticulocyte count+;High light scatter reticulocyte percentage+;Nervous feelings+;Reticulocyte count+;Reticulocyte percentage+;Systolic blood pressure, automated reading+
9. Hair colour (natural, before greying): Blonde-;Hair colour (natural, before greying): Dark brown+
10. Hair colour (natural, before greying): Blonde-;Hair colour (natural, before greying): Dark brown+
11. 3mm strong meridian (left)+;3mm weak meridian (left)+;6mm strong meridian (left)+;6mm weak meridian (left)+;Age high blood pressure diagnosed-;Arm fat-free mass (left)-;Arm fat-free mass (right)-;Arm predicted mass (left)-;Arm predicted mass (right)-;Basal metabolic rate-;Basophil count+;Birth weight of first child-;Birth weight-;Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor: Hayfever, allergic rhinitis or eczema-;Blood clot, DVT, bronchitis, emphysema, asthma, rhinitis, eczema, allergy diagnosed by doctor: None of the above+;Comparative height size at age 10-;Coronary atherosclerosis+;Diagnoses

- main ICD10: I21 Acute myocardial infarction+;Diastolic blood pressure, automated reading+;Diseases of the circulatory system+;Duration to first press of snap-button in each round+;Eosinophill count+;Eosinophill percentage+;Ever smoked+;Haematocrit percentage+;Haemoglobin concentration+;High light scatter reticulocyte count+;High light scatter reticulocyte percentage+;Hip circumference-;Illnesses of father: Heart disease+;Illnesses of siblings: High blood pressure+;Illnesses of siblings: None of the above (group 1)-;Immature reticulocyte fraction+;Impedance of arm (left)+;Impedance of arm (right)+;Impedance of leg (right)+;Impedance of whole body+;Ischaemic heart disease, wide definition+;Leg fat-free mass (left)-;Leg fat-free mass (right)-;Leg predicted mass (left)-;Leg predicted mass (right)-;Long-standing illness, disability or infirmity+;Lymphocyte count+;Lymphocyte percentage+;Major coronary heart disease event excluding revascularizations+;Major coronary heart disease event+;Mean corpuscular haemoglobin+;Mean corpuscular volume+;Mean spheroid cell volume+;Mean time to correctly identify matches+;Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: Blood pressure medication+;Medication for cholesterol, blood pressure, diabetes, or take exogenous hormones: None of the above-;Medication for cholesterol, blood pressure or diabetes: Blood pressure medication+;Medication for cholesterol, blood pressure or diabetes: None of the above-;Monocyte count+;Myocardial infarction+;Myocardial infarction, strict+;Neutrophill count+;Neutrophill percentage+;Non-cancer illness code, self-reported: hypertension+;Non-cancer illness code, self-reported: psoriasis+;Number of self-reported non-cancer illnesses+;Past tobacco smoking+;Platelet count+;Platelet crit+;Platelet distribution width+;Red blood cell (erythrocyte) count+;Reticulocyte count+;Reticulocyte percentage+;Smoking status: Never-;Smoking status: Previous+;Standing height+;Systolic blood pressure, automated reading+;Taking other prescription medications+;Trunk fat-free mass+;Trunk predicted mass+;Weight+;White blood cell (leukocyte) count+;Whole body fat-free mass+;Whole body water mass-
- 12. Comparative body size at age 10-;Impedance of leg (left)+;Impedance of leg (right)+;Lymphocyte percentage-;Monocyte count+;Monocyte percentage+;Neutrophill count+;White blood cell (leukocyte) count+
- 13. Arm fat mass (left)-;Arm fat mass (right)-;Arm fat percentage (right)-;Basal metabolic rate-;Body mass index (BMI)-;Comparative body size at age 10-;Duration to first press of snap-button in each round+;Eosinophill percentage+;Haematocrit percentage+;Haemoglobin concentration-;High light scatter reticulocyte count-;High light scatter reticulocyte percentage-;Impedance of leg (left)+;Leg fat-free mass (left)-;Leg fat-free mass (right)-;Leg fat mass (left)-;Leg fat mass (right)-;Leg predicted mass (left)-;Leg predicted mass (right)-;Lymphocyte percentage+;Mean corpuscular haemoglobin+;Mean corpuscular volume+;Mean platelet (thrombocyte) volume+;Mean reticulocyte volume+;Mean spheroid cell volume+;Mean time to correctly identify matches+;Nap during day+;Neutrophill count-;Neutrophill percentage-;Platelet count-;Red blood cell (erythrocyte) count-;Red blood cell (erythrocyte) distribution width-;Reticulocyte count-;Waist circumference-;Weight-;Whole body fat mass-
- 14. Arm fat mass (right)-;Arm fat percentage (right)-;Body mass index (BMI)-;Comparative body size at age 10-;Duration to first press of snap-button in each round+;Eosinophill percentage+;Haematocrit percentage+;Haemoglobin concentration-;High light scatter reticulocyte count-;High light scatter reticulocyte percentage-;Lymphocyte percentage+;Mean corpuscular haemoglobin+;Mean corpuscular volume+;Mean platelet (thrombocyte) volume+;Mean reticulocyte volume+;Mean spheroid cell volume+;Mean time to correctly identify matches+;Nap during day+;Neutrophill count-;Neutrophill percentage-;Platelet count-;Red blood cell (erythrocyte) count-;Red blood cell (erythrocyte) distribution width-;Reticulocyte count-;Reticulocyte percentage-;Sitting height+;Waist circumference-;White blood cell (leukocyte) count-
- 15. Mean platelet (thrombocyte) volume-;Platelet count+;Platelet distribution width-



CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
1:76.1-76.4	BP76111796	KHV (EAS) ITU (SAS) STU (SAS)	$6.8 \times 10^{-10}$ $1.2 \times 10^{-8}$ $3.5 \times 10^{-9}$	SLC44A5	(MSH4)			0.03	0.21	0.39	0.63	<a href="#">[1]</a> ; Fay & Wu's H
1:236.5-236.6	rs76420343	FIN (EUR) BEB (SAS) GIH (SAS) ITU (SAS) PJI (SAS)	$1 \times 10^{-9}$ $1.4 \times 10^{-8}$ $2.6 \times 10^{-9}$ $4.2 \times 10^{-8}$ $4.8 \times 10^{-8}$	<b>EDARADD</b>	EDARADD			0.08	0.52	0.38	0.22	The EDARADD protein is known to directly interact with the EDAR protein <a href="#">[2]</a> .
	rs79881482	STU (SAS)	$3 \times 10^{-9}$	<b>EDARADD</b>	EDARADD			0.06	0.52	0.38	0.22	
2:108.9-109.6	rs11123695	CDX (EAS) CHB (EAS) CHS (EAS) KHV (EAS)	$7.4 \times 10^{-10}$ $1.2 \times 10^{-12}$ $5.7 \times 10^{-12}$ $8 \times 10^{-12}$	<b>GCC2</b>	(GCC2)	a	EDAR	0	0.01	0.01	0.85	<a href="#">[1]</a> <a href="#">[3]</a> ; iHS, XPEHH
2:135.6-136.9	rs6730157 rs1375131 rs56369224	FIN (EUR) GBR (EUR) CEU (EUR)	$1.4 \times 10^{-8}$ $4.3 \times 10^{-10}$ $3.3 \times 10^{-9}$	<b>RAB3GAP1</b> <b>ZRANB3</b> <b>R3HDM1</b>	<b>MCM6</b> <b>MCM6</b> MCM6	<b>1</b> ,b <b>1</b> ,b <b>2</b> ,b	(RAB3GAP1) (RAB3GAP1) (RAB3GAP1)	0 0 0	0.49 0.49 0.49	0.12 0.12 0.12	0 0 0	<a href="#">[3]</a> <a href="#">[1]</a> <a href="#">[4]</a> <a href="#">[5]</a> <a href="#">[6]</a> <a href="#">[7]</a> iHS, XPEHH Fu & Li's D, $\alpha$
2:168.4-168.5	rs150960584	CEU (EUR) GBR (EUR) IBS (EUR) TSI (EUR)	$1.7 \times 10^{-9}$ $2.8 \times 10^{-8}$ $7.3 \times 10^{-9}$ $2.5 \times 10^{-9}$	<b>U7</b>				0.02	0.67	0.2	0.04	—
3:48.7-50.5	rs139083518 rs201632611	KHV (EAS) CDX (EAS) CHS (EAS)	$3.6 \times 10^{-9}$ $3 \times 10^{-8}$ $1.3 \times 10^{-11}$	<b>DAG1</b> <b>GNAI2</b>	(NAT6) NAT6	(3)	(C3orf45)	0 0	0.01 0.01	0.02 0.01	0.6 0.56	<a href="#">[7]</a> <a href="#">[1]</a>
4:107.6-107.8	rs1364808 rs817146 rs704049 rs3111706	ESN (AFR) MSL (AFR) LWK (AFR) GWD (AFR)	$3.1 \times 10^{-8}$ $9.7 \times 10^{-9}$ $5 \times 10^{-10}$ $1 \times 10^{-8}$	<b>DKK2</b> <b>DKK2</b> <b>DKK2</b> <b>DKK2</b>				0.86 0.93 0.92 0.93	0.87 0.87 0.87 0.87	0.95 0.95 0.95 0.96	1 1 1 1	<a href="#">[1]</a> ; Fu & Li's D, Fu & Li's F

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
4:107.8-108	rs6829139 rs17037205	YRI (AFR) LWK (AFR) MSL (AFR)	$7.8 \times 10^{-10}$ $1.3 \times 10^{-8}$ $8.7 \times 10^{-9}$	<b>DKK2</b> <b>DKK2</b>				0.9 0.9	0.97 0.97	0.99 0.99	1 1	in close physical proximity with previous region
5:65.2-65.3	rs59755544	GIH (SAS) PJL (SAS) STU (SAS)	$1.7 \times 10^{-8}$ $9.3 \times 10^{-10}$ $3.5 \times 10^{-8}$	<b>ERBB2IP</b>				0.44	0.85	0.86	0.77	–
5:178.2-178.3	BP178258803	TSI (EUR) BEB (SAS) ITU (SAS)	$2.5 \times 10^{-9}$ $1.3 \times 10^{-9}$ $1.7 \times 10^{-9}$	RP11-281O15.3	<b>AACSP1</b>			0.13	0.28	0.32	0.25	Fu & Li's D
7:19.5-19.6	rs71530658	CEU (EUR) GBR (EUR) IBS (EUR) TSI (EUR)	$1.4 \times 10^{-8}$ $1.3 \times 10^{-10}$ $2.8 \times 10^{-8}$ $2.8 \times 10^{-9}$	<b>AC007091.1</b>		(4)		0.03	0.42	0.15	0.01	$F_{ST}$
7:98.9-99.1	BP98971118 rs10229886 BP98971173 BP98971260	GWD (AFR) ESN (AFR) YRI (AFR) LWK (AFR)	$1.2 \times 10^{-12}$ $9.2 \times 10^{-9}$ $1.1 \times 10^{-10}$ $2.1 \times 10^{-13}$	ARPC1A <b>ARPC1A</b> ARPC1A (ARPC1A)	ARPC1B (ARPC1B) (ARPC1B) (GS1-259H13.2)	(5) (5) (5) (5)		0.5 0.37 0.37 0.37	0.07 0.03 0.03 0.03	0.06 0.02 0.02 0.04	0 0 0 0	–
8:38-38.3	rs59911155 rs3739252	ESN (AFR) GWD (AFR) LWK (AFR)	$4.3 \times 10^{-8}$ $3.4 \times 10^{-8}$ $4.2 \times 10^{-9}$	<b>LSM1</b> <b>DDHD2</b>	<b>DDHD2</b> <b>DDHD2</b>	<b>6,(c)</b> <b>7,d</b>	(DDHD2) (DDHD2)	0.75 0.78	0.74 0.74	0.91 0.91	0.68 0.68	–
9:99.4-99.5	BP99411763 BP99411801	YRI (AFR) GWD (AFR) MSL (AFR)	$4.2 \times 10^{-8}$ $5.4 \times 10^{-9}$ $1.2 \times 10^{-9}$					0.29 0.33	0 0	0 0	0 0	–
9:107.3-107.4	rs112315552	ESN (AFR) GWD (AFR) MSL (AFR) YRI (AFR)	$3.8 \times 10^{-8}$ $2.4 \times 10^{-10}$ $2.1 \times 10^{-9}$ $2.3 \times 10^{-9}$	<b>OR13C5</b>	NIPSNAP3A		OR13C2	0.6	0.16	0.38	0.55	–

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
10:0.4-0.6	rs77347335	ESN (AFR) GWD (AFR) LWK (AFR) MSL (AFR)	$2.4 \times 10^{-10}$ $2.5 \times 10^{-8}$ $1.1 \times 10^{-11}$ $4 \times 10^{-8}$	<b>DIP2C</b>				0.43	0.01	0.03	0.02	–
10:104.6-105	rs11191469	LWK (AFR) MSL (AFR) YRI (AFR)	$8.4 \times 10^{-9}$ $6.8 \times 10^{-10}$ $5.6 \times 10^{-9}$	<b>CNNM2</b>	C10orf32	<b>8</b>		0.66	0.38	0.25	0.27	<a href="#">[7]</a>
10:117.2-117.5	rs150028049	CHB (EAS) CHS (EAS) KHV (EAS)	$4.3 \times 10^{-8}$ $1.9 \times 10^{-12}$ $1.2 \times 10^{-8}$	<b>ATRNL1</b>				0.1	0.23	0.34	0.62	$F_{ST}$
10:122.8-123	rs2246730 rs1873446 rs10886862	IBS (EUR) ITU (SAS) STU (SAS)	$3.8 \times 10^{-8}$ $1.7 \times 10^{-8}$ $4.8 \times 10^{-8}$	<b>RP11-159H3.2</b> <b>RP11-159H3.2</b> <b>RP11-159H3.2</b>				0.39 0.37 0.42	0.97 0.97 0.97	0.96 0.95 0.95	0.65 0.64 0.61	Fay & Wu's H
11:65.1-65.2	rs188162087	BEB (SAS) GIH (SAS) PJL (SAS)	$1.1 \times 10^{-8}$ $1.7 \times 10^{-11}$ $1.1 \times 10^{-8}$	<b>DPF2</b>	(AP003068.18)			0	0.29	0.24	0.12	–
11:91.8-92	rs11019805	CHS (EAS) BEB (SAS) ITU (SAS)	$3.7 \times 10^{-8}$ $6.5 \times 10^{-9}$ $4.1 \times 10^{-8}$	<b>FAT3</b>	(FAT3)			0.08	0.31	0.54	0.59	<a href="#">[6]</a> for GBR
12:79.7-80.2	rs10778678 rs12316084 rs7306681	LWK (AFR) YRI (AFR) MSL (AFR)	$3.6 \times 10^{-9}$ $7.3 \times 10^{-9}$ $2.5 \times 10^{-8}$	<b>RP11-359M6.1</b> <b>PAWR</b> <b>PAWR</b>	<b>RP11-530C5.2</b> RP11-530C5.2 <b>RP11-530C5.2</b>			0.72 0.54 0.59	0.01 0 0	0.08 0 0	0.21 0.01 0.01	<a href="#">[3]</a> <a href="#">[1]</a> ; iHS
12:83-83.1	rs11115333	ESN (AFR) GWD (AFR) LWK (AFR)	$3.9 \times 10^{-10}$ $9.2 \times 10^{-10}$ $6.9 \times 10^{-9}$	<b>TMTC2</b>				0.8	0.77	0.81	0.71	–
12:87.3-87.4	rs11104181 rs11503304 rs2406741 rs7309012	GWD (AFR) ESN (AFR) LWK (AFR) MSL (AFR)	$4 \times 10^{-8}$ $4.8 \times 10^{-9}$ $1.8 \times 10^{-10}$ $7.4 \times 10^{-9}$	<b>RP11-202H2.1</b> <b>RP11-202H2.1</b> <b>RP11-202H2.1</b> <b>RP11-202H2.1</b>		(9) (9) <b>10</b>		0.87 0.82 0.85 0.84	0.43 0.66 0.66 0.67	0.48 0.54 0.54 0.52	0.77 0.78 0.78 0.61	<a href="#">[7]</a> (KITLG for CEU)

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
12:111.7-113	rs7137828	GBR (EUR) IBS (EUR) TSI (EUR)	$3.3 \times 10^{-11}$ $1.2 \times 10^{-9}$ $9.9 \times 10^{-11}$	<b>ATXN2</b>	ALDH2	<b>11,e</b>	SH2B3	0	0.46	0.07	0	<a href="#">5</a> ; iHS
13:28.6-28.7	rs9554250	CDX (EAS) CHB (EAS) CHS (EAS) KHV (EAS)	$7.1 \times 10^{-10}$ $1.9 \times 10^{-10}$ $9.2 \times 10^{-11}$ $4.1 \times 10^{-10}$	<b>FLT3</b>	<b>FLT3</b>	<b>12</b>		0.28	0.55	0.45	0.7	–
14:32.9-33	rs7153204 rs10138310 rs11628486	MSL (AFR) LWK (AFR) GWD (AFR) YRI (AFR)	$5.6 \times 10^{-9}$ $1.5 \times 10^{-8}$ $1.9 \times 10^{-8}$ $1.4 \times 10^{-8}$	<b>AKAP6</b> <b>AKAP6</b> <b>AKAP6</b>	(AKAP6) (AKAP6) (AKAP6)			0.88 0.86 0.86	0.08 0.1 0.09	0.12 0.21 0.2	0.29 0.3 0.26	–
16:22.9-23.1	rs16974808 rs1604799 rs8063811	YRI (AFR) ESN (AFR) MSL (AFR)	$4.8 \times 10^{-9}$ $3.1 \times 10^{-8}$ $9.5 \times 10^{-9}$	<b>HS3ST2</b> <b>HS3ST2</b> <b>RP11-20G6.2</b>				0.79 0.79 0.79	0.06 0.06 0.06	0.11 0.11 0.11	0.01 0.01 0.01	<a href="#">1</a>
16:59.5-59.7	rs9929021 rs9937266	ESN (AFR) GWD (AFR) YRI (AFR)	$3.1 \times 10^{-8}$ $2.1 \times 10^{-8}$ $4.2 \times 10^{-8}$	<b>U4</b> <b>U4</b>				0.76 0.76	0.32 0.35	0.46 0.45	0.29 0.32	–
16:81.4-81.5	rs310010	ESN (AFR) GWD (AFR) MSL (AFR) YRI (AFR)	$2.3 \times 10^{-8}$ $3.3 \times 10^{-8}$ $5.1 \times 10^{-10}$ $6.3 \times 10^{-10}$	<b>CMIP</b>				0.5	0.06	0.13	0.1	–
17:44-44.9	BP44262496 BP44363740	STU (SAS) ITU (SAS) PJL (SAS)	$2.6 \times 10^{-8}$ $1.9 \times 10^{-9}$ $2.4 \times 10^{-10}$	KANSL1 (KANSL1)	<b>RP11-798G7.5</b> <b>RP11-798G7.5</b>	<b>13</b> <b>14</b>	(KANSL1)	0.03 0.03	0.31 0.31	0.6 0.63	0.02 0.04	<a href="#">6</a> (MYL4 in GBR); Fay & Wu’s H
18:37.7-37.8	rs7236000 rs1943603	GWD (AFR) MSL (AFR) ESN (AFR)	$1.1 \times 10^{-9}$ $5.7 \times 10^{-10}$ $4.5 \times 10^{-9}$	<b>RP11-653G8.2</b> <b>RP11-653G8.2</b>				0.62 0.63	0.15 0.15	0.26 0.27	0.62 0.59	–

CHR:REGION	ID/BP	population	logP	gene by r2	eQTL	GWAS	NSM	AFR	EUR	SAS	EAS	notes
19:31.7-31.8	BP31733665 rs62101246	LWK (AFR) ESN (AFR) MSL (AFR)	$3.3 \times 10^{-11}$ $7.4 \times 10^{-10}$ $3.2 \times 10^{-9}$	TSHZ3 TSHZ3				0.44 0.47	0.13 0.14	0.2 0.24	0.18 0.19	–
19:45.8-45.9	rs12609631 rs10853773	MSL (AFR) ESN (AFR) YRI (AFR)	$2.9 \times 10^{-8}$ $3.5 \times 10^{-8}$ $3.5 \times 10^{-8}$	KLC3 KLC3		15 15		0.71 0.72	0.28 0.27	0.21 0.21	0.35 0.34	3
21:17.5-17.7	rs2823681	GWD (AFR) LWK (AFR) MSL (AFR) YRI (AFR)	$7 \times 10^{-9}$ $3.4 \times 10^{-10}$ $2.8 \times 10^{-8}$ $4 \times 10^{-8}$	LINC00478				0.47	0	0	0	–
22:23-23.2	rs2003444	CEU (EUR) IBS (EUR) TSI (EUR) BEB (SAS) GIH (SAS) ITU (SAS) PJL (SAS) STU (SAS)	$6.2 \times 10^{-9}$ $1.8 \times 10^{-8}$ $2 \times 10^{-9}$ $8.3 \times 10^{-10}$ $3.8 \times 10^{-10}$ $6.6 \times 10^{-11}$ $9.4 \times 10^{-9}$ $1.4 \times 10^{-9}$	IGLV3-16	IGLV3-12			0.49	0.62	0.65	0.64	Fu & Li's D

## Bibliography

[1] B. F. Voight, S. Kudaravalli, X. Wen, and J. K. Pritchard. A map of recent positive selection in the human genome. *PLoS Biology*, 4:e72, 2006.

[2] A. Sadier, L. Viriot, S. Pantalacci, and V. Laudet. The ectodysplasin pathway: from diseases to adaptations. *Trends in Genetics*, 30:24–31, 2014.

[3] P. C. Sabeti et al. Genome-wide detection and characterization of positive selection in human populations. *Nature*, 449:913–918, 2007.

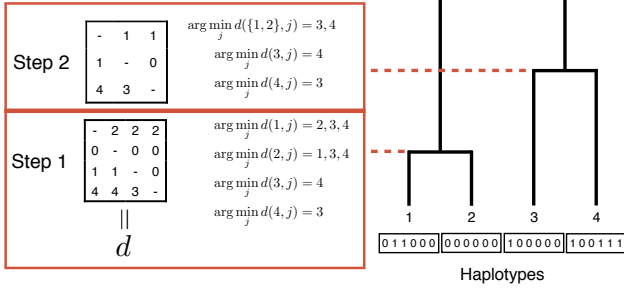
[4] Y. Field et al. Detection of human adaptation during the past 2000 years. *Science*, 354:760–764, 2016.

[5] I. Mathieson et al. Genome-wide patterns of selection in 230 ancient Eurasians. *Nature*, 528:499–503, 2015.

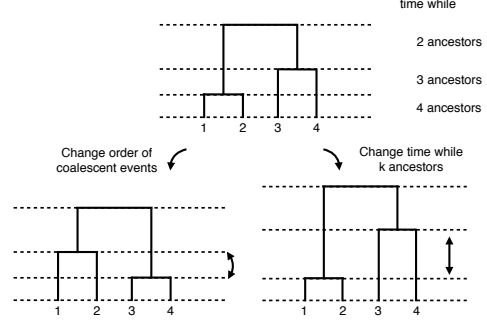
[6] P. F. Palamara, J. Terhorst, Y. S. Song, and A. L. Price. High-throughput inference of pairwise coalescence times identifies signals of selection and enriched disease heritability. *Nature Genetics*, 50:1311–1317, 2018.

[7] H. Chen, J. Hey, and M. Slatkin. A hidden Markov model for investigating recent positive selection through haplotype structure. *Theoretical Population Biology*, 99:18–30, 2015.

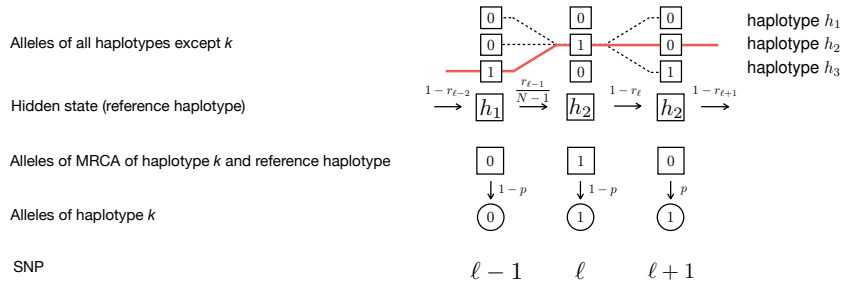
### a Tree builder



### b MCMC for branch length estimation

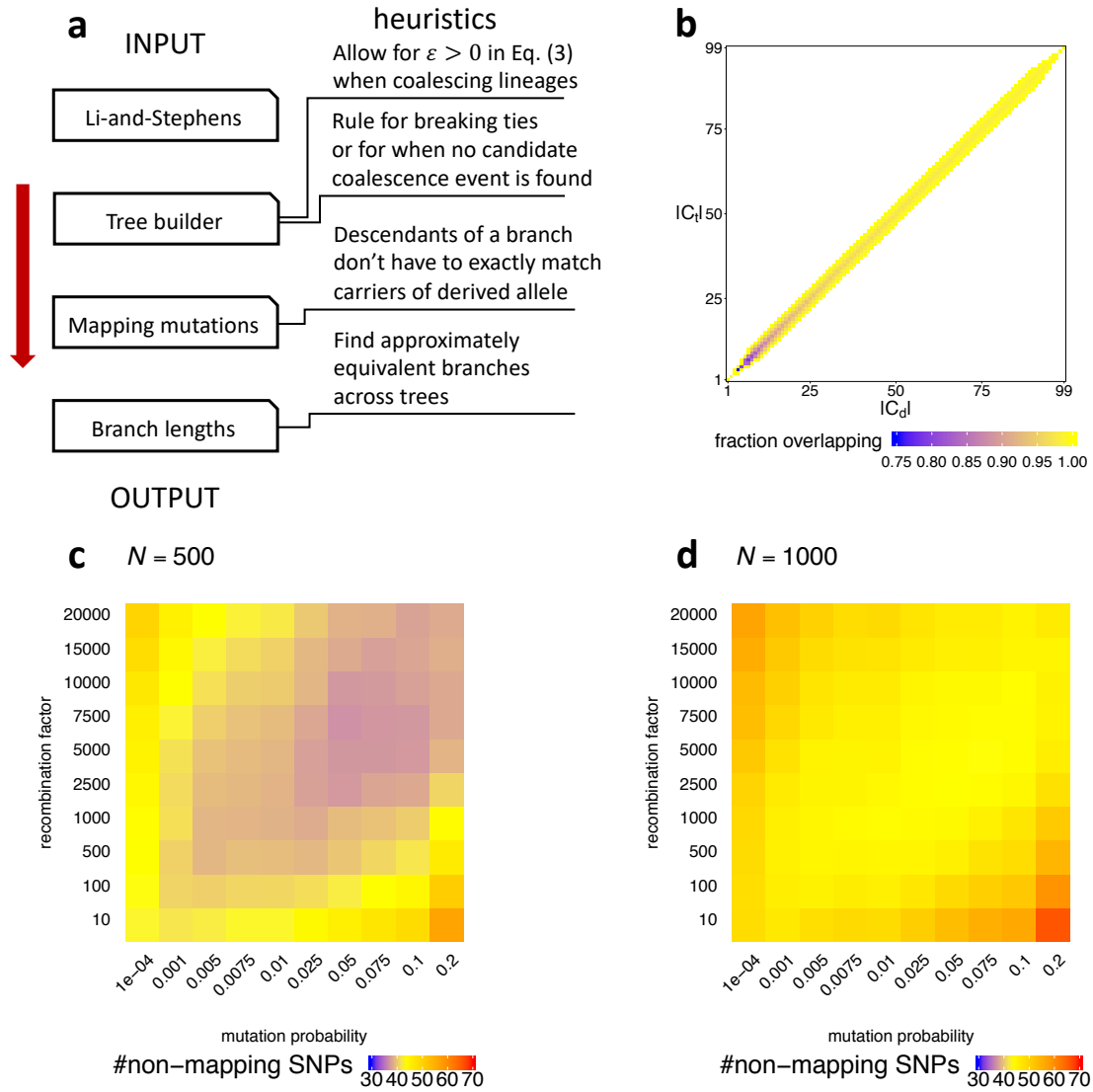


### c Modified Li-and-Stephens hidden Markov model



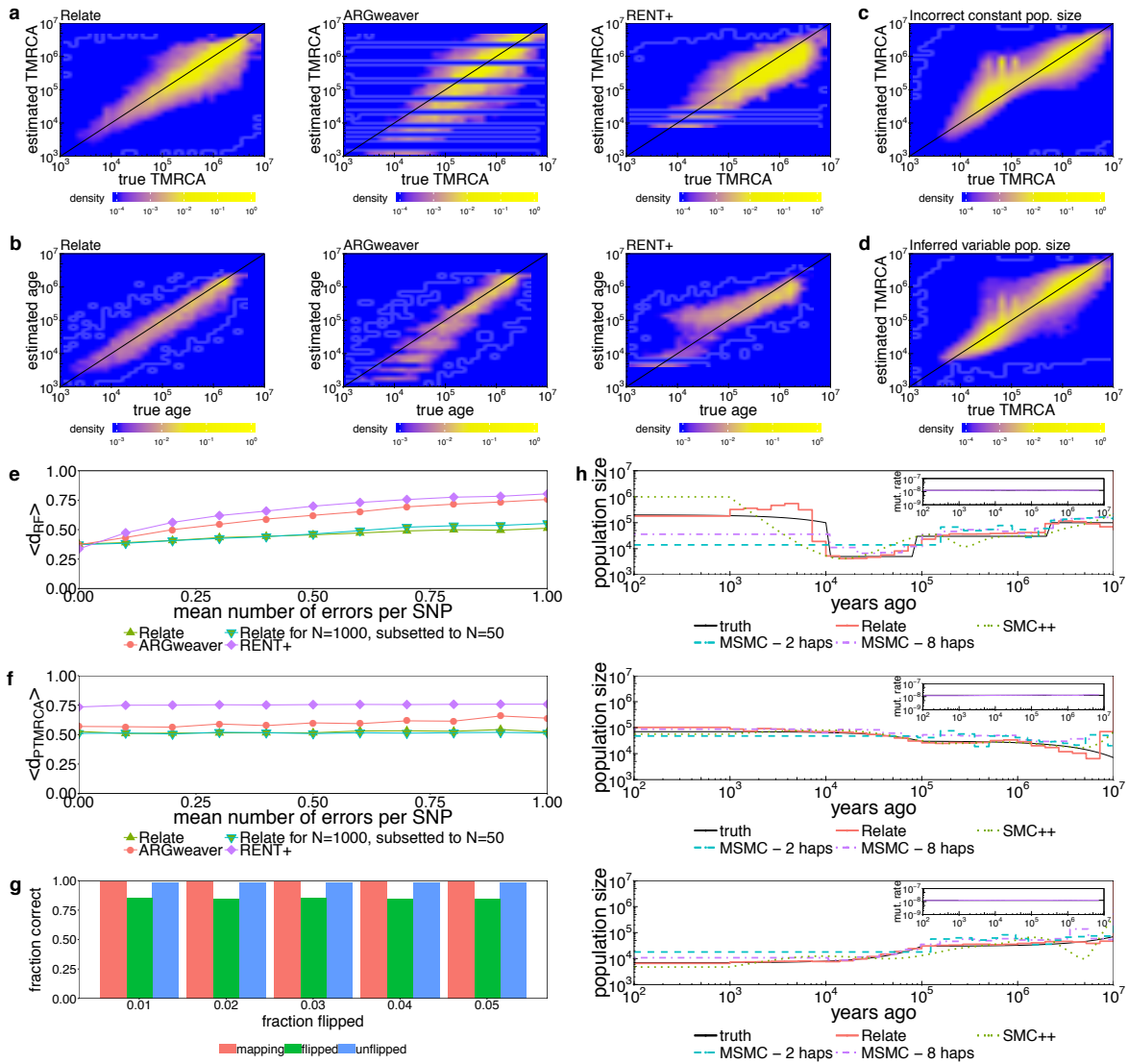
## Supplementary Figure 1: Schematics of the tree builder, branch length estimator, and modified Li-and-Stephens hidden Markov model.

**a**, Schematic of the hierarchical clustering algorithm for estimating tree topology. In the case of no recombination, the algorithm obtains matrix  $d$  containing the number of derived mutations as input. Row  $i$  of this matrix determines the order in which haplotype  $i$  coalesced with other haplotypes. Using Eq. (1), the algorithm finds the pair that coalesces with each other before coalescing with any other sequence. In the example shown here, we can coalesce haplotypes 0 and 2 or haplotypes 3 and 4. We choose to coalesce haplotypes 1 and 2 first because the symmetrised distance is smaller for this pair, however this choice does not affect tree topology in this case. The resulting tree topology is consistent with the gene tree describing the data. In contrast, when the hierarchical clustering algorithm is applied to the symmetrised matrix  $(d(i, j) + d(j, i))_{i, j=1, \dots, N}$ , haplotypes 2 and 3 are coalesced first and the constructed tree topology is wrong. This is equivalent to applying the UPGMA algorithm to the symmetrised matrix of derived mutations (Sokal R., Michener C. University of Kansas Science Bulletin 38, 1409-1438, 1958). **b**, Schematic of possible proposal moves in the MCMC algorithm for estimating branch lengths. We propose either a change in the order of coalescence events or a change in the time while  $k$  ancestors remain. **c**, Schematic of the modified Li-and-Stephens hidden Markov model (HMM) applied to haplotype  $k$ , which has alleles 0, 1, 1 at loci  $\ell - 1$ ,  $\ell$ ,  $\ell + 1$ . The emission and transition probabilities shown correspond to the path indicated by the red solid line. At SNP  $\ell - 1$ , the reference haplotype is  $h_1$  which has allele 1. Because the allele of haplotype  $k$  is 0, the allele of the MRCA with  $h_1$  is also 0 assuming that every mutation is unique in history. Therefore, the emission probability equals  $1 - p$ , where  $p$  is the probability of a mutation. At SNP  $\ell$ , the reference haplotype has changed to  $h_2$ . The alleles of haplotype  $k$  and  $h_2$  are 1. Therefore, the MRCA has allele 1 and the emission probability is given by  $1 - p$ . At SNP  $\ell + 1$ , haplotype  $k$  has allele 1. The allele of the reference haplotype  $h_2$  is 0 and so is that of the MRCA, such that the emission probability equals  $p$ . Using this HMM, we calculate the likelihood  $P_m(H_\ell = j | D^{(k)})$ . This is the likelihood of copying from reference haplotype  $j$  at SNP  $\ell$ , conditional on observing  $D^{(k)}$ . We notice that  $P_m(H_\ell = j | D^{(k)})$  is obtained as the sum of all possible paths when  $H_\ell = j$  is fixed (indicated by the dashed lines).



**Supplementary Figure 2: Mapping rule for mutations and sensitivity of the modified Li-and-Stephens HMM to parameter choice.**

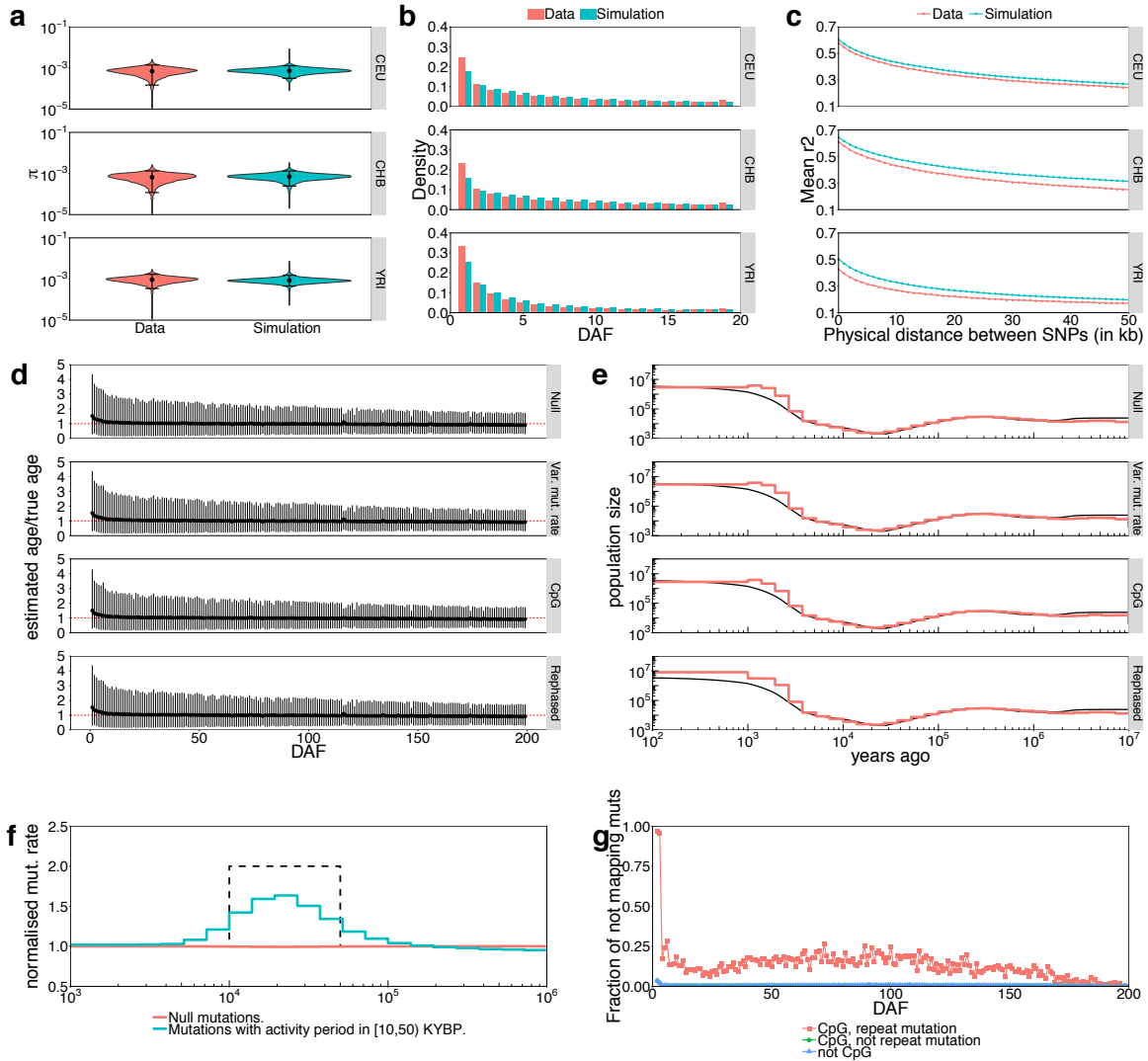
**a**, Schematic illustrating which parts of Relate use heuristic approaches. **b**, Heatmap showing the necessary and sufficient overlap between the set of descendants of a branch ( $C_t$ ) and the set of carriers of the derived allele ( $C_d$ ), given  $|C_t|$  and  $|C_d|$ , with  $N = 100$ , as determined by Eqs. (4) and (5). Colours show  $|C_t \cap C_d| / \min\{|C_t|, |C_d|\}$ , where white indicates that a mutation can never be mapped for the corresponding combination of  $|C_t|$  and  $|C_d|$ . **c**, **d**, Number of non-mapping SNPs for different values of  $p$  (horizontal axis) and  $R$  (vertical axis) for  $N = 500$  (**c**) and  $N = 1000$  (**d**). The subsets of haplotypes are chosen uniformly at random from all haplotypes. We calculated the mean over 50 randomly chosen subregions of length 1200 SNPs on chromosome 20. In our implementation, we fixed  $p = 0.025$  and  $R = 2500$ .



**Supplementary Figure 3: Performance of Relate on simulated data.**

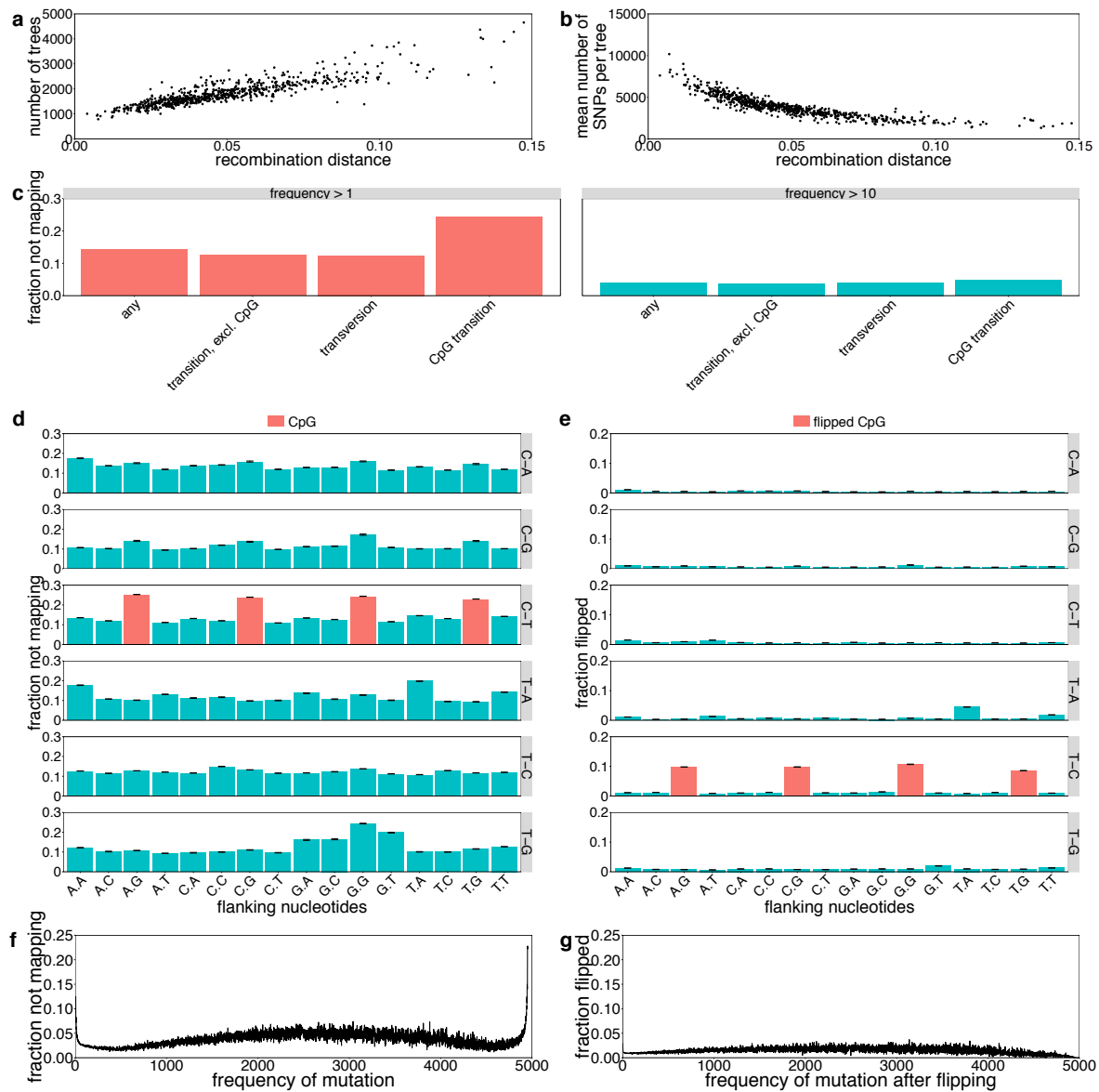
**a**, Estimated times to most recent common ancestors (TMRCA) between pairs of haplotypes compared to the truth for Relate, ARGweaver, and RENT+. **b**, Estimated ages of mutations plotted against the true age for Relate, ARGweaver, and RENT+. We determine the age of a mutation by placing it at the midpoint of the branch onto which it maps. In **a** and **b**, we simulate  $N = 200$  haplotypes with  $2N_e = 40,000$ . **c**, TMRCA between pairs of haplotypes compared to the truth for a simulated data set with  $N = 200$  haplotypes and a population bottleneck resembling that of Europeans, where branch lengths are estimated using a constant population size of  $2N_e = 30,000$ . **d**, Estimated TMRCA compared to the truth for the same example as in **c**, where branch lengths and population size history are jointly inferred. **e**, Robinson-Foulds distance and **f**, pairwise TMRCA distance averaged over 2.4Mb for Relate, ARGweaver, and RENT+. We estimate genealogies for  $N = 50$  haplotypes at different number of errors. In addition, we show the accuracy of the genealogy corresponding to  $N = 50$  haplotypes, embedded in an estimated genealogy for  $N = 1000$  haplotypes (see Supplementary Note: Simulations, Section 2.1 for details). **g**, Robustness of Relate with respect to randomly introduced flipped mutations. We show the fraction of SNPs mapping to a unique branch, fraction of correctly flipped SNPs, and fraction of correctly unflipped SNPs for Relate. We exclude SNPs at frequency 1, which always map to the tree. We simulate 2.5Mb for  $N = 200$  haplotypes with  $2N_e = 30,000$ . **h**, Population size estimates for simulations with a discrete bottleneck, an increasing trend, and a decreasing trend in populations size. Estimates using Relate are shown by the blue solid line. We apply SMC++ to the same data set and we also apply MSMC2 with 2 and 8 haplotypes. In the inset, we show the mutation rate over time estimated by Relate. For each scenario, we simulate 200Mb for  $N = 200$  haplotypes. In all simulations, the mutation rate is set to  $1.25 \times 10^{-8}$  and recombination rates are taken from the 1000 Genomes Project map for chromosome 1.





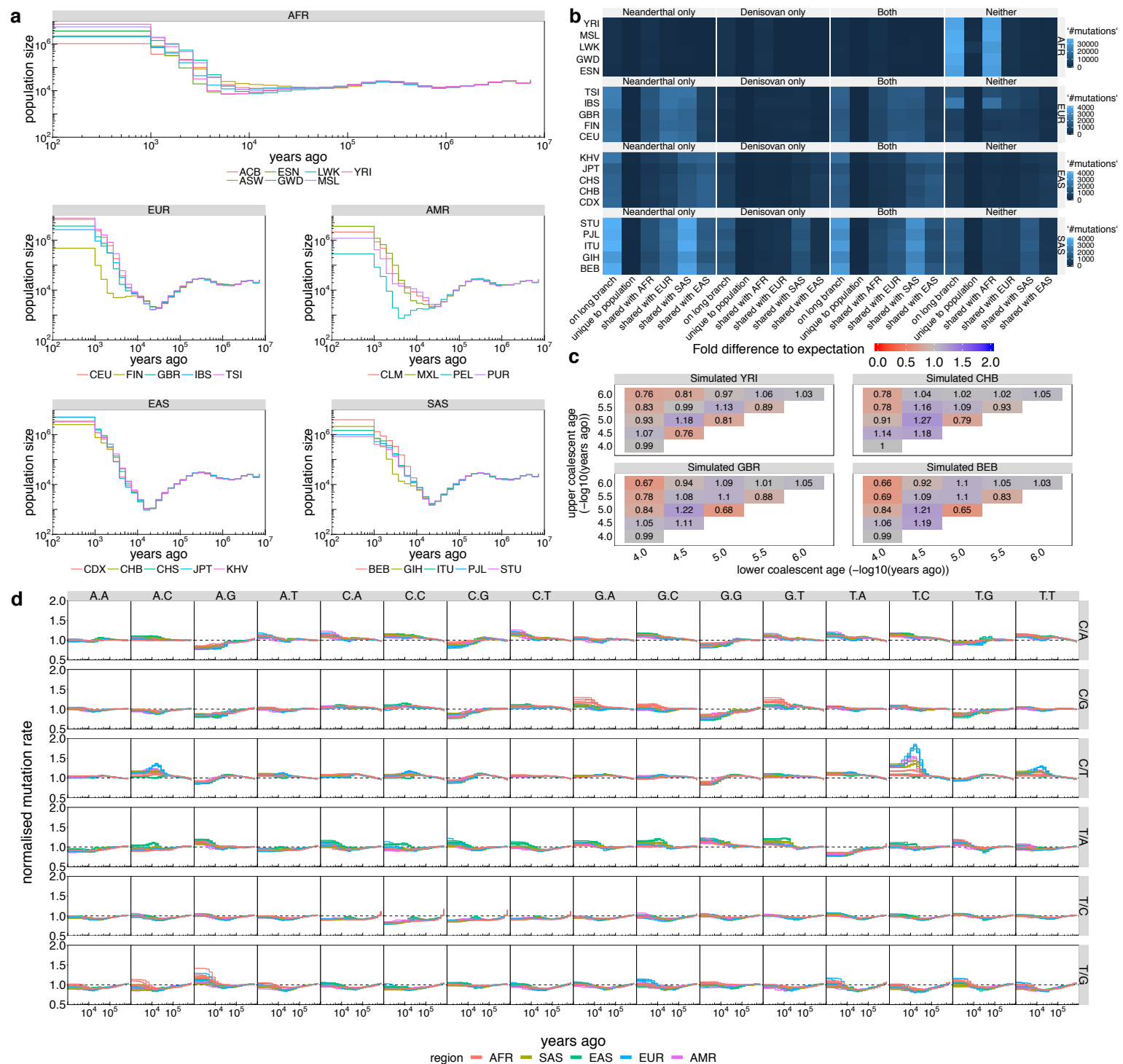
**Supplementary Figure 4: Accuracy under perturbations from infinite-sites, constant mutation rate, or perfect phase.**

**a**, Expected heterozygosity ( $\pi$ ) calculated for 20,000 randomly chosen 100kb windows. Circles show the mean and bars indicate the 2.5th and 97.5th percentiles. **b** Derived allele frequencies and **c**, LD decay patterns. For **a**, **b**, and **c**, we used ten 1000 Genomes Project individuals, and simulated 20 haplotypes using the demographic histories estimated by Relate. Each statistic is calculated using chromosome 1 (see Supplementary Note: Simulations, Section 3 for details). **d**, Ratio of estimated and true age of a mutation, estimated as the mean of the lower and upper ends of the branch onto which the mutation maps, as a function of DAF. Circles show the mean ratio and bars indicate the 2.5th and 97.5th percentiles. Base-line simulation assumes infinite-sites and a constant mutation rate of  $1.25 \times 10^{-8}$ . We introduce perturbations, such as a variable mutation rate to a subset of sites, hypermutable base-pair positions emulating CpG dinucleotides, and inferred phase (see Suppl. Note Simulations, Sec. 4 for details). **e**, Accuracy of Relate-estimated population sizes on the same simulations as in (a). **f**, Normalised mutation rate for null mutations with a constant mutation rate of  $1.25 \times 10^{-8}$  and a mutation category with an activity period in [10, 50) YBP during which the mutation rate doubled (dashed lines). **g**, Fraction of not mapping mutations as a function of DAF for the simulation with CpG-like mutations, categorised by whether the CpG-like site mutated once or more than once.



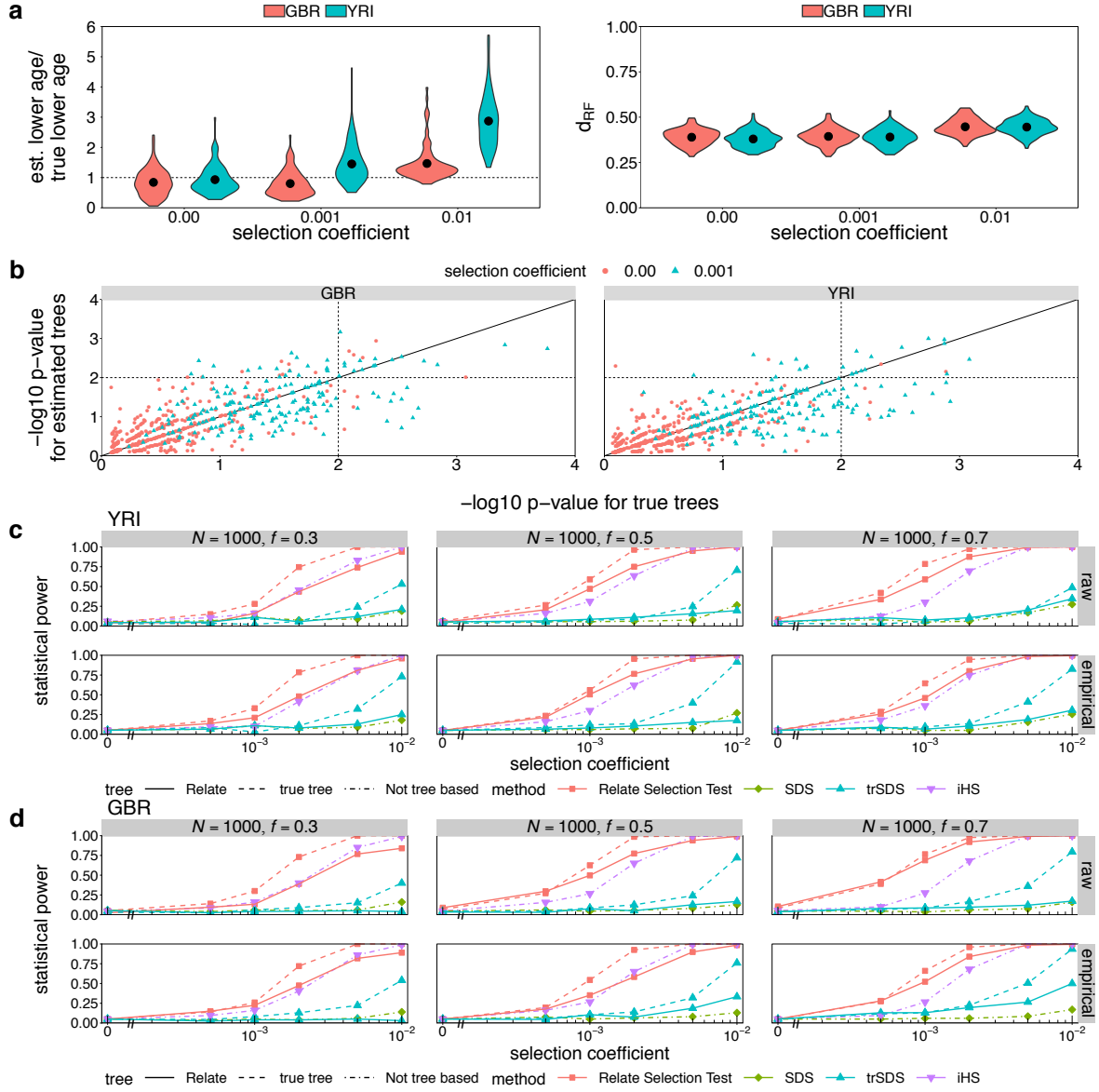
**Supplementary Figure 5: Properties of the genealogy constructed for the 1000 Genomes Project data set.**

**a**, Number of trees built versus the recombination distance for all 22 chromosomes. **b**, Mean number of SNPs that map to a unique branch versus the recombination distance in that bin. Every point represents a subregion of  $10^5$  SNPs. **c**, Fraction of SNPs that could not be mapped to a unique branch for SNPs excluding singletons (left) and SNPs with derived allele frequencies larger than 10 (right). **d**, Fraction of SNPs that could not be mapped to a unique branch for all 96 possible triplet mutations, excluding singletons. **e**, Fraction of SNPs that were flipped for all 96 possible triplet mutations, excluding singletons. In **d** and **e**, CpG transitions are indicated in red. The 95% confidence intervals are indicated by black brackets. **f**, Fraction of non-mapping SNPs by derived allele frequency of the mutation in the sample. For each frequency, we divide the number of non-mapping mutations of that frequency by the number of mutations of that frequency. **g**, Fraction of flipped SNPs by derived allele frequency of the mutation after flipping. For each frequency, we divide the number of flipped SNPs of that frequency (after flipping) by the number of SNPs of that frequency.



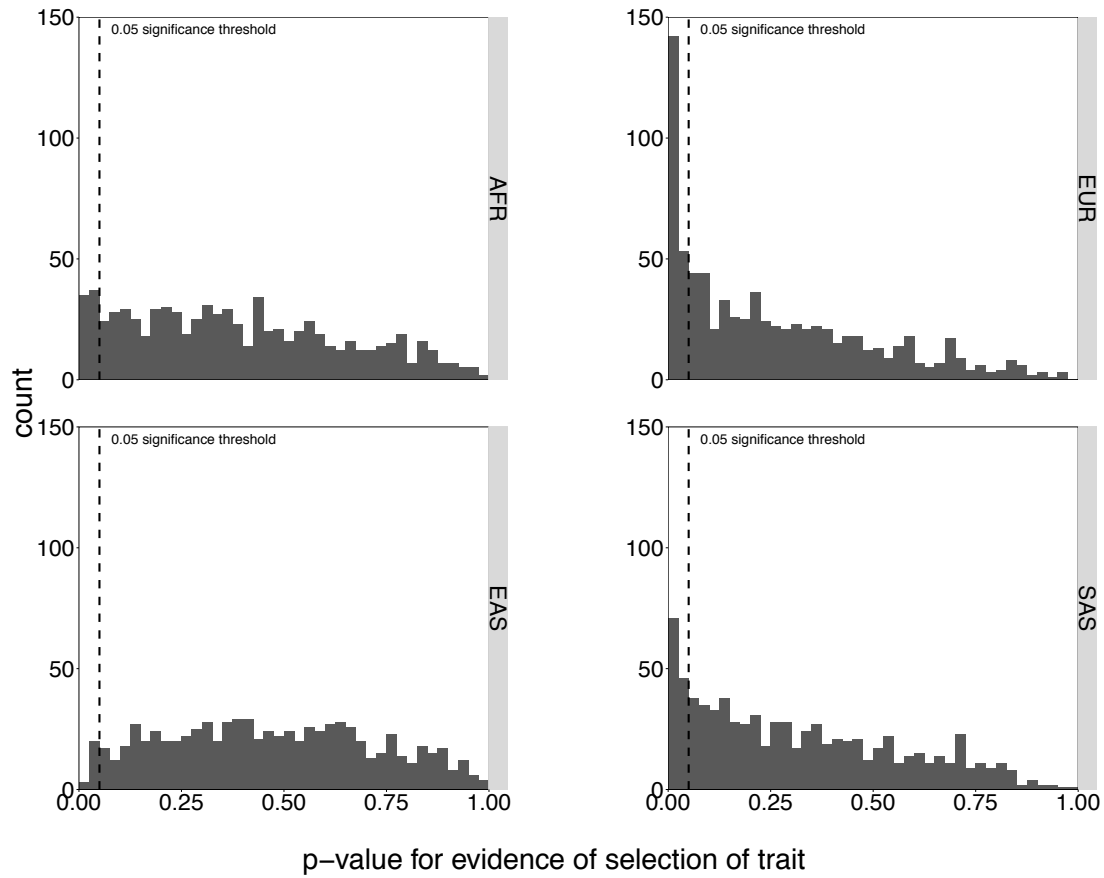
**Supplementary Figure 6: Historical effective population sizes and evidence of introgression, mutation rate trends for 96 triplet mutations.**

**a**, Historical effective population sizes for all 26 populations in the 1000 Genomes Project dataset. For each population, we first extract the genealogy corresponding to that population. We then estimate the population size using this genealogy. **b**, Number of mutations on branches with an upper end older than 1M YBP and lower end younger than 30,000 YBP, categorised by whether the mutation is additionally found only in Neanderthals, only in Denisovans, both, or neither. For each category, we also distinguish whether the mutation is unique to the population of interest or shared with other populations in AFR, EUR, SAS, or EAS. **c**, Number of mutations binned by age of upper and lower coalescence event, relative to the expected number of mutations when randomising topology while fixing ages of coalescence events for four simulated data sets (**Methods**). We simulated  $3 \times 10^9$  bases with population size histories of YRI, CHB, GBR, and BEB. **d**, Normalised mutation rate of triplet mutations for all 96 possible categories. Analogous to Figure 4a.



**Supplementary Figure 7: Power simulations for selection test.**

**a**, Ratio of estimated and true lower-end ages of the branches onto which a mutation with present-day DAF of 0.5 maps. This mutation has a selection coefficient of 0, 0.001, or 0.01 and is positioned at 10Mb of a 20Mb simulated genomic region with Relate-estimated population size histories for GBR and YRI. We simulated 100 realisations of  $N = 200$  haplotypes. Circles indicate the mean ratios. **b**, P-values for selection evidence in simulations calculated using true trees (horizontal axis) and estimated trees (vertical axis) for the same simulation scenario as in **a**. We plot p-values for loci under no selection (circles) and loci under weak selection (triangles). **c**, **d**, Power simulations with  $N = 1000$  haplotypes and present-day derived allele frequencies of 0.3, 0.5, and 0.7. We assume a population size history estimated for YRI (**c**), and GBR (**d**), respectively. The significance threshold is 0.05. We show power estimates using the p-values for trees estimated by Relate, as well as those for the true trees. In both cases, we estimate power using raw p-values of our test statistic (top row) and empirical p-values given the distribution of raw p-values in the neutral case (bottom row). For iHS, SDS, and trSDS, power is estimated by standardising raw scores by the frequency specific mean and standard deviation under the null. In the top row, we assume a standard normal distribution of the standardised score and in the bottom row, we calculate empirical p-values by determining a critical score corresponding to the 0.05 significance level in the neutral case.



**Supplementary Figure 8: Histograms of p-values for evidence of selection on complex traits.** We aggregated both effect directions of 84 considered traits, as well as populations in each of the four considered geographic regions (AFR, EAS, EUR, SAS).